UNIVERSITÉ PARIS SUD | FACULTÉ DES SCIENCES D'ORSAY

ECOLE NORMALE SUPERIEURE

THÈSE

présentée pour obtenir

LE GRADE DE DOCTEUR EN SCIENCES
DE L'UNIVERSITÉ PARIS-SUD 11

Spécialité : Mathématiques

par

Sébastien GERCHINOVITZ

Prédiction de suites individuelles et cadre statistique classique : étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation

---

dirigée par Gilles STOLTZ

Rapporteurs :

M. Arnak	**DALALYAN**	CREST et ENSAE
M. Claudio	**GENTILE**	Università degli Studi dell'Insubria, Varèse

Soutenue le 12 décembre 2011 à l'École normale supérieure devant le jury composé de

M. Pierre	**ALQUIER**	CREST et Université Paris 7	Examinateur
M. Olivier	**CATONI**	CNRS et École normale supérieure	Examinateur
M. Arnak	**DALALYAN**	CREST et ENSAE	Rapporteur
M. Pascal	**MASSART**	Université Paris-Sud 11	Examinateur
M. Gilles	**STOLTZ**	CNRS et École normale supérieure	Directeur
M. Alexandre	**TSYBAKOV**	CREST, ENSAE et Université Paris 6	Examinateur

Prédiction de suites individuelles et cadre statistique classique : étude de quelques liens autour de la régression parcimonieuse et des techniques d'agrégation

Résumé : Cette thèse s'inscrit dans le domaine de l'apprentissage statistique. Le cadre principal est celui de la prévision de suites déterministes arbitraires (ou *suites individuelles*), qui recouvre des problèmes d'apprentissage séquentiel où l'on ne peut ou ne veut pas faire d'hypothèses de stochasticité sur la suite des données à prévoir. Cela conduit à des méthodes très robustes. Dans ces travaux, on étudie quelques liens étroits entre la théorie de la prévision de suites individuelles et le cadre statistique classique, notamment le modèle de régression avec *design* aléatoire ou fixe, où les données sont modélisées de façon stochastique. Les apports entre ces deux cadres sont mutuels : certaines méthodes statistiques peuvent être adaptées au cadre séquentiel pour bénéficier de garanties déterministes ; réciproquement, des techniques de suites individuelles permettent de calibrer automatiquement des méthodes statistiques pour obtenir des bornes adaptatives en la variance du bruit. On étudie de tels liens sur plusieurs problèmes voisins : la régression linéaire séquentielle parcimonieuse en grande dimension (avec application au cadre stochastique), la régression linéaire séquentielle sur des boules ℓ^1, et l'agrégation de modèles non linéaires dans un cadre de sélection de modèles (régression avec *design* fixe). Enfin, des techniques stochastiques sont utilisées et développées pour déterminer les vitesses minimax de divers critères de performance séquentielle (regrets interne et *swap* notamment) en environnement déterministe ou stochastique.

Mots-clés : Apprentissage statistique, prévision séquentielle, suites individuelles, agrégation PAC-bayésienne, pondération exponentielle, régression parcimonieuse, grande dimension, calibration automatique, vitesses minimax, regret externe, regret interne, sélection de modèles.

Prediction of individual sequences and prediction in the statistical framework: some links around sparse regression and aggregation techniques

Abstract: The topics addressed in this thesis lie in statistical machine learning. Our main framework is the prediction of arbitrary deterministic sequences (or *individual sequences*). It includes online learning tasks for which we cannot make any stochasticity assumption on the data to be predicted, which requires robust methods. In this work, we analyze several connections between the theory of individual sequences and the classical statistical setting, e.g., the regression model with fixed or random design, where stochastic assumptions are made. These two frameworks benefit from one another: some statistical methods can be adapted to the online learning setting to satisfy deterministic performance guarantees. Conversely, some individual-sequence techniques are useful to tune the parameters of a statistical method and to get risk bounds that are adaptive to the unknown variance. We study such connections for several connected problems: high-dimensional online linear regression under a sparsity scenario (with an application to the stochastic setting), online linear regression on ℓ^1-balls, and aggregation of nonlinear models in a model selection framework (regression on a fixed design). We also use and develop stochastic techniques to compute the minimax rates of game-theoretic online measures of performance (e.g., internal and swap regrets) in a deterministic or stochastic environment.

Keywords: Statistical learning, online learning, machine learning, individual sequences, regret bounds, PAC-Bayesian aggregation, exponential weighting, high-dimensional regression, sparsity, parameter tuning, minimax rates, external regret, internal regret, swap regret, model selection.

AMS Classification: 68Q32, 62J02, 62J05, 62C20.

Remerciements

Je voudrais tout d'abord te remercier, Gilles, pour ton encadrement inestimable. Ton dynamisme, ta rigueur, ton intuition, ton sens exceptionnel de l'organisation et ton enthousiasme continu font de toi un directeur de thèse exemplaire. J'espère pouvoir continuer à travailler avec toi sur ce sujet passionnant et aux multiples facettes qu'est la prévision de suites individuelles. Mille mercis donc.

Je suis très honoré qu'Arnak Dalayan et Claudio Gentile aient accepté de rapporter ma thèse. Merci pour l'attention que vous avez portée à la lecture de ce manuscrit.

Je tiens ensuite à remercier Pascal Massart, non seulement pour m'avoir initié aux joies de la concentration de la mesure et de la sélection de modèles en cours de M2, mais aussi pour ses précieux conseils tant sur les plans mathématiques qu'académiques ; c'est notamment grâce à toi que j'ai réalisé ma thèse avec Gilles. Merci vivement à Alexandre Tsybakov et Olivier Catoni pour l'intérêt qu'ils ont porté à mes travaux en acceptant de faire partie du jury ; cette thèse doit beaucoup à vos travaux respectifs. Enfin, Pierre, je suis également très heureux de te voir parmi le jury, et je serais ravi de pouvoir profiter de ton dynamisme au cours d'une collaboration future.

Avant de remercier mes collègues de travail, je souhaiterais exprimer ma reconnaissance envers plusieurs professeurs de mathématiques qui ont guidé mon parcours scolaire puis universitaire. Je pense en particulier à Jean-Paul Courant (en Terminale S) et à Denis Choimet (en MPSI2 à Clémenceau) : c'est vous qui m'avez véritablement donné le goût des mathématiques. Je tiens aussi à remercier Erick Herbin et Lionel Gabet, à l'École Centrale, pour avoir chaleureusement facilité ma transition vers le monde académique. Merci également à Jean-François Le Gall et Elisabeth Gassiat qui, tout comme Pascal, m'ont prodigué de précieux conseils dès mon arrivée à Orsay ; vos cours de M2 sont un excellent tremplin vers le monde des probabilités et statistiques. Je tiens enfin à remercier Vivien Mallet pour son co-encadrement hors du commun lors de mon stage de M2 : ta collaboration avec Gilles promet un très bel avenir aux techniques de suites individuelles dans le domaine de la prévision de la qualité de l'air !

Préparer ma thèse au sein du Département de Mathématiques et Applications de l'ENS fut un vrai plaisir. Je ne pourrai remercier tout le monde, tant la liste est longue. Je remercierai en particulier mes collègues de bureau : Christophe, Léo et Ben au 302, puis Jia Yuan, Thomas, Pierre et Émilien au 102. Merci aussi à Marie, Amandine, Max et Ting Yu pour avoir partagé leurs bureaux et de courts (mais chaleureux) moments pendant les travaux au DMA. Merci également à Vincent, Nicolas, Laure, Augusto, Gérard, Thierry, Zhan et tous les autres qui j'espère se reconnaîtront entre les lignes. Ah si, je tiens à remercier particulièrement Zaïna et Bénédicte pour leur gentillesse et leur efficacité exemplaire : le DMA vous doit beaucoup.

Je tiens également à remercier tous les collègues et amis que j'ai côtoyés à Orsay. Je ne pourrai en citer que quelques-uns, car la liste est encore longue : Jérémie et Shweta, au 108, mais aussi Aurélien, Laure, Olivier, Caroline et Thierry pour les bons moments passés en salle de thé ou autour d'un jeu de cartes à Fréjus ! Caroline, merci également pour ta relecture attentive de mon introduction. Un grand merci aussi aux anciens, avec une mention spéciale pour Robin : tes conseils et ton enthousiasme m'ont été très précieux. Enfin, Valérie, que serait l'École Doctorale

sans vous ? Merci pour votre aide et votre disponibilité.

Je souhaiterais par ailleurs saluer toutes les personnes que j'ai pu croiser au cours de séminaires ou conférences, pour les discussions mathématiques que nous avons pu entretenir, mais aussi pour les bons moments passés au soleil (ou sous la pluie) : Joseph, Odalric, Sébastien, Sylvain, Christophe, Antoine, Guillaume, pour n'en mentionner que quelques-uns. Et une mention spéciale pour Mohamed : un grand merci pour ta relecture minitieuse de mon introduction et pour les conseils que tu as pu m'apporter au cours de ma thèse. Merci aussi à Nicolò pour son accueil chaleureux à Milan : les quelques jours de recherche avec vous en mars dernier ont été très stimulants !

Enfin, je souhaiterais remercier tous mes proches pour leur bienveillance et tous les bons moments passés ensemble. Je pense évidemment à la petite troupe de Centrale : Benoît, Christophe, Manu, Toutoune et Thibaut ; merci d'avoir supporté mes digressions mathématiques ! Merci aussi à Cédric, Wafaa et Fabrizio ; les sorties ciné et crêperie à Montparnasse ont été un régal ! François, je suis content que nos parcours parallèles depuis le collège nous aient conduits tous deux à Orsay ; nous aurons même soutenu à deux semaines d'intervalle ! Pour finir, je tiens à remercier chaleureusement mes parents, mes grands-parents et ma sœur pour le soutien inconditionnel dont ils ont fait preuve tout au long de cette thèse ; ce travail vous est entièrement dédié.

Table des matières

Chapitre 1

Vue d'ensemble des résultats

Ce chapitre est une exposition des principaux résultats de cette thèse. On présente d'abord très brièvement le cadre, qui est celui de la prévision séquentielle de suites déterministes arbitraires (ou *suites individuelles*) ainsi que ses liens étroits avec des cadres statistiques plus classiques comme le modèle de régression avec un plan d'expérience aléatoire ou fixe. (Une présentation plus étoffée est proposée au chapitre 2.) Nous détaillons ensuite les contributions principales de chaque chapitre (sections 1.2 à 1.5) et concluons en présentant plusieurs axes de recherche futurs.

Contents

1.1 Prévision de suites individuelles et cadre statistique classique

Dans cette thèse, on s'intéresse à deux types de problèmes d'apprentissage, tous deux du domaine de la prévision :

- Le cadre principal de cette thèse est celui de la prévision de suites déterministes arbitraires (ou *suites individuelles*) : il recouvre des problèmes d'apprentissage séquentiel où l'on ne peut ou ne veut pas faire d'hypothèses de stochasticité sur la suite des données à prévoir. Les algorithmes séquentiels qui en résultent bénéficient de garanties déterministes – valables dans le pire des cas – et sont donc en ce sens très robustes.
- Nous nous sommes également intéressés aux liens étroits entre la prévision de suites individuelles et des cadres satistiques plus classiques comme le modèle de régression avec *design* fixe ou aléatoire, où les données observées sont cette fois modélisées de façon stochastique.

Dans ce chapitre, nous introduisons brièvement le cadre de la prévision de suites individuelles et décrivons les liens qu'il nourrit avec le cadre statistique classique. Nous exposons ensuite les contributions principales de cette thèse dans les sections 1.2 à 1.5, lesquelles correspondent aux chapitres centraux, i.e., les chapitres 3 à 6. On clôt ce chapitre par un bref exposé des perspectives de recherche (section 1.6). Une introduction plus mathématique aux prérequis nécessaires à la lecture de cette thèse est proposée au chapitre 2.

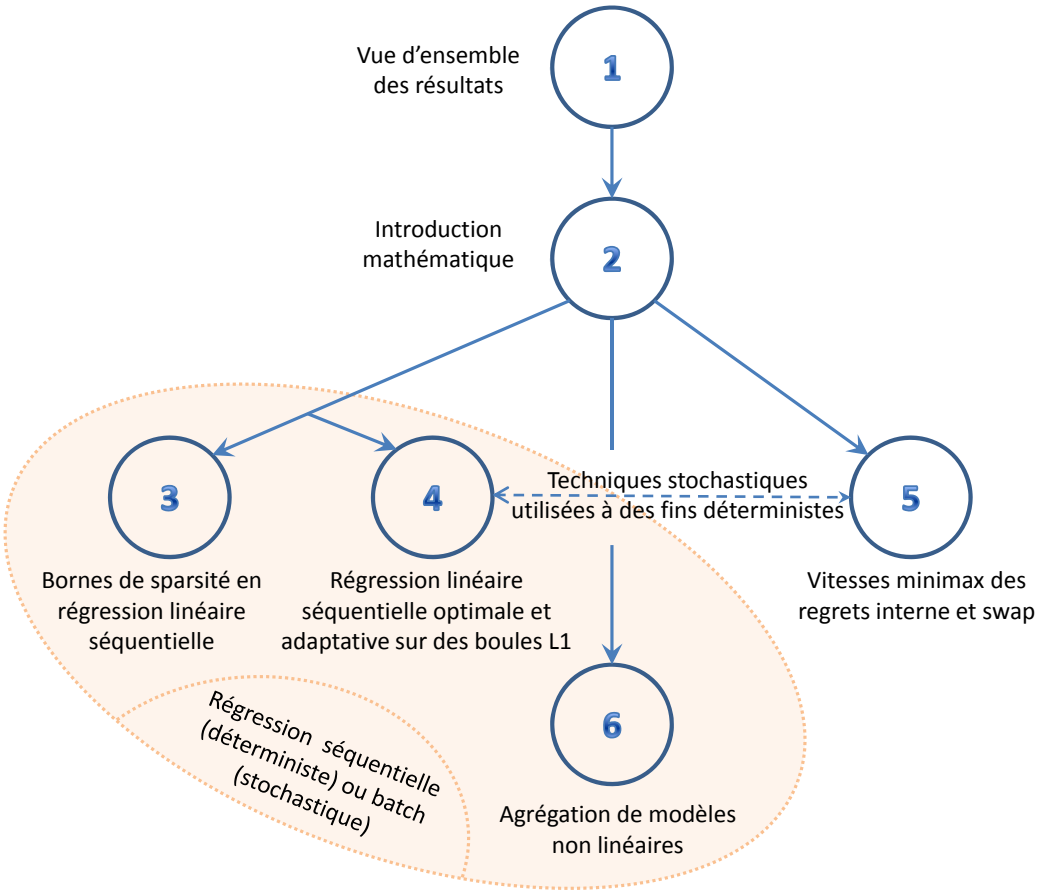


FIGURE 1.1 – Structure générale de cette thèse : dépendences entre les chapitres 1 à 6.

1.1.1 Prévision de suites individuelles

Considérons la tâche de prévision séquentielle suivante. Un statisticien cherche à prévoir tour après tour les valeurs inconnues d'une suite d'observations $y_1, y_2, \ldots \in \mathcal{Y}$ à partir de prévisions (ou décisions) $\widehat{a}_1, \widehat{a}_2, \ldots \in \mathcal{D}$. (Les espaces d'observation $\mathcal{Y}$ et de décision $\mathcal{D}$ peuvent différer.) En théorie statistique classique de prévision séquentielle, il est d'usage de supposer que la suite $y_1, y_2, \ldots$ est la réalisation d'un certain processus stochastique, par exemple ergodique stationnaire. De telles hypothèses permettent d'estimer séquentiellement les caractéristiques du processus sous-jascent, et ainsi de construire des méthodes de prévision performantes quand le modèle statistique choisi décrit bien les données en jeu. Cela peut en revanche s'avérer irréaliste dans certaines situations difficilement modélisables de façon statistique, par exemple, lorsque la suite $y_1, y_2, \ldots$ évolue et réagit aux décisions $\widehat{a}_1, \widehat{a}_2, \ldots$ comme c'est le cas pour la détection de courriels frauduleux ou pour l'investissement sur le marché boursier.

Dans la théorie dite de prévision de suites individuelles, aucune hypothèse de stochasticité n'est faite sur la façon dont est générée la suite des observations $y_1, y_2, \ldots$. Toutes les suites possibles sont considérées et des garanties théoriques sont disponibles pour chacune d'elles – d'où le nom de prévision de *suites individuelles*.

Dans un cadre aussi général, il est irréaliste de chercher à prévoir correctement l'observation y_t à chaque date t et sur le seul fondement des observations passées. En revanche, si le statisticien dispose à chaque instant t de prévisions de base (ou avis d'experts) $a_{\theta,t} \in \mathcal{D}$, $\theta \in \Theta$, alors un but raisonnable consiste à prévoir presque aussi bien que le meilleur des experts sur le long terme. Ce problème générique, qualifié de *prévision avec avis d'experts*, est celui considéré dans cette thèse. Une description sous la forme d'un jeu répété entre le statisticien et l'environnement est donnée en figure 1.2.

Paramètres : espace de décision convexe $\mathcal{D}$, espace d'observation $\mathcal{Y}$, fonction de perte $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$, et ensemble Θ des indices d'experts.

A chaque date $t \in \mathbb{N}^* \triangleq \{1, 2, \ldots\}$,

1. l'environnement choisit les avis d'experts $a_{\theta,t} \in \mathcal{D}$ pour tout $\theta \in \Theta$; ils sont révélés au statisticien ;
2. le statisticien prend une décision $\widehat{a}_t \in \mathcal{D}$, qu'il garde confidentielle ou révèle[a] à l'environnement ;
3. l'environnement choisit et révèle l'observation $y_t \in \mathcal{Y}$;
4. le statisticien encourt la perte $\ell\big(\widehat{a}_t, y_t\big)$ et chaque expert $\theta \in \Theta$ encourt la perte $\ell\big(a_{\theta,t}, y_t\big)$.

[a] Si l'environnement n'a pas accès aux décisions $\widehat{a}_t$ du statisticien, il est qualifié d'*oublieux*. Si, à l'inverse, l'environnement peut réagir aux décisions passées du statisicien, il est qualifié d'*antagoniste*. Ces deux cadres sont équivalents lorsque l'algorithme de prévision utilisé est déterministe ; cf. section 2.3.1.

FIGURE 1.2 – Prévision avec avis d'experts.

Dans ce cadre, la qualité des prévisions du statisticien après T tours de prévision est mesurée par sa perte cumulée $\sum_{t=1}^{T} \ell(\widehat{a}_t, y_t)$. Un objectif classique est alors de faire en sorte que, malgré la contrainte de prévision séquentielle, cette perte cumulée soit presque aussi petite que celle du meilleur expert a posteriori. Cela correspond à minimiser la différence

$$\sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) - \inf_{\theta\in\Theta} \sum_{t=1}^{T} \ell(a_{\theta,t}, y_t) \; .$$

Cette différence est appelée *regret (externe)*. D'autres formes de regret en lien avec la théorie des jeux (par ex., regret interne, regret *swap*) sont considérées au chapitre 5 — cf. section 1.4.

Dans l'essentiel de cette thèse, nous suivons l'approche des suites individuelles, i.e., nous étudions des stratégies du statisticien dont le regret (externe) est "petit" uniformément en toutes les suites $y_1, y_2, \ldots \in \mathcal{Y}$. Par "petit", il convient d'entrendre sous-linéaire en T (puisqu'une vitesse linéaire en T est triviale quand ℓ est bornée). Cela correspond à un regret moyen dans le pire des cas qui est asymptotiquement négatif quand $T \to +\infty$, i.e.,

$$\sup_{\substack{y_1,\ldots,y_T\in\mathcal{Y} \\ (a_{\theta,1})_\theta,\ldots,(a_{\theta,T})_\theta\in\mathcal{D}^\Theta}} \left\{ \frac{1}{T}\sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) - \inf_{\theta\in\Theta} \frac{1}{T}\sum_{t=1}^{T} \ell(a_{\theta,t}, y_t) \right\} \leqslant o(1) \qquad \text{quand} \quad T \to +\infty \; .$$

Une telle garantie indique qu'en moyenne, le statisticien prévoit presque aussi bien que le meilleur des experts a posteriori. Quand Θ est fini de cardinal K, des ordres de grandeur typiques pour le regret moyen dans le pire des cas sont $\sqrt{(\ln K)/T}$ lorque la perte ℓ est bornée et convexe ou $(\ln K)/T$ lorsque la perte ℓ est exp-concave.

On suppose ci-après que $\Theta = \{1, \ldots, K\}$. Un exemple classique et fondamental d'algorithme séquentiel atteignant les vitesses mentionnées ci-dessus est le *prédicteur par pondération exponentielle* introduit en *machine learning* par [LW94] et [Vov90]. A chaque date $t \geqslant 1$, la prévision de cet algorithme est donnée par la combinaison convexe $\widehat{a}_t \triangleq \sum_{j=1}^{K} p_{j,t}\, a_{j,t}$, où $(p_{1,1}, \ldots, p_{K,1}) = (1/K, \ldots, 1/K)$ et où, pour tout $t \geqslant 2$,

$$p_{i,t} \triangleq \frac{\exp\Big(-\eta \sum_{s=1}^{t-1} \ell(a_{i,s}, y_s)\Big)}{\sum_{j=1}^{K} \exp\Big(-\eta \sum_{s=1}^{t-1} \ell(a_{j,s}, y_s)\Big)} \; , \quad 1 \leqslant i \leqslant K \; ,$$

où $\eta > 0$ est un paramètre de l'algorithme. Le théorème suivant indique que pour une calibration judicieuse de η, le regret de cet algorithme est au plus de l'ordre de $\sqrt{T \ln K}$ ou de $\ln(K)$ selon que la fonction de perte ℓ est convexe bornée ou exp-concave. Ce résultat est prouvé au chapitre 2 aux théorèmes 2.1 et 2.2, qui sont dus respectivement à [CB99] (cf. aussi [CBL06, théorème 2.2] et [CBFH+97, CBL99]) et à [KW99].

Théorème 1.1. *Supposons que l'une des deux hypothèses suivantes soit vérifiée :*

(A1) La fonction $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ est convexe en son premier argument et est bornée à valeurs $[B_1, B_2]$, où $B_1 < B_2 \in \mathbb{R}$.

(A2) La fonction $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ est η_0-exp-concave en son premier argument pour un certain $\eta_0 > 0$, i.e., la fonction $a \mapsto e^{-\eta_0 \ell(a,y)}$ est concave sur $\mathcal{D}$ pour tout $y \in \mathcal{Y}$.

Alors, pour tout $T \in \mathbb{N}^*$ et toute suite d'avis d'experts $a_{i,t} \in \mathcal{D}$ et d'observations $y_t \in \mathcal{Y}$, le regret du prédicteur par pondération exponentielle calibré avec $\eta > 0$ vérifie :

- *Sous (A1),*

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \leqslant \frac{\ln K}{\eta} + \frac{\eta T (B_2 - B_1)^2}{8} \, .$$

  Cette borne, minimisée en $\eta = (B_2 - B_1)^{-1}\sqrt{8(\ln K)/T}$, devient $(B_2 - B_1)\sqrt{(T/2)\ln K}$.

- *Sous (A2) et lorsque $\eta \in (0, \eta_0]$,*

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \leqslant \frac{\ln K}{\eta} \, .$$

La calibration du paramètre η est un problème crucial. En effet, les valeurs suggérées pour η dépendent de quantités potentiellement inconnues au début de la tâche de prévision comme l'étendue des pertes $B_2 - B_1$ et l'horizon de prévision T sous l'hypothèse (A1). Sous l'hypothèse (A2), la valeur optimale suggérée pour η est η_0, qui est également inconnue en général — par exemple, la perte carrée $\ell : [-B, B] \times [-B, B] \to \mathbb{R}$ définie par $\ell(a, y) = (y - a)^2$ est $1/(8B^2)$-exp-concave en son premier argument ; l'amplitude des observations et des avis d'experts B est généralement inconnue.

Il est possible de calibrer *séquentiellement* η de façon totalement automatique, tout en garantissant des bornes de regret quasiment identiques (à de petits facteurs multiplicatif et additif près). Une technique générale due à [ACBG02] puis à [CBMS07] consiste à redéfinir à chaque date t les poids exponentiels $(p_{1,t}, \ldots, p_{K,t})$ à l'aide d'un paramètre η_t choisi en fonction des observations passées y_s et des avis d'experts passés $a_{i,s}$, $s = 1, \ldots, t-1$. De telles procédures de calibration séquentielle sont décrites en détail au chapitre 2 (cf. section 2.2.2). Nous en développons aux chapitres 3 et 4 pour la perte carrée (cf. sections 1.2 et 1.3).

1.1.2 Liens avec le cadre statistique classique

On décrit ci-après des liens qu'entretient la prévision de suites individuelles avec des cadres satistiques plus classiques comme le modèle de régression avec plan d'expérience (*design*) fixe ou aléatoire, où les données observées sont cette fois modélisées de façon stochastique.

Considérons le problème générique de prévision suivant. Soit $\mathcal{D}$ un espace de décision convexe, $\mathcal{Z}$ un espace d'observation[1], et $\ell : \mathcal{D} \times \mathcal{Z} \to \mathbb{R}$ une fonction de perte convexe en son premier argument. Au début de la tâche de prévision, le statisticien observe T copies indépendantes $Z_1, \ldots, Z_T$ de $Z \in \mathcal{Z}$, de loi commune inconnue. Le but du statisticien est de prévoir l'observation suivante[2] $Z_{T+1} \sim Z$ presque aussi bien que le meilleur élément (constant) d'un ensemble $\Theta \subset \mathcal{D}$. Plus précisément, il s'agit de construire une décision $\widehat{a}_T \in \mathcal{D}$ mesurable en l'échantillon $(Z_1, \ldots, Z_T)$

[1] On utilise la notation $\mathcal{Z}$ au lieu de $\mathcal{Y}$ pour éviter toute ambiguité avec le modèle de régression avec *design* aléatoire, où Y_t désigne uniquement la sortie alors que le statisticien observe le couple $Z_t = (X_t, Y_t) \in \mathcal{X} \times \mathbb{R}$. Dans ce cadre, $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$.

[2] La variable aléatoire $Z_{T+1} \in \mathcal{Z}$ est indépendente de $(Z_1, \ldots, Z_T)$ et de même loi que Z.

de sorte à minimiser une quantité appelée *excès de risque en espérance*[3]

$$\mathbb{E}\big[\ell\big(\widehat{a}_T, Z\big)\big] - \inf_{a\in\Theta} \mathbb{E}\big[\ell\big(a, Z\big)\big] \ ,$$

où l'espérance de gauche est prise par rapport à $(Z_1, \dots, Z_T)$ et Z. Un exemple classique de telle tâche de prévision est donnée par le modèle de régression avec *design* aléatoire.

Exemple 1.1 (Agrégation dans le modèle de régression avec *design* aléatoire)**.**
Soit $(\mathcal{X}, \mathcal{B})$ un espace mesurable et Θ un ensemble de fonctions mesurables de $\mathcal{X}$ vers $\mathbb{R}$. Le statisticien observe T copies indépendantes $(X_1, Y_1), \dots, (X_T, Y_T)$ d'un couple aléatoire $(X, Y) \in \mathcal{X} \times \mathbb{R}$ de loi inconnue, avec $\mathbb{E}[Y^2] < \infty$. Dans ce cadre, un objectif de prévision consiste à estimer la fonction de régression $f : x \in \mathcal{X} \mapsto \mathbb{E}[Y|X = x]$ presque aussi bien que le meilleur élément de Θ. Plus précisément, il s'agit de construire un estimateur $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ à partir de l'échantillon $(X_1, Y_1), \dots, (X_T, Y_T)$ de sorte à minimiser l'excès de risque en espérance

$$\mathbb{E}\Big[\big(f(X) - \widehat{f}_T(X)\big)^2\Big] - \inf_{g\in\Theta} \mathbb{E}\Big[\big(f(X) - g(X)\big)^2\Big] \ ,$$

où les espérances sont prises par rapport à $(X_1, Y_1), \dots, (X_T, Y_T)$ et X. Or, par de simples manipulations (en développant les carrés et en conditionnant par $(X_1, Y_1), \dots, (X_T, Y_T), X$), l'excès de risque précédent est égal à

$$\mathbb{E}\Big[\big(Y - \widehat{f}_T(X)\big)^2\Big] - \inf_{g\in\Theta} \mathbb{E}\Big[\big(Y - g(X)\big)^2\Big] \ .$$

Par conséquent, ce problème d'agrégation correspond au cadre décrit ci-dessus avec $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$, avec $\mathcal{D}$ égal à l'ensemble des fonctions mesurables de $\mathcal{X}$ vers $\mathbb{R}$, et avec $\ell : \mathcal{D} \times \mathcal{Z} \to \mathbb{R}$ définie par $\ell\big(g, (x, y)\big) = \big(y - g(x)\big)^2$.

La tâche de prévision décrite précédemment n'est pas séquentielle – on la qualifie de *batch* en anglais car toutes les observations sont disponibles d'emblée. Cela n'interdit pas en revanche de traiter l'échantillon de façon séquentielle. On rappelle ci-dessous une technique standard qui permet de convertir un algorithme séquentiel encourant un faible regret pour des suites individuelles en une méthode stochastique encourant un petit excès de risque en espérance pour des suites i.i.d..

Soit $(\widetilde{a}_t)_{t\geqslant 1}$ un algorithme séquentiel, i.e., dans ce cadre, une suite de fonctions mesurables $\widetilde{a}_t : \mathcal{Z}^{t-1} \to \mathcal{D}$ ($\widetilde{a}_1$ est déterministe). L'échantillon $Z_{1:T} \triangleq (Z_1, \dots, Z_T)$ est traité de façon séquentielle de la date 1 à la date T : l'algorithme $(\widetilde{a}_t)_{t\geqslant 1}$ produit séquentiellement les décisions $\widetilde{a}_t(Z_{1:t-1}) \in \mathcal{D}$ mesurables en $Z_{1:t-1} \triangleq (Z_1, \dots, Z_{t-1})$, $t = 1, \dots, T$. Le résultat suivant est dû à [CBCG04] (cf. aussi [Lit89]) ; nous le reprouvons au chapitre 2, proposition 2.5. Il indique qu'une façon simple de convertir l'algorithme séquentiel $(\widetilde{a}_t)_{t\geqslant 1}$ en méthode stochastique est de considérer la moyenne

$$\widehat{a}_T(Z_{1:T}) = \frac{1}{T}\sum_{t=1}^{T} \widetilde{a}_t(Z_{1:t-1}) \ . \tag{1.1}$$

Proposition 1.1 (Conversion *online to batch*, cf. proposition 2.5)**.**
Soit $\mathcal{D}$ un espace de décision convexe, $\mathcal{Z}$ un espace d'observation, et $\ell : \mathcal{D} \times \mathcal{Z} \to \mathbb{R}$ une fonction de perte convexe en son premier argument. Soit $(\widetilde{a}_t)_{t\geqslant 1}$ un algorithme séquentiel et $(R_T)_{T\geqslant 1}$ une

[3] Le *risque en espérance* de la procédure $\widehat{a}_T$ correspond quant à lui à la quantité $\mathbb{E}\big[\ell\big(\widehat{a}_T, Z\big)\big]$.

suite de réels telle que, pour tout $T \geqslant 1$ *et tous* $z_1, \dots, z_T \in \mathcal{Z}$*,*

$$\sum_{t=1}^{T} \ell(\widetilde{a}_t, z_t) - \inf_{a \in \Theta} \sum_{t=1}^{T} \ell(a, z_t) \leqslant R_T \, .$$

Alors la conversion (1.1) *appliquée à l'algorithme* $(\widetilde{a}_t)_{t \geqslant 1}$ *donne une procédure* $\widehat{a}_T$ *telle que, pour tout échantillon i.i.d.* $(Z_1, \dots, Z_T) \in \mathcal{Z}^T$*,*

$$\mathbb{E}\big[\ell\big(\widehat{a}_T, Z\big)\big] - \inf_{a \in \Theta} \mathbb{E}\big[\ell\big(a, Z\big)\big] \leqslant \frac{R_T}{T} \, ,$$

où les espérances sont prises par rapport à $(Z_1, \dots, Z_T, Z)$*, avec* $Z \in \mathcal{Z}$ *une variable aléatoire indépendante de* $(Z_1, \dots, Z_T)$ *et de même loi que* Z_1*.*

Sans surprise, la proposition précédente montre que tout algorithme séquentiel qui bénéficie de garanties déterministes peut être converti en une méthode statistique bénéficiant de garanties en espérance. Des bornes avec grande probabilité ont également été obtenues par [CBG08] dans un cadre non nécessairement convexe et par [Zha05, KT09] dans un cadre "très convexe" (perte carrée ou pertes fortement convexes). En régression, la conversion précédente est adaptée au modèle de régression avec *design* aléatoire (cf. exemple 1.1). Le cas du *design* fixe peut, dans une certaine mesure, être traité avec des techniques similaires ; voir la section 3.4.2.

La conversion précédente – qualifiée de *online to batch* en anglais – établit un lien de la prévision de suites individuelles vers le cadre statistique classique. Il s'avère que les apports de ces deux domaines sont en fait réciproques.

- Des méthodes statistiques classiques, conçues et étudiées sous des hypothèses stochastiques, peuvent aussi, moyennant quelques adaptations, s'avérer performantes dans un cadre de suites individuelles. C'est le cas de la méthode de régression *ridge*[4] de [HK70], initialement analysée dans le modèle de régression avec *design* fixe, qui a ensuite été adaptée et étudiée pour des suites individuelles par [AW01] et [Vov01]. Les algorithmes de descente de gradient stochastique bénéficient également de garanties déterministes comme l'ont montré, par ex., [CBLW96, Zin03]. C'est le cas également du prédicteur par pondération exponentielle, qui a été étudié parallèlement en *machine learning* et en statistique[5], des travaux fondateurs étant respectivement [LW94, Vov90] et [Cat99, Yan00, Yan01]. Dans tous les cas, analyser des méthodes statistiques dans un cadre de suites individuelles permet d'en comprendre le cœur déterministe et d'en évaluer la robustesse (i.e., de jauger à quel point les hypothèses stochastiques sont nécessaires). Au chapitre 3, notre algorithme séquentiel SeqSEW est inspiré de la méthode statistique *Sparse Exponential Weighting* [DT08, DT11]. Ainsi, notre analyse déterministe de l'algorithme SeqSEW indique que la méthode de [DT11] fonctionne essentiellement pour des raisons déterministes.

- D'après la remarque précédente, la théorie de la prévision de suites individuelles hérite d'idées fructueuses venant du cadre statistique classique, puisque ce dernier lui fournit de sérieux candidats pour la conception de nouveaux algorithmes séquentiels.

[4] Rappelons qu'il s'agit d'une méthode de régression des moindres carrés régularisés par la norme ℓ^2.

[5] Dans le modèle de régression avec *design* aléatoire, la méthode statistique résultant d'un prédicteur par pondération exponentielle via la conversion (1.1) est qualifiée de *progressive mixture rule* en anglais ; cf. [Cat04].

- En retour, les algorithmes séquentiels nouvellement conçus peuvent être rapatriés dans le cadre statistique classique via la conversion *online to batch* (1.1). Cela permet de construire des méthodes statistiques calibrées automatiquement en fonction des données (par des techniques de suites individuelles) et qui sont adaptatives. Nous illustrons cet intérêt au chapitre 3 où l'on déduit des bornes de risque en *design* aléatoire similaires à [DT11], mais qui sont adaptatives en la variance inconnue du bruit (à un facteur logarithmique près) quand ce dernier est gaussien.

Nous venons d'évoquer des liens algorithmiques entre la prévision de suites individuelles et le cadre statistique classique. Cela induit notamment des similarités au niveau des techniques de preuve. Par exemple, la formule de dualité pour la divergence de Kullback-Leibler rappelée en annexe A.1 est un outil clé tant dans le cadre déterministe (chapitres 3 et 4) que dans le cadre stochastique (chapitre 6, où l'on considère le modèle de régression avec *design* fixe). Au chapitre 4, nous adaptons également un argument statistique classique connu sous le nom d'*argument à la Maurey*, qui permet de déterminer la qualité de l'approximation d'une discrétisation adéquate du simplexe en dimension quelconque. Comme en témoigne la preuve du théorème 4.2, cet argument s'adapte directement au cadre déterministe.

D'autres similarités dans les techniques de preuve apparaissent aussi pour l'obtention de bornes inférieures. Une façon d'obtenir des bornes inférieures non asymptotiques en suites individuelles repose en effet sur l'utilisation d'outils de théorie de l'information comme le lemme de Fano ou l'inégalité de Pinsker (cf. annexe A.7), comme en statistique classique. La vitesse minimax du regret externe peut ainsi être obtenue via une variante du lemme de Fano ; l'analyse correspondante est due à [ACBFS02, CBLS05] et rappelée en section 2.3.2. En s'inspirant de ces techniques et de [Sto05, théorème 3.3], nous obtenons au chapitre 5 une borne inférieure sur le regret *swap* via l'inégalité de Pinsker.

Enfin, des techniques stochastiques peuvent être utilisées à des fins purement déterministes. C'est le cas de la randomisation, que nous exploitons pour l'argument à la Maurey mentionné ci-dessus ou pour obtenir des bornes inférieures (cf. section 2.1.3 pour plus de détails). C'est le cas aussi des inégalités de concentration, telle l'inégalité de Hoeffding ou l'inégalité de Bernstein, qui permettent de déduire des bornes de regret pour des suites *déterministes* — cf. section 2.2.1. Au chapitre 5, nous développons également une technique stochastique qui permet de majorer le regret minimax pour des suites individuelles (relativement à diverses formes de regret, par ex., externe, interne, *swap*). Cette technique repose en partie sur l'utilisation d'une inégalité élémentaire de concentration de martingales, l'inégalité de Hoeffding-Azuma (cf. annexe A.5).

Les liens entre suites individuelles et cadre statistique classique évoqués précédemment sont partiellement représentés en figure 1.1. Nous détaillons ci-après les contributions principales de cette thèse, qui correspondent aux chapitres 3 à 6.

1.2 Bornes de parcimonie en régression linéaire séquentielle

Au cours de la dernière décennie, le phénomène de parcimonie – ou *sparsité* – a fait l'objet de nombreux travaux dans le cadre statistique classique. Parmi les outils introduits à cet effet, la notion d'*inégalité oracle de sparsité* – ou *sparsity oracle inequality* en anglais – joue un rôle fondamental. En régression linéaire, de telles bornes impliquent que la tâche consistant à prévoir

presque aussi bien qu'un vecteur inconnu de grande dimension est statistiquement faisable pourvu que ce vecteur ait peu de coordonnées non nulles.

Au chapitre 3, on introduit un équivalent séquentiel déterministe de la notion d'inégalité oracle de sparsité. Nous prouvons de telles bornes pour un algorithme séquentiel appelé *SeqSEW* qui procède par pondération exponentielle et par troncature dépendante des données. Dans un second temps seulement, on applique une version totalement automatique de cet algorithme au cas particulier de suites i.i.d.. Les bornes de risque obtenues sont similaires à celles de [DT11] mais répondent à deux questions soulevées par les auteurs. En particulier, nos bornes sont adaptatives en la variance inconnue du bruit (à un facteur logarithmique près) si ce dernier est gaussien. Nous traitons aussi le cas du *design* fixe comme dans [DT08].

Les contributions principales du chapitre 3 sont détaillées ci-après.

1.2.1 Cadre et enjeux

Contexte : régression linéaire séquentielle pour des suites individuelles

Le cadre principal du chapitre 3 est celui de la *régression linéaire séquentielle pour des suites individuelles*. Il s'agit d'un cas particulier du problème de la prévision avec avis d'experts décrit en figure 1.2. Un statisticien doit prévoir de façon séquentielle, à chaque tour $t = 1, 2, \ldots$, la valeur $y_t \in \mathbb{R}$ d'une suite inconnue d'observations en fonction d'une valeur d'entrée $x_t \in \mathcal{X}$ et de prédicteurs de base $\varphi_j : \mathcal{X} \to \mathbb{R}$, $1 \leqslant j \leqslant d$, à partir desquels il formule sa propre prévision $\widehat{y}_t \in \mathbb{R}$ (la famille $(\varphi_j)_{1 \leqslant j \leqslant d}$ est qualifiée de *dictionnaire*). La qualité des prévisions est évaluée avec la perte carrée. L'objectif du statisticien est de prévoir presque aussi bien que le meilleur prédicteur linéaire $\boldsymbol{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^{d} u_j \varphi_j$, où $\boldsymbol{u} \in \mathbb{R}^d$, i.e., de satisfaire, uniformément sur toutes les suites individuelles $(x_t, y_t)_{1 \leqslant t \leqslant T}$, une borne de regret de la forme

$$\sum_{t=1}^{T} \bigl(y_t - \widehat{y}_t\bigr)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \bigl(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\bigr)^2 + \Delta_{T,d}(\boldsymbol{u}) \right\} ,$$

pour un terme de regret $\Delta_{T,d}(\boldsymbol{u})$ aussi petit que possible et, en particulier, sous-linéaire en T. (Par soucis de clarté, on omet les dépendances de $\Delta_{T,d}(\boldsymbol{u})$ en les amplitudes $\max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty$ et $\max_{1 \leqslant t \leqslant T} |y_t|$.)

Hypothèse de parcimonie

Dans le cadre décrit ci-dessus, une variante[6] de l'algorithme séquentiel *ridge* étudiée par [AW01] et [Vov01] assure, lorsqu'elle est calibrée illégalement comme suggéré en section 2.4.2, un regret d'ordre au plus $d \ln T$. Quand la dimension ambiante d est bien plus grande que le nombre de tours de prévision T, cette dernière borne de regret est bien supérieure à T et est donc en quelque sorte triviale. Puisque la borne $d \ln T$ est optimale en un certain sens (cf. [Vov01, théorème 2]), des hypothèses supplémentaires sont nécessaires pour garantir des performances théoriques intéressantes.

Une hypothèse naturelle, qui a déjà été maintes fois étudiée dans le cadre stochastique, est qu'il existe une combinaison linéaire parcimonieuse $\boldsymbol{u}^*$ (*sparse* en anglais, i.e., avec $s \ll T/(\ln T)$ coordonnées non nulles) dont la perte cumulée est petite. Si le statisticien connaissait à l'avance le support $J(\boldsymbol{u}^*) \triangleq \{j : u_j^* \neq 0\}$ de $\boldsymbol{u}^*$, il pourrait appliquer le même algorithme de prévision

[6] Ce prédicteur séquentiel est rappelé au chapitre 2 ; cf. (2.26) en section 2.4.2.

séquentielle que précédemment mais seulement au sous-espace vectoriel de dimension s donné par $\{\boldsymbol{u} \in \mathbb{R}^d : \forall j \notin J(\boldsymbol{u}^*), u_j = 0\}$. Le regret de cet "oracle" serait alors au plus de l'ordre de $s \ln T$ et donc sous-linéaire en T. Sous cette hypothèse de parcimonie, un regret sous-linéaire semble donc possible, même si, bien sûr, la borne de regret $s \ln T$ peut seulement être utilisée comme une borne idéale de référence (puisque le support de $\boldsymbol{u}^*$ est inconnu).

Au chapitre 3, on montre qu'il est possible d'atteindre une borne de regret proportionnelle à s (à un facteur logarithmique près). On prouve ainsi en corollaire 3.1 (cf. proposition 1.2 ci-dessous) et ses raffinements (cf., par ex., proposition 1.3 ci-dessous) des bornes de regret de la forme

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + \big(\|\boldsymbol{u}\|_0 + 1\big)\, g_{T,d}\big(\|\boldsymbol{u}\|_1, \|\boldsymbol{\varphi}\|_\infty\big) \right\}, \qquad (1.2)$$

où $\|\boldsymbol{u}\|_0$ désigne le nombre de coordonnées non nulles de $\boldsymbol{u}$ et où g est croissante mais croît au plus logarithmiquement en T, d, $\|\boldsymbol{u}\|_1 \triangleq \sum_{j=1}^{d} |u_j|$, et $\|\boldsymbol{\varphi}\|_\infty \triangleq \sup_{x \in \mathcal{X}} \max_{1 \leqslant j \leqslant d} |\varphi_j(x)|$. Nous appellerons *bornes de regret de sparsité* – ou *bornes de parcimonie* – les bornes de regret de la forme précédente.

Travaux connexes dans les cadres stochastique et déterministe

La borne de regret (1.2) peut être vue comme un équivalent séquentiel déterministe des *inégalités oracle de sparsité* introduites dans le cadre statistique classique au cours de la dernière décennie. Un exemple typique de telles bornes de risque est

$$R(\widehat{\boldsymbol{u}}_T) \leqslant (1+a) \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ R(\boldsymbol{u}) + C(a)\, \frac{\|\boldsymbol{u}\|_0 \ln d + 1}{T} \right\} \qquad (1.3)$$

en espérance ou avec grande probabilité, où $R(\boldsymbol{u})$ désigne le risque L^2 de $\boldsymbol{u} \cdot \boldsymbol{\varphi}$ si le *design* est aléatoire (i.e., $R(\boldsymbol{u}) = \mathbb{E}\big[\big(f(X) - \boldsymbol{u} \cdot \boldsymbol{\varphi}(X)\big)^2\big]$) ou le risque empirique de $\boldsymbol{u} \cdot \boldsymbol{\varphi}$ si le *design* est fixe (i.e., $R(\boldsymbol{u}) = T^{-1} \sum_{t=1}^{T} \big(f(x_t) - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2$ sur le *design* $(x_1, \ldots, x_T)$). Ainsi, les inégalités oracle de sparsité expriment un compromis entre le risque $R(\boldsymbol{u})$ et le nombre de coordonnées non-nulles $\|\boldsymbol{u}\|_0$ de tout vecteur $\boldsymbol{u} \in \mathbb{R}^d$. De telles bornes ont été obtenues par [BM01a] via des arguments de sélection de modèles et ont ensuite été développées, entre autres, par [BM07a, BTW07a] dans le modèle de régression avec *design* fixe et par [BTW04] dans le modèle de régression avec *design* aléatoire. Une introduction plus détaillée avec de plus amples références est proposée au chapitre 2 (section 2.6).

Mentionnons néanmoins que, récemment, depuis les travaux de [DT08], des inégalités oracle de sparsité avec constante 1 devant l'infimum[7] ont été prouvées sans presque aucune hypothèse sur le dictionnaire $(\varphi_j)_j$, et pour des méthodes pouvant être approchées numériquement à un coût algorithmique raisonnable pour de grandes valeurs de la dimension ambiante d. Ces méthodes procèdent par pondération exponentielle ; cf. [DT07, DT08, RT11, AL11] pour le modèle de régression avec *design* fixe et [DT11, AL11] pour le modèle de régression avec *design* aléatoire.

Quant au cadre séquentiel déterministe, à notre connaissance, les propositions 1.2 et 1.3 ci-dessous (cf. aussi théorème 3.1 au chapitre 3) fournissent les premiers exemples de borne de regret

[7] Un exemple de telles bornes est donné par (1.3) avec $a = 0$. Ces bornes permettent de majorer les excès de risque $R(\widehat{\boldsymbol{u}}_T) - \inf_{\{\|\boldsymbol{u}\|_0 \leqslant s\}} R(\boldsymbol{u})$ pour tout $s \in \{0, \ldots, d\}$.

de sparsité au sens de (1.2). De récents travaux [LLZ09, SST09, Xia10, DSSST10] en optimisation convexe séquentielle ont certes abordé la question de la sparsité, mais sous un tout autre angle. Dans le cas de la régularisation ℓ^1 sous la perte carrée, ces travaux proposent des algorithmes qui prédisent comme une combinaison linéaire parcimonieuse $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{\varphi}(x_t)$ des prévisions de base (i.e., $\|\widehat{\boldsymbol{u}}_t\|_0$ est petit), alors que de telles garanties ne semblent pas pouvoir être montrées pour notre algorithme SeqSEW. En revanche, ces travaux prouvent des bornes sur le regret ℓ^1-régularisé de la forme

$$\sum_{t=1}^{T}\Big((y_t - \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t)^2 + \lambda \left\|\widehat{\boldsymbol{u}}_t\right\|_1\Big) \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\Big((y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 + \lambda \left\|\boldsymbol{u}\right\|_1\Big) + \widetilde{\Delta}_{T,d}(\boldsymbol{u})\right\} ,$$

pour un terme de regret $\widetilde{\Delta}_{T,d}(\boldsymbol{u})$ qui croît beaucoup plus rapidement (comme une puissance et non logarithmiquement) en la dimension ambiante d, en la norme $\|\boldsymbol{u}\|_1$ ou en T. Les bornes prouvées pour ces algorithmes sont donc sous-optimales dans le cadre qui nous intéresse ici (prévision sur des boules ℓ^0 de petit diamètre). Cela contraste avec les bornes de regret de la forme (1.2) que vérifie, par exemple, notre algorithme SeqSEW.

On reprend ci-après à grands traits les algorithmes et résultats principaux du chapitre 3, d'abord dans le cadre déterministe (section 1.2.2) puis dans le cadre stochastique (section 1.2.3).

1.2.2 Bornes de sparsité en suites individuelles

Pour simplifier l'analyse, on suppose d'abord que, au début du jeu, le statisticien a accès au nombre T de tours de prévision, à une borne B_y sur l'amplitude des observations $|y_1|, \dots, |y_T|$ et à une borne B_Φ sur la trace de la matrice de Gram empirique, i.e.,

$$y_1, \dots, y_T \in [-B_y, B_y] \qquad \text{et} \qquad \sum_{j=1}^{d}\sum_{t=1}^{T} \varphi_j^2(x_t) \leqslant B_\Phi \, .$$

La première version de notre algorithme est définie en figure 1.3. Nous l'appelons *SeqSEW* puisqu'il s'agit d'une variante adaptée aux suites individuelles de l'algorithme *Sparse Exponential Weighting* introduit dans le cadre statistique classique par [DT07, DT08].

En utilisant un lemme PAC-Bayésien déterministe dû à [Aud09] et la forme particulière du prior π_τ (à queue lourde), on montre que cet algorithme vérifie la borne de regret suivante.

Proposition 1.2 (cf. corollaire 3.1)**.** *Supposons que, pour des constantes connues $B_y, B_\Phi > 0$, les $(x_1, y_1), \dots, (x_T, y_T)$ sont tels que $y_1, \dots, y_T \in [-B_y, B_y]$ et $\sum_{j=1}^{d}\sum_{t=1}^{T} \varphi_j^2(x_t) \leqslant B_\Phi$. Alors, l'algorithme $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$ calibré avec $B = B_y$, $\eta = 1/(8B_y^2)$ et $\tau = \sqrt{16B_y^2/B_\Phi}$ vérifie*

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + 32\, B_y^2 \left\|\boldsymbol{u}\right\|_0 \ln\left(1 + \frac{\sqrt{B_\Phi}\left\|\boldsymbol{u}\right\|_1}{4\, B_y \left\|\boldsymbol{u}\right\|_0}\right)\right\} + 16B_y^2 \, .$$

Remarquons que, si $\|\boldsymbol{\varphi}\|_\infty \triangleq \sup_{x\in\mathcal{X}} \max_{1\leqslant j\leqslant d} |\varphi_j(x)|$ est fini, alors la proposition précédente fournit une *borne de regret de sparsité* au sens de (1.2). En effet, dans ce cas, on peut prendre $B_\Phi = d\,T\,\|\boldsymbol{\varphi}\|_\infty^2$, ce qui donne une borne de regret proportionnelle à $\|\boldsymbol{u}\|_0$ et qui croît logarithmiquement en d, T, $\|\boldsymbol{u}\|_1$ et $\|\boldsymbol{\varphi}\|_\infty$.

Paramètres : seuil $B > 0$, température inverse $\eta > 0$ et résolution $\tau > 0$ à laquelle on associe la loi a priori π_τ sur $\mathbb{R}^d$ défini par

$$\pi_\tau(\mathrm{d}\boldsymbol{u}) \triangleq \prod_{j=1}^{d} \frac{(3/\tau)\,\mathrm{d}u_j}{2\big(1+|u_j|/\tau\big)^4} \,.$$

Initialisation : $p_1 \triangleq \pi_\tau$.

A chaque tour de prévision $t \geqslant 1$,

1. Recevoir la donnée x_t et prévoir $\widehat{y}_t \triangleq \displaystyle\int_{\mathbb{R}^d} \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_B \, p_t(\mathrm{d}\boldsymbol{u})$,
   où $[x]_B \triangleq \max\big\{-B, \min\{B,x\}\big\}$;
2. Recevoir l'observation y_t et calculer la probabilité a posteriori p_{t+1} sur $\mathbb{R}^d$ via l'expression (W_{t+1} est une constante de renormalisation)

$$p_{t+1}(\mathrm{d}\boldsymbol{u}) \triangleq \frac{1}{W_{t+1}} \exp\left(-\eta \sum_{s=1}^{t} \Big(y_s - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_s)\big]_B\Big)^2\right) \pi_\tau(\mathrm{d}\boldsymbol{u}) \,. \tag{1.4}$$

FIGURE 1.3 – Définition de l'algorithme $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$.

Si le statisticien n'a pas accès à une borne a priori B_y sur les observations, il peut s'adapter séquentiellement à cette borne inconnue en tronquant les prévisions $\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)$ de façon dépendante des données. L'algorithme plus sophistiqué SeqSEW_τ^* produit ainsi les prévisions

$$\widehat{y}_t \triangleq \int_{\mathbb{R}^d} \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t} \, p_t(\mathrm{d}\boldsymbol{u}) \,, \qquad \text{où} \quad B_t \triangleq \inf\left(\Big\{\sqrt{2^k}; k\in\mathbb{Z}\Big\} \cap \Big[\max_{1\leqslant s\leqslant t-1} |y_s|, +\infty\Big)\right) \,,$$

et où la probabilité a posteriori p_t sur $\mathbb{R}^d$ est définie comme précédemment mais en remplaçant la température η par $\eta_t \triangleq 1/(8B_t^2)$ et le seuil B par B_s pour chaque indice s de la somme dans (1.4). Une analyse PAC-Bayésienne plus approfondie (cf. lemme 3.2) conduit à la borne suivante.

Proposition 1.3 (cf. proposition 3.2)**.** *Pour tout $\tau > 0$, l'algorithme SeqSEW_τ^* précédent vérifie*

$$\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t-\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + 64\left(\max_{1\leqslant t\leqslant T} y_t^2\right)\|\boldsymbol{u}\|_0 \ln\left(1+\frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0\,\tau}\right)\right\}$$
$$+\,\tau^2\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t) \,+ 32\max_{1\leqslant t\leqslant T} y_t^2 \,.$$

Au vu de la dernière proposition, la calibration de τ requiert encore la connaissance a priori d'une borne B_Φ sur $\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t)$. Cela peut être évité au moyen d'une technique classique appelée *doubling trick*, qui donne lieu à une borne similaire (cf. théorème 3.1 et corollaire 3.4) à un facteur logarithmique près, mais pour un algorithme cette fois totalement automatique.

1.2.3 Adaptativité en la variance pour des données i.i.d.

Dans cette sous-section, on applique l'algorithme SeqSEW au modèle de régression avec *design* aléatoire (le cas du *design* fixe peut, dans une moindre mesure, être traité avec des techniques simi-

laires, cf. section 3.4.2). Le statisticien a accès à T copies indépendantes $(X_1, Y_1), \dots, (X_T, Y_T)$ de $(X, Y) \in \mathcal{X} \times \mathbb{R}$ de loi inconnue. On suppose que $\mathbb{E}[Y^2] < \infty$; l'objectif du statisticien est d'estimer la fonction de régression $f : \mathcal{X} \to \mathbb{R}$ définie par $f(x) \triangleq \mathbb{E}[Y|X = x]$ pour tout $x \in \mathcal{X}$. On pose aussi $\|h\|_{L^2} \triangleq (\mathbb{E}[h(X)^2])^{1/2}$ pour toute fonction mesurable $h : \mathcal{X} \to \mathbb{R}$.

On emploie la conversion *online to batch* décrite en proposition 1.1. L'échantillon $(X_t, Y_t)_{t=1}^T$ est traité de façon séquentielle, en appliquant l'algorithme SeqSEW^*_τ de la date 1 à la date T avec $\tau = 1/\sqrt{dT}$. L'estimateur $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ retenu est défini par

$$\widehat{f}_T(x) \triangleq \frac{1}{T} \sum_{t=1}^{T} \int_{\mathbb{R}^d} \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x)\big]_{B_t} p_t(\mathrm{d}\boldsymbol{u}) \, .$$

Contrairement à de nombreux travaux en statistique comme [Cat04, BN08, DT11], l'estimateur $\widehat{f}_T$ est totalement automatique : il ne dépend d'aucune connaissance a priori sur la loi inconnue de (X, Y) telle que la variance du bruit $\mathbb{E}\big[(Y - f(X))^2\big]$ ou les normes $\|\varphi_j\|_\infty$ ou $\|f - \varphi_j\|_\infty$ (nous ne supposons d'ailleurs pas que ces dernières quantités sont finies). Nous prouvons au chapitre 3 une borne de risque pour $\widehat{f}_T$ valable sous de faibles hypothèses sur la loi de Y (cf. théorème 3.2 et corollaire 3.5). Nous mentionnons seulement le cas sous-gaussien ci-dessous.

Proposition 1.4 (cf. corollaire 3.6)**.** *Supposons que $\|f\|_\infty < +\infty$ et que, pour une constante $\sigma^2 > 0$ inconnue, $\mathbb{E}\Big[e^{\lambda(Y - f(X))} \,\Big|\, X\Big] \leqslant e^{\lambda^2\sigma^2/2}$ p.s.. Alors, pour tout $T \geqslant 2$,*

$$\begin{aligned}
&\mathbb{E}\left[\left\|f - \widehat{f}_T\right\|_{L^2}^2\right] \\
&\quad\leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \|f - \boldsymbol{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 + 128\Big(\|f\|_\infty^2 + 2\sigma^2 \ln(2eT)\Big) \frac{\|\boldsymbol{u}\|_0}{T} \ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right) \right\} \\
&\qquad + \frac{1}{dT} \sum_{j=1}^{d} \|\varphi_j\|_{L^2}^2 + \frac{64}{T}\Big(\|f\|_\infty^2 + 2\sigma^2 \ln(2eT)\Big) \, .
\end{aligned}$$

Cette borne est comparable à la proposition 1 prouvée par Dalalyan et Tsybakov [DT11]. Elle vaut néanmoins sur $\mathbb{R}^d$ tout entier au lieu de boules ℓ^1 de rayons finis, ce qui résout une question laissée ouverte dans [DT11, section 4.2]. Par ailleurs, notre algorithme ne requiert pas la connaissance a priori du facteur de variance $\sigma^2 > 0$ du bruit, ce qui résout une seconde question soulevée dans [DT11, section 5.1, remarque 6].

1.3 Régression linéaire séquentielle optimale et adaptative sur des boules ℓ^1

Au chapitre 4, nous abordons un problème proche de celui du chapitre 3 : la régression linéaire séquentielle sur des boules ℓ^1. Nous en détaillons ci-après le cadre et nos principales contributions.

1.3.1 Cadre et objectif de prévision

On considère la formulation suivante[8] du problème de régression linéaire séquentielle pour des suites individuelles (cf. section 2.4 pour une introduction à ce cadre). Au début de la tâche de

[8] La description du cadre de régression linéaire séquentielle diffère très légèrement de celle de la section 1.2 : le vecteur des prévisions de base à l'instant t est $\boldsymbol{x}_t$ alors qu'il s'agissait de $\boldsymbol{\varphi}(x_t)$ en section 1.2. Les deux descriptions

prévision, l'environnement choisit une suite d'observations $(y_t)_{t\geqslant 1}$ dans $\mathbb{R}$ et une suite de vecteurs de prévisions de base $(\boldsymbol{x}_t)_{t\geqslant 1}$ dans $\mathbb{R}^d$, toutes deux initialement cachées au statisticien. A chaque date $t \in \mathbb{N}^* = \{1, 2, \ldots\}$, l'environnement révèle le vecteur $\boldsymbol{x}_t \in \mathbb{R}^d$, le statisticien formule ensuite sa propre prévision $\widehat{y}_t \in \mathbb{R}$, puis l'environnement révèle l'observation $y_t \in \mathbb{R}$; le statisticien encourt alors la perte carrée $(y_t - \widehat{y}_t)^2$.

Objectif de prévision

Etant donné un rayon $U > 0$ et un horizon de prévision $T \geqslant 1$, l'objectif du statisticien est ici de prévoir presque aussi bien que le meilleur prédicteur linéaire $\boldsymbol{x} \in \mathbb{R}^d \mapsto \boldsymbol{u} \cdot \boldsymbol{x} \triangleq \sum_{j=1}^d u_j x_j$ tel que $\|\boldsymbol{u}\|_1 \triangleq \sum_{j=1}^d |u_j| \leqslant U$. Autrement dit, il s'agit de minimiser le regret sur la boule ℓ^1 $B_1(U) \triangleq \left\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U\right\}$ défini par

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \min_{\boldsymbol{u}\in B_1(U)} \left\{\sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2\right\}.$$

Cet objectif de prévision généralise la tâche d'agrégation convexe ; il est d'ailleurs possible d'être compétitif vis-à-vis de toutes les boules $B_1(U)$, simultanément pour tout $U > 0$ (cf. fin de la section 1.3.3). Cette tâche peut s'avérer utile quand les observations y_t sont correctement approchées par une combinaison linéaire $\boldsymbol{u} \in \mathbb{R}^d$ des prévisions de base $x_{j,t}$, $j = 1, \ldots, d$, avec une petite norme $\|\boldsymbol{u}\|_1$ — ce qui peut être le cas, par exemple, si $\boldsymbol{u}$ est approximativement parcimonieuse. Notons enfin que la dimension ambiante d peut être petite ou grande relativement à l'horizon de prévision T : on considère tous les cas.

Dans la suite, on présente des algorithmes et des bornes sur leur regret qui valent uniformément en toutes les individuelles[9] $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T}$ telles que $\|\boldsymbol{x}_t\|_\infty \leqslant X$ and $|y_t| \leqslant Y$ pour tout $t = 1, \ldots, T$, où $X, Y > 0$. Ces bornes de regret dépendent de quatre quantités importantes : U, X, Y et T, lequelles peuvent être connues ou inconnues du statisticien.

On présente ci-après les principales contributions du chapitre 4 : la première concerne la détermination de la vitesse minimax du regret (section 1.3.2), la seconde a trait à l'adaptation en les quantités X, Y, T et U lorsqu'elles sont inconnues (section 1.3.3), et la troisième consiste en un raffinement des bornes de regret via une technique appelée *lipschitzification des pertes* (section 1.3.4).

1.3.2 Vitesse optimale

Notre première contribution consiste en la détermination de l'ordre de grandeur du regret minimax sur $B_1(U)$ pour des données bornées par X et Y défini par

$$\inf_{(\widehat{y}_t)_{t\geqslant 1}} \sup_{\substack{\|\boldsymbol{x}_1\|_\infty,\ldots,\|\boldsymbol{x}_T\|_\infty\leqslant X \\ |y_1|,\ldots,|y_T|\leqslant Y}} \left\{\sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \min_{\|\boldsymbol{u}\|_1\leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2\right\}, \tag{1.5}$$

sont en fait équivalentes ; celle choisie ici est plus classique en suites individuelles, alors que celle avec φ était plus adaptée pour le passage au *design* aléatoire.

[9] En fait, nos résultats sont aussi valables quand $(\boldsymbol{x}_t, y_t)_{t\geqslant 1}$ est engendrée par un environnement antagoniste puisque nous ne considérons que des algorithmes déterministes. Cf. section 2.3.1 pour de plus amples détails.

où le supremum est pris sur toutes les suites individuelles $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_T, y_T) \in \mathbb{R}^d \times \mathbb{R}$ telles que $\|\boldsymbol{x}_t\|_\infty \leqslant X$ et $|y_t| \leqslant Y$ pour tout $t = 1, \dots, T$, et où l'infimum est pris sur tous les prédicteurs séquentiels $(\widehat{y}_t)_{t \geqslant 1}$, i.e., toutes les suites de fonctions $\widehat{y}_t : (\mathbb{R}^d \times \mathbb{R})^{t-1} \times \mathbb{R}^d \to \mathbb{R}$ associant aux données passées $(\boldsymbol{x}_s, y_s)$, $1 \leqslant s \leqslant t-1$, et à la donnée d'entrée $\boldsymbol{x}_t$ la prévision au temps t, encore notée $\widehat{y}_t$ par un léger abus de notation. Le regret minimax (1.5) correspond donc à la meilleure performance possible d'un prédicteur séquentiel, lorsque cette performance est évaluée en termes du regret sur $B_1(U)$ dans le pire des cas.

Notre premier résultat est une borne supérieure sur le regret minimax qui, selon la valeur de U, améliore légèrement la borne de regret de l'algorithme séquentiel EG$^\pm$ de [KW97] ou du prédicteur séquentiel *ridge* de [AW01, Vov01]. En particulier, la deuxième borne améliore la borne de l'algorithme EG$^\pm$ d'un facteur au plus de l'ordre de $\ln d$. Cette borne découle d'un *argument à la Maurey*, qui consiste à discrétiser la boule $B_1(U)$ et à montrer – par le biais d'une randomisation auxiliaire – qu'il est sensiblement équivalent de minimiser le regret sur $B_1(U)$ ou sur une discrétisation judicieuse. Cet argument est classique en statistique et a été utilisé, par exemple, par [Nem00, Tsy03, BN08, SSSZ10] ; la preuve du résultat suivant montre qu'il s'adapte directement au cadre déterministe.

Théorème 1.2 (cf. théorème 4.1)**.** *Soit $d, T \geqslant 1$ et $U, X, Y > 0$. Le regret minimax sur $B_1(U)$ pour des données bornées par X et Y vérifie*

$$\inf_{(\widehat{y}_t)_{t\geqslant 1}} \sup_{\substack{\|\boldsymbol{x}_1\|_\infty, \dots, \|\boldsymbol{x}_T\|_\infty \leqslant X \\ |y_1|, \dots, |y_T| \leqslant Y}} \left\{ \sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \right\}$$
$$\leqslant \begin{cases} 3\,UXY\sqrt{2T\ln(2d)} & \text{si} \quad U < \frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}}\,, \\ 26\,UXY\sqrt{T\ln\left(1+\frac{2dY}{\sqrt{T}UX}\right)} & \text{si} \quad \frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}} \leqslant U \leqslant \frac{2dY}{\sqrt{T}X}\,, \\ 32\,dY^2\ln\left(1+\frac{\sqrt{T}UX}{dY}\right) + dY^2 & \text{si} \quad U > \frac{2dY}{X\sqrt{T}}\,. \end{cases}$$

La borne de regret précédente peut être réécrite en termes de d, Y et d'une quantité intrinsèque $\kappa \triangleq \sqrt{T}UX/(2dY)$ qui relie la dimension ambiante d à $\sqrt{T}UX/(2Y)$. On obtient :

$$\inf_{(\widehat{y}_t)_{t\geqslant 1}} \sup_{\substack{\|\boldsymbol{x}_1\|_\infty, \dots, \|\boldsymbol{x}_T\|_\infty \leqslant X \\ |y_1|, \dots, |y_T| \leqslant Y}} \left\{ \sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \right\}$$
$$\leqslant \begin{cases} 6\,dY^2\kappa\sqrt{2\ln(2d)} & \text{si} \quad \kappa < \frac{\sqrt{\ln(1+2d)}}{2d\sqrt{\ln 2}}\,, \\ 52\,dY^2\kappa\sqrt{\ln(1+1/\kappa)} & \text{si} \quad \frac{\sqrt{\ln(1+2d)}}{2d\sqrt{\ln 2}} \leqslant \kappa \leqslant 1\,, \\ 32\,dY^2\big(\ln(1+2\kappa)+1\big) & \text{si} \quad \kappa > 1\,. \end{cases} \tag{1.6}$$

En petite dimension, on remarque une transition d'une borne de regret de l'ordre de $\sqrt{T}$ à une borne de l'ordre de $\ln T$ autour du point $\kappa = 1$. Cette transition est absente en grande dimension : pour $d \geqslant \omega T$, avec $\omega \triangleq \big(32(\ln(3)+1)\big)^{-1}$, la borne de regret $32\,dY^2\big(\ln(1+2\kappa)+1\big)$ est supérieure à la borne triviale TY^2 pour tout $\kappa \geqslant 1$.

Quand $\kappa \geqslant \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$, la borne (1.6) correspond (à une division par T près) à la vitesse optimale d'agrégation convexe dans le modèle de régression gaussienne avec *de-*

sign aléatoire [Tsy03] et à la vitesse optimale d'estimation sur des boules ℓ^1 dans le modèle de régression gaussienne avec *design* fixe et matrice identité[10] [BM01a] (cf. aussi [DJ94b] et [RWY11]). Ce fait indique que la régression linéaire sur des boules ℓ^1 n'est pas plus difficile dans un cadre de suites individuelles que dans un cadre statistique classique.

Ces deux tâches de prévision (stochastique et déterministe) sont en fait de même complexité (à des facteurs logarithmiques près) : d'après la conversion *online to batch* décrite à la proposition 1.1, la borne inférieure de [Tsy03] pour l'agrégation convexe en *design* aléatoire implique une borne inférieure en suites individuelles (on exploite aussi la bornitude d'un bruit gaussien avec grande probabilité). Le résultat suivant indique que pour tout $d \in \mathbb{N}^*$, $Y > 0$ et $\kappa \geqslant \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$, la borne (1.6) ne peut être améliorée de plus d'un facteur logarithmique. Cette borne inférieure étend celles de [CB99, KW97], qui valent seulement pour κ petit de l'ordre de $1/d$.

Théorème 1.3 (cf. théorème 4.2)**.** *Pour tous $d \in \mathbb{N}^*$, $Y > 0$ et $\kappa \geqslant \frac{\sqrt{\ln(1+2d)}}{2d\sqrt{\ln 2}}$, il existe $T \geqslant 1$, $U > 0$ et $X > 0$ tels que $\sqrt{T}UX/(2dY) = \kappa$ et*

$$\inf_{(\widehat{y}_t)_{t\geqslant 1}} \sup_{\substack{\|\boldsymbol{x}_1\|_\infty,\dots,\|\boldsymbol{x}_T\|_\infty \leqslant X \\ |y_1|,\dots,|y_T|\leqslant Y}} \left\{ \sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\}$$
$$\geqslant \begin{cases} \frac{c_1}{\ln\left(2+16d^2\right)} dY^2 \kappa \sqrt{\ln\left(1+1/\kappa\right)} & \text{si} \quad \frac{\sqrt{\ln(1+2d)}}{2d\sqrt{\ln 2}} \leqslant \kappa \leqslant 1\,, \\ \frac{c_2}{\ln\left(2+16d^2\right)} dY^2 & \text{si} \quad \kappa > 1\,, \end{cases}$$

où $c_1, c_2 > 0$ sont des constantes absolues.

1.3.3 Adaptation aux paramètres du problème

Certains prédicteurs séquentiels utilisés dans la preuve du théorème 1.2 n'admettent pas de mise en œuvre algorithmique efficace en grande dimension d (par exemple, celui utilisé conjointement avec l'argument à la Maurey), et tous utilisent la connaissance a priori de X, Y, T ou U.

Une façon de surmonter ces limites est fournie par le *self-confident p-norm algorithm* de [ACBFS02] ; pour $p = 2\ln d$, le regret de ce prédicteur sur $B_1(U)$ est borné par[11]

$$8UXY\sqrt{eT\ln d} + 32e\,U^2X^2\ln d\,.$$

Cet algorithme est efficace et notre borne inférieure montre qu'il est optimal à un facteur logarithmique près dans le régime $\kappa \leqslant 1$, i.e., en dimension $d \geqslant \sqrt{T}UX/(2Y)$, et ce sans utiliser de connaissance a priori sur X, Y et T (voir une remarque ci-après pour le régime $\kappa > 1$).

Notre deuxième contribution consiste à montrer que des propriétés d'adaptativité similaires peuvent être obtenues par pondération exponentielle, et ce pour un même coût algorithmique (linéaire en d). Plus précisément, on étudie une variante du prédicteur séquentiel EG$^\pm$ de [KW97] calibré avec un paramètre évolutif η_t choisi en fonction des données.

Cet algorithme séquentiel, que nous appelons *algorithme EG$^\pm$ adaptatif*, dépend du rayon U de la boule considérée. A chaque date $t \geqslant 1$, sa prévision est donnée par $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t$ avec

[10] Ce modèle est connu sous le nom de *Gaussian sequence framework* en anglais.

[11] Ce prédicteur vérifie en fait une borne plus fine, du même type que celle du théorème 1.4.

$\widehat{\boldsymbol{u}}_t \in B_1(U)$ défini par

$$\widehat{\boldsymbol{u}}_t \triangleq \sum_{i=1}^{d} \left(p_{i,t}^{+} \left(U \boldsymbol{e}_i \right) + p_{i,t}^{-} \left(-U \boldsymbol{e}_i \right) \right) ,$$

où $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_d)$ désigne la base canonique de $\mathbb{R}^d$ et où les poids $p_{i,t}^{+}$ et $p_{i,t}^{-}$ sont définis par

$$p_{i,t}^{\gamma} \triangleq \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \gamma U \nabla_i \ell_s(\widehat{\boldsymbol{u}}_s) \right)}{\sum_{\substack{1 \leqslant j \leqslant d \\ \mu \in \{+,-\}}} \exp\left(-\eta_t \sum_{s=1}^{t-1} \mu U \nabla_j \ell_s(\widehat{\boldsymbol{u}}_s) \right)} , \quad 1 \leqslant i \leqslant d , \quad \gamma \in \{+,-\} ,$$

avec $\nabla_i \ell_t(\boldsymbol{u}) = -2(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t) x_{i,t}$ pour la perte carrée[12]. Lorsque l'on choisit la suite $(\eta_t)_{t \geqslant 1}$ selon la calibration automatique (fonction de la variance cumulée du prédicteur) introduite par [CBMS07] et définie précisément en section 2.4.3, l'algorithme EG$^{\pm}$ adaptatif vérifie la borne de regret suivante. Cette dernière indique avec les théorèmes 1.2 et 1.3 que l'algorithme EG$^{\pm}$ adaptatif est effectivement adaptatif (à un facteur logarithmique près) en les paramètres X, Y et T dans le régime $\kappa \leqslant 1$, i.e., en dimension $d \geqslant \sqrt{T} U X / (2Y)$.

Théorème 1.4 (cf. corollaire 2.2)**.** *Soit $U > 0$. L'algorithme EG$^{\pm}$ adaptatif calibré avec U et la suite $(\eta_t)_{t \geqslant 1}$ préconisée par [CBMS07] vérifie, pour toute suite $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T) \in \mathbb{R}^d \times \mathbb{R}$,*

$$\begin{aligned} \sum_{t=1}^{T} (y_t - \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t)^2 - \min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 & \\ \leqslant 8UX \sqrt{L_T^* \ln(2d)} + \left(137 \ln(2d) + 24 \right) \left(UXY + U^2 X^2 \right) & \\ \leqslant 8UXY \sqrt{T \ln(2d)} + \left(137 \ln(2d) + 24 \right) \left(UXY + U^2 X^2 \right) , & \qquad (1.7) \end{aligned}$$

où $L_T^* \triangleq \min_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$, $X \triangleq \max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty$ et $Y \triangleq \max_{1 \leqslant t \leqslant T} |y_t|$ sont inconnus du statisticien.

Cette borne de regret est similaire à celle de l'algorithme EG$^{\pm}$ de [KW97], mais est obtenue sans la connaissance a priori de X, Y, ou T. Elle est également de la même forme que la borne du *self-confident p-norm algorithm* de [ACBG02], ce qui corrobore la proximité déjà observée par [Gen03] entre l'algorithme *p-norm* et le prédicteur EG$^{\pm}$ (avant calibration adaptative).

Au chapitre 4, nous ne détaillons l'adaptation en X, Y et T que dans le régime $\kappa \leqslant 1$ (i.e., $d \geqslant \sqrt{T} U X / (2Y)$). Il est néanmoins également possible d'obtenir une borne adaptative (à un facteur logarithmique près) dans le régime $\kappa > 1$. Pour ce faire, il suffit de modifier l'algorithme de prévision *ridge* séquentiel défini en section 2.4.2 en tronquant ses prévisions et en le calibrant au moyen d'une technique générique appelée *doubling trick*. Cet algorithme peut être mis en oeuvre avec une complexité algorithmique égale à celle de l'algorithme *ridge* séquentiel, donc au plus de

[12] De même que dans [CB99] et [CBL06, section 2.5], l'algorithme EG$^{\pm}$ adaptatif est générique et peut être utilisé avec toute suite $(\ell_t)_{t \geqslant 1}$ de fonctions de pertes convexes et différentiables sur $\mathbb{R}^d$. Nous l'utilisons dans un premier temps avec les fonctions de perte $\ell_t : \boldsymbol{u} \in \mathbb{R}^d \mapsto (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$.

l'ordre de d^3 à chaque tour de prévision.

Remarquons que la connaissance de U est encore requise par l'algorithme EG$^{\pm}$ adaptatif. Il est en fait possible de s'adapter à U, i.e., d'atteindre approximativement le regret minimax simultanément sur toutes les boules $B_1(U), U > 0$. A cette fin, on suppose d'abord pour simplifier que X, Y et T sont connus. Il suffit alors d'agréger plusieurs instances de l'algorithme EG$^{\pm}$ adaptatif calibrées avec différentes valeurs de U — celles associées à une grille exponentielle de la forme $\{U_r = U_0\, 2^r : r = 0, \dots, R\}$, où $U_0 \triangleq Y/\big(X\sqrt{T\ln(2d)}\big)$; cf. théorème 4.4. Nous expliquons ensuite, en section 4.5, comment étendre cette méthode pour construire un algorithme totalement adaptatif, i.e., qui vérifie une borne de regret similaire à (1.7) pour tout $U > 0$ et qui ne dépend d'aucune connaissance a priori sur U, X, Y ou T.

1.3.4 Une amélioration : la lipschitzification des pertes

Notre troisième contribution consiste en l'introduction d'une technique générique appelée *lipschitzification des pertes*. On transforme les fonctions de perte $\boldsymbol{u} \mapsto (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$ (ou $\boldsymbol{u} \mapsto \big|y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big|^{\alpha}$ si les prévisions sont évaluées avec la perte ℓ^{α}, $\alpha \geqslant 2$) en des fonctions $\widetilde{\ell}_t : \mathbb{R}^d \to \mathbb{R}$ convexes et lipschitziennes sur $\mathbb{R}^d$. A chaque date $t \geqslant 1$, la lipschitzification est effectuée le long de la direction de $\boldsymbol{x}_t$ et au-delà d'un seuil adaptatif $B_t \triangleq \big(2^{\lceil \log_2 \max_{1\leqslant s\leqslant t-1} y_s^2\rceil}\big)^{1/2} \approx \max_{1\leqslant s\leqslant t-1}|y_s|$, que nous avions déjà utilisé auparavant (cf. section 1.2.2). Plus précisément, si $y_t \notin [-B_t, B_t]$, alors on pose $\widetilde{\ell}_t \equiv 0$, sinon, on définit $\widetilde{\ell}_t$ comme étant la plus petite fonction convexe coïncidant avec la perte carrée $(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$ lorsque $|\boldsymbol{u}\cdot\boldsymbol{x}_t| \leqslant B_t$ (elle est donc affine en dehors de $[-B_t, B_t]$). Ce dernier cas correspond graphiquement à la courbe en pointillés sur la figure 1.4.

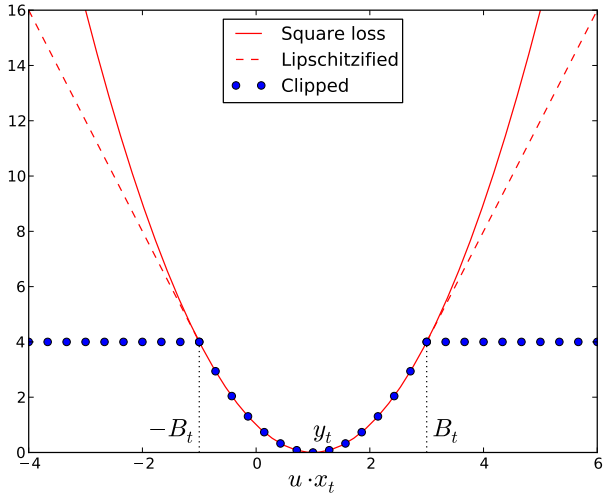


FIGURE 1.4 – La perte carrée $(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$, sa version tronquée $\big(y_t - [\boldsymbol{u}\cdot\boldsymbol{x}_t]_{B_t}\big)^2$ – *clipped* en anglais – et sa version lipschitzifiée $\widetilde{\ell}_t(\boldsymbol{u})$ sont tracées en fonction de $\boldsymbol{u}\cdot\boldsymbol{x}_t$.

L'intérêt de la lipschitzification des pertes peut être illustré avec l'algorithme EG$^{\pm}$ adaptatif. En effet, lorsque l'on applique ce prédicteur séquentiel aux fonctions de perte lipschitzifiées $\widetilde{\ell}_t$, il

vérifie la borne de regret suivante (cf. théorème 4.3) :

$$\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 \leqslant \inf_{\|\boldsymbol{u}\|_1\leqslant U}\sum_{t=1}^{T}\widetilde{\ell}_t(\boldsymbol{u}) + 8UX\sqrt{\left(\inf_{\|\boldsymbol{u}\|_1\leqslant U}\sum_{t=1}^{T}\widetilde{\ell}_t(\boldsymbol{u})\right)\ln(2d)} + \big(153\ln(2d)+58\big)\big(UXY+U^2X^2\big)+12Y^2\,.$$

Les deux termes principaux de cette borne améliorent légèrement ceux de la borne obtenue sans lipschitzification des pertes, puisqu'on a toujours $\widetilde{\ell}_t(\boldsymbol{u}) \leqslant (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$ pour tout $\boldsymbol{u}\in\mathbb{R}^d$.

L'intérêt de la lipschitzification est plus clair pour des fonctions de perte de plus grande courbure, par exemple, $x\mapsto|y_t-x|^\alpha$ avec $\alpha>2$. Dans ce cas, l'algorithme EG$^\pm$ adaptatif avec pertes lipschitzifiées a un regret qui croît au plus linéairement en U, alors que la borne de l'algorithme EG$^\pm$ adaptatif sans lipschitzification donne (au premier abord du moins) une borne naïve en $U^{\alpha/2}$. Voir la remarque 4.1 pour de plus amples détails.

1.4 Vitesses minimax des regrets interne et *swap*

Au chapitre 5, on étudie une instance du protocole de prévision avec avis d'experts correspondant à des pertes linéaires sur le simplexe. Ce problème de décision séquentielle est dû à [FS97] et peut être décrit comme suit. A chaque date $t\in\mathbb{N}^*=\{1,2,\dots\}$, le statisticien choisit un vecteur de poids $\boldsymbol{p}_t=(p_{1,t},\dots,p_{K,t})$ sur $K\geqslant 2$ actions différentes, i.e., $\boldsymbol{p}_t$ appartient au simplexe

$$\mathcal{X}_K \triangleq \left\{\boldsymbol{x}\in\mathbb{R}_+^K,\quad \sum_{i=1}^{K}x_i=1\right\}\,.$$

L'environnement révèle ensuite le vecteur de pertes $\boldsymbol{\ell}_t\triangleq(\ell_{i,t})_{1\leqslant i\leqslant K}\in[0,1]^K$; chaque action $i\in\{1,\dots,K\}$ encourt la perte $\ell_{i,t}$ et le statisticien encourt la perte linéaire (ou perte moyenne) $\boldsymbol{p}_t\cdot\boldsymbol{\ell}_t=\sum_{i=1}^K p_{i,t}\,\ell_{i,t}$. Après $T\geqslant 1$ pas de temps, la perte cumulée du statisticien vaut $\sum_{t=1}^T\boldsymbol{p}_t\cdot\boldsymbol{\ell}_t$, et son objectif premier est de la minimiser.

On suppose que la suite $(\boldsymbol{\ell}_t)_{t\geqslant1}$ est fixée à l'avance par l'environnement, et on étudie les deux situations suivantes : $(\boldsymbol{\ell}_t)_{t\geqslant1}$ est déterministe arbitraire[13] (i.e., c'est une suite individuelle), ou $(\boldsymbol{\ell}_t)_{t\geqslant1}$ est aléatoire i.i.d. de loi inconnue.

Les vecteurs de poids $\boldsymbol{p}_t$ sont choisis en fonction des vecteurs de pertes passées et peuvent donc être vus comme des valeurs de fonctions $\boldsymbol{p}_t(\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_{t-1})$. On appelle *stratégie (du statisticien)* toute suite $(\boldsymbol{p}_t)_{t\geqslant1}$ de fonctions boréliennes $\boldsymbol{p}_t:[0,1]^{K(t-1)}\to\mathcal{X}_K$. Pour simplifier les notations, on omettra souvent les dépendences et $\boldsymbol{p}_t(\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_{t-1})$ sera simplement noté $\boldsymbol{p}_t$.

Regret interne et regret *swap*

Jusqu'à présent (chapitres 3 et 4), nous avons évalué la qualité d'un prédicteur séquentiel par son regret externe. Dans le cadre considéré ici, le regret externe d'une stratégie $S=(\boldsymbol{p}_t)_{t\geqslant1}$ pour une suite $\boldsymbol{\ell}_{1:T}\triangleq(\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T)$ est défini par

[13] En fait, dans le cadre de suites individuelles, on pourrait plus généralement supposer que les pertes $\boldsymbol{\ell}_t$ sont choisies par un environnement antagoniste, i.e., qui réagit aux décisions du statisticien (cf. section 2.3.1).

$$R_T^{\text{ext}}(S, \boldsymbol{\ell}_{1:T}) \triangleq \sum_{t=1}^{T} \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{1 \leqslant j \leqslant K} \sum_{t=1}^{T} \ell_{j,t} \,. \tag{1.8}$$

Au chapitre 5, on étudie deux autres notions de regret qui jouent un rôle important en théorie des jeux : les regrets interne et *swap*. L'ensemble des stratégies auxquelles est comparée la stratégie S n'est plus externe comme dans (1.8) (où les stratégies de référence sont les stratégies constantes $(\boldsymbol{\delta}_i)_{t \geqslant 1}$, avec $\boldsymbol{\delta}_i$ la masse de Dirac en $i \in \{1, \dots, K\}$) ; au contraire, il est composé de modifications de la stratégie S elle-même.

La notion de regret interne a été introduite et étudiée par [FV97, FV98, FV99] (cf. aussi [FL99, HMC00, HMC01]). Pour une stratégie $S = (\boldsymbol{p}_t)_{t \geqslant 1}$ et une suite finie $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$, le *regret interne* $R_T^{\text{int}}(S, \boldsymbol{\ell}_{1:T})$ d'une stratégie S associé à $\boldsymbol{\ell}_{1:T} \triangleq (\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T)$ est défini par

$$R_T^{\text{int}}(S, \boldsymbol{\ell}_{1:T}) \triangleq \sum_{t=1}^{T} \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^{T} \boldsymbol{p}_t^{i \to j} \cdot \boldsymbol{\ell}_t \,, \tag{1.9}$$

où le vecteur de poids modifié $\boldsymbol{p}_t^{i \to j} \in \mathcal{X}_K$ est obtenu à partir de $\boldsymbol{p}_t$ en remplaçant l'action i par l'action j. Plus précisément, pour tout $k = 1, \dots, K$, la k-ème composante de $\boldsymbol{p}_t^{i \to j}$ est définie par

$$(p_t^{i \to j})_k = \begin{cases} 0 & \text{si } k = i, \\ p_{i,t} + p_{j,t} & \text{si } k = j, \\ p_{k,t} & \text{si } k \notin \{i, j\}. \end{cases}$$

Ainsi, le regret interne mesure le "regret" qu'encourt le statisticien à n'avoir pas choisi l'action j à chaque fois qu'il a choisi l'action i, et ce pour tous les couples (i, j) possibles, $i \neq j$. Intuitivement, si le statisticien minimise son regret interne, alors il bénéficie de propriétés de stabilité. Cela a été illustré en théorie des jeux : [FV97, FV99] ont montré que, dans un jeu répété randomisé entre un nombre fini de joueurs, si tous les joueurs suivent une stratégie dont le regret interne est sous-linéaire en T, la distribution empirique jointe de leurs actions converge vers un ensemble d'équilibres appelé l'ensemble des équilibres corrélés du jeu (cf. aussi [FL95, HMC00, SL07]). Le regret interne a aussi des liens historiques avec une autre branche de la théorie des jeux appelée *calibration* en anglais (à ne pas confondre avec *parameter tuning*). Ainsi, l'existence de stratégies assurant un regret interne sous-linéaire en T implique l'existence d'algorithmes de prévision bien calibrés (*calibrated forecasters*, cf. [FV98] par ex.).

La notion de regret *swap* a été introduite par [BM07b] (voir aussi [GJ03] pour la notion plus générale de Φ-regret). Le *regret swap* $R_T^{\text{sw}}(S, \boldsymbol{\ell}_{1:T})$ d'une stratégie $S = (\boldsymbol{p}_t)_{t \geqslant 1}$ pour une suite finie $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$ est défini par

$$R_T^{\text{sw}}(S, \boldsymbol{\ell}_{1:T}) \triangleq \sum_{t=1}^{T} \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F \in \mathcal{F}_K} \sum_{t=1}^{T} \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t \,,$$

où $\mathcal{F}_K$ désigne l'ensemble des fonctions $F : \{1 \dots, K\} \to \{1 \dots, K\}$ et où le vecteur de poids modifié $\boldsymbol{p}_t^F \in \mathcal{X}_K$ est obtenu à partir de $\boldsymbol{p}_t$ en remplaçant chaque action i par l'action $F(i)$, i.e., sa j-ème composante est définie par $(p_t^F)_j = \sum_{i : F(i) = j} p_{i,t}$, $1 \leqslant j \leqslant K$.

En particulier, le regret *swap* d'une stratégie est plus grand que ses regrets externe et interne.

Bornes existantes sur les regrets interne et *swap*

Dans la suite, on détaille les contributions principales du chapitre 5, qui concernent principalement les vitesses minimax des regrets interne et *swap* en environnement stochastique ou déterministe.

Pour le regret interne, on s'intéresse aux deux quantités minimax suivantes. D'après la borne inférieure de [Sto05, théorème 3.3] et la borne supérieure de [SL05], on sait que[14] le *regret interne minimax pour des vecteurs de pertes i.i.d.* (à gauche ci-dessous) et le *regret interne minimax pour des suites individuelles* [SL05, théorème 3] (à droite ci-dessous) vérifient l'encadrement suivant :

$$\sqrt{T}/\big(64\sqrt{3}\big) \leqslant \inf_S \sup_Q \mathbb{E}_{Q^{\otimes T}}\Big[R_T^{\mathrm{int}}(S, \boldsymbol{\ell}_{1:T})\Big] \leqslant \inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in [0,1]^K} R_T^{\mathrm{int}}(S, \boldsymbol{\ell}_{1:T}) \leqslant \sqrt{T \ln K}\,,$$

où la première inégalité vaut pour tout $T \geqslant K^2/192$ (les deux autres inégalités sont valides pour tout $T \geqslant 1$), où les deux infima sont pris sur toutes les stratégies $S = (\boldsymbol{p}_t)_{t\geqslant 1}$, où le supremum $\sup_Q$ s'étend sur toutes les probabilités sur $[0,1]^K$ (muni de sa tribu borélienne), et où dans l'espérance, les vecteurs de pertes $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$ sont supposés i.i.d. de loi Q. On remarque un facteur $\sqrt{\ln K}$ manquant entre les bornes inférieure et supérieure.

Quant au regret *swap*, [BM07b] ont construit une stratégie dont le regret *swap* est majoré par $\sqrt{(T/2)K\ln K}$ uniformément sur toutes les suites individuelles (cf. aussi [SL05]). Ainsi, le *regret swap minimax pour des suites individuelles* est majoré comme suit :

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in [0,1]^K} R_T^{\mathrm{sw}}(S, \boldsymbol{\ell}_{1:T}) \leqslant \sqrt{(T/2)K\ln K}\,.$$

Une borne inférieure sur le regret *swap* en suites individuelles de l'ordre de $\sqrt{TK}$ a été exhibée par [BM07b], mais seulement dans un sens assez faible : leur borne inférieure est prouvée dans un cadre randomisé antagoniste et pour une quantité plus grande que le regret *swap stricto sensu* ; de plus, elle n'est prouvée que lorsque T est sous-exponentiel en K.

1.4.1 Vitesse minimax du regret interne dans un environnement stochastique

La première contribution du chapitre 5 est la détermination de la vitesse minimax du regret interne en environnement stochastique, qui est de l'ordre de $\sqrt{T}$ et est donc indépendante de la dimension ambiante K.

Théorème 1.5 (cf. corollaire 5.1)**.** *Il existe des constantes absolues $c_1, c_2, c_3 > 0$ telles que, pour tout $K \geqslant 2$ et tout $T \geqslant c_1 K^2$, le regret interne minimax pour des vecteurs de pertes i.i.d. vérifie*

$$c_2\sqrt{T} \leqslant \inf_S \sup_Q \mathbb{E}_{Q^{\otimes T}}\Big[R_T^{\mathrm{int}}(S, \boldsymbol{\ell}_{1:T})\Big] \leqslant c_3\sqrt{T}\,.$$

En particulier, le résultat est vrai pour $c_1 = 1/192$ et $c_2 = 1/(64\sqrt{3})$; la constante c_3 peut être calculée explicitement et directement à partir des fins des preuves des théorème 5.2 et corollaire 5.1, mais sa valeur n'a pas été optimisée.

La borne supérieure est obtenue de façon constructive : on montre en section 5.3 qu'une stratégie procédant par pondération exponentielle et par estimation séquentielle des espérances

[14] Voir l'introduction du chapitre 5 pour de plus amples références.

des pertes encourt un regret interne au plus de l'ordre de $\sqrt{T}$ avec grande probabilité (via des outils élémentaires de concentration, par ex., l'inégalité de Hoeffding). La forme des poids choisie impose une répartition uniforme de la masse entre des pertes d'espérances proches ; cela s'avère clé pour supprimer la dépendance du regret interne minimax en la dimension ambiante K.

1.4.2 Borne inférieure sur le regret *swap* pour des suites individuelles

Notre deuxième contribution consiste en l'obtention d'une borne inférieure de l'ordre de $\sqrt{TK}$ sur le regret *swap* minimax pour des suites individuelles. La preuve du théorème suivant s'appuie sur l'inégalité de Pinsker (cf. annexe A.7) et repose en partie sur des techniques de borne inférieure séquentielle développées par [ACBG02, CBLS05] et [Sto05, Theorem 3.3].

Théorème 1.6 (cf. théorème 5.3)**.** *Il existe une constante absolue $c > 0$ telle que, pour tous $K \geqslant 2$ et $T \geqslant \max\{128c^2K^5, K\}$, le regret swap minimax pour des suites individuelles vérifie*

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \left\{ \sum_{t=1}^T \boldsymbol{p}_t\cdot\boldsymbol{\ell}_t - \min_{F\in\mathcal{F}_K} \sum_{t=1}^T \boldsymbol{p}_t^F\cdot\boldsymbol{\ell}_t \right\} \geqslant c\sqrt{TK}\,.$$

En particulier, on prouve le théorème pour $c = 1/\big(16\sqrt{128\ln(4/3)}\big)$.

Cette borne inférieure est plus forte que celle de [BM07b, théorème 9], puisqu'elle vaut pour le regret *swap* lui-même plutôt qu'une variante randomisée dans un cadre antagoniste (qui rend l'obtention de la borne inférieure plus simple). Cela résout ainsi un problème ouvert de [BM07b, section 9]. De plus, nous nous sommes affranchis de l'hypothèse d'un horizon T sous-exponentiel en K.

Comme remarqué en section 5.4.2, notre borne inférieure de l'ordre de $\sqrt{TK}$ pointe une différence essentielle entre les regrets externe et *swap*. En effet, alors que le regret externe minimax est du même ordre de grandeur pour des pertes i.i.d. ou pour des suites individuelles (en l'occurence, $\sqrt{T\ln K}$, cf. chapitre 2), le regret *swap* est bien plus difficile à minimiser pour des suites individuelles que pour des vecteurs de pertes i.i.d. (comparer la borne inférieure en $\sqrt{TK}$ ci-dessus avec la vitesse minimax en $\sqrt{T\ln K}$ pour des pertes i.i.d. prouvée en section 5.4.2).

1.4.3 Une technique stochastique pour des majorations en suites individuelles

La troisième contribution du chapitre 5 est le développement d'une technique stochastique pour majorer dans un cadre de suites individuelles (déterministes) une forme généralisée de regret incluant les regrets externe, interne et *swap*, et définie comme suit.

Définition 1.1. *Soit E un espace vectoriel réel, $\psi = (\psi_t)_{t\geqslant 1}$ une suite de fonctions convexes $\psi_t : E \to \mathbb{R}$, et $\varphi : \mathbb{R}^K \times \mathbb{R}^K \to E$ une fonction bi-affine au sens où $\varphi(\boldsymbol{u},\cdot)$ and $\varphi(\cdot,\boldsymbol{v})$ sont affines pour tous $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^K$. On appelle* (ψ,φ)-regret *d'une stratégie $S = (\boldsymbol{p}_t)_{t\geqslant 1}$ sur une suite finie de pertes $\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K$ la quantité $\psi_T\Big(\sum_{t=1}^T \varphi(\boldsymbol{p}_t,\boldsymbol{\ell}_t)\Big)$.*

La quantité minimax associée au (ψ,φ)-regret peut être ré-interprétée de façon stochastique à l'aide du théorème minimax suivant.

Théorème 1.7 (cf. théorème 5.4)**.** *Soit E un espace vectoriel réel, $\psi = (\psi_t)_{t\geqslant 1}$ une suite de fonctions convexes $\psi_t : E \to \mathbb{R}$, et $\varphi : \mathbb{R}^K \times \mathbb{R}^K \to E$ une fonction bi-affine. Alors, le (ψ, φ)-regret vérifie la formule de dualité suivante :*

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \psi_T\left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t,\boldsymbol{\ell}_t)\right) = \sup_{\mathbb{Q}\in\mathcal{M}_1^+([0,1]^{KT})} \inf_S \mathbb{E}_{\mathbb{Q}}\left[\psi_T\left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t,\boldsymbol{\ell}_t)\right)\right],$$

où les deux infima sont pris sur toutes les stratégies $S = (\boldsymbol{p}_t)_{t\geqslant 1}$, où $\mathcal{M}_1^+\big([0,1]^{KT}\big)$ désigne l'ensemble de toutes les probabilités sur $[0,1]^{KT}$, et où l'espérance $\mathbb{E}_{\mathbb{Q}}[\,\cdot\,]$ est prise par rapport aux variables aléatoires $\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in [0,1]^K$ supposées de loi jointe $\mathbb{Q}$.

Le théorème ci-dessus fournit un moyen non constructif de majorer le regret minimax (à gauche) ; la quantité maximin (à droite) est d'apparence plus simple à majorer car la stratégie S choisie peut dépendre de la loi jointe $\mathbb{Q}$ des pertes. Cette technique permet de retrouver de façon stochastique les meilleures bornes supérieures connues sur les regrets externe, interne et *swap* minimax, à savoir $\sqrt{(T/2)\ln K}$, $\sqrt{T\ln K}$ et $\sqrt{(T/2)K\ln K}$ (cf. proposition 5.4). On l'utilise également à la proposition 5.5 pour prouver une borne supérieure de l'ordre de $\sqrt{T\ln K}$ sur le regret *makespan* (utile pour modéliser des problèmes de planification de tâches ou de répartition de charges), améliorant ainsi la borne d'ordre $\ln(K)\sqrt{T}$ obtenue par [EDKMM09].

Mentionnons qu'une technique similaire a été étudiée indépendamment par [RST11]. Puisque nous travaillons dans un cadre beaucoup plus restreint, il nous est possible d'obtenir des constantes explicites (et même optimales dans le cas du regret externe). Notre preuve du théorème 1.7 s'appuie sur des arguments simples comme la technique de bernoullisation de [Sch03] — qui permet de recourir à une version du théorème minimax de von Neumann sans considérations topologiques fines. La majoration de la quantité maximin (à droite) repose sur des outils élémentaires de concentration de martingales comme l'inégalité de Hoeffding-Azuma, que nous combinons avec une inégalité maximale pour des variables aléatoires sous-gaussiennes. Nous renvoyons le lecteur à la section 5.5.1 pour une comparaison plus détaillée avec la littérature.

Notons que quelques questions importantes restent encore ouvertes. Tout d'abord, même si la technique stochastique décrite précédemment est utile pour mieux comprendre le problème de prévision sous-jascent (puisqu'elle permet de majorer le regret minimax associé), elle n'est pas constructive — tout comme dans [RST11]. Il est donc important, pour la suite, d'exhiber des algorithmes explicites qui atteignent les bornes supérieures nouvellement prouvées (par ex., existe-t-il un algorithme efficace dont le regret *makespan* est au plus de l'ordre de $\sqrt{T\ln K}$?). Par ailleurs, la question du facteur logarihmique manquant $\sqrt{\ln K}$ entre les bornes inférieure et supérieure des regrets interne et *swap* est toujours partiellement ouverte. On a prouvé que le facteur $\sqrt{\ln K}$ n'était pas nécessaire pour le regret interne en environnement stochastique, mais la question de savoir si cela est aussi le cas pour des suites individuelles n'est pas encore résolue.

1.5 Agrégation de modèles non linéaires

Au chapitre 6, on étudie un problème de régression en considérant des estimateurs reposant sur des techniques d'agrégation, comme aux chapitres 3 et 4, mais dans un cadre de sélection de modèles. Ce chapitre est un travail en cours.

1.5.1 Cadre et objectif de prévision

Le cadre considéré au chapitre 6 est un modèle linéaire gaussien généralisé introduit par [BM01a] et qui inclut les modèles de régression avec *design* fixe et le modèle de bruit blanc gaussien. Par souci de simplicité, on se concentre ci-après sur le cas fini-dimensionnel, mais tous les résultats énoncés sont aussi prouvés dans le cas où $\mathbb{R}^n$ est remplacé par un espace de Hilbert séparable.

Considérons ainsi le modèle de régression gaussienne avec *design* fixe : le statisticien observe le vecteur $(Y_1, \dots, Y_n) \in \mathbb{R}^n$ donné par

$$Y_i = s_i + \sigma \xi_i \in \mathbb{R} \,, \quad 1 \leqslant i \leqslant n \,,$$

où les variables aléatoires $\xi_1, \dots, \xi_n$ sont i.i.d. de loi $\mathcal{N}(0,1)$, où $\sigma > 0$ est le niveau de bruit supposé connu, et où $s = (s_1, \dots, s_n) \in \mathbb{R}^n$ est un vecteur déterministe inconnu.

L'objectif du statisticien est d'estimer s en fonction de $Y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$. La performance d'un estimateur $\widetilde{s} \in \mathbb{R}^n$ est évaluée via son risque quadratique (empirique) $\|\widetilde{s} - s\|_n^2$, où l'on pose $\|u\|_n^2 \triangleq n^{-1} \sum_{i=1}^n u_i^2$ pour tout $u \in \mathbb{R}^n$.

Afin d'estimer s, le statisticien a accès à une famille au plus dénombrable $(S_m)_{m \in \mathcal{M}}$ de parties non vides de $\mathbb{R}^n$ (appelées *modèles*[15] ci-après) ; il dispose alors des estimateurs des moindres carrés[16]

$$\widehat{s}_m \in \operatorname*{argmin}_{t \in S_m} \|Y - t\|_n^2 \,, \quad m \in \mathcal{M} \,. \tag{1.10}$$

La tâche de prévision consiste à construire un estimateur $\widetilde{s}$ de s presque aussi bon que le meilleur des estimateurs parmi $\{\widehat{s}_m : m \in \mathcal{M}\}$. Par exemple, on dit que c'est le cas lorsque

$$\mathbb{E}_s \Big[\|\widetilde{s} - s\|_n^2 \Big] \leqslant C \inf_{m \in \mathcal{M}} \mathbb{E}_s \Big[\|\widehat{s}_m - s\|_n^2 \Big] \,,$$

où $C \geqslant 1$ est une constante (qui peut dépendre de la "taille" de $\mathcal{M}$) et où $\mathbb{E}_s$ désigne l'espérance prise par rapport à Y (dont la loi dépend de s). La borne de risque précédente est qualifiée d'*inégalité oracle* selon la terminologie de [DJ94a, BM01a].

1.5.2 Sélection et agrégation de modèles linéaires

On suppose dans cette sous-section que les modèles S_m sont *linéaires*, i.e., qu'il s'agit de sous-espaces vectoriels de S_m. On rappelle ci-après — à très grands traits[17] — deux approches alternatives : sélection de modèles et agrégation de modèles. Une caractéristique commune de ces deux approches est que les estimateurs $\widehat{s}_m$ sont combinés (ou sélectionnés) via les mêmes données que celles ayant servi à leur construction ; on ne fait pas de *sample splitting*, qui n'est pas adapté au cas du *design* fixe.

[15] Le terme *modèle* recouvre plusieurs significations : il est utilisé à la fois pour désigner le cadre (*modèle* de régression) et une partie de $\mathbb{R}^n$.

[16] On suppose pour simplifier que de tels estimateurs existent ; en toute généralité, on peut considérer des estimateurs approchés des moindres carrés — cf. section 6.2.2.

[17] Une introduction plus détaillée (avec bien plus de références) est proposée au chapitre 6.

La procédure de *sélection de modèles par pénalisation* de [BM01a] estime s avec $\widetilde{s} = \widehat{s}_{\widehat{m}}$, où l'indice $\widehat{m}$ sélectionné est défini par

$$\widehat{m} \in \operatorname*{argmin}_{m\in\mathcal{M}} \Big\{ \|Y - \widehat{s}_m\|_n^2 + \mathrm{pen}(m) \Big\}\,, \qquad \text{avec} \quad \mathrm{pen}(m) \overset{(\star)}{\geqslant} K\frac{\sigma^2 D_m}{n}\left(1+\sqrt{2L_m}\right)^2\,;$$

dans l'expression ci-dessus, $D_m \triangleq \dim(S_m)$ désigne la dimension de S_m, et $K > 1$ ainsi que $(L_m)_{m\in\mathcal{M}}$ sont des paramètres de la procédure tels que $\Sigma \triangleq \sum_{m:D_m>0} e^{-L_m D_m} < \infty$. Comme l'ont montré [BM01a], la procédure précédente vérifie l'inégalité oracle

$$\begin{aligned}\mathbb{E}_s\Big[\|\widehat{s}_{\widehat{m}} - s\|_n^2\Big] &\leqslant C_K\left(\inf_{m\in\mathcal{M}}\big\{d^2(s,S_m) + \mathrm{pen}(m)\big\} + \frac{\sigma^2}{n}(\Sigma+1)\right)\\ &\leqslant \big(1+\sup_{m\in\mathcal{M}} L_m\big)\, C'_{K,\Sigma}\, \inf_{m\in\mathcal{M}} \mathbb{E}_s\Big[\|\widehat{s}_m - s\|_n^2\Big] \quad \text{si égalité dans } (\star),\end{aligned}$$

où $d^2(s,S_m) \triangleq \inf_{t\in S_m}\|s-t\|_n^2$ et où $C_K, C'_{K,\Sigma} > 1$ sont deux constantes dépendant uniquement de K et (K,Σ) respectivement. Ces résultats et la procédure de sélection de modèles associée ont ensuite été étendus par [Mas07] au cas de modèles non linaires via une notion de dimension généralisée (cf. section 1.5.3).

Plus récemment, [LB06] ont étudié une variante bayésienne de la procédure de sélection de modèles. Au lieu de retenir $\widehat{s}_{\widehat{m}}$, où $\widehat{m} \in \operatorname{argmin}_{m\in\mathcal{M}}\big\{\|Y-\widehat{s}_m\|_n^2 + \mathrm{pen}(m)\big\}$, ils considèrent la *combinaison convexe* $\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)}\widehat{s}_m$, où

$$\widehat{\rho}_m^{(\eta)} = \frac{\exp\Big[-\eta\big(\|Y-\widehat{s}_m\|_n^2 + \mathrm{pen}^{(\eta)}(m)\big)\Big]}{\sum_{m'\in\mathcal{M}}\exp\Big[-\eta\big(\|Y-\widehat{s}_{m'}\|_n^2 + \mathrm{pen}^{(\eta)}(m')\big)\Big]}\,, \quad m\in\mathcal{M}\,, \tag{1.11}$$

avec une pénalité $\mathrm{pen}^{(\eta)}$ qui peut maintenant dépendre de η (afin de prendre en compte une probabilité a priori sur les modèles). Comme l'ont montré [LB06], si $\eta \leqslant n/(4\sigma^2)$ et $\mathrm{pen}^{(\eta)}(m) = 2\sigma^2 D_m/n + x_m/\eta$, où les $x_m \geqslant 0$ sont tels que $\Sigma \triangleq \sum_{m\in\mathcal{M}} e^{-x_m} < +\infty$, alors

$$\mathbb{E}_s\left[\Big\|\sum_{m\in\mathcal{M}}\widehat{\rho}_m^{(\eta)}\widehat{s}_m - s\Big\|_n^2\right] \leqslant \inf_{m\in\mathcal{M}}\left\{\mathbb{E}_s\Big[\|\widehat{s}_m - s\|_n^2\Big] + \frac{x_m}{\eta}\right\} + \frac{\ln\Sigma}{\eta}\,. \tag{1.12}$$

Lorsque $\eta = n/(4\sigma^2)$ et $\sup_m x_m < \infty$, la borne de risque précédente est une inégalité oracle *exacte*, i.e., avec constante 1 devant l'infimum.

1.5.3 Agrégation de modèles non linéaires : contributions

La borne de risque (1.12) de [LB06] a été obtenue sous l'hypothèse que les modèles $S_m \subset \mathbb{R}^n$ sont linéaires et que les $\widehat{s}_m$ sont les estimateurs des moindres carrés associés (i.e., les projecteurs orthogonaux de $Y\in\mathbb{R}^n$ sur les S_m). Ces travaux ont été étendus dans deux directions. D'une part, le cas de la variance inconnue a été traité par [Gir08]. D'autre part, [DS11] ont remplacé la famille de projecteurs orthogonaux $\big(\widehat{s}_m\big)_{m\in\mathcal{M}}$ par une famille quasi-arbitraire d'estimateurs affines ; cette large classe d'estimateurs inclut, par ex., les filtres diagonaux et la régression *ridge* à noyau.

Au chapitre 6, on étend les travaux de [LB06] dans une troisième direction : on considère toujours des estimateurs par projection (cf. (1.10)), mais les modèles $S_m \subset \mathbb{R}^n$ peuvent être quasi-arbitraires (ou *non linéaires*). Dans une telle généralité, l'emploi de la formule d'estimation

sans biais du risque de Stein [Ste81] à la manière de [LB06, DT08, DS11] semble difficile. On suit à la place l'approche par concentration de [Mas07] pour obtenir des inégalités de type oracle avec grande probabilité (mais avec une constante devant l'infimum supérieure à 1).

Dans le même esprit que [LB06], on procède par pondération exponentielle : on considère l'estimateur de type Gibbs $\widetilde{s}^{(\eta)} \triangleq \sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} \widehat{s}_m$ défini en (1.10) – (1.11). La pénalité $\mathrm{pen}^{(\eta)}(m)$ est quant à elle choisie en fonction d'une dimension généralisée D_m du modèle S_m introduite par [Mas07] : D_m est la solution dans $\mathbb{R}_+^*$ de l'équation $\varphi_m\big(\tau_m\sigma\sqrt{D_m/n}\big) = \sigma D_m/\sqrt{n}$, où $\tau_m \triangleq 1$ si S_m est fermé et convexe et $\tau_m \triangleq 2$ sinon, et où $\varphi_m : \mathbb{R}_+ \to \mathbb{R}_+$ vérifie l'hypothèse suivante.

Hypothèse 1.1. *$\varphi_m : \mathbb{R}_+ \to \mathbb{R}_+$ est croissante, continue et telle que $x \in \mathbb{R}_+^* \mapsto x^{-1}\varphi_m(x)$ est décroissante et, en posant $\xi \triangleq (\xi_1, \dots, \xi_n)$ et $<u, v>_n \triangleq n^{-1}\sum_{i=1}^n u_i v_i$ pour tous $u, v \in \mathbb{R}^n$,*

$$\forall u \in S_m, \quad \forall x > 0, \qquad 2\sqrt{n}\,\mathbb{E}\left[\sup_{t\in S_m}\left(\frac{<\xi, t>_n - <\xi, u>_n}{\|t-u\|_n^2 + x^2}\right)\right] \leqslant x^{-2}\varphi_m(x)\ .$$

Comme l'a montré [Mas07], D_m mesure la taille du modèle S_m (D_m est liée à la notion d'entropie métrique). Par exemple, si S_m est linéaire, alors on peut choisir $D_m = \dim(S_m)$; si S_m est fini de cardinal $|S_m|$, alors on peut choisir $D_m = 8\ln|S_m|$; enfin, si $S_m = \big\{\sum_{j=1}^d u_j\varphi_j : \boldsymbol{u} \in \mathbb{R}^d, \|\boldsymbol{u}\|_1 \leqslant U_m\big\}$ pour un dictionnaire $\varphi_1, \dots, \varphi_d \in \mathbb{R}^n$, alors on peut choisir D_m proportionnel à U_m [MM11].

En combinant l'analyse par concentration de [Mas07] avec une formule de dualité sur la divergence de Kullback-Leibler (déjà utilisée au chapitre 3), on obtient l'inégalité de type oracle avec grande probabilité suivante. Comme à la section précédente, on pose $d^2(s, S_m) \triangleq \inf_{t\in S_m}\|s-t\|_n^2$ pour tout $m \in \mathcal{M}$.

Théorème 1.8 (cf. théorème 6.2 et remarque 6.1)**.** *Soit $\eta > 0$, $K > 1$, et $(x_m)_{m\in\mathcal{M}} \in \mathbb{R}_+^{\mathcal{M}}$ tel que $\Sigma \triangleq \sum_{m\in\mathcal{M}} e^{-x_m} < \infty$. Fixons $\mathrm{pen}^{(\eta)} : \mathcal{M} \to \mathbb{R}_+$ telle que*

$$\forall m \in \mathcal{M}, \quad \mathrm{pen}^{(\eta)}(m) \geqslant \frac{K\sigma^2}{n}\left(\sqrt{D_m} + \sqrt{2x_m}\right)^2 + \frac{x_m}{\eta}\ .$$

Alors, pour une constante $C_K > 1$ dépendant uniquement de K, l'estimateur $\widetilde{s}^{(\eta)} = \sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)}\widehat{s}_m$ défini en (1.10) – (1.11) vérifie, pour tout $s \in \mathbb{R}^n$ et tout $z > 0$, avec probabilité au moins égale à $1 - \Sigma^2 e^{-z}$,

$$\left\|\widetilde{s}^{(\eta)} - s\right\|_n^2 \leqslant C_K \inf_{m\in\mathcal{M}}\left\{d^2(s, S_m) + \mathrm{pen}^{(\eta)}(m) + \frac{\ln\Sigma}{\eta} + \frac{\sigma^2}{n}(z+1)\right\} - \mathcal{J}(\widehat{\rho}^{(\eta)})\ , \quad (1.13)$$

où $\mathcal{J}(\rho) \triangleq \sum_{m\in\mathcal{M}} \rho_m \|\widehat{s}_m - s\|_n^2 - \left\|\sum_{m\in\mathcal{M}} \rho_m \widehat{s}_m - s\right\|_n^2 \geqslant 0$.

En intégrant la borne précédente, on peut en déduire une borne en espérance. Le théorème 6.2 fournit une autre borne en espérance un peu plus fine, de type PAC-bayésien ; cf. (6.21).

Le théorème précédent pointe un lien naturel entre agrégation de modèles et sélection de modèles : notre inégalité de type oracle est valide pour un continuum d'estimateurs $\big\{\widetilde{s}^{(\eta)} : \eta > 0\big\}$ qui s'étend de l'agrégation de modèles classique (où η est au plus de l'ordre de n/σ^2) à la sélection de modèles (où $\eta = +\infty$).

En particulier, notre estimateur agrégé $\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} \widehat{s}_m$ converge presque sûrement vers l'estimateur sélectionné $\widehat{s}_{\widehat{m}}$ de [Mas07] quand $\eta \to \infty$ (si $\widehat{m}$ est unique), et on retrouve la même borne de risque que celle de [Mas07, théorème 4.18] lorsqu'on passe à la limite dans (1.13) quand $\eta \to +\infty$ (cf. corollaire 6.1). Par ailleurs, pour η assez grand (au moins de l'ordre de n/σ^2), la borne du théorème 1.8 est du même ordre de grandeur que celle de [Mas07, théorème 4.18]. Les mêmes applications que celles traitées par [Mas07, MM11] peuvent donc être considérées : par ex., modèles linéaires, modèles finis, ellipsoïdes de Besov, boules ℓ^1. On en traite quelques-unes en section 6.4.1.

Nous n'avons pas encore eu le temps d'étudier en détails si l'agrégation possède de meilleures performances que la sélection de modèles pour des modèles non linéaires classiques comme ceux cités précédemment. En revanche, la borne du théorème 1.8 suggère que cela puisse être le cas à cause de la présence du terme positif $\mathcal{J}(\widehat{\rho}^{(\eta)})$, qui est une différence dans une inégalité de Jensen. Une autre motivation en faveur de l'agrégation est que, même dans le cas simple de modèles linéaires, il existe des situations de fort biais pour lesquelles l'agrégation est plus robuste que la sélection en termes d'excès de risque. On prouve ainsi en proposition 6.1 le fait suivant : il existe une collection de modèles linéaires $(S_m)_{m\in\mathcal{M}}$ avec $|\mathcal{M}| = 2$ telle que, pour tout $n \geqslant 16/(\sqrt{2}-1)^2$,

$$\forall s \in \mathbb{R}^n, \quad \mathbb{E}_s\left[\left\| \sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} \widehat{s}_m - s \right\|_n^2 \right] \leqslant \inf_{m\in\mathcal{M}} \mathbb{E}_s\left[\left\| \widehat{s}_m - s \right\|_n^2 \right] + \frac{4\ln(2)\,\sigma^2}{n}\,, \tag{1.14}$$

$$\forall \widehat{m}, \quad \exists s \in \mathbb{R}^n, \quad \mathbb{E}_s\left[\left\| \widehat{s}_{\widehat{m}} - s \right\|_n^2 \right] \geqslant \inf_{m\in\mathcal{M}} \mathbb{E}_s\left[\left\| \widehat{s}_m - s \right\|_n^2 \right] + \frac{\sigma^2}{4\sqrt{n}}\,, \tag{1.15}$$

où $\widehat{\rho}_m^{(\eta)}$ est défini en (1.11) avec $\eta = n/(4\sigma^2)$ et $\mathrm{pen}^{(\eta)}(m) = 2\dim(S_m)\sigma^2/n$ (il s'agit de l'estimateur de [LB06] avec la plus grande température inverse autorisée, cf. (1.12)), et où (1.15) est valide pour toute fonction de sélection $\widehat{m} : \mathbb{R}^n \to \mathcal{M}$ mesurable en les données.

Les modèles linéaires $S_1, S_2 \subset \mathbb{R}^n$ et le vecteur $s \in \mathbb{R}^n$ exhibés dans la preuve de (1.15) sont tels que les estimateurs des moindres carrés $\widehat{s}_1$ et $\widehat{s}_2$ associés à S_1 et S_2 possèdent un fort biais (de l'ordre de la variance du bruit σ^2), ont un risque $\mathbb{E}_s\big[\| \widehat{s}_m - s \|_n^2 \big]$ proche et sont suffisamment séparés l'un de l'autre. Les deux bornes (1.14) et (1.15) ci-dessus indiquent que l'estimateur agrégé de [LB06] a un excès de risque au plus de l'ordre de $1/n$ uniformément en s, alors que toute méthode de sélection de modèles encourt dans au moins une situation de fort biais (et telle que décrite précédemment) un excès de risque au moins de l'ordre de $1/\sqrt{n}$. En ce sens, l'agrégation de modèles est plus robuste que la sélection de modèles.

La borne inférieure précédente est prouvée avec des modèles linéaires, mais sa simplicité suggère que l'agrégation de modèles pourrait bénéficier d'une propriété de robustesse similaire pour des modèles non linéaires classiques ; cette question ouverte sera abordée prochainement.

1.5.4 Travaux futurs

Comme mentionné précédemment, ce chapitre est un travail en cours. En particulier, d'importantes questions restent ouvertes :

- Nos inégalités de type oracle ont une constante devant l'infimum strictement supérieure à 1. Est-ce une conséquence de l'approche par concentration — qui donne en revanche des

bornes avec grande probabilité — ou de la généralité des modèles ? En particulier, quand les modèles sont linéaires, il pourrait être intéressant de retrouver via une analyse unifiée les bornes plus fines de [LB06] et de [BM07a] obtenues respectivement pour l'agrégation et la sélection de modèles.

- La question importante de la calibration du paramètre η est ouverte. Est-il possible d'identifier — au moins pour des problèmes classiques — un choix optimal de η ? Si tel est le cas, peut-on calibrer η de façon automatique et quasi-optimale ?

- Enfin, l'étude d'exemples classiques de modèles non linéaires (par ex., ellipsoïdes de Besov, boules ℓ^1, réseaux de neurones) pourrait permettre de mieux comparer la procédure de sélection de modèles de [Mas07] avec les méthodes d'agrégation.

1.6 Perspectives de recherche dans la droite lignée des travaux de cette thèse

Ces travaux de thèse soulèvent plusieurs questions que nous projetons d'aborder par la suite ; nous les présentons brièvement ci-après. Ces problèmes sont à la frontière entre l'apprentissage séquentiel de suites individuelles et l'apprentissage dans un cadre statistique plus classique.

Régression linéaire séquentielle parcimonieuse

Au chapitre 3, nous importons la notion d'*inégalité oracle de sparsité* dans un cadre de suites déterministes arbitraires et traitons des problèmes d'adaptativité (dans un cadre déterministe dans un premier temps, puis, en corollaire, dans un cadre statistique classique). Ces résultats pourraient être prolongés de la façon suivante.

Peut-on modifier l'algorithme séquentiel SeqSEW pour produire des combinaisons linéaires parcimonieuses ? Le prédicteur séquentiel SeqSEW construit au chapitre 3 vérifie des bornes de regret de sparsité, mais ses prévisions séquentielles $\widehat{y}_t$ ne sont en général pas – au premier abord du moins – des combinaisons linéaires parcimonieuses des prévisions de base, i.e., des prévisions de la forme $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t$ avec $\|\widehat{\boldsymbol{u}}_t\|_0 \ll T$. En fait, les prévisions de l'algorithme SeqSEW ont une forme un peu plus élaborée car elles font intervenir l'opérateur de troncature avant le mélange convexe ; elles sont en effet de la forme :

$$\widehat{y}_t \triangleq \int_{\mathbb{R}^d} \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big]_{B_t} \, p_t(\mathrm{d}\boldsymbol{u}) \; . \tag{1.16}$$

En grande dimension, produire des prévisions qui sont des combinaisons linéaires parcimonieuses des prévisions de base pourrait pourtant être utile d'un point de vue statistique (à des fins de sélection de variables) et algorithmique (pour diminuer l'espace mémoire nécessaire). On pourrait envisager de modifier notre prédicteur SeqSEW en remarquant que la probabilité a priori $\pi_\tau(\mathrm{d}\boldsymbol{u})$ choisie sur $\mathbb{R}^d$ (et donc, dans une moindre mesure, les probabilités a posteriori $p_t(\mathrm{d}\boldsymbol{u})$ associées) charge(nt) davantage les combinaisons linéaires $\boldsymbol{u}$ approximativement parcimonieuses. Dans le modèle de régression avec *design* fixe, [DT09] remarquent ainsi sur des simulations que leur algorithme exponentiel sélectionne correctement les variables pertinentes (pourvu qu'une troncature

raisonnable soit appliquée aux composantes de la combinaison linéaire produite) ; voir [DT09, section 5.2.1]. Dans notre cadre séquentiel, on pourrait envisager d'étudier si de telles propriétés sont vraies (d'un point de vue pratique ou théorique) pour une modification appropriée de l'algorithme SeqSEW. Autrement dit, peut-on approcher les prévisions $\widehat{y}_t$ définies par (1.16) par des prévisions de la forme $\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t$ avec $\|\widehat{\boldsymbol{u}}_t\|_0$ petit (par ex., $\|\widehat{\boldsymbol{u}}_t\|_0 \ll t$) et telles que la perte cumulée encourue soit proche ?

Peut-on prouver des bornes de parcimonie pour des algorithmes séquentiels parcimonieux ? Une autre piste de recherche, actuellement en cours, consiste à tenter de prouver des bornes de parcimonie pour des algorithmes séquentiels dont on sait qu'ils produisent des combinaisons parcimonieuses. Un exemple de tel algorithme est donné par une variante séquentielle de l'estimateur Lasso [Tib96, DJ94a] ; cette variante produit la prévision $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t$, où $\widehat{\boldsymbol{u}}_t$ est donné par

$$\widehat{\boldsymbol{u}}_t \in \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} (y_s - \boldsymbol{u}\cdot\boldsymbol{x}_s)^2 + \lambda \|\boldsymbol{u}\|_1 \right\}$$

pour un paramètre de régularisation $\lambda > 0$ à calibrer judicieusement. En plus de produire des combinaisons parcimonieuses, l'algorithme précédent a l'avantage de pouvoir être implémenté[18] avec un coût algorithmique faible. Cela contraste ainsi avec notre prédicteur théorique SeqSEW, qui pourrait certes être approché numériquement par des méthodes de Langevin Monte-Carlo étudiées par [DT09] dans le cadre stochastique, mais qui ne jouit pour l'instant pas de garanties théoriques quant à la précision de cette approximation.

Notons qu'on pourrait également tenter de prouver des bornes de parcimonie pour les algorithmes séquentiels de [LLZ09, SST09, Xia10, DSSST10] mentionnés en section 1.2.1.

Regret interne

Comme précisé en section 1.4.3, plusieurs questions relatives au regret interne et au regret *swap* sont encore ouvertes. Ainsi, la question du facteur logarithmique manquant $\sqrt{\ln K}$ entre les bornes inférieure et supérieure des regrets interne et *swap* nécessite sans doute des techniques plus fines que celles utilisées jusqu'à présent. On décrit en section 5.6 des pistes de majoration ou de minoration.

Par ailleurs, la technique stochastique développée à la fin du chapitre 5 est utile d'un point de vue théorique (puisqu'elle permet de majorer le regret minimax), mais elle n'est pas constructive. Nous souhaiterions donc nous pencher sur la construction d'algorithmes $(\boldsymbol{p}_t)_{t\geqslant 1}$ explicites (et efficaces) atteignant les bornes supérieures nouvellement prouvées, par exemple pour le regret *makespan*, dont on a majoré la valeur minimax par une quantité de l'ordre de $\sqrt{T \ln K}$. Une construction générique traitant d'emblée le regret généralisé défini en section 5.5.1 (lequel inclut regrets externe, interne, *swap* et *makespan*) serait idéale.

Agrégation de modèles non linéaires

Le dernier chapitre de la thèse présente des travaux en cours sur l'agrégation de modèles non linéaires. Comme précisé en section 1.5.4, plusieurs questions importantes sont encore ouvertes. Nous projetons ainsi d'étudier la possibilité d'obtenir des inégalités de type oracle *exactes* pour

[18] De surcroît, une implémentation du type LARS [EHJT04] permet de calculer le chemin entier de régularisation, ce qui est utile à des fins de calibration.

des modèles S_m non linéaires (au moins sur des exemples classiques comme, par exemple, les ellipsoïdes de Besov et les réseaux de neurones). Ces exemples pourraient permettre de mieux comparer les procédures d'agrégation de modèles et de sélection de modèles.

Enfin, nous souhaiterions aborder la question – cruciale en pratique – de la calibration du paramètre η. Il s'agira probablement d'étudier l'existence d'une calibration optimale – au moins sur des exemples classiques – puis de chercher à imiter cette calibration à l'aide des données seulement. On pourrait par exemple mêler des arguments de calibration séquentielle proches de ceux du chapitre 3 et des idées propres à l'heuristique de pente introduite par [BM07a, AM09].

Autres liens entre suites individuelles et sélection de modèles

Dans le cadre de la prévision avec un nombre fini d'avis d'experts (cf. figure 1.2 avec $\Theta = \{1, \dots, K\}$), et pour des fonctions de perte convexes et bornées, on connaît depuis plus d'une décennie des procédures d'agrégation optimales au sens minimax. En particulier, le regret dans le pire des cas des stratégies optimales correspondantes ne peut pas être amélioré (même d'un facteur multiplicatif, cf. remarque 2.3 page 61). En revanche, les travaux plus récents de [FS97, ACBG02, ANN04, CBMS07, HK08] ont montré qu'il existe des algorithmes qui, dans des cas favorables (donc loin du pire des cas considéré pour la quantité minimax), possèdent des performances bien meilleures. Les bornes associées ont été qualifiées de bornes du premier ou second ordre (on en présente une introduction en section 2.2.2).

Dans ce cadre, un problème encore ouvert – formulé par [CBMS07] – consiste en l'obtention de bornes de regret de type oracle, i.e., des bornes de regret du second ordre qui sont un analogue séquentiel des inégalités de type oracle en sélection de modèles. Plus précisément, pour des fonctions de perte $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ bornées et convexes en leur premier argument, il s'agirait de prouver des bornes de regret de la forme

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) \leqslant \min_{1 \leqslant i \leqslant K} \left\{ \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) + \gamma_1 \sqrt{Q_{i,T} \ln K} \right\} + \gamma_2 \, E \ln K \;, \tag{1.17}$$

où $\gamma_1, \gamma_2 > 0$ sont des constantes, où $E \triangleq \max_{1 \leqslant t \leqslant T} \max_{1 \leqslant i,j \leqslant K} \big| \ell(a_{i,t}, y_t) - \ell(a_{j,t}, y_t) \big|$ désigne l'étendue des pertes jusqu'à la date T, et où $Q_{i,T}$ est une quantité du second ordre, par exemple, $Q_{i,T} = \sum_{t=1}^{T} \ell^2(a_{i,t}, y_t)$ ou, mieux, un terme de variance empirique

$$Q_{i,T} = \sum_{t=1}^{T} \big(\ell(a_{i,t}, y_t) - \mu_{i,T} \big)^2 \;, \quad \text{avec} \quad \mu_{i,T} \triangleq \frac{1}{T} \sum_{t=1}^{T} \ell(a_{i,t}, y_t) \;.$$

Une borne de la forme (1.17) permettrait de réaliser un compromis de type biais-variance entre les experts : la perte cumulée $\sum_{t=1}^{T} \ell(a_{i,t}, y_t)$ du i-ème expert joue le rôle d'une erreur d'approximation, alors que la quantité $\gamma_1 \sqrt{Q_{i,T} \ln K}$ est une mesure de la difficulté séquentielle d'estimation (et joue donc le rôle d'un terme de variance).

Les exemples de quantités du second ordre $Q_{i,T}$ mentionnés ci-dessus furent introduits par [CBMS07, HK08], mais ces deux travaux ne prouvent une borne de la forme (1.17) qu'au prix d'une très forte connaissance a priori sur la suite des données à prévoir (en l'occurrence, pour obtenir la borne (1.17), il convient de calibrer leurs algorithmes en fonction de la quantité $Q_{i_T^*,T}$, où i_T^* réalise le minimum dans (1.17) ; en l'absence d'un tel a priori, leurs bornes sont plus faibles). Notre objectif est donc de prouver une borne du type (1.17) pour un algorithme n'utilisant pas un tel fort a priori. Il est vraisemblable que de nouvelles techniques de calibration soient nécessaires.

Chapter 2

Mathematical introduction

This chapter is a mathematical introduction to the content of this manuscript. We present the basics of the theory of prediction of individual sequences, some of its connections with the stochastic setting, and explain the main motivations under the notion of sparsity oracle inequalities in the stochastic setting. Part of the material below is based on the monograph [CBL06] as well as on recent lectures given by Gilles Stoltz at Paris-Sud XI University (cf. [Sto10b]) and by Peter Bartlett at IHP (cf. [Bar11]).

Contents

2.1 Introduction

In this thesis we study sequential prediction problems that can all be cast into the following setting. A decision-maker – or forecaster – has to predict in a sequential fashion the values of an unknown sequence $y_1, y_2, \dots$ of elements of an outcome space $\mathcal{Y}$. His decisions $\widehat{a}_t$ – or predictions – belong to a decision space $\mathcal{D}$, which we assume to be a convex subset of a vector space. Even if the case when $\mathcal{D} = \mathcal{Y}$ is easier to interpret, $\mathcal{D}$ may be different from $\mathcal{Y}$. The prediction task is sequential: the outcomes are only revealed one after another; at time t, the forecaster guesses the next outcome y_t right before it is revealed.

In the classical statistical theory of sequential prediction, some stochastic assumptions are made on the way the sequence $y_1, y_2, \dots$ is generated. For example, it may be assumed to be the realization of an ergodic stationary process. Such assumptions enable to sequentially estimate the properties of the underlying stochastic process and therefore to design statistical methods that work well when the statistical model properly describes the data at hand. This however may be unrealistic in practical problems where the process is hard to model from a statistical viewpoint and may even react to the forecaster's decisions – the last situation occurs, e.g., in computer security and computational finance.

The theory of prediction of individual sequences addresses the sequential prediction problem from a quite different angle. No stochastic assumptions whatsoever are made on the sequence of outcomes $y_1, y_2, \dots$ to be predicted. Therefore, all outcome sequences are considered and we look for prediction methods that are robust in the sense that they work well even in the worst case. The name *individual sequences* comes from the fact that performance guarantees are proved for any arbitrary deterministic sequence $y_1, y_2, \dots \in \mathcal{Y}$.

Without any stochastic model, it is not immediately clear how the prediction problem can be made meaningful and which goals are reasonable. One popular possibility is to measure the performance of a decision-maker by the loss he has accumulated in the long run, where the losses are scored by a loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$. The goal of the forecaster is to minimize his cumulative loss $\sum_{t=1}^{T} \ell(\widehat{a}_t, y_t)$.

Since no stochastic assumptions are made on the outcome sequence, a classical approach consists in comparing the forecaster's performance to that of reference forecasters – also called *experts*. Namely, we assume that at each time t, the forecaster has access to base forecasts $a_{\theta,t} \in \mathcal{D}$, $\theta \in \Theta$, where Θ is a fixed index set — the $a_{\theta,t}$ are called the *experts' predictions* or *expert advice*. Then we look for methods that guarantee that the regret

$$\sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) - \inf_{\theta \in \Theta} \sum_{t=1}^{T} \ell(a_{\theta,t}, y_t)$$

is small uniformly over all outcome sequences $y_1, y_2, \dots \in \mathcal{Y}$ and all expert advice sequences $(a_{\theta,1})_{\theta\in\Theta}, (a_{\theta,2})_{\theta\in\Theta}, \dots \in \mathcal{D}^{\Theta}$. Further comments on the regret are made in Section 2.1.2 below.

The expert advice can be of quite different nature. They can correspond to statistical methods designed under different assumptions on the underlying stochastic process. Minimizing the regret

above then ensures that, in the long run, the predictions of the forecaster are almost as good as that of the method associated with the unknown best statistical model (in this respect, regret minimization can be seen as a meta-statistical problem). The expert advice can also be truly deterministic predictions based on scientific modelling. Such situations occur, e.g., in daily ozone forecasting, where the experts may be numerical simulations computed from chemico-physical PDE models (electricity consumption forecasting is another fruitful example). Even worse, the experts can also be malicious opponents that react to the forecaster's decisions – as, e.g., in finance or in spam email detection problems. Since the prediction guarantees of the forecasting methods presented in the sequel hold uniformly over all individual sequences, all the examples above can be handled by the theory at hand.

A few notations

Throughout this chapter, $\mathbb{N} = \{0, 1, \ldots\}$ and $\mathbb{N}^* \triangleq \{1, 2, \ldots, \}$ denote the sets of nonnegative and positive integers respectively, and $e \triangleq \exp(1)$ denotes Euler's number. Vectors are denoted by bold letters. Additional notations will be stated explicitely when necessary.

2.1.1 Prediction with expert advice: main framework

The problem of prediction with expert advice mentioned in the introductory paragraphs can be formulated as a repeated game between the forecaster and the environment; see Figure 2.1.

Parameters: convex decision space $\mathcal{D}$, outcome space $\mathcal{Y}$, loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$, and set Θ of expert indices.

At each time round $t \in \mathbb{N}^*$,

1. the environment chooses the expert advice $a_{\theta,t} \in \mathcal{D}$ for all $\theta \in \Theta$; they are revealed to the forecaster;
2. the forecaster chooses a point $\widehat{a}_t \in \mathcal{D}$, which may be kept secret or revealed[a] to the environment;
3. the environment chooses and reveals the outcome $y_t \in \mathcal{Y}$;
4. the forecaster incurs the loss $\ell\big(\widehat{a}_t, y_t\big)$ and each expert $\theta \in \Theta$ incurs the loss $\ell\big(a_{\theta,t}, y_t\big)$.

[a] If the environment does not have access to the forecaster's predictions $\widehat{a}_t$, then we say that it is *oblivious* to the forecaster's predictions. But if the environment can react to the forecaster's past moves, then we say that it is *adversarial*. See Section 2.3.1 for further comments.

Figure 2.1: Prediction with expert advice.

This formulation as a repeated game is convenient to make clear all the dependences between the quantities at hand. For example, the prediction $\widehat{a}_t \in \mathcal{D}$ of the forecaster at time t is a function of the past expert advice $\big(a_{\theta,s}\big)_{\theta\in\Theta} \in \mathcal{D}^\Theta$ and outcomes $y_s \in \mathcal{Y}$, $1 \leqslant s \leqslant t-1$, and of the current expert advice $\big(a_{\theta,t}\big)_{\theta\in\Theta} \in \mathcal{D}^\Theta$. More formally, in this setting, we call *strategy of the forecaster* any sequence $(\widehat{a}_t)_{t\geqslant 1}$ of functions $\widehat{a}_t : \big(\mathcal{D}^\Theta \times \mathcal{Y}\big)^{t-1} \times \mathcal{D}^\Theta \to \mathcal{D}$. Though we most often omit

these dependences for notational convenience, keeping them in mind is crucial to properly define the problem and the associated optimal performance guarantees such as the minimax regret (see Section 2.3).

Next we give some examples of decision spaces, outcome spaces, and loss functions that have been extensively studied in the online prediction protocol of Figure 2.1.

Example 2.1. *Typical examples of loss functions include (note that $\mathcal{D}$ and $\mathcal{Y}$ may be different):*

- *the* square loss *for bounded outcomes and predictions, which corresponds to $\mathcal{D} = \mathcal{Y} = [-B, B]$ (for some $B > 0$) and $\ell(a, y) = (y - a)^2$;*
- *the* relative entropy loss*, which corresponds to $\mathcal{D} = \mathcal{Y} = [0, 1]$ and $\ell(a, y) = y \ln(y/a) + (1 - y) \ln\big((1 - y)/(1 - a)\big)$; this loss is called the* logarithmic loss *when $\mathcal{Y} = \{0, 1\}$;*
- *the* Hellinger loss*, which corresponds to $\mathcal{D} = \mathcal{Y} = [0, 1]$ and $\ell(a, y) = (1/2)(\sqrt{a} - \sqrt{y})^2 + (1/2)(\sqrt{1 - a} - \sqrt{1 - y})^2$;*
- *the* absolute loss*, which corresponds to $\mathcal{D} = [0, 1]$, $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = [0, 1]$, and $\ell(a, y) = |y - a|$.*
- *the* linear loss *(or* mixture loss*), which corresponds to $\mathcal{D} = \mathcal{X}_K$, $\mathcal{Y} = [0, 1]^K$ (for some $K \in \mathbb{N}^*$), and $\ell(\boldsymbol{a}, \boldsymbol{y}) = \sum_{i=1}^K a_i y_i$, where $\mathcal{X}_K$ denotes the simplex of order K:*

$$\mathcal{X}_K \triangleq \left\{ \boldsymbol{x} \in \mathbb{R}_+^K : \sum_{i=1}^K x_i = 1 \right\} . \tag{2.1}$$

2.1.2 A performance criterion: the (external) regret

Since the sequences of outcomes and expert advice can be totally arbitrary, it is in general unrealistic for the forecaster to try to incur at each time t the smallest possible loss $\inf_{a \in \mathcal{D}} \ell(a, y_t)$ or even the loss $\inf_{\theta \in \Theta} \ell(a_{\theta,t}, y_t)$ of the best expert at time t (which may change at each round t). However, in the long run (i.e., if the forecaster and the experts are scored through their *cumulative* loss), predicting almost as well as the best *fixed* expert in hindsight is a realistic goal. Stated otherwise, it corresponds to minimizing the *regret*

$$\sum_{t=1}^T \ell(\widehat{a}_t, y_t) - \inf_{\theta \in \Theta} \sum_{t=1}^T \ell(a_{\theta,t}, y_t) ,$$

where the time horizon T may be known or unknown to the forecaster. Regret is sometimes called *external regret* in contrast with other forms of regret such as internal and swap regrets (the last two performance criteria are studied in Chapter 5). It can be thought of as the regret that the forecaster feels after T time steps for not following the advice of the best expert in hindsight.

The notion of regret can be interpreted as an estimation error in the statistical terminology. As

noted in [Sto10a], the cumulative loss of the forecaster up to time T can be decomposed as

$$\sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) = \underbrace{\inf_{\theta\in\Theta} \sum_{t=1}^{T} \ell(a_{\theta,t}, y_t)}_{\sim \text{ approximation error}} + \underbrace{\sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) - \inf_{\theta\in\Theta} \sum_{t=1}^{T} \ell(a_{\theta,t}, y_t)}_{\sim \text{ estimation error}} \ .$$

The first term is an online counterpart of an *approximation error* given by the cumulative loss incurred by the best expert in hindsight, while the second quantity — the regret — is an online counterpart of an *estimation error*, which measures the difficulty of the forecaster to mimic the best expert in hindsight while being compelled to output predictions in a sequential fashion.

To minimize his cumulative loss, the forecaster should control both the approximation and the estimation error terms. Minimizing the approximation error is an important problem both in pratice and in theory: the experts should be carefully chosen for their approximation properties (they can be, e.g., statistical estimators associated to different functional bases with various approximation properties, or numerical simulations associated with different physical models or different numerical approximation schemes). In this thesis, we focus on the other quantity — the estimation error — and study methods whose regret is small uniformly over all outcome sequences $y_1, y_2, \ldots \in \mathcal{Y}$ and all expert advice sequences $(a_{\theta,1})_{\theta\in\Theta}, (a_{\theta,2})_{\theta\in\Theta}, \ldots \in \mathcal{D}^{\Theta}$.

When the loss function ℓ is nonnegative and bounded, the regret grows at most linearly in the number of time rounds T. Therefore, a first reasonable goal is to ensure a sublinear regret, i.e., to guarantee a vanishing worst-case per-round regret

$$\sup_{\substack{y_1,\ldots,y_T \\ (a_{\theta,1})_\theta,\ldots,(a_{\theta,T})_\theta}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \ell(\widehat{a}_t, y_t) - \inf_{\theta\in\Theta} \frac{1}{T} \sum_{t=1}^{T} \ell(a_{\theta,t}, y_t) \right\} \leqslant o(1) \qquad \text{as} \quad T \to +\infty \ ,$$

where the supremum is taken over all outcome sequences $y_1, y_2, \ldots$ in $\mathcal{Y}$ and over all expert advice sequences $(a_{\theta,1})_{\theta\in\Theta}, \ldots, (a_{\theta,T})_{\theta\in\Theta}$ in $\mathcal{D}^{\Theta}$ (the outcomes and the expert advice can be chosen adversarially, see Section 2.3.1). The above guarantee indicates that, on the average, the forecaster predicts almost as well as the best fixed expert in hindsight. As we show in Section 2.2, when Θ is finite with cardinality $|\Theta|$, typical rates of the per-round regret are $\sqrt{(\ln|\Theta|)/T}$ and $(\ln|\Theta|)/T$.

The fact that the forecaster's and experts' predictions are scored through their cumulative losses contrasts with the stochastic batch setting (where the forecaster is given an i.i.d. sample $(y_1, \ldots, y_T)$ from an unknown distribution, as, e.g., in the regression model with random design). In the latter setting, the performance of a statistical predictor constructed on $(y_1, \ldots, y_T)$ are instead assessed on a new outcome $y_{T+1} \in \mathcal{Y}$. The usual criterion to be minimized is the *risk*, i.e., the expected loss of the predictor on the next outcome (where the expectation is with respect to the distribution of the outcome). The two performance criteria – cumulative loss or risk – are of different nature. However, a close connection exists between them: we explain in Section 2.5 how to convert an online forecaster into a method suitable for a stochastic batch setting.

2.1.3 On the use of randomization

In this thesis all the strategies of the forecaster that we consider are deterministic. This is sufficient since both the decision space $\mathcal{D}$ and the functions $\ell(\cdot, y_t)$ are assumed to be convex. Such assumptions are natural in many applications: online linear regression, classification with the absolute loss, sequential probability assignement, online portfolio optimization, among other examples. Next we mention important situations where these convexity assumptions are however not satisfied; in such cases, randomization is useful since it is a way to convexify the problem.

An example where $\mathcal{D}$ is not convex is given by the binary classification problem, where $\mathcal{D} = \mathcal{Y} = \{0, 1\}$ and where $\ell(a, y) = \mathbb{I}_{\{a \neq y\}}$. Assume that at each time $t = 1, \dots, T$, the forecaster is given two expert advice: $a_{1,t} = 0$ and $a_{2,t} = 1$. Note that whatever the forecaster plays, there is always an individual sequence $y_1, \dots, y_T \in \{0, 1\}$ such that $\sum_{t=1}^T \ell(\widehat{a}_t, y_t) = T$ (it is given by $y_t = 1 - \widehat{a}_t$). Since in addition one of the two experts predicts correctly for at least half of the rounds, we get that $\min_{1 \leqslant i \leqslant 2} \sum_{t=1}^T \ell(a_{i,t}, y_t) \leqslant T/2$. Therefore, the regret of the forecaster is lower bounded by

$$\sum_{t=1}^T \ell(\widehat{a}_t, y_t) - \min_{1 \leqslant i \leqslant 2} \sum_{t=1}^T \ell(a_{i,t}, y_t) \geqslant \frac{T}{2} \,.$$

Thus, in this very simple but non-convex setting, a regret sublinear in T cannot be achieved. Had the decision space (and the loss function) been convex, e.g., had we considered $\widetilde{\mathcal{D}} = [0, 1]$ and $\ell(a, y) = |a - y|$ instead, then the forecaster would have been allowed to choose his decisions $\widehat{a}_t$ as convex combinations of the expert advice $a_{1,t}$ and $a_{2,t}$ (since $[a_{1,t}, a_{2,t}] \subset \widetilde{\mathcal{D}}$). Such weighted average predictions are key to get a sublinear regret (see the next sections), but are forbidden in the non-convex setting $\mathcal{D} = \{0, 1\}$ described above.

A way to compensate for the lack of convexity of $\mathcal{D}$ or $\ell(\cdot, y_t)$ is to resort to a randomized strategy: at each time t, the forecaster chooses a probability distribution p_t on Θ, draws an expert index $\widehat{\theta}_t \in \Theta$ at random from p_t, and outputs the decision $\widehat{a}_t = a_{\widehat{\theta}_t, t} \in \mathcal{D}$. The environment has access to p_t before choosing the outcome y_t, but only gets to see the decision $\widehat{a}_t$ after revealing y_t (contrary to the prediction protocol of Figure 2.1). This way, even if the environment uses the knowledge of p_t to react to the forecaster's decisions, the forecaster can counteract the environment's possible diabolic movements thanks to randomization; in many applications, it yields a regret sublinear in T with high probability. This is essentially because the conditional expected loss

$$\mathbb{E}\big[\ell(\widehat{a}_t, y_t) \,|\, \widehat{a}_1, \dots, \widehat{a}_{t-1}\big] = \int_\Theta \ell(a_{\theta,t}, y_t)\, p_t(\mathrm{d}\theta)$$

is linear and thus convex in the probability distribution p_t: randomization is a way to convexify the problem. (Unsurprisingly, many randomized strategies are based on the ideas presented in this chapter for deterministic prediction with convex decision spaces and loss functions; the variability introduced by randomization is then handled via martingale concentration inequalities.)

The setting of *randomized prediction with expert advice* (with finite Θ) has been extensively studied since the seminal works of [Bla56] and [Han57]; see, eg, [FMG92, FV99, CBL99] and [CBL06, Chapter 4] for a thorough introduction. More recently it has been analysed under various restrictions on the information available to the forecaster. Well-known problems include:

- bandit games, where the forecaster has only access to the loss $\ell(a_{\widehat{\theta}_t,t}, y_t)$ of his own decision, but not to that of the other experts $a_{\theta,t}$, $\theta \neq \widehat{\theta}_t$; see, e.g., [Rob52, ACBF02, ACBFS02, AB09, BMSS11] and [CBL06, Chapter 6] for a detailed overview;
- label-efficient prediction, where the forecaster has only access to the outcomes y_t at a small number of rounds; see, eg, [HP97, CBLS05];
- sequential prediction under partial monitoring, where the forecaster does not have access to the past outcomes y_t but only to a feedback signal; see, e.g., [Rus99, CBLS06, LMS08].

We also refer the reader to [AB10] for a detailed account on minimax strategies under (combinations of) some of the above restrictive assumptions.

In this thesis, we only consider the full information setting: at the beginning of each time round $t \geqslant 1$, the whole history $\big((a_{\theta,1})_{\theta\in\Theta}, y_1\big), \ldots, \big((a_{\theta,t-1})_{\theta\in\Theta}, y_{t-1}\big)$ is available to the forecaster. Moreover, since we focus on cases where both the decision space $\mathcal{D}$ and the functions $\ell(\cdot, y_t)$ are convex, it is enough to consider only deterministic strategies.

However, even if our setting and our strategies are deterministic, we sometimes use randomization for the sake of mathematical analysis. Combined with Fano's lemma or Pinsker's inequality, randomization indeed turns out to be useful to derive lower bounds on the minimax regret (cf. Section 2.3 for the external regret and Chapter 5, Section 5.4 for the swap regret). We also use ideas based on randomization to derive upper bounds (cf. Chapter 4, Section 4.2 where we use a Maurey-type argument and Chapter 5, Section 5.5 where we restrict our attention to Bernoulli losses through a simple randomization argument).

2.2 Prediction with expert advice

In this section we present some basic results on the theory of prediction with expert advice. We consider the prediction protocol of Figure 2.1. For the sake of clarity, we assume thereafter that Θ is finite. Therefore, up to a one-to-one relabelling, we have $\Theta = \{1, \ldots, K\}$ for some $K \in \mathbb{N}^*$. For all $t \geqslant 1$, we index the expert advice $a_{i,t}$ with $i \in \{1, \ldots, K\}$.

In the sequel, for all $t \geqslant 1$ and all $i \in \{1, \ldots, K\}$, we denote by $\ell_{i,t} \triangleq \ell(a_{i,t}, y_t)$ the loss of expert i at time t and by $L_{i,t} \triangleq \sum_{s=1}^{t} \ell_{i,s}$ its cumulative loss up to time t (by convention, we also set $L_{i,0} \triangleq 0$).

In the next subsections, we focus on the celebrated exponentially weighted average forecaster and its refined variants. This forecaster benefits from interesting properties both in the online and stochastic settings. For other algorithms belonging to the more general family of weighted average forecasters or related to more specific problems, we refer the reader to the monograph [CBL06].

2.2.1 The exponentially weighted average forecaster

Next we recall one of the most famous algorithms in prediction with expert advice called the *exponentially weighted average forecaster*. In machine learning theory this algorithm was introduced by [LW94] and [Vov90].

The statement of this algorithm is given in Figure 2.2. Note that the initial weight vector $\boldsymbol{p}_1 = (1/K, \ldots, 1/K)$ is the uniform weight vector and that the weight $p_{i,t}$ assigned to expert i at each time $t \geqslant 2$ is a smooth nonincreasing function of its past cumulative loss $L_{i,t-1} \triangleq \sum_{s=1}^{t-1} \ell_{i,s}$.

Parameter: $\eta > 0$.

Initialization: $L_{i,0} \triangleq 0$ for all $i = 1, \ldots, K$.

At each time round $t \geqslant 1$,

1. Receive the expert advice $a_{1,t}, \ldots, a_{K,t} \in \mathcal{D}$;
2. Compute the weight vector $\boldsymbol{p}_t = \big(p_{i,t}\big)_{1 \leqslant i \leqslant K} \in \mathcal{X}_K$ defined by
$$p_{i,t} \triangleq \frac{e^{-\eta L_{i,t-1}}}{\sum_{j=1}^K e^{-\eta L_{j,t-1}}}\,, \quad 1 \leqslant i \leqslant K\,.$$
3. Output the prediction $\widehat{a}_t \triangleq \sum_{i=1}^K p_{i,t} a_{i,t} \in \mathcal{D}$;
4. Receive the outcome $y_t \in \mathcal{Y}$, compute the instantaneous losses $\ell_{i,t} \triangleq \ell(a_{i,t}, y_t)$, and update the cumulative losses $L_{i,t} \triangleq L_{i,t-1} + \ell_{i,t}$ for all $i = 1, \ldots, K$.

Figure 2.2: The exponentially weighted average forecaster.

The next theorem bounds the regret of the exponentially weighted average forecaster when the loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ is bounded and convex in its first argument. It is based on the work of [CB99] and can be found, e.g., in [CBL06, Theorem 2.2] for $[B_1, B_2] = [0, 1]$. See also [CBFH+97, CBL99] for the particular case of binary prediction with the absolute loss.

Theorem 2.1. *Assume that the loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ is convex in its first argument and takes its values in $[B_1, B_2]$ for some constants $B_1 < B_2 \in \mathbb{R}$. Then, for all $T \in \mathbb{N}^*$ and all $\eta > 0$, and for all sequences of expert advice $a_{i,t} \in \mathcal{D}$ and of outcomes $y_t \in \mathcal{Y}$, the regret of the exponentially weighted average forecaster with fixed parameter η is upper bounded by*

$$\sum_{t=1}^T \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^T \ell\big(a_{i,t}, y_t\big) \leqslant \frac{\ln K}{\eta} + \frac{\eta T (B_2 - B_1)^2}{8}\,.$$

This bound is minimized at $\eta = (B_2 - B_1)^{-1}\sqrt{8(\ln K)/T}$ and becomes $(B_2 - B_1)\sqrt{(T/2)\ln K}$.

Proof: We set $W_t \triangleq (1/K)\sum_{i=1}^K e^{-\eta L_{i,t-1}}$ for all $t = 1, \ldots, T+1$ (recall that $L_{i,0} \triangleq 0$ for all $i = 1, \ldots, K$ by convention, so that $W_1 = 1$). Next we bound the key quantity $\ln(W_{T+1}/W_1)$ from below and above. On the one hand,

$$\ln\left(\frac{W_{T+1}}{W_1}\right) = \ln\left(\sum_{i=1}^K e^{-\eta L_{i,T}}\right) - \ln K \geqslant \ln\left(\max_{1 \leqslant i \leqslant K} e^{-\eta L_{i,T}}\right) - \ln K = -\eta \min_{1 \leqslant i \leqslant K} L_{i,T} - \ln K\,. \tag{2.2}$$

On the other hand, we can rewrite $\ln(W_{T+1}/W_1)$ as a telescopic sum and get

$$\ln\left(\frac{W_{T+1}}{W_1}\right) = \sum_{t=1}^{T} \ln\left(\frac{W_{t+1}}{W_t}\right) = \sum_{t=1}^{T} \ln\left(\frac{\sum_{i=1}^{K} e^{-\eta L_{i,t-1}} e^{-\eta \ell_{i,t}}}{\sum_{i=1}^{K} e^{-\eta L_{i,t-1}}}\right) = \sum_{t=1}^{T} \ln\left(\sum_{i=1}^{K} p_{i,t} e^{-\eta \ell_{i,t}}\right), \tag{2.3}$$

where we used the definition of $p_{i,t}$ in Figure 2.2. But, by Hoeffding's lemma (see Lemma A.4 in Appendix A.5) and by the fact that the loss function ℓ is $[B_1, B_2]$-valued, we get that, for all $t = 1, \dots, T$,

$$\ln\left(\sum_{i=1}^{K} p_{i,t} e^{-\eta \ell_{i,t}}\right) \leqslant -\eta \sum_{i=1}^{K} p_{i,t} \ell_{i,t} + \frac{\eta^2 (B_2 - B_1)^2}{8}.$$

Substituting the last inequality in (2.3), we get

$$\ln\left(\frac{W_{T+1}}{W_1}\right) \leqslant -\eta \sum_{t=1}^{T} \sum_{i=1}^{K} p_{i,t} \ell_{i,t} + \frac{\eta^2 T (B_2 - B_1)^2}{8}. \tag{2.4}$$

Combining the last inequality with (2.2) and dividing by η yields

$$\sum_{t=1}^{T} \sum_{i=1}^{K} p_{i,t} \ell_{i,t} - \min_{1 \leqslant i \leqslant K} L_{i,t} \leqslant \frac{\ln K}{\eta} + \frac{\eta T (B_2 - B_1)^2}{8}.$$

We conclude the proof by noting that $\ell\big(\widehat{a}_t, y_t\big) \leqslant \sum_{i=1}^{K} p_{i,t} \ell\big(a_{i,t}, y_t\big) = \sum_{i=1}^{K} p_{i,t} \ell_{i,t}$ for all $t = 1, \dots, T$ (by definition of $\widehat{a}_t$ and by convexity of ℓ in its first argument). □

The above theorem shows that for all loss functions that are bounded and convex in their first argument, the regret of the exponentially weighted average forecaster is at most of order $\sqrt{T \ln K}$. The next theorem shows that if the loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ is exp-concave in its first argument (which implies convexity, but not necessarily boundedness), then the regret of the same forecaster is at most of the order of $\ln K$ (with a properly chosen η).

More precisely, following Appendix A.2, we say that a loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ is η_0-exp-concave in its first argument for some $\eta_0 > 0$ if the function $a \mapsto e^{-\eta_0 \ell(a,y)}$ is concave on $\mathcal{D}$ for all $y \in \mathcal{Y}$.

Among the loss functions listed in Example 2.1 above, the following are η_0-exp-concave: the square loss on $[-B, B] \times [-B, B]$ (with $\eta_0 = 1/(8B^2)$), the relative entropy loss (with $\eta_0 = 1$), and the Hellinger loss (with $\eta_0 = 1$). On the contrary, the linear loss and the absolute loss are not η_0-exp-concave for any value of $\eta_0 > 0$ and therefore do not satisfy the assumptions of the following theorem. We refer the reader to [Vov98, Vov01] and [HKW98, KW99] for further details on exp-concavity. Finally, note that though exp-concavity implies convexity (cf. Appendix A.2), it does not necessarily imply boundedness (e.g., the relative entropy is not bounded).

Theorem 2.2. *Assume that for some $\eta_0 > 0$, the loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ is η_0-exp-concave in its first argument. Then, for all $T \geqslant 1$, the exponentially weighted average forecaster tuned with any $\eta \in (0, \eta_0]$ satisfies, uniformly over all sequences of expert advice $a_{i,t} \in \mathcal{D}$ and of outcomes $y_t \in \mathcal{Y}$,*

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \leqslant \frac{\ln K}{\eta}.$$

The above theorem is due to [KW99] (though stated in a slightly different form). See also [CBL03] for an analysis in terms of the exponential potential $\boldsymbol{u} \in \mathbb{R}^K \mapsto \eta^{-1} \ln\bigl(\sum_{i=1}^K e^{\eta u_i}\bigr)$. The following proof is however closer in spirit to [Vov01].

Proof: We follow the same lines as for Theorem 2.1, i.e., we start from (2.2) and (2.3) and simply replace the call to Hoeffding's lemma by a sharper argument. Indeed, we upper bound the right-hand side of (2.3) by noting that, for all $t = 1, \dots, T$,

$$\sum_{i=1}^K p_{i,t} e^{-\eta \ell\left(a_{i,t}, y_t\right)} \leqslant \exp\left(-\eta\, \ell\left(\sum_{i=1}^K p_{i,t} a_{i,t}, y_t\right)\right) = \exp\Bigl(-\eta\, \ell\bigl(\widehat{a}_t, y_t\bigr)\Bigr)\,, \tag{2.5}$$

where the inequality follows by concavity of $a \mapsto e^{-\eta \ell(a, y_t)}$ on $\mathcal{D}$ (since, by assumption, ℓ is η_0-exp-concave in its first argument and therefore η-exp-concave for all $\eta \in (0, \eta_0]$; see Appendix A.2).

Taking the logarithms of both sides of the last inequality, summing it over $t = 1, \dots, T$, and combining it with (2.2) and (2.3), we get

$$-\eta \min_{1 \leqslant i \leqslant K} L_{i,T} - \ln K \leqslant -\eta \sum_{t=1}^T \ell\bigl(\widehat{a}_t, y_t\bigr)\,.$$

Dividing the last inequality by η and rearranging terms, we conclude the proof. □

Remark 2.1 (The Aggregating Algorithm)**.** *In the proof above, the only place where the particular form of $\widehat{a}_t = \sum_{i=1}^K p_{i,t} a_{i,t}$ is used is in (2.5). In particular, the bound of Theorem 2.2 would have remained true for any $\widehat{a}_t \in \mathcal{D}$ such that (2.5) holds, i.e., such that*

$$\forall y_t \in \mathcal{Y}\,, \qquad \ell\bigl(\widehat{a}_t, y_t\bigr) \leqslant -\frac{1}{\eta} \ln\left(\sum_{i=1}^K p_{i,t} e^{-\eta \ell\left(a_{i,t}, y_t\right)}\right)\,.$$

Any algorithm that outputs predictions $\widehat{a}_t \in \mathcal{D}$ satisfying the above inequality is called an aggregating algorithm (see [Vov90, Vov98, Vov01]). An example is given by the exponentially weighted average forecaster when the loss function ℓ is η-exp-concave in its first argument. But for some loss functions, such an algorithm may exist even for values of $\eta > 0$ for which the loss function is not η-exp-concave (e.g., for the square loss on $[-B, B]$, there is an aggregating algorithm with $\eta = 1/(2B^2)$ while the square loss is only $1/(8B^2)$-exp concave). The existence of an aggregating algorithm is ensured by a weaker assumption than exp-concavity that is called mixability *in [CBL06, Sections 3.5 and 3.6]; see [Vov90, Vov98, HKW98, KW99, Vov01] for further details.*

In the subsequent chapters, we could sometimes directly address the more general aggregating algorithm instead of studying only the exponentially weighted average forecaster. This is the case, e.g., in Chapter 3 where, for the square loss, we could also use the aggregating algorithm to get similar bounds (with actually a leading constant better by a factor of 4, and without any additional difficulties). We however chose to focus on the exponentially weighted average forecaster for its popularity, its wide use in practice, its nice theoretical performance, and the various parameter tunings that have already been proposed so far.

2.2.2 Parameter tuning techniques

The value of the parameter η minimizing the upper bound of Theorem 2.1 depends on the range $B_2 - B_1$ and on the time horizon T, which may be unknown in practice. Similarly, Theorem 2.2 suggests to take $\eta = \eta_0$, which may depend on quantities that are not known beforehand. For example, the square loss on $\mathcal{D} \times \mathcal{Y} = [-B, B]^2$ is $1/(8B^2)$-exp-concave in its first argument (see Appendix A.2); in this case, $\eta_0 = 1/(8B^2)$ depends on the possibly unknown range B of the outcomes and the expert advice.

Next we introduce tuning techniques to choose η in an adaptive way, i.e., that do not require any a priori knowledge on the data to be predicted (or less knowledge), while still ensuring regret bounds of the same order of magnitude.

The doubling trick

We consider the setting of Theorem 2.1; we also assume for the moment that the range $B_2 - B_1$ is known beforehand. A way to adapt to the unknown time horizon T is the so-called *doubling trick*, whose first precise analysis in machine learning theory can probably be dated back to [CBFH+97] and [Vov98].

The idea underlying the doubling trick is to partition time into periods (or *regimes*) of exponentially increasing lengths. Traditionnally we take regimes with doubling lengths, i.e., time intervals of the form $\{2^r, \dots, 2^{r+1} - 1\}$, $r \in \mathbb{N}$. At the beginning of each regime r, the exponentially weighted average forecaster is run with η tuned optimally as a function of the length of the period (i.e., $\eta = (B_2 - B_1)^{-1}\sqrt{8(\ln K)/2^r}$). When the regime ends, the algorithm is re-initialized[1] and run on the next regime with a new value of $\eta = (B_2 - B_1)^{-1}\sqrt{8(\ln K)/2^{r+1}}$.

Theorem 2.3 (Adaptation to T via a doubling trick in T)**.**
Assume that the loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ is convex in its first argument and takes its values in $[B_1, B_2]$ for some known constants $B_1 < B_2 \in \mathbb{R}$. Then, the doubling version of the exponentially weighted average forecaster described above satisfies, for all $T \in \mathbb{N}^*$ and for all choices of the experts' predictions $a_{i,t} \in \mathcal{D}$ and outcomes $y_t \in \mathcal{Y}$,

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \leqslant \frac{\sqrt{2}}{\sqrt{2}-1}\,(B_2 - B_1)\sqrt{\frac{T}{2}\ln K} + (B_2 - B_1)\,.$$

A version of the above theorem can be found, e.g., in [Sto05, Theorem 2.2]. As mentioned therein, the classical choice of 2 for the regimes' lengths ratio is not optimal but close to the optimum. Note that it leads to a regret bound whose main term is within a factor of $\sqrt{2}/(\sqrt{2}-1)$ of the best bound of Theorem 2.1.

The main idea underlying the proof of Theorem 2.3 as well as all other applications of the doubling trick is the following. By superadditivity of the minimum, the regret on $\{1, \dots, T\}$ is smaller than the sum of the regrets on each regime $\{2^r, \dots, 2^{r+1} - 1\} \cap [1, T]$ such that $2^r \leqslant T$, i.e., such that $r \leqslant R \triangleq \lfloor \log_2 T \rfloor$. Therefore, by Theorem 2.1, the regret on $\{1, \dots, T\}$ is at most of the order of $\sum_{r=0}^{R} \sqrt{2^r \ln K}$. The last sum is geometric and is therefore of the order of $\sqrt{2^R \ln K}$, which is smaller than $\sqrt{T \ln K}$ by definition of R. This yields the desired result.

[1] In particular, the experts' losses on the past regimes are no longer used and the weight vector of the forecaster is reset to $(1/K, \dots, 1/K)$.

Time-varying tunings

Next we present an improved tuning technique that does not require that the exponentially weighted average forecaster be restarted repeatedly, which is more desirable in practice. This technique consists in tuning the exponentially weighted average forecaster with a time-varying parameter η_t that can depend on t but also on the whole information available to the forecaster at the beginning of the t-th round.

Parameter: sequence of functions[a] $(\eta_t)_{t\geqslant 2}$ where $\eta_t : \big(\mathcal{D}^K \times \mathcal{Y}\big)^{t-1} \times \mathcal{D}^K \to (0, +\infty)$.

Initialization: $L_{i,0} \triangleq 0$ for all $i = 1, \dots, K$.

At each time round $t \geqslant 1$,

1. Access the experts' advice $a_{1,t}, \dots, a_{K,t} \in \mathcal{D}$;
2. Compute the weight vector $\boldsymbol{p}_t = \big(p_{i,t}\big)_{1\leqslant i\leqslant K} \in \mathcal{X}_K$ defined by
$$p_{i,t} \triangleq \frac{e^{-\eta_t L_{i,t-1}}}{\sum_{j=1}^K e^{-\eta_t L_{j,t-1}}}\,, \quad 1 \leqslant i \leqslant K\,.$$
3. Output the prediction $\widehat{a}_t \triangleq \sum_{i=1}^K p_{i,t} a_{i,t} \in \mathcal{D}$;
4. Receive the outcome $y_t \in \mathcal{Y}$, compute the instantaneous losses $\ell_{i,t} \triangleq \ell(a_{i,t}, y_t)$, and update the cumulative losses $L_{i,t} \triangleq L_{i,t-1} + \ell_{i,t}$ for all $i = 1, \dots, K$.

[a] The parameter $\eta_t > 0$ can be chosen as a function of t and of the information available to the forecaster at the beginning of the t-th round.

Figure 2.3: The exponentially weighted average forecaster with time-varying parameter.

The corresponding algorithm is stated in Figure 2.3. The following lemma upper bounds its regret when the sequence $(\eta_t)_{t\geqslant 2}$ is nonincreasing.

Lemma 2.1 (Time-varying parameter)**.**
Let $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ be any loss function and $(\eta_t)_{t\geqslant 2}$ be any nonincreasing sequence of positive real numbers (possibly chosen as a function of the past). Then, the exponentially weighted average forecaster tuned with η_t as in Figure 2.3 satisfies, for all $T \in \mathbb{N}^*$, for all choices of the experts' predictions $a_{i,t} \in \mathcal{D}$ and outcomes $y_t \in \mathcal{Y}$, and for all $\eta_1 \geqslant \eta_2$,

$$\sum_{t=1}^T \ell\big(\widehat{a}_t, y_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^T \ell\big(a_{i,t}, y_t\big) \leqslant \frac{\ln K}{\eta_{T+1}} + \sum_{t=1}^T \frac{1}{\eta_t} \ln\left(\sum_{i=1}^K p_{i,t} e^{-\eta_t\left[\ell_{i,t} - \ell(\widehat{a}_t, y_t)\right]}\right)\,. \tag{2.6}$$

Moreover, if ℓ is convex in its first argument, then, setting $\bar{\ell}_t \triangleq \sum_{i=1}^K p_{i,t}\ell_{i,t}$ for all $t = 1, \dots, T$, the regret can be further upper bounded by

$$\sum_{t=1}^T \ell\big(\widehat{a}_t, y_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^T \ell\big(a_{i,t}, y_t\big) \leqslant \frac{\ln K}{\eta_{T+1}} + \sum_{t=1}^T \frac{1}{\eta_t} \ln\left(\sum_{i=1}^K p_{i,t} e^{-\eta_t\left(\ell_{i,t} - \bar{\ell}_t\right)}\right)\,. \tag{2.7}$$

A general way to use the above lemma is to choose a nonincreasing sequence $(\eta_t)_{t\geqslant 2}$ such that $(\ln K)/\eta_{T+1}$ is of the order the desired bound for all $T \geqslant 1$. As for the remaining sum, note that the logarithm in (2.7) is a log-moment generating function. In the case of general convex and bounded loss functions, it can be upper bounded via Hoeffding's lemma (see Proposition 2.1 below) or via Bernstein's inequality (see Theorem 2.4 below). In the case of exp-concave loss functions, it is preferable to use (2.6) since the corresponding logarithm is nonnegative if $\ell(\cdot, y_t)$ is η_t-exp concave. In Chapter 3 we detail for the square loss how to choose η_t so that $\ell(\cdot, y_t)$ is indeed η_t-exp-concave for most time rounds t (the other ones only accounting for a small regret).

Lemma 2.1 above is essentially due to an argument of [ACBG02], which was then adapted by [CBMS07]. However, it slightly improves on [CBMS07, Lemma 3] in two ways. First, the term $(\ln K)/\eta_{T+1}$ above replaces the quantity $(2/\eta_{T+1} - 1/\eta_1)\ln K$ of [CBMS07]. Our term is always smaller (since $(\eta_t)_{t\geqslant 1}$ is nonincreasing) and can be up to twice as small. See also Remark 2.2 below for a consequence of this fact. Second, the following proof, which is essentially due to [GO07, Lemma 1], is much shorter and simply relies on Jensen's inequality (see also [Aud06, Theorem D.1] for a similar result in the batch stochastic setting under very generic assumptions).

Proof: We adapt the beginning of the proof of Theorem 2.1 to handle the case of a time-varying parameter. More precisely, instead of controlling the telescopic sum $\sum_{t=1}^T \ln(W_{t+1}/W_t)$ and dividing the resulting bounds by η, we directly control $\sum_{t=1}^T \big[(\ln W_{t+1})/\eta_{t+1} - (\ln W_t)/\eta_t\big]$, where $W_t \triangleq (1/K)\sum_{i=1}^K e^{-\eta_t L_{i,t-1}}$ for all $t = 1, \dots, T+1$ (recall that $L_{i,0} \triangleq 0$ for all $i = 1, \dots, K$ by convention, so that $W_1 = 1$). One the one hand, we get as in (2.2) that

$$\frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} = \frac{1}{\eta_{T+1}} \ln\left(\sum_{i=1}^K e^{-\eta_{T+1} L_{i,T}}\right) - \frac{\ln K}{\eta_{T+1}} \geqslant - \min_{1\leqslant i\leqslant K} L_{i,T} - \frac{\ln K}{\eta_{T+1}} \,. \tag{2.8}$$

On the other hand, we can rewrite $(\ln W_{T+1})/\eta_{T+1} - (\ln W_1)/\eta_1$ as a telescopic sum and get

$$\frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} = \sum_{t=1}^T \left(\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W_t}{\eta_t}\right) = \sum_{t=1}^T \Bigg(\underbrace{\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W'_{t+1}}{\eta_t}}_{\triangleq a_t} + \underbrace{\frac{1}{\eta_t} \ln \frac{W'_{t+1}}{W_t}}_{\triangleq b_t} \Bigg) \,, \tag{2.9}$$

where W'_{t+1} is obtained from W_{t+1} by replacing η_{t+1} with η_t, i.e., $W'_{t+1} \triangleq (1/K)\sum_{i=1}^K e^{-\eta_t L_{i,t}}$.

Let $t \in \{1, \dots, T\}$. As noted in [GO07, Lemma 1], the first term a_t is non-positive by Jensen's inequality. Indeed, by concavity of $x \mapsto x^{\eta_{t+1}/\eta_t}$ on $\mathbb{R}_+^*$ (since $0 < \eta_{t+1} \leqslant \eta_t$ by assumption), we get that

$$W_{t+1} \triangleq \frac{1}{K}\sum_{i=1}^K e^{-\eta_{t+1} L_{i,t}} = \frac{1}{K}\sum_{i=1}^K \left(e^{-\eta_t L_{i,t}}\right)^{\eta_{t+1}/\eta_t} \leqslant \left(\frac{1}{K}\sum_{i=1}^K e^{-\eta_t L_{i,t}}\right)^{\frac{\eta_{t+1}}{\eta_t}} \triangleq \left(W'_{t+1}\right)^{\frac{\eta_{t+1}}{\eta_t}} .$$

Taking the logarithms of both sides of the last inequality and dividing it by η_{t+1}, we get that $(\ln W_{t+1})/\eta_{t+1} \leqslant (\ln W'_{t+1})/\eta_t$, so that $a_t \leqslant 0$. As for the second term b_t, we get as in (2.3) that

$$b_t \triangleq \frac{1}{\eta_t} \ln\left(\frac{W'_{t+1}}{W_t}\right) = \frac{1}{\eta_t} \ln\left(\frac{\sum_{i=1}^K e^{-\eta_t L_{i,t-1}} e^{-\eta_t \ell_{i,t}}}{\sum_{i=1}^K e^{-\eta_t L_{i,t-1}}}\right) = \frac{1}{\eta_t} \ln\left(\sum_{i=1}^K p_{i,t} e^{-\eta_t \ell_{i,t}}\right) \,,$$

where the last equality follows by definition of $p_{i,t}$ in Figure 2.3. Therefore, substituting the last upper bounds on a_t and b_t in (2.9), and combining the latter inequality with (2.8), we get that

$$-\min_{1\leqslant i\leqslant K} L_{i,T} - \frac{\ln K}{\eta_{T+1}} \leqslant \sum_{t=1}^{T} \frac{1}{\eta_t} \ln\left(\sum_{i=1}^{K} p_{i,t} e^{-\eta_t \ell_{i,t}}\right) .$$

Adding $\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big)$ to both sides of the last inequality and rearranging terms yields (2.6). As for (2.7), it follows from the fact that $\ell(\widehat{a}_t, y_t) \leqslant \bar{\ell}_t$ by definition of $\widehat{a}_t$ and by convexity of $\ell(\cdot, y_t)$. This concludes the proof. □

The above lemma can be used for several adaptation purposes. Next we derive a result similar to Theorem 2.3 but for the exponentially weighted average forecaster tuned with a time-varying parameter η_t. In view of Theorem 2.1, this parameter is chosen as $\eta_t = (B_2 - B_1)^{-1}\sqrt{c\ln(K)/t}$ for some constant $c > 0$ (we have in mind $c = 8$, but it turns out that $c = 4$ leads to a better bound).

The next proposition is a variant of [CBL06, Theorem 2.3]. Our only modification is that we use Lemma 2.1 instead of [CBMS07, Lemma 3]. This yields an improvement of a $\sqrt{2}$ multiplicative factor (see the comments below).

Proposition 2.1 (Adaptation to T via a time-varying parameter)**.**
Let $c > 0$. Assume that the loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ is convex in its first argument and takes its values in $[B_1, B_2]$ for some known constants $B_1 < B_2 \in \mathbb{R}$. Then, the exponentially weighted average forecaster with time-varying parameter $\eta_t = (B_2 - B_1)^{-1}\sqrt{c\ln(K)/t}$ of Figure 2.3 satisfies, for all $T \in \mathbb{N}^*$ and for all sequences of expert advice $a_{i,t} \in \mathcal{D}$ and of outcomes $y_t \in \mathcal{Y}$,

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \leqslant (B_2 - B_1)\left(\left(\frac{1}{\sqrt{c}} + \frac{\sqrt{c}}{4}\right)\sqrt{T\ln K} + \sqrt{\frac{\ln K}{c}}\right) .$$

This upper bound is approximately minimized with $c = 4$ and becomes $(B_2 - B_1)\sqrt{T\ln K} + (B_2 - B_1)\sqrt{\ln K}/2$.

Note that the main term $(B_2 - B_1)\sqrt{T\ln K}$ of the above regret bound with $c = 4$ is within a mutiplicative factor of $\sqrt{2}$ of the best bound of Theorem 2.1 (which is minimax optimal; cf. Remark 2.3 in Section 2.3). Therefore, adaptation to the unknown time horizon T is possible at the price of a multiplicative factor at most of $\sqrt{2}$. In particular, the above bound improves on Theorem 2.3 obtained via a doubling trick, where the price was a factor of $\sqrt{2}/(\sqrt{2}-1) \approx 3.41$ (more importantly, the forecaster is no longer repeatedly re-initialized, which may lead to better performance in practice). It also improves on the best bound known so far[2] of [CBL06, Theorem 2.3], where the price was a factor of 2.

Proof: By (2.7) in Lemma 2.1, we have

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \leqslant \frac{\ln K}{\eta_{T+1}} + \sum_{t=1}^{T} \frac{1}{\eta_t} \ln\left(\sum_{i=1}^{K} p_{i,t} e^{-\eta_t\left(\ell_{i,t} - \bar{\ell}_t\right)}\right) , \qquad (2.10)$$

[2] We compare existing bounds in the case when $B_2 - B_1$ is known but T is unknown.

where $\bar{\ell}_t \triangleq \sum_{i=1}^K p_{i,t}\ell_{i,t}$ for all $t = 1, \dots, T$. But, as in Theorem 2.1, by Hoeffding's lemma and by the fact that $\ell_{i,t} \in [B_1, B_2]$ by assumption,

$$\sum_{t=1}^T \frac{1}{\eta_t} \ln\left(\sum_{i=1}^K p_{i,t} e^{-\eta_t\left(\ell_{i,t} - \bar{\ell}_t\right)}\right) \leqslant \sum_{t=1}^T \frac{1}{\eta_t} \frac{\eta_t^2 (B_2 - B_1)^2}{8} = \frac{(B_2 - B_1)\sqrt{c \ln K}}{8} \sum_{t=1}^T \frac{1}{\sqrt{t}}$$
$$\leqslant \frac{(B_2 - B_1)\sqrt{c\, T \ln K}}{4},$$

where the first line follows by definition of $\eta_t = (B_2 - B_1)^{-1}\sqrt{c \ln(K)/t}$, and where the last inequality follows from $\sum_{t=1}^T 1/\sqrt{t} \leqslant 2\sqrt{T}$. Substituting the last upper bound in (2.10), we get

$$\sum_{t=1}^T \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^T \ell\big(a_{i,t}, y_t\big) \leqslant (B_2 - B_1)\left(\sqrt{\frac{(T+1)\ln K}{c}} + \frac{\sqrt{c\, T \ln K}}{4}\right)$$
$$\leqslant (B_2 - B_1)\left(\left(\frac{1}{\sqrt{c}} + \frac{\sqrt{c}}{4}\right)\sqrt{T \ln K} + \sqrt{\frac{\ln K}{c}}\right),$$

where we used the upper bound $\sqrt{T+1} \leqslant \sqrt{T} + 1$. This concludes the proof. □

Lemma 2.1 can also be used to adapt simultaneously to the unknown time horizon T and the unknown range $B_2 - B_1$. Time-varying tunings achieving this task have been proposed by at least two papers so far[3]: [ACBG02] and then [CBMS07]. The key idea in both papers is to use a sharper inequality than Hoeffding's lemma to upper bound the log-moment generating function appearing in (2.7). Next we recall the result of [CBMS07] who use a Bernstein-type inequality to upper bound the log-moment generating function.

The most sophisticated tuning of [CBMS07] for the exponentially weighted average forecaster relies on the following two time-varying quantities. First, for all $t \geqslant 1$, the effective range of the losses $\ell_{i,s}$ up to time t is approximated (and upper bounded) by

$$\widehat{E}_t \triangleq \inf\left\{2^k : k \in \mathbb{Z},\ 2^k \geqslant \max_{1 \leqslant s \leqslant t} \max_{1 \leqslant i,j \leqslant K} \big|\ell_{i,s} - \ell_{j,s}\big|\right\}.$$

Second, the authors keep track of the cumulative variance of the forecaster up to time t defined by

$$V_t \triangleq \sum_{s=1}^t \sum_{i=1}^K p_{i,s}\left(\ell_{i,s} - \sum_{j=1}^K p_{j,s}\ell_{j,s}\right)^2.$$

Then, setting $C \triangleq \sqrt{2(\sqrt{2}-1)/(e-2)}$, the time-varying parameter η_t is chosen for all $t \geqslant 2$ as

$$\eta_t \triangleq \min\left\{\frac{1}{\widehat{E}_{t-1}},\ C\sqrt{\frac{\ln K}{V_{t-1}}}\right\}. \tag{2.11}$$

Note that η_t depends on the forecaster's past predictions (through V_{t-1}) and is totally parameter-

[3] Time-varying tunings have also been designed for other frameworks or for other types of algorithms than the exponentially weighted average forecaster: see, e.g., [BHR08, MS10, DHS10] for time-varying tunings in online convex optimization.

free. The next theorem bounds the regret of the corresponding exponentially weighted average forecaster in terms of the cumulative variance V_T.

Theorem 2.4 (Theorem 6 and Corollary 1 of [CBMS07])**.**
Assume that the loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ is convex in its first argument. Then, the exponentially weighted average forecaster with time-varying parameter η_t defined by Figure 2.3 and (2.11) satisfies, for all $T \in \mathbb{N}^*$ and for all sequences of expert advice $a_{i,t} \in \mathcal{D}$ and of outcomes $y_t \in \mathcal{Y}$,

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \leqslant 4\sqrt{V_T \ln K} + 4E \ln K + 6E \,, \tag{2.12}$$

where $E \triangleq \max_{1 \leqslant t \leqslant T} E_t$ is the maximum value of the effective ranges $E_t \triangleq \max_{1 \leqslant i,j \leqslant K} \big|\ell_{i,t} - \ell_{j,t}\big|$. As a consequence, the regret is upper bounded by

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \leqslant 2\sqrt{\left(\sum_{t=1}^{T} E_t^2\right) \ln K} + 4E \ln K + 6E \tag{2.13}$$

$$\leqslant 2E\sqrt{T \ln K} + 4E \ln K + 6E \,. \tag{2.14}$$

Remark 2.2 (A slight improvement in the constants)**.** *The regret bound (2.12) can actually be slightly improved. To do so, it suffices to follow the proof of [CBMS07, Theorem 6] and to use Lemma 2.1 instead of [CBMS07, Lemma 3]. The improvement is in the leading constant: the bound $4\sqrt{V_T \ln K} + 4E \ln K + 6E$ is replaced by*

$$2\sqrt{(e-2)(\sqrt{2}+1)}\sqrt{V_T \ln K} + 2E \ln K + 6E \leqslant 2.64\, E\sqrt{V_T \ln K} + 2E \ln K + 6E \,.$$

Note from (2.14) that if ℓ is convex in its first argument and takes its values in $[B_1, B_2]$, then, without knowing neither T nor $B_2 - B_1$, the above algorithm satisfies the regret bound $(B_2 - B_1)\sqrt{(T/2) \ln K}$ of Theorem 2.1 up to a multiplicative factor[4] of $2\sqrt{2}$ and small remaining terms (since $E \leqslant B_2 - B_1$).

Moreover, the regret bound in (2.12) may improve significantly over the worst-case bound $(B_2 - B_1)\sqrt{(T/2) \ln K}$ of Theorem 2.1. Though the latter is minimax optimal for some loss functions (see Remark 2.3 in Section 2.3), the bound in (2.12) can be much smaller in situations that are more favorable than the worst case, and in particular, when the cumulative variance V_T of the forecaster is small. Note that this property is natural: if the forecaster is confident enough to rapidly concentrate its mass around the experts of smallest losses (which corresponds to a small cumulative variance), then its regret should be small. This is close in spirit to the self-confident forecasters of [ACBG02].

More generally, regret bounds that are minimax optimal (up to multiplicative factors) and that can be significantly smaller than the worst-case bound in some favorable situations are called *re-*

[4] By Remark 2.2 above, the constant $2\sqrt{2} \approx 2.83$ can actually be improved. More precisely, the mutiplicative price to pay for adaptation to T and $B_2 - B_1$ is smaller than $2.64/2 = 1.87$.

fined regret bounds. Following the terminology of [CBMS07], the bound (2.12) is called a *second-order* regret bound (since the main term depends on second-order quantities like the squared losses). Previous refined regret bounds were obtained by [FS97, ACBG02] and [ANN04]. These papers provide *first-order* regret bounds, i.e., regret bounds expressed in terms of the quantities $\sum_{t=1}^{T} |\ell_{i,t}|$, $1 \leqslant i \leqslant K$.

In particular, in the case of nonnegative losses $0 \leqslant \ell_{i,t} \leqslant E$, [FS97] showed that (a properly tuned version of) the exponentially weighted average forecaster satisfies an *improvement for small losses*, i.e., a regret bound of the order of $\sqrt{EL_T^* \ln K} + E \ln K$, where $L_T^* \triangleq \min_{1 \leqslant i \leqslant K} L_{i,T}$ is the smallest cumulative loss up to time T. If L_T^* is much smaller than TE, then the latter regret bound is much smaller than the *zero-order* bound $E\sqrt{T \ln K}$ of Theorem 2.1 (hence the name of the improvement).

An improvement for small losses can actually also be derived from (2.12) above. Indeed, [CBMS07] show in Corollary 3 therein that for nonnegative losses $\ell_{i,t} \in [0, E]$,

$$\sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \leqslant 4\sqrt{EL_T^* \ln K} + 39E \max\{1, \ln K\} \,. \tag{2.15}$$

2.2.3 An online PAC-Bayesian-style analysis

In the last section we mentioned several refined regret bounds satisfied by the exponentially weighted average forecaster when the latter is properly tuned. It turns out that, even for the basic exponentially weighted average forecaster (with constant parameter $\eta > 0$), another type of refinement in the regret bounds is possible — see Proposition 2.2 below. It uses the notion of Kullback-Leibler divergence and resembles risk bounds from the PAC-Bayesian literature.

We first need the following definitions. Given a measurable space $(E, \mathcal{B})$, we denote by $\mathcal{M}_1^+(E)$ the set of all probability distributions on $(E, \mathcal{B})$. Moreover, for all $\rho, \pi \in \mathcal{M}_1^+(E)$, the Kullback-Leibler divergence $\mathcal{K}(\rho, \pi)$ between ρ and π is defined by

$$\mathcal{K}(\rho, \pi) \triangleq \begin{cases} \displaystyle\int_E \ln\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) \mathrm{d}\rho & \text{if } \rho \text{ is absolutely continuous with respect to } \pi; \\ +\infty & \text{otherwise,} \end{cases}$$

where $\frac{\mathrm{d}\rho}{\mathrm{d}\pi}$ denotes the Radon-Nikodym derivative of ρ with respect to π.

Example 2.2. *If E is finite, say $E = \{1, \ldots, K\}$, then the Kullback-Leibler divergence $\mathcal{K}(\boldsymbol{p}, \boldsymbol{q})$ between two elements $\boldsymbol{p} = (p_1, \ldots, p_K)$ and $\boldsymbol{q} = (q_1, \ldots, q_K)$ of the simplex $\mathcal{X}_K$ reads:*

$$\mathcal{K}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{i=1}^{K} p_i \ln\left(\frac{p_i}{q_i}\right) \,,$$

where by convention $0 \ln(0/x) = 0$ for all $x \geqslant 0$ and $x \ln(x/0) = +\infty$ for all $x > 0$.

As recalled in Appendix A.1, the Kullback-Leibler divergence satisfies the following key duality formula (see, e.g., [Cat04, pp. 159–160] for a proof of it): for all functions $h : E \to [a, +\infty)$

lower bounded by some constant $a \in \mathbb{R}$,

$$-\ln \int_E e^{-h} \mathrm{d}\pi = \inf_{\rho \in \mathcal{M}_1^+(E)} \left\{ \int_E h \, \mathrm{d}\rho \, + \, \mathcal{K}(\rho, \pi) \right\} . \tag{2.16}$$

Moreover, the last infimum is achieved at $\rho = \pi_{-h}^{\exp}$, where $\pi_{-h}^{\exp} \in \mathcal{M}_1^+(E)$ is absolutely continuous with respect to π and is given by

$$\mathrm{d}\pi_{-h}^{\exp} \triangleq \frac{e^{-h}}{\int_E e^{-h} \mathrm{d}\pi} \, \mathrm{d}\pi \, . \tag{2.17}$$

The elementary equality (2.16) proves to be very useful in most papers of the PAC-Bayesian literature (see, e.g., the monographs [Cat04, Cat07, Aud04a] and the references therein for the stochastic batch setting; see also [Aud09, Section 4.2] for the online deterministic setting). We use it as a key tool in Chapters 3 and 6 for online and batch purposes respectively.

An improvement over Theorem 2.1

The above duality formula can be used to refine the regret bound of Theorem 2.1 for general convex and bounded loss functions. The next upper bound will be (somewhat abusively) called a *PAC-Bayesian bound*.

Proposition 2.2. *Assume that the loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ is convex in its first argument and takes its values in $[B_1, B_2]$ for some constants $B_1 < B_2 \in \mathbb{R}$. Then, for all $T \in \mathbb{N}^*$ and all $\eta > 0$, and for all sequences of expert advice $a_{i,t} \in \mathcal{D}$ and outcomes $y_t \in \mathcal{Y}$, the exponentially weighted average forecaster with fixed parameter η satisfies*

$$\sum_{t=1}^T \ell\big(\widehat{a}_t, y_t\big) \leqslant \inf_{\boldsymbol{q} \in \mathcal{X}_K} \left\{ \sum_{i=1}^K q_i \sum_{t=1}^T \ell\big(a_{i,t}, y_t\big) + \frac{\mathcal{K}(\boldsymbol{q}, \boldsymbol{p}_1)}{\eta} \right\} + \frac{\eta T (B_2 - B_1)^2}{8} \, ,$$

where $\boldsymbol{p}_1 = (1/K, \ldots, 1/K) \in \mathcal{X}_K$ is the initial weight vector of the forecaster.

The above proposition (together with the next remarks) is essentially due to [FSSW97] (see also [KW99, CB99]), whose analysis is based on a telescopic argument involving the progress $\mathcal{K}(\boldsymbol{q}, \boldsymbol{p}_t) - \mathcal{K}(\boldsymbol{q}, \boldsymbol{p}_{t+1})$ for any vector $\boldsymbol{q} \in \mathcal{X}_K$.

More recently, [Aud09] proved a PAC-Bayesian result on individual sequences for general losses and prediction sets. Combined with Hoeffding's lemma, [Aud09, Theorem 4.6] also yields the above proposition. As in [Aud09], the next proof relies on the duality formula (2.16). Our analysis is however slightly simpler since we only work in a particular case of [Aud09, Theorem 4.6].

Proof: The improvement over Theorem 2.1 appears at the beginning of the proof: instead of lower bounding the sum $\sum_{i=1}^K e^{-\eta L_{i,T}}$ by $\max_{1 \leqslant i \leqslant K} e^{-\eta L_{i,T}}$ in (2.2), we use the duality formula (2.16) with $E = \{1, \ldots, K\}$ and the prior $\boldsymbol{p}_1 = (1/K, \ldots, 1/K) \in \mathcal{X}_K$ to get

$$\ln\left(\frac{W_{T+1}}{W_1}\right) = \ln\left(\frac{1}{K} \sum_{i=1}^K e^{-\eta L_{i,T}}\right) = - \inf_{\boldsymbol{q} \in \mathcal{X}_K} \left\{ \eta \sum_{i=1}^K q_i L_{i,t} + \mathcal{K}(\boldsymbol{q}, \boldsymbol{p}_1) \right\} .$$

Combing the above equality with (2.4) in the proof of Theorem 2.1 and rearranging terms, we conclude the proof. □

The bound of Proposition 2.2 improves on the regret bound of Theorem 2.1. To see this, it suffices to take the Dirac probability distribution $\boldsymbol{q} = \boldsymbol{\delta}_{i^*}$ at $i^* \in \operatorname{argmin}_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big)$ and to use that $\mathcal{K}(\boldsymbol{q}, \boldsymbol{p}_1) = \ln K$ since $\boldsymbol{p}_1 = (1/K, \dots, 1/K)$.

Since the bound of Theorem 2.1 is minimax optimal for some loss functions (see Section 2.3), the improvement of Proposition 2.2 over Theorem 2.1 is not significant in all cases. However, it indicates that the regret term $(\ln K)/\eta$ of Theorem 2.1 can actually be made smaller when several experts have a cumulative loss close to the minimal one. For example, assume that the set $\mathcal{J}^*$ of optimal experts (i.e., the experts whose cumulative loss up to time T is minimal) contains at least $k \geqslant 2$ experts. Then, taking $\boldsymbol{q} = \big(\mathbb{I}_{\{i \in \mathcal{J}^*\}}/k\big)_{1 \leqslant i \leqslant K}$ as the uniform weight vector over $\mathcal{J}^*$, we can see from Proposition 2.2 that the term $\ln(K)/\eta$ can be replaced with the smaller term $\ln(K/k)/\eta$. Therefore, the PAC-Bayesian bound of Proposition 2.2 better reflects the complexity of the family of experts.

Another interesting consequence of the duality formula (2.16) and of the form of the minimizer (2.17) is that, at each time $t \geqslant 1$, the exponentially weighted average forecaster is seen to choose exactly the convex combination $\boldsymbol{p}_t \in \mathcal{X}_K$ that minimizes the upper bound of Proposition 2.2 at time $t-1$, i.e., it satisfies that

$$\boldsymbol{p}_t \in \operatorname*{argmin}_{\boldsymbol{q} \in \mathcal{X}_K} \left\{ \sum_{i=1}^{K} q_i \sum_{s=1}^{t-1} \ell\big(a_{i,s}, y_s\big) + \frac{\mathcal{K}(\boldsymbol{q}, \boldsymbol{p}_1)}{\eta} \right\} .$$

(Put differently, $\boldsymbol{p}_t$ minimizes the linearized past cumulative loss $\boldsymbol{q} \mapsto \sum_{i=1}^{K} q_i L_{i,t-1}$ regularized by the Kullback-Leibler divergence.)

Note also that the above analysis obviously remains the same if we allow the initial vector of the exponentially weighted average forecaster to be arbitrary (instead of $\boldsymbol{p}_1 = (1/K, \dots, 1/K)$). More precisely, for any prior $\boldsymbol{\pi} \in \mathcal{X}_K$, if we define the weights $p_{i,t}$ by

$$p_{i,t} \triangleq \frac{\pi_i e^{-\eta L_{i,t-1}}}{\sum_{j=1}^{K} \pi_j e^{-\eta L_{j,t-1}}} \quad \text{instead of} \quad p_{i,t} = \frac{(1/K) e^{-\eta L_{i,t-1}}}{\sum_{j=1}^{K} (1/K) e^{-\eta L_{j,t-1}}} ,$$

then the bound of Proposition 2.2 remains true with the initial weight vector $\boldsymbol{p}_1 = \boldsymbol{\pi}$.

Other improvements

Following the same lines as in the proof of Proposition 2.2 above, it is also possible to improve the regret bounds of Theorem 2.2 and Lemma 2.1. The resulting improvements consist in replacing in the upper bounds the quantity

$$\min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) + \frac{\ln K}{\eta}$$

by the smaller quantity

$$\inf_{\boldsymbol{q}\in\mathcal{X}_K}\left\{\sum_{i=1}^{K} q_i \sum_{t=1}^{T}\ell\big(a_{i,t},y_t\big)+\frac{\mathcal{K}(\boldsymbol{q},\boldsymbol{p}_1)}{\eta}\right\}.$$

(For Lemma 2.1, the above bound is obtained for $\eta=\eta_{T+1}$.)

Such improvements are not new. The improvement over Theorem 2.2 mentioned above is a consequence of [Aud09, Theorem 4.6]. As for the improvement over Lemma 2.1, a similar result in the stochastic batch setting (and that can be straightforwardly adapted to our deterministic online setting) can be found in [Aud06, Theorem D.1].

We also note that the aforementioned PAC-Bayesian-type upper bounds readily extend to the case where Θ is an arbitrary measurable space (possibly uncountably infinite); we only dealt with the finite case to be consistent with the previous sections.

2.3 Minimax regret

In this section we first define two notions of minimax regret — associated with adversarial or oblivious environments respectively — and show that these quantities are actually equal. We then prove a lower bound on the minimax regret with the linear loss that matches the upper bound of Theorem 2.1.

Let $\mathcal{D}$ be a decision space, $\mathcal{Y}$ be an outcome space, and $\ell:\mathcal{D}\times\mathcal{Y}\to\mathbb{R}$ be any loss function. We consider the same setting as in Section 2.2, i.e., the prediction protocol of Figure 2.1 with a finite set of experts Θ (we use the same notations). We consider the next two definitions, which are associated with adversarial or oblivious environments respectively.

Definition 2.1. *We call* minimax regret with an adversarial environment *the quantity*

$$\inf_{S}\sup_{A}\left\{\sum_{t=1}^{T}\ell\big(\widehat{a}_t,y_t\big)-\min_{1\leqslant i\leqslant K}\sum_{t=1}^{T}\ell\big(a_{i,t},y_t\big)\right\}, \tag{2.18}$$

where the infimum is taken over all strategies $S=(\widehat{a}_t)_{t\geqslant 1}$ of the forecaster and where the supremum is taken over all strategies $A=\big((a_{i,t})_{1\leqslant i\leqslant K},y_t\big)_{t\geqslant 1}$ of the environment. More precisely, the functions $\widehat{a}_t:\big(\mathcal{D}^K\times\mathcal{Y}\big)^{t-1}\times\mathcal{D}^K\to\mathcal{D}$ associate[5] with the past expert advice and outcomes and with the current expert advice the prediction of the forecaster at time t. The experts $a_{i,t}:\mathcal{D}^{t-1}\to\mathcal{D}$ associate to the past predictions of the forecaster their advice at time t. The outcomes are chosen as functions $y_t:\mathcal{D}^t\to\mathcal{Y}$ of the forecaster's past and current predictions. In this case, the environment is said to be adversarial *(i.e., it can react adversarially to the forecaster's predictions).*

[5] We do not consider any dependence between the forecaster's current prediction and its past predictions: this is useless since the forecaster does not randomize and can thus at each time t re-compute all its past predictions (at least theoretically). A similar comment holds for the environment's moves, that only depend on the forecaster's predictions.

Definition 2.2. *We call* minimax regret on individual sequences *the quantity*

$$\inf_{S} \sup_{a_{i,t},y_t} \left\{ \sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^{T} \ell\big(a_{i,t}, y_t\big) \right\} , \tag{2.19}$$

where the infimum is taken over all strategies $S = (\widehat{a}_t)_{t\geqslant 1}$ of the forecaster and where the supremum is taken over all sequences of elements $(a_{i,1})_{1\leqslant i\leqslant K}, \dots, (a_{i,T})_{1\leqslant i\leqslant K} \in \mathcal{D}^K$ and $y_1, \dots, y_T \in \mathcal{Y}$. In this case, the environment does not react to the forecaster's past moves: it is said oblivious *to the forecaster's predictions. The sequences of fixed-in-advance elements $\big((a_{i,t})_{1\leqslant i\leqslant K}, y_t\big)_{t\geqslant 1}$ are called* individual sequences.

2.3.1 On the equivalence between oblivious and adversarial environments

In Proposition 2.3 below, which is now folklore knowledge in the theory of prediction with expert advice, we show that the quantities (2.18) and (2.19) are equal. In other words, in our setting, adversarial environments are not harder to beat than oblivious ones in a minimax sense. (This is no longer true in general when the forecaster is allowed to resort to randomization, see below.)

For notational convenience, we write $\boldsymbol{a}_t = (a_{i,t})_{1\leqslant i\leqslant K}$ for all $t = 1, \dots, T$. In the first claim below, we also write explicitely the dependencies of the predictions $\widehat{a}_t$ of the forecaster on the available data $(\boldsymbol{a}'_s, y'_s)_{s\leqslant t-1}$ and $\boldsymbol{a}'_t$. However, in the second claim, we make some slight abuse of notations by dropping these dependencies for the sake of readability.

Proposition 2.3 (Oblivious and adversarial environments are equivalent in deterministic games)**.** *Consider the prediction protocol of Figure 2.1. Let $S = (\widehat{a}_t)_{t\geqslant 1}$ be any strategy of the forecaster. Then the following claims hold true.*

- *For all strategies $A = \big((a_{i,t})_{1\leqslant i\leqslant K}, y_t\big)_{t\geqslant 1}$ of the environment, the regret of S against A equals the regret of S on a particular individual sequence of expert advice and outcomes $(\boldsymbol{a}'_1, y'_1), \dots, (\boldsymbol{a}'_T, y'_t) \in \mathcal{D}^K \times \mathcal{Y}$, i.e.,*

$$\sum_{t=1}^{T} \ell\Big(\widehat{a}_t\big((\boldsymbol{a}'_s, y'_s)_{s\leqslant t-1}, \boldsymbol{a}'_t\big), y'_t\Big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^{T} \ell\big(a'_{i,t}, y'_t\big) ,$$

  where the quantities $\boldsymbol{a}'_t = (a'_{i,t})_{1\leqslant i\leqslant K} \in \mathcal{D}^K$ and $y'_t \in \mathcal{Y}$ are the values of the functions $(a_{i,t})_{1\leqslant i\leqslant K}$ and y_t evaluated at the forecasters' past predictions (see (2.20) and (2.21)).

- *As a consequence, the worst-case regrets of S against adversarial environments and on individual sequences are equal:*

$$\sup_{A} \left\{ \sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^{T} \ell(a_{i,t}, y_t) \right\} = \sup_{a_{i,t},y_t} \left\{ \sum_{t=1}^{T} \ell\big(\widehat{a}_t, y_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^{T} \ell(a_{i,t}, y_t) \right\} ,$$

  where the first supremum is taken over all strategies of the environment while the second supremum is restricted to the set of all individual sequences (see Definitions 2.1 and 2.2).

- *Therefore, the minimax quantities (2.18) and (2.19) are equal. Their common value may be simply refered to as the* minimax regret.

Proof: The third claim is straightforward, so that we only prove the first two ones.

First claim.
The first claim is just a matter of rewriting things properly, i.e., with all dependencies on the past data. More formally, the sequence $(\boldsymbol{a}'_1, y'_1), \dots, (\boldsymbol{a}'_T, y'_t) \in \mathcal{D}^K \times \mathcal{Y}$ is defined by $\boldsymbol{a}'_1 \triangleq \boldsymbol{a}_1$ and $y'_1 \triangleq y_1\big(\widehat{a}_1(\boldsymbol{a}'_1)\big)$ and, by induction, for all $t \in \{2, \dots, T\}$,

$$a'_{i,t} \triangleq a_{i,t}\Big(\widehat{a}_1(\boldsymbol{a}'_1), \dots, \widehat{a}_{t-1}((\boldsymbol{a}'_s, y'_s)_{s\leqslant t-2}, \boldsymbol{a}'_{t-1})\Big)\,, \quad 1 \leqslant i \leqslant K\,, \tag{2.20}$$

$$y_t \triangleq y'_t\Big(\widehat{a}_1(\boldsymbol{a}'_1), \dots, \widehat{a}_t\big((\boldsymbol{a}'_s, y'_s)_{s\leqslant t-1}, \boldsymbol{a}'_t\big)\Big)\,. \tag{2.21}$$

The first claim then follows by definition of the regret.

Second claim.
Since the set of all strategies of the environment is larger than the set of all individual sequences, we only need to prove that

$$\sup_A \left\{ \sum_{t=1}^T \ell\big(\widehat{a}_t, y_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^T \ell\big(a_{i,t}, y_t\big) \right\} \leqslant \sup_{a_{i,t}, y_t} \left\{ \sum_{t=1}^T \ell\big(\widehat{a}_t, y_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^T \ell\big(a_{i,t}, y_t\big) \right\}. \tag{2.22}$$

For this purpose, let $A = \big((a_{i,t})_{1\leqslant i\leqslant K}, y_t\big)_{t\geqslant 1}$ be any strategy of the environment (i.e., a sequence of functions). But, by the first claim, the regret of S against the environment's strategy A satisfies

$$\sum_{t=1}^T \ell\big(\widehat{a}_t, y_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^T \ell\big(a_{i,t}, y_t\big) \leqslant \sup_{a'_{i,t}, y'_t} \left\{ \sum_{t=1}^T \ell\big(\widehat{a}_t, y'_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^T \ell\big(a'_{i,t}, y'_t\big) \right\},$$

where we made a slight abuse of notation by not writing all dependencies explicitely. The last inequality yields (2.22), which concludes the proof. □

In all subsequent chapters, we only consider individual sequences (i.e., the prediction game is described as if the environment were oblivious to the forecaster's predictions). By the above proposition, this is actually not a restriction and we could just as well assume that the environment were adversarial. Our choice however leads to a simpler presentation.

We stress that this equivalence is due to the fact that we only consider deterministic strategies of the forecaster. Indeed, if the forecaster were allowed to resort to randomization (cf. Section 2.1.3), then its worst-case *expected* regret could differ whether it were computed against adversarial environements or against individual sequences. More formally, if at each time t, the forecaster picks $I_t \in \{1, \dots, K\}$ at random according to a probability distribution $\boldsymbol{p}_t \in \mathcal{X}_K$ built on the past data and predicts as $a_{I_t,t}$, then there are situations for which

$$\sup_A \mathbb{E}\left[\sum_{t=1}^T \ell\big(a_{I_t,t}, y_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^T \ell(a_{i,t}, y_t) \right] > \sup_{a_{i,t}, y_t} \mathbb{E}\left[\sum_{t=1}^T \ell\big(a_{I_t,t}, y_t\big) - \min_{1\leqslant i\leqslant K} \sum_{t=1}^T \ell(a_{i,t}, y_t) \right]$$

where the first supremum is taken over all strategies of the environment[6] while the second supre-

[6] As recalled in Section 2.1.3, in this randomized setting, the environment has access to p_t before choosing the

mum is restricted to the set of all individual sequences.

2.3.2 Lower bound on the minimax regret

Next we show that the upper bound of Theorem 2.1 cannot be improved for the linear loss. In other words, the minimax regret associated with the linear loss is exactly of the order of $\sqrt{T \ln K}$.

More precisely, we consider the prediction protocol of Figure 2.1 with the linear loss function $\ell : \mathcal{X}_K \times [0,1]^K \to \mathbb{R}$ defined by $\ell(\boldsymbol{a}, \boldsymbol{y}) = \sum_{i=1}^K a_i y_i$ (see Example 2.1). In the sequel the standard inner product between $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^K$ is denoted by $\boldsymbol{u} \cdot \boldsymbol{v}$. We also denote by $\boldsymbol{\delta}_i = (\mathbb{I}_{\{j=i\}})_{1 \leqslant j \leqslant K} \in \mathcal{X}_K$ the Dirac probability distribution at $i \in \{1, \dots, K\}$.

Note that ℓ is convex in its first argument and takes its values in $[0,1]$. Therefore, by Theorem 2.1, the minimax regret for the linear loss is at most of $\sqrt{(T/2) \ln K}$. In the next theorem we show that this upper bound cannot be improved by more than a constant factor (see also Remark 2.3 below about the tightness of the constant c_2).

Theorem 2.5 (Minimax lower bound for the linear loss)**.**
There exist two absolute constants $c_1, c_2 > 0$ such that the following holds true. Let $K \geqslant 1$ and $T \geqslant c_1 \ln K$. Consider the prediction protocol of Figure 2.1 with $\mathcal{D} = \mathcal{X}_K$, $\mathcal{Y} = [0,1]^K$, and the linear loss $\ell(\boldsymbol{a}, \boldsymbol{y}) = \boldsymbol{a} \cdot \boldsymbol{y}$. Then, the minimax regret for the linear loss is lower bounded by

$$\inf_S \sup_{\boldsymbol{a}_{i,t}, \boldsymbol{y}_t} \left\{ \sum_{t=1}^T \ell\big(\widehat{\boldsymbol{a}}_t, \boldsymbol{y}_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^T \ell\big(\boldsymbol{a}_{i,t}, \boldsymbol{y}_t\big) \right\} \geqslant c_2 \sqrt{\frac{T}{2} \ln K}\,, \tag{2.23}$$

where the infimum is taken over all strategies of the forecaster and where the supremum is taken over all individual sequences such that $\boldsymbol{a}_{i,t} \in \mathcal{X}_K$ and $\boldsymbol{y}_t \in [0,1]^K$. In particular, we prove the theorem for $c_1 \triangleq 40e/(2e+1) \in [16.8, 16.9]$ and $c_2 \triangleq [2/(2e+1)]\sqrt{e/[5(2e+1)]} \in [0.09; 0.1]$.

The above theorem is essentially due to [CBLS05] and uses techniques of [ACBFS02]. It relies on a probabilistic method: we lower bound the supremum of the regret over all individual sequences $(\boldsymbol{y}_t)_{t \geqslant 1}$ by the expected regret on a suitably chosen i.i.d. random sequence $(\boldsymbol{y}_t)_{t \geqslant 1}$, which is at least of $c_2 \sqrt{(T/2) \ln K}$; see Lemma 2.2 below.

Since the upper bound of Theorem 2.1 was obtained for individual sequences, our lower bound on i.i.d. sequences mentioned above indicates that, surprisingly at first sight, minimizing the regret on individual sequences is just as hard as minimizing it on i.i.d. sequences. As we show in Chapter 5, this property in no longer true for refined notions of regret such as swap regret.

The first lower bounds on the minimax regret were also derived through a probabilistic method but were asymptotic; e.g., for the absolute loss defined in Example 2.1, [CBFH+97] proved that

$$\liminf_{K \to +\infty} \liminf_{T \to +\infty} \left(\frac{1}{\sqrt{(T/2) \ln K}} \inf_S \sup_{\boldsymbol{a}_{i,t}, \boldsymbol{y}_t} \left\{ \sum_{t=1}^T \ell\big(\widehat{\boldsymbol{a}}_t, \boldsymbol{y}_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^T \ell\big(\boldsymbol{a}_{i,t}, \boldsymbol{y}_t\big) \right\} \right) \geqslant 1\,.$$

outcome y_t, but only gets to see the decision $a_{I_t,t}$ after revealing y_t.

Remark 2.3 (On the tightness of the constant $1/\sqrt{2}$)**.**
The asymptotic lower bound above indicates that the constant $1/\sqrt{2}$ of the upper bound of Theorem 2.1 is asymptotically tight for the absolute loss. Since the minimax regret associated with the linear loss is at least as large as that associated with the absolute loss (by convexity), it implies that the constant c_2 in (2.23) can be chosen as close to 1 as desired provided that K and T are large enough.

We refer the reader to [HKW98] for asymptotic lower bounds with other loss functions (via a probabilistic method) and for other deterministic techniques to derive lower bounds on individual sequences (e.g., by induction). We also refer to [CBL99] and [CBL06, Chapter 8] for lower bounds associated with the absolute loss and particular families of experts (via tools from empirical process theory). See also [RST10] for generic lower bounds on the regret (in a randomized game) in terms of combinatorial parameters or sequential Rademacher averages.

Theorem 2.5 is a straightforward consequence of the following lemma, the proof of which is essentially due to [CBLS05, Sto10b] and is postposed to Section 2.A. It is yet another application of Fano's lemma (see Appendix A.7), which has already proved very useful in nonparametric statistics.

Lemma 2.2. *There exist two absolute constants $c_1, c_2 > 0$ such that the following holds true. Let $K \geqslant 1$ and $T \geqslant c_1 \ln K$. Consider the prediction protocol of Figure 2.1 with $\mathcal{D} = \mathcal{X}_K$, $\mathcal{Y} = [0,1]^K$, and the linear loss $\ell(\boldsymbol{a}, \boldsymbol{y}) = \boldsymbol{a} \cdot \boldsymbol{y}$. Then, for constant expert advice given by $\big(\boldsymbol{a}_{i,t}\big)_{t \geqslant 1} = \big(\boldsymbol{\delta}_i\big)_{t \geqslant 1}$ for all $i \in \{1, \dots, K\}$, we have*

$$\inf_S \sup_{(\boldsymbol{Y}_t)_t \text{ i.i.d.}} \mathbb{E}\left[\sum_{t=1}^T \ell\big(\widehat{\boldsymbol{a}}_t, \boldsymbol{Y}_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^T \ell\big(\boldsymbol{\delta}_i, \boldsymbol{Y}_t\big)\right] \geqslant c_2 \sqrt{\frac{T}{2} \ln K}\,, \tag{2.24}$$

where in the last expectation, $\big(\boldsymbol{Y}_t\big)_{1 \leqslant t \leqslant T}$ is an i.i.d. random sequence in $\{0,1\}^K$, and where the supremum is taken over all possible distributions for $\boldsymbol{Y}_1$ (i.e., over all probability distributions on $\{0,1\}^K$). In particular, we prove the lemma with the constants $c_1 \triangleq 40e/(2e+1) \in [16.8, 16.9]$ and $c_2 \triangleq [2/(2e+1)]\sqrt{e/[5(2e+1)]} \in [0.09; 0.1]$.

Proof (of Theorem 2.5): The proof follows straightforwardly from Lemma 2.2. Indeed, for any strategy $S = \big(\widehat{\boldsymbol{a}}_t\big)_{t \geqslant 1}$ of the forecaster, the worst-case regret of S over all individual sequences is lower bounded by

$$\begin{aligned}
&\sup_{\boldsymbol{a}_{i,t}, \boldsymbol{y}_t} \left\{\sum_{t=1}^T \ell\big(\widehat{\boldsymbol{a}}_t, \boldsymbol{y}_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^T \ell\big(\boldsymbol{a}_{i,t}, \boldsymbol{y}_t\big)\right\} \geqslant \sup_{\boldsymbol{y}_t} \left\{\sum_{t=1}^T \ell\big(\widehat{\boldsymbol{a}}_t, \boldsymbol{y}_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^T \ell\big(\boldsymbol{\delta}_i, \boldsymbol{y}_t\big)\right\} \\
&\qquad \geqslant \mathbb{E}\left[\sum_{t=1}^T \ell\big(\widehat{\boldsymbol{a}}_t, \boldsymbol{Y}_t\big) - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^T \ell\big(\boldsymbol{\delta}_i, \boldsymbol{Y}_t\big)\right]\,,
\end{aligned}$$

where in the last expectation, $(\boldsymbol{Y}_t)_{t \geqslant 1}$ is any i.i.d. sequence in $\{0,1\}^K$. We conclude the proof by using Lemma 2.2. □

Remark 2.4. *In the above proof, Lemma 2.2 provides an i.i.d. sequence such that the expected regret is at least of the order of $\sqrt{T \ln K}$. The distribution of $\boldsymbol{Y}_1$ may depend on the strategy of the forecaster (note that the* sup *is after the* inf *in the statement of Lemma 2.2). However, it is also possible to construct explicitely a random sequence $(\boldsymbol{Y}_t)_{1 \leqslant t \leqslant T}$ yielding a similar lower bound but whose distribution is independent of the forecaster (but the $\boldsymbol{Y}_T$ may be no longer i.i.d.). For further details, see Remark 2.5 in Section 2.A.*

Theorem 2.5 indicates that the upper bound of order $\sqrt{T \ln K}$ of Theorem 2.1 cannot be improved uniformly over all convex and bounded loss functions. This does not mean that the rate $\sqrt{T \ln K}$ is minimax optimal for any bounded and convex loss function. For instance, by Theorem 2.2, the minimax regret associated with exp-concave and bounded loss functions (which are in particular convex and bounded) is at most of the order of $\ln K$ and is therefore much smaller. For such losses, lower bounds of the order of $\ln K$ can usually be derived — see, e.g., [HKW98, Theorems 3.19 and 3.22] for the square loss, the relative entropy loss, and the Hellinger loss.

2.4 Online linear regression

In this section we introduce the setting of *online linear regression*, which we study in Chapter 3 under a sparsity scenario and in Chapter 4 for the problem of aggregation over ℓ^1-balls.

In the sequel, $\boldsymbol{u} \cdot \boldsymbol{v}$ denotes the standard inner product between $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, and we set $\|\boldsymbol{u}\|_\infty \triangleq \max_{1 \leqslant j \leqslant d} |u_j|$ and $\|\boldsymbol{u}\|_p \triangleq \left(\sum_{j=1}^d |u_j|^p\right)^{1/p}$ for all $p \in [1, +\infty)$.

2.4.1 Framework

The online linear regression framework, also known as *prediction with side information* under the square loss (cf. [CBL06, Chapter 11]), is a particular case of the framework of prediction with expert advice that unfolds as follows. A forecaster has to predict in a sequential fashion the values $y_t \in \mathbb{R}$ of an unknown sequence of observations given some input data $\boldsymbol{x}_t \in \mathbb{R}^d$. At each time $t \geqslant 1$, on the basis of the newly revealed input data $\boldsymbol{x}_t$ and on the past information $(\boldsymbol{x}_s, y_s)_{1 \leqslant s \leqslant t-1}$, he outputs a prediction $\widehat{y}_t \in \mathbb{R}$, which is finally compared to the new observation y_t through the square loss. A precise description of this repeated game is given in Figure 2.4.

In this setting the goal of the forecaster is to predict almost as well as the best linear forecaster $\boldsymbol{x} \in \mathbb{R}^d \mapsto \boldsymbol{u} \cdot \boldsymbol{x}$, where $\boldsymbol{u} \in \mathbb{R}^d$, i.e., to satisfy, uniformly over all individual sequences $(\boldsymbol{x}_t, y_t)_{1 \leqslant t \leqslant T}$, a regret bound of the form

$$\sum_{t=1}^T \left(y_t - \widehat{y}_t\right)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T \left(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\right)^2 + \Delta_{T,d}(\boldsymbol{u}) \right\} ,$$

for some regret term $\Delta_{T,d}(\boldsymbol{u})$ that should be as small as possible and, in particular, sublinear in T (actually, $\Delta_{T,d}(\boldsymbol{u})$ may also depend on the amplitudes of the individual sequence $(\boldsymbol{x}_t, y_t)_{1 \leqslant t \leqslant T}$ such as $\max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty$ and $\max_{1 \leqslant t \leqslant T} |y_t|$). Sublinearity in T ensures that the regret bound is non trivial: it is indeed easy to ensure a regret of TB_y^2 if the observations all lie in a bounded interval $[-B_y, B_y]$ — this regret is achieved by, e.g., the constant predictions $\widehat{y}_t = 0$. Moreover,

Initial step: the environment chooses a sequence of observations $(y_t)_{t\geqslant 1}$ in $\mathbb{R}$ and a sequence of input data $(\boldsymbol{x}_t)_{t\geqslant 1}$ in $\mathbb{R}^d$ but the forecaster has not access to them.

At each time round $t \in \mathbb{N}^*$,

1. The environment reveals the input data $\boldsymbol{x}_t \in \mathbb{R}^d$.
2. The forecaster chooses a prediction $\widehat{y}_t \in \mathbb{R}$
   (possibly as a linear function of $\boldsymbol{x}_t$, but this is not necessary).
3. The environment reveals the observation $y_t \in \mathbb{R}$.
4. Each linear forecaster $\boldsymbol{x} \mapsto \boldsymbol{u} \cdot \boldsymbol{x}$, for $\boldsymbol{u} \in \mathbb{R}^d$, incurs the loss $\big(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\big)^2$ and the forecaster incurs the loss $(y_t - \widehat{y}_t)^2$.

Figure 2.4: The online linear regression setting.

dividing both sides by T, the above regret bound becomes

$$\frac{1}{T}\sum_{t=1}^{T}\big(y_t - \widehat{y}_t\big)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\frac{1}{T}\sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2 + \frac{\Delta_{T,d}(\boldsymbol{u})}{T}\right\}.$$

Therefore, sublinearity of $\Delta_{T,d}(\boldsymbol{u})$ in T implies that, on the average, the loss of the forecaster is smaller than that of each linear forecaster $\boldsymbol{x} \mapsto \boldsymbol{u}\cdot\boldsymbol{x}$ up to a vanishing remainder term $\Delta_{T,d}(\boldsymbol{u})/T$. The similarity of the above regret bound with risk bounds in the stochastic batch setting is exploited in Section 2.5.

The next two comments are of qualitative nature; they aim to at better comparing the different frameworks considered in this manuscript. Therefore, the reader only interested in online linear regression can skip them and go directly to Section 2.4.2.

A first comment

The setting of Figure 2.4 is a particular case of the prediction protocol of Figure 2.1 (cf. page 41) with decision and outcome spaces[7] $\mathcal{D} = \mathcal{Y} = \mathbb{R}$, with the square loss function $(a, y) \mapsto (y - a)^2$, and with experts indexed by $\Theta = \mathbb{R}^d$ and predicting $a_{\boldsymbol{u},t} = \boldsymbol{u}\cdot\boldsymbol{x}_t$ at each time $t \geqslant 1$ for all $\boldsymbol{u} \in \Theta$.

Another way to cast online linear regression into the prediction protocol of Figure 2.1 is the following. At each round $t \geqslant 1$, the prediction $\widehat{y}_t$ is chosen as a function of the new input $\boldsymbol{x}_t$ and the past data $(\boldsymbol{x}_s, y_s)_{1\leqslant s\leqslant t-1}$, so that the forecaster can be thought of, before observing $\boldsymbol{x}_t$, as choosing a function $\widetilde{f}_t : \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R})^{t-1} \to \mathbb{R}$; its t-th prediction is then given by $\widehat{y}_t = \widetilde{f}_t\big(\boldsymbol{x}_t; (\boldsymbol{x}_s, y_s)_{1\leqslant s\leqslant t-1}\big)$. Therefore, another way to cast online linear regression into the prediction protocol of Figure 2.1 is to consider $\mathcal{Y} = \mathbb{R}^d \times \mathbb{R}$ (the set of all pairs $(\boldsymbol{x}, y)$), $\mathcal{D} = \mathbb{R}^{\mathbb{R}^d}$ (the set of all functions from $\mathbb{R}^d$ to $\mathbb{R}$), $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}$ defined by $\ell\big(f, (\boldsymbol{x}, y)\big) \triangleq \big(y - f(\boldsymbol{x})\big)^2$,

[7] In the sequel, we may restrict $\mathcal{Y}$ to a bounded interval $[-B_y, B_y]$ to emphasize the fact that the performance of the online algorithm under analysis are assessed for bounded observations $y_t \in [-B_y, B_y]$.

$\Theta = \left\{\boldsymbol{x} \in \mathbb{R}^d \mapsto \boldsymbol{u} \cdot \boldsymbol{x} : \boldsymbol{u} \in \mathbb{R}^d\right\}$, and constant experts' advice $a_{\theta,t} = \theta$. This description is closer to works from the stochastic setting. We use a similar description in Section 2.5.2 for the online to batch conversion.

A second comment

The setting of Figure 2.4 is studied in Chapter 4. In Chapter 3 we consider the following generalized variant (more suited for a comparison with the stochastic setting, see Section 2.5.2 and Chapter 3). Instead of repeatedly observing pairs $(\boldsymbol{x}_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$, the forecaster observes input-output pairs $(x_t, y_t) \in \mathcal{X} \times \mathbb{R}$ where $\mathcal{X}$ is an arbitrary measurable set. At the beginning of the game, the forecaster is also given a dictionary $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_d)$ of base forecasters $\varphi_j : \mathcal{X} \to \mathbb{R}$, $1 \leqslant j \leqslant d$ (the φ_j can be, e.g., elements of a suitably chosen functional basis or estimators associated with different statistical models). The goal of the forecaster is then to predict almost as well as the best forecaster $\boldsymbol{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^d u_j \varphi_j$ for $\boldsymbol{u} \in \mathbb{R}^d$. The last setting is clearly a generalization of the prediction protocol of Figure 2.4 (consider the particular case where $\mathcal{X} = \mathbb{R}^d$ and $\boldsymbol{\varphi}$ is the identity function). However, if the input data x_t are only used through the base predictions $\boldsymbol{\varphi}(x_t) \in \mathbb{R}^d$, then the two settings are equivalent.

Among the many papers that addressed the online linear regression framework, the first individual sequence analyses can be dated back to [Fos91, LLW95, CBLW96, KW97]. Next we recall a few basic algorithms in a non-chronological order together with their regret guarantees.

2.4.2 The sequential ridge regression forecaster

In the online linear regression framework described above, we can consider the following online analogue of the ridge regression method that [HK70] introduced in the stochastic setting (linear regression model with fixed design). Its predictions are of the form $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t$, where $\widehat{\boldsymbol{u}}_1 = \mathbf{0} \in \mathbb{R}^d$ and where, at each time $t \geqslant 2$, the linear combination $\widehat{\boldsymbol{u}}_t \in \mathbb{R}^d$ is the solution of the following optimization problem:

$$\widehat{\boldsymbol{u}}_t \in \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} (y_s - \boldsymbol{u} \cdot \boldsymbol{x}_s)^2 + \lambda \|\boldsymbol{u}\|_2^2 \right\} . \tag{2.25}$$

In the above equation, $\lambda > 0$ is a parameter of the algorithm. The regularization term $\lambda \|\boldsymbol{u}\|_2^2$ ensures that the solution $\widehat{\boldsymbol{u}}_t$ is unique and, more importantly, that it cannot be too far away from the null vector $\mathbf{0}$. This shrinking property (or, to see it from an online convex optimization perspective, the strong convexity of $\|\cdot\|_2^2$ with respect to the norm $\|\cdot\|_2$) is important to get non-trivial regret guarantees.

The following theorem was proved by [AW01] via a key telescoping argument involving linear algebra calculations (see also [CBL06, Theorem 11.7], which combines such arguments with other ideas of [For99]).

Theorem 2.6 (Theorem 4.6 of [AW01])**.**
For all $\lambda > 0$, the online algorithm described above satisfies

$$\sum_{t=1}^{T} \big(y_t - \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t\big)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\big)^2 + \lambda \|\boldsymbol{u}\|_2^2 \right\} + 4B^2 d \ln\left(1 + \frac{TB_x^2}{\lambda}\right) ,$$

where $B_x = \max\big\{|x_{j,t}| : 1 \leqslant j \leqslant d,\ 1 \leqslant t \leqslant T\big\}$ and $B = \max\big\{|y_t|, |\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t| : 1 \leqslant t \leqslant T\big\}$.

A drawback of the previous bound is that the quantity B depends on the amplitude of the predictions of the algorithm $|\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t|$. It is not difficult to see that $|\widehat{\boldsymbol{u}}_T \cdot \boldsymbol{x}_T|^2$ can be as large as $d(T-1)/(4\lambda)$, so that the above regret bound only implies regret bounds that have a large dependence in T (e.g., a slow rate $\sqrt{T}$ with λ chosen of the order of $\sqrt{T}$, instead of a fast rate $\ln T$ as in Theorem 2.7 below). Such a situation occurs, e.g., when

$$y_1 = \ldots, y_{T-1} = 1 , \quad \boldsymbol{x}_1, \ldots, \boldsymbol{x}_{T-1} = (\alpha, \ldots, \alpha) \in \mathbb{R}^d ,$$

and

$$y_T = 0 , \quad \boldsymbol{x}_T = (1, \ldots, 1) \in \mathbb{R}^d .$$

Indeed, we can see by (2.25) and by symmetry of the problem that

$$\widehat{\boldsymbol{u}}_T = \frac{(T-1)\alpha}{\lambda + (T-1)d\alpha^2} (1, \ldots, 1) .$$

Therefore, choosing $\alpha = \sqrt{\lambda/\big(d(T-1)\big)}$ ensures that all observations y_t and base predictions $x_{i,t}$ lie in $[-1, 1]$ as soon as $T \geqslant \lambda/d + 1$, and that $(\widehat{\boldsymbol{u}}_T \cdot \boldsymbol{x}_T)^2 = d(T-1)/(4\lambda)$.

Fortunately it turns out that a key modification of the previous algorithm no longer suffers from this drawback. The next algorithm is due to [AW01] and [Vov01]. We call[8] it the *sequential ridge regression forecaster* throughout this manuscript; it should not be confused with the algorithm defined by (2.25). Its predictions are of the form $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t$, where $\widehat{\boldsymbol{u}}_1 = \boldsymbol{0} \in \mathbb{R}^d$ and where, at each time $t \geqslant 2$, the linear combination $\widehat{\boldsymbol{u}}_t \in \mathbb{R}^d$ is the solution of the following optimization problem:

$$\widehat{\boldsymbol{u}}_t \in \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{s=1}^{t-1} (y_s - \boldsymbol{u} \cdot \boldsymbol{x}_s)^2 + (\boldsymbol{u} \cdot \boldsymbol{x}_t)^2 + \lambda \|\boldsymbol{u}\|_2^2 \right\} . \tag{2.26}$$

Note that the modification consists in adding $(\boldsymbol{u} \cdot \boldsymbol{x}_t)^2$. This quantity can be interpreted as the proxy loss at time t of the linear forecaster $\boldsymbol{x} \mapsto \boldsymbol{u} \cdot \boldsymbol{x}$, where the unknown observation y_t is replaced by 0.

Theorem 2.7 (Theorem 5.6 of [AW01] and Theorem 1 of [Vov01])**.**
For all $\lambda > 0$, the sequential ridge regression forecaster defined in (2.26) satisfies

$$\sum_{t=1}^{T} \big(y_t - \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t\big)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\big)^2 + \lambda \|\boldsymbol{u}\|_2^2 \right\} + B_y^2 d \ln\left(1 + \frac{TB_x^2}{\lambda}\right) ,$$

where $B_x = \max\big\{|x_{j,t}| : 1 \leqslant j \leqslant d,\ 1 \leqslant t \leqslant T\big\}$ and $B_y = \max\big\{|y_t| : 1 \leqslant t \leqslant T\big\}$.

[8] This algorithm is also called the *Vovk-Azoury-Warmuth forecaster* in [CBL06, Section 11.8].

The above theorem was proved independently by [AW01] and by [Vov01] via quite different arguments (see also [For99]). The proof of [AW01, Theorem 5.6] uses a key telescopic lemma combined with linear algebra calculations. As for the analysis of [Vov01, Theorem 1], it consists in interpreting the sequential ridge regression forecaster as an aggregating algorithm with continuous weights on $\mathbb{R}^d$ and a Gaussian prior. The regret of this aggregating algorithm is then upper bounded via an analysis close to the online PAC-Bayesian-style analysis carried out in Section 2.2.3. But instead of upper bounding the right-hand side of the duality formula (2.16) via the choice of a suitable probability distribution $\rho \in \mathcal{M}_+^1(\mathbb{R}^d)$, Vovk uses exact calculations in [Vov01, Appendix A.2] to compute the log-moment generating function appearing on the left-hand side.

The optimal a posteriori tuning of the ridge regression forecaster

We end this subsection with a comment on the tuning of the sequential ridge regression forecaster. We explain below that an (ideal) tuning of λ leads to a regret bound that depends logarithmically — as opposed to linearly — in $\|\boldsymbol{u}\|_2^2$. This tuning is first carried out in an illegal way (i.e., it depends the whole data sequence $(\boldsymbol{x}_t, y_t)_{1 \leqslant t \leqslant T}$); we explain in the last two paragraphs of this subsection how to overcome this limitation.

In the sequel we set, for all $\lambda > 0$ and all $\boldsymbol{u} \in \mathbb{R}^d$,

$$B_\lambda(\boldsymbol{u}) \triangleq \sum_{t=1}^T \big(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\big)^2 + \lambda \|\boldsymbol{u}\|_2^2 + B_y^2 d \ln\left(1 + \frac{TB_x^2}{\lambda}\right) ,$$

so that the upper bound of Theorem 2.7 reads $\inf_{\boldsymbol{u} \in \mathbb{R}^d} B_\lambda(\boldsymbol{u})$. To minimize it over $\lambda \in \mathbb{R}_+^*$, we first note that

$$\begin{aligned}\inf_{\lambda > 0} \inf_{\boldsymbol{u} \in \mathbb{R}^d} B_\lambda(\boldsymbol{u}) &= \inf_{\boldsymbol{u} \in \mathbb{R}^d} \inf_{\lambda > 0} B_\lambda(\boldsymbol{u}) \\ &= \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T \big(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\big)^2 + \inf_{\lambda > 0} \left[\lambda \|\boldsymbol{u}\|_2^2 + B_y^2 d \ln\left(1 + \frac{TB_x^2}{\lambda}\right) \right] \right\} .\end{aligned}$$

By elementary calculations (i.e., derivation at the first order), the last infimum over $\lambda > 0$ can be seen to be achieved at $\lambda = \lambda^*(\boldsymbol{u})$, where

$$\lambda^*(\boldsymbol{u}) \triangleq \frac{TB_x^2}{2} \left(-1 + \sqrt{1 + \frac{4dB_y^2}{T \|\boldsymbol{u}\|_2^2 B_x^2}} \right) .$$

Substituting the last expression into the previous equality, we can see that the optimal a posteriori (and therefore illegal) choice of λ leads to the ideal upper bound:

$$\begin{aligned}\inf_{\lambda > 0} \inf_{\boldsymbol{u} \in \mathbb{R}^d} B_\lambda(\boldsymbol{u}) = \inf_{\boldsymbol{u} \in \mathbb{R}^d} \Bigg\{ &\sum_{t=1}^T \big(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\big)^2 + \frac{TB_x^2 \|\boldsymbol{u}\|_2^2}{2} \left(-1 + \sqrt{1 + \frac{4dB_y^2}{T \|\boldsymbol{u}\|_2^2 B_x^2}} \right) \\ &+ B_y^2 d \ln\left(1 + \frac{2}{-1 + \sqrt{1 + (4dB_y^2)/(T \|\boldsymbol{u}\|_2^2 B_x^2)}} \right) \Bigg\} .\end{aligned}$$

The above upper bound does not depend linearly in $\|\boldsymbol{u}\|_2^2$ as in Theorem 2.7 but only logarithmi-

cally when $\|\boldsymbol{u}\|_2 \to +\infty$ (because of the equivalent $\sqrt{1+x} \sim 1 + x/2$ when $x \to 0$). To see it perhaps more simply, we can use the suboptimal but simpler tuning $\widetilde{\lambda}(\boldsymbol{u}) \triangleq B_y^2 d / \|\boldsymbol{u}\|_2^2$ to get the ideal (illegal) bound:

$$\inf_{\lambda>0} \inf_{\boldsymbol{u}\in\mathbb{R}^d} B_\lambda(\boldsymbol{u}) \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} B_{\widetilde{\lambda}(\boldsymbol{u})}(\boldsymbol{u})$$
$$= \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2 + B_y^2 d + B_y^2 d \ln\left(1 + \frac{T B_x^2 \|\boldsymbol{u}\|_2^2}{B_y^2 d}\right) \right\} . \quad (2.27)$$

Therefore, if the linear forecaster $\boldsymbol{u} \in \mathbb{R}^d$ with smallest cumulative loss has a "small" ℓ^2-norm (say, at most of the order of T^γ for some $\gamma > 0$), then the regret on $\mathbb{R}^d$ of the sequential ridge regression forecaster with optimal a posteriori tuning is roughly upper bounded by $dB_y^2 \ln(T)$ — note that a linear dependence in $\|\boldsymbol{u}\|_2^2$ would yield a much worse regret bound. The bound $dB_y^2 \ln(T)$ is used in Chapter 3 to motivate the notion of sparsity regret bound.

As we already mentioned at the beginning of this subsection, the upper bound (2.27) is ideal since the optimal tuning of λ depends on the sequence $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T}$. (To be more precise, to get (2.27), it suffices to choose $\lambda = \widetilde{\lambda}(\widetilde{\boldsymbol{u}}) = B_y^2 d / \|\widetilde{\boldsymbol{u}}\|_2^2$, where $\widetilde{\boldsymbol{u}}$ minimizes the right-hand side of (2.27) over $\boldsymbol{u} \in \mathbb{R}^d$; this tuning however still depends on the data sequence through $\|\widetilde{\boldsymbol{u}}\|_2$.) One way around this is to take a grid $\{U_r = 2^{2^r} : r = 0, 1, \ldots\}$ of $\mathbb{R}_+$ and to associate with each ℓ^2-ball $\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_2 \leqslant U_r\}$ a sequential ridge regression forecaster $\big(\widehat{\boldsymbol{u}}_t^{(r)}\big)_{t\geqslant 1}$ tuned with the quasi-optimal parameter $\lambda_r = B_y^2 d / U_r^2$. Then, by Theorem 2.7, we get, for all $r \geqslant 0$,

$$\sum_{t=1}^T \big(y_t - \widehat{\boldsymbol{u}}_t^{(r)} \cdot \boldsymbol{x}_t\big)^2 \leqslant \inf_{\|\boldsymbol{u}\|_2 \leqslant U_r} \left\{ \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2 + B_y^2 d + B_y^2 d \ln\left(1 + \frac{T B_x^2 U_r^2}{B_y^2 d}\right) \right\} .$$

Using an exponentially weighted average forecaster (with a so-called clipping technique) to combine the sub-algorithms $\big(\widehat{\boldsymbol{u}}_t^{(r)}\big)_{t\geqslant 1}$, we can construct a single algorithm that almost achieves the last upper bounds uniformly over all $r \in \mathbb{N}$, and that therefore satisfies a bound similar to (2.27). The main argument is that the square loss is exp-concave on bounded intervals. For further details, see Chapter 4, Section 4.4, where a similar double mixture is carried out for adaptation purposes.

2.4.3 The Exponentiated Gradient forecaster

Various gradient-based forecasters have been proposed for online linear regression, and, more generally, for online convex optimization: the gradient-descent algorithm [WH60, CBLW96, KW97, CB99], the Exponentiated Gradient forecaster [KW97, CB99], the p-norm algorithms[9] [GL99, Gen03], and unifying forecasters such as the general additive algorithms[10] [WJ98, KW01], the mirror descent algorithm [NY83, BT03], and the composite objective mirror descent algorithm [DSSST10]. Next we recall the basic properties of the Exponentiated Gradient forecaster, which will be used in Chapter 4.

The Exponentiated Gradient forecaster was designed by [KW97] to be competitive against any vector of the simplex $\mathcal{X}_d$ — or, by a simple trick detailed later, of ℓ^1-balls of arbitrary radii.

[9] See also [GLS01] in the classification context.
[10] Same comment.

We present below a generic version of this algorithm suited not only for the square loss but for general convex and differentiable loss functions[11]: at each time t, the forecaster chooses a linear combination $\widehat{\boldsymbol{u}}_t \in \mathbb{R}^d$ (denoted below by $\boldsymbol{p}_t$ since it belongs to $\mathcal{X}_d$), then the environment chooses and reveals a convex and differentiable loss function $\ell_t : \mathbb{R}^d \to \mathbb{R}$, and the forecaster incurs the loss $\ell_t(\widehat{\boldsymbol{u}}_t)$. In online linear regression, the loss functions are given by $\ell_t(\boldsymbol{u}) = (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$.

Let $(\eta_t)_{t \geqslant 2}$ be a sequence of nonnegative parameters that can be chosen in a sequential fashion (strictly speaking, η_t is a function of $\ell_1, \dots, \ell_{t-1}$). Then, the Exponentiated Gradient forecaster tuned with $(\eta_t)_{t \geqslant 2}$ predicts at each time t as $\widehat{y}_t = \boldsymbol{p}_t \cdot \boldsymbol{x}_t$, where the weight vector $\boldsymbol{p}_t = (p_{i,t})_{1 \leqslant i \leqslant d} \in \mathcal{X}_d$ is defined by $\boldsymbol{p}_1 \triangleq (1/d, \dots, 1/d)$, and, for all $t \geqslant 2$, by

$$p_{i,t} \triangleq \frac{\exp\Bigl(-\eta_t \sum_{s=1}^{t-1} \nabla_i \ell_s(\boldsymbol{p}_s)\Bigr)}{\sum_{j=1}^{d} \exp\Bigl(-\eta_t \sum_{s=1}^{t-1} \nabla_j \ell_s(\boldsymbol{p}_s)\Bigr)} \,, \quad 1 \leqslant i \leqslant d \,, \tag{2.28}$$

where $\nabla_i \ell_t(\boldsymbol{u})$ denotes the first-order partial derivative of $\ell_t : \mathbb{R}^d \to \mathbb{R}$ in its i-th variable at the point $\boldsymbol{u}$ — e.g., for the square loss $\ell_t(\boldsymbol{u}) = (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$, we have $\nabla_i \ell_t(\boldsymbol{u}) = -2(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t) x_{i,t}$.

An automatic tuning for the Exponentiated Gradient algorithm with general loss functions

Several regret bounds have been derived for the Exponentiated Gradient algorithm. Originally analysed with the square loss by [KW97], it was later studied for more general loss functions by [CB99]. A general and simple analysis for arbitrary differentiable convex loss functions can also be found in [CBL06, Section 2.5]. The latter analysis probably gives the best intuition on the Exponentiated Gradient algorithm. It relies on the fact that this algorithm is nothing but an exponentially weighted average forecaster applied to the loss vectors $\nabla \ell_t(\boldsymbol{p}_t) \triangleq \bigl(\nabla_i \ell_t(\boldsymbol{p}_t)\bigr)_{1 \leqslant i \leqslant d} \in \mathbb{R}^d$, $t \geqslant 1$. In the next corollary we use the fully automatic exponentially weighted average forecaster of [CBMS07]; it yields an Exponentiated Gradient algorithm for which $(\eta_t)_{t \geqslant 2}$ is tuned in a fully automatic way. More precisely, replacing the losses with the gradients of the losses in (2.11), we set, for all $t \geqslant 2$,

$$\eta_t \triangleq \min\left\{ \frac{1}{\widehat{E}_{t-1}}, C\sqrt{\frac{\ln K}{V_{t-1}}} \right\} \,, \tag{2.29}$$

where $C \triangleq \sqrt{2(\sqrt{2}-1)/(\mathrm{e}-2)}$ and where

$$\widehat{E}_{t-1} \triangleq \inf_{k \in \mathbb{Z}} \left\{ 2^k : 2^k \geqslant \max_{1 \leqslant s \leqslant t-1} \max_{1 \leqslant j,k \leqslant d} \bigl|\nabla_j \ell_s(\boldsymbol{p}_s) - \nabla_k \ell_s(\boldsymbol{p}_s)\bigr| \right\} \,,$$
$$V_{t-1} \triangleq \sum_{s=1}^{t-1} \sum_{j=1}^{d} p_{j,s} \left(\nabla_j \ell_s(\boldsymbol{p}_s) - \sum_{k=1}^{d} p_{k,s} \nabla_k \ell_s(\boldsymbol{p}_s) \right)^2 .$$

The next proposition is a direct consequence of [CBMS07, Corollary 1] (see Theorem 2.4). We denote the gradient of ℓ_t at any $\boldsymbol{u} \in \mathbb{R}^d$ by $\nabla \ell_t(\boldsymbol{u}) \triangleq \bigl(\nabla_1 \ell_t(\boldsymbol{u}), \dots, \nabla_d \ell_t(\boldsymbol{u})\bigr)$. The reason why we do not bound $\|\nabla \ell_t(\boldsymbol{p}_t)\|_\infty^2$ uniformly over all $t = 1, \dots, T$ within the square root will become clear in Corollary 2.2 (application to the square loss).

[11] This corresponds to the online convex optimization setting.

Proposition 2.4 (The Exponentiated Gradient algorithm with automatic tuning)**.**
Assume that the forecaster is repeatedly given a convex differentiable[12] *loss function* $\ell_t : \mathbb{R}^d \to \mathbb{R}$*, and that he uses the Exponentiated Gradient algorithm defined in* (2.28) *with* $(\eta_t)_{t\geqslant 2}$ *given by Equation* (2.29)*. Then, for all* $T \geqslant 1$ *and all sequences* $\ell_1, \dots, \ell_T$*,*

$$\sum_{t=1}^{T} \ell_t(\boldsymbol{p}_t) - \min_{\boldsymbol{q}\in\mathcal{X}_d} \sum_{t=1}^{T} \ell_t(\boldsymbol{q}) \leqslant 4\sqrt{\left(\sum_{t=1}^{T} \|\nabla\ell_t(\boldsymbol{p}_t)\|_\infty^2\right) \ln d} + (8\ln d + 12) \max_{1\leqslant t\leqslant T} \|\nabla\ell_t(\boldsymbol{p}_t)\|_\infty \ .$$

Proof: Since $\ell_t : \mathbb{R}^d \to \mathbb{R}$ is differentiable and convex for all $t = 1, \dots, T$, we get that

$$\begin{aligned}
\sum_{t=1}^{T} \ell_t(\boldsymbol{p}_t) - \min_{\boldsymbol{q}\in\mathcal{X}_d} \sum_{t=1}^{T} \ell_t(\boldsymbol{q}) &= \max_{\boldsymbol{q}\in\mathcal{X}_d} \sum_{t=1}^{T} \big(\ell_t(\boldsymbol{p}_t) - \ell_t(\boldsymbol{q})\big) \leqslant \max_{\boldsymbol{q}\in\mathcal{X}_d} \sum_{t=1}^{T} \nabla\ell_t(\boldsymbol{p}_t)\cdot(\boldsymbol{p}_t - \boldsymbol{q}) \\
&= \max_{1\leqslant i\leqslant d} \sum_{t=1}^{T} \nabla\ell_t(\boldsymbol{p}_t)\cdot(\boldsymbol{p}_t - \boldsymbol{e}_i) \qquad (2.30) \\
&= \sum_{t=1}^{T}\sum_{i=1}^{d} p_{i,t}\nabla_i\ell_t(\boldsymbol{p}_t) - \min_{1\leqslant i\leqslant d} \sum_{t=1}^{T} \nabla_i\ell_t(\boldsymbol{p}_t)\ , \qquad (2.31)
\end{aligned}$$

where (2.30) follows from the fact that $\boldsymbol{q} \mapsto \sum_{t=1}^{T} \nabla\ell_t(\boldsymbol{p}_t)\cdot(\boldsymbol{p}_t - \boldsymbol{q})$ is affine (convexity is sufficient) on the polytope $\mathcal{X}_d$, the vertices of which are denoted by $\boldsymbol{e}_i \triangleq \big(\mathbb{I}_{\{j=i\}}\big)_{1\leqslant j\leqslant d}$. But, by definition of $\boldsymbol{p}_t$ and η_t above, we can apply Theorem 2.4 with the linear loss (cf. Example 2.1 on page 42), the observations $\boldsymbol{y}_t = \nabla\ell_t(\boldsymbol{p}_t) \in \mathbb{R}^d$, and the constant expert advice $\boldsymbol{a}_{i,t} = \boldsymbol{e}_i \in \mathcal{X}_d$ to get that

$$\begin{aligned}
\sum_{t=1}^{T}\sum_{i=1}^{d} p_{i,t}\nabla_i\ell_t(\boldsymbol{p}_t) - \min_{1\leqslant i\leqslant d} \sum_{t=1}^{T} \nabla_i\ell_t(\boldsymbol{p}_t) \leqslant 4\sqrt{\left(\sum_{t=1}^{T} \|\nabla\ell_t(\boldsymbol{p}_t)\|_\infty^2\right) \ln d} \\
+ (4\ln d + 6)\left(2\max_{1\leqslant t\leqslant T} \|\nabla\ell_t(\boldsymbol{p}_t)\|_\infty\right)\ ,
\end{aligned}$$

where we used the fact that, in our case, the effective ranges $E_t \triangleq \max_{1\leqslant i,j\leqslant d}\big|\nabla_i\ell_t(\boldsymbol{p}_t) - \nabla_j\ell_t(\boldsymbol{p}_t)\big|$ in (2.13) are upper bounded as $E_t \leqslant 2\,\|\nabla\ell_t(\boldsymbol{p}_t)\|_\infty$. This concludes the proof. □

Extension to ℓ^1-balls

In the previous paragraphs, we showed that the Exponentiated Gradient algorithm is competitive against the whole simplex $\mathcal{X}_d$. A limitation is that the vectors of $\mathcal{X}_d$ are restricted to have nonnegative components and an ℓ^1-norm bounded by 1. Next, we overcome this limitation via a trick due to [KW97] and that transforms the Exponentiated Gradient forecaster into an algorithm which is competitive against all vectors of the ℓ^1-ball $B_1(U) \triangleq \big\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U\big\}$ for a given $U > 0$.

In the general case of convex and differentiable loss functions $\ell_t : \mathbb{R}^d \to \mathbb{R}$, the trick of [KW97] unfolds as follows. First note that the polytope $B_1(U)$ is the convex hull of its $2d$ vertices $\pm U\boldsymbol{e}_i$, $i = 1, \dots, d$ (recall that $\boldsymbol{e}_i \triangleq \big(\mathbb{I}_{\{j=i\}}\big)_{1\leqslant j\leqslant d}$). Therefore, for all $\boldsymbol{u} \in B_1(U)$, there

[12] If the convex loss functions $\ell_t : \mathbb{R}^d \to \mathbb{R}$ are not differentiable, gradients can be replaced with subgradients.

exists[13] a convex combination $\boldsymbol{p} = (p_1^+, p_1^-, \ldots, p_d^+, p_d^-) \in \mathcal{X}_{2d}$ such that

$$\boldsymbol{u} = \sum_{i=1}^{d} \left[p_i^+ \, (U\boldsymbol{e}_i) + p_i^- \, (-U\boldsymbol{e}_i) \right] = U \sum_{i=1}^{d} \left(p_i^+ - p_i^- \right) \boldsymbol{e}_i \, .$$

The last remark suggests to apply the Exponentiated Gradient algorithm defined in (2.28)–(2.29) to the augmented loss function $\ell_t^{(U)} : \mathbb{R}^{2d} \to \mathbb{R}$ defined for all $\boldsymbol{v} = (v_1^+, v_1^-, \ldots, v_d^+, v_d^-) \in \mathbb{R}^{2d}$ by

$$\ell_t^{(U)}(\boldsymbol{v}) \triangleq \ell_t \left(U \sum_{i=1}^{d} \left(v_i^+ - v_i^- \right) \boldsymbol{e}_i \right) \, . \tag{2.32}$$

Denote the resulting weight vectors by $\boldsymbol{p}_t = (p_{1,t}^+, p_{1,t}^-, \ldots, p_{d,t}^+, p_{d,t}^-) \in \mathcal{X}_{2d}$. Then, we call *adaptive EG$^\pm$ algorithm on* $B_1(U)$ the forecaster that outputs the linear combinations $\widehat{\boldsymbol{u}}_t \in B_1(U)$ given by

$$\widehat{\boldsymbol{u}}_t = U \sum_{i=1}^{d} \left(p_{i,t}^+ - p_{i,t}^- \right) \boldsymbol{e}_i \, , \quad t = 1, \ldots, T \, .$$

A formal definition of the adaptive EG$^\pm$ algorithm on $B_1(U)$ is given in Figure 2.5. Note that this forecaster takes as input parameter the radius U of the ℓ^1-ball $B_1(U)$. The form of the update (2.33) follows from the fact that, for all $t = 1, \ldots, T$ and all $\boldsymbol{v} = (v_1^+, v_1^-, \ldots, v_d^+, v_d^-) \in \mathbb{R}^{2d}$,

$$\frac{\mathrm{d}\ell_t^{(U)}}{\mathrm{d}v_j^+}(\boldsymbol{v}) = U \nabla_j \ell_t \left(U \sum_{i=1}^{d} (v_i^+ - v_i^-) \boldsymbol{e}_i \right) \quad \text{and} \quad \frac{\mathrm{d}\ell_t^{(U)}}{\mathrm{d}v_j^-}(\boldsymbol{v}) = -U \nabla_j \ell_t \left(U \sum_{i=1}^{d} (v_i^+ - v_i^-) \boldsymbol{e}_i \right) \, ,$$

so that, by definition of $\widehat{\boldsymbol{u}}_t$ above, we have, for all $j \in \{1, \ldots, d\}$ and all $\gamma \in \{+, -\}$,

$$\frac{\mathrm{d}\ell_t^{(U)}}{\mathrm{d}v_j^\gamma}(\boldsymbol{p}_t) = \gamma U \nabla_j \ell_t(\widehat{\boldsymbol{u}}_t) \, , \tag{2.34}$$

where, by a slight abuse of notation, the symbols "$+$" and "$-$" also denote the values $+1$ and -1 respectively (e.g., γU should be understood as $-U$ if $\gamma = -$). We can use Proposition 2.4 to bound the regret of the adaptive EG$^\pm$ algorithm as follows.

Corollary 2.1 (The adaptive EG$^\pm$ algorithm for general convex and differentiable loss functions)**.** *Let $U > 0$. Then, the adaptive EG$^\pm$ algorithm on $B_1(U)$ defined in Figure 2.5 satisfies, for all $T \geqslant 1$ and all sequences of convex and differentiable[14] loss functions $\ell_1, \ldots, \ell_T : \mathbb{R}^d \to \mathbb{R}$,*

$$\sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{u}}_t) - \min_{\boldsymbol{u} : \|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \ell_t(\boldsymbol{u})$$
$$\leqslant 4U \sqrt{\left(\sum_{t=1}^{T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty^2 \right) \ln(2d)} + U \left(8 \ln(2d) + 12 \right) \max_{1 \leqslant t \leqslant T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty \, .$$

In particular, the regret is bounded by $4U \left(\max_{1 \leqslant t \leqslant T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty \right) \left(\sqrt{T \ln(2d)} + 2 \ln(2d) + 3 \right)$.

[13] The corresponding vector $\boldsymbol{p} \in \mathcal{X}_{2d}$ is not unique. An example is given by $p_i^+ = (u_i)_+ / U + \left(U - \|\boldsymbol{u}\|_1 \right) / (2dU)$ and by $p_i^- = (u_i)_- / U + \left(U - \|\boldsymbol{u}\|_1 \right) / (2dU)$.

[14] Again, gradients can be replaced with subgradients in case of non-differentiability.

Parameter: radius $U > 0$.

Initialization: $\boldsymbol{p}_1 = (p_{1,1}^+, p_{1,1}^-, \dots, p_{d,1}^+, p_{d,1}^-) \triangleq \big(1/(2d), \dots, 1/(2d)\big) \in \mathbb{R}^{2d}$.

At each time round $t \geqslant 1$,

1. Output the linear combination $\widehat{\boldsymbol{u}}_t \triangleq U \sum_{j=1}^{d} \big(p_{j,t}^+ - p_{j,t}^-\big)\, \boldsymbol{e}_j \in B_1(U)$;

2. Get the (convex and differentiable) loss function $\ell_t : \mathbb{R}^d \to \mathbb{R}$ and define

$$z_{j,s}^+ \triangleq U \nabla_j \ell_s(\widehat{\boldsymbol{u}}_s) \quad \text{and} \quad z_{j,s}^- \triangleq -U \nabla_j \ell_s(\widehat{\boldsymbol{u}}_s)\,, \quad j = 1, \dots, d, \quad s = 1, \dots, t\,,$$

$$\widehat{E}_t \triangleq \inf_{k \in \mathbb{Z}} \left\{ 2^k : 2^k \geqslant \max_{1 \leqslant s \leqslant t} \max_{\substack{1 \leqslant j,k \leqslant d \\ \gamma,\mu \in \{+,-\}}} \big| z_{j,s}^{\gamma} - z_{k,s}^{\mu} \big| \right\}\,,$$

$$V_t \triangleq \sum_{s=1}^{t} \sum_{\substack{1 \leqslant j \leqslant d \\ \gamma \in \{+,-\}}} p_{j,s}^{\gamma} \left(z_{j,s}^{\gamma} - \sum_{\substack{1 \leqslant k \leqslant d \\ \mu \in \{+,-\}}} p_{k,s}^{\mu} z_{k,s}^{\mu} \right)^2 \,;$$

3. Update the parameter η_{t+1} according to

$$\eta_{t+1} \triangleq \min \left\{ \frac{1}{\widehat{E}_t}, C \sqrt{\frac{\ln K}{V_t}} \right\}\,, \qquad \text{where} \qquad C \triangleq \sqrt{2(\sqrt{2}-1)/(\mathrm{e}-2)}\,;$$

4. Update the weight vector $\boldsymbol{p}_{t+1} = (p_{1,t+1}^+, p_{1,t+1}^-, \dots, p_{d,t+1}^+, p_{d,t+1}^-) \in \mathcal{X}_{2d}$ defined for all $j = 1, \dots, d$ and $\gamma \in \{+,-\}$ by

$$p_{j,t+1}^{\gamma} \triangleq \frac{\exp\left(-\eta_{t+1} \sum_{s=1}^{t} \gamma U \nabla_j \ell_s(\widehat{\boldsymbol{u}}_s) \right)}{\sum_{\substack{1 \leqslant k \leqslant K \\ \mu \in \{+,-\}}} \exp\left(-\eta_{t+1} \sum_{s=1}^{t} \mu U \nabla_k \ell_s(\widehat{\boldsymbol{u}}_s) \right)}\,. \tag{2.33}$$

Figure 2.5: The adaptive EG$^{\pm}$ algorithm on $B_1(U)$ for general convex and differentiable loss functions (cf. Corollary 2.1).

Proof: The result follows straightforwardly from Proposition 2.4 by noting that, on the one hand, by definitions of $\ell_t^{(U)}$, $\boldsymbol{p}_t$, and $\widehat{\boldsymbol{u}}_t$,

$$\sum_{t=1}^{T} \ell_t(\widehat{\boldsymbol{u}}_t) = \sum_{t=1}^{T} \ell_t^{(U)}(\boldsymbol{p}_t) \quad \text{and} \quad \min_{\boldsymbol{u}:\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \ell_t(\boldsymbol{u}) = \min_{\boldsymbol{q} \in \mathcal{X}_{2d}} \sum_{t=1}^{T} \ell_t^{(U)}(\boldsymbol{q}) \;,$$

and, on the other hand, $\left\| \nabla \ell_t^{(U)}(\boldsymbol{p}_t) \right\|_\infty = U \left\| \nabla \ell_t(\widehat{\boldsymbol{u}}_t) \right\|_\infty$ for all $t = 1, \ldots, T$ (by (2.34)). □

An improvement for small losses under the square loss

In the particular case of the square loss $\ell_t(\boldsymbol{u}) = (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$, the gradients are given by $\nabla \ell_t(\boldsymbol{u}) = -2(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)\, \boldsymbol{x}_t$ for all $\boldsymbol{u} \in \mathbb{R}^d$. Applying Corollary 2.1, we get the following improvement for small losses.

Corollary 2.2 (An improvement for small losses under the square loss)**.**
Let $U > 0$. Consider the online linear regression setting. Then, the adaptive EG$^\pm$ algorithm on $B_1(U)$ defined in Figure 2.5 with the loss functions $\ell_t : \boldsymbol{u} \mapsto (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$ satisfies, for all sequences of $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T) \in \mathbb{R}^d \times \mathbb{R}$,

$$\sum_{t=1}^{T} (y_t - \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t)^2 \leqslant L_T^* + 8UX\sqrt{L_T^* \ln(2d)} + \big(137 \ln(2d) + 24\big)\left(UXY + U^2X^2\right) ,$$

where the quantities $L_T^* \triangleq \min_{\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U\}} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$, $X \triangleq \max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty$, and $Y \triangleq \max_{1 \leqslant t \leqslant T} |y_t|$ are unknown to the forecaster.

We point out that the large constants 137 and 24 above can be improved. This can be done, e.g., by using the tighter bound of Remark 2.2 on page 54, instead of the original bound of [CBMS07]. Note that this also reduces the constant 8 of the main regret term to $8 \times 1.32/2 = 5.28$.

The above bound is comparable to the bound $2UX\sqrt{2B \ln(2d)} + 2U^2X^2 \ln(2d)$ implied by [KW97, Theorem 5.11], where B is a known upper bound on L_T^*. Note that our main term is larger than that of [KW97] by a multiplicative factor of $2\sqrt{2}$; our lower-order term is also larger by (quite large[15]) multiplicative factors and by an additional term of the order of $UXY \ln(2d)$. However, to get their bound, [KW97] tuned the EG$^\pm$ algorithm as a function of B and of two known upper bounds X and Y on the input data and the observations (cf. the choice of $\eta = (\sqrt{\ln(2d)})/(UX\sqrt{2B} + 2U^2X^2\sqrt{\ln(2d)})$ therein). On the contrary, the version of the EG$^\pm$ algorithm we use does not require the knowledge of B, X, and Y.

Thus, Corollary 2.2 is of the same flavor as the regret bound of [ACBG02, Theorem 3.1] for the self-confident p-norm algorithm. Indeed, for a given parameter $U > 0$ and for $p = 2 \ln d$, [ACBG02] show that the cumulative loss $\widehat{L}_T$ of the self-confident p-norm algorithm[16] satisfies,

[15] See the comment on the constants 137 and 24 above.

[16] The vectors output by the self-confident p-norm algorithm with parameter U lie in the ℓ^q-ball $\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_q \leqslant U\}$, where q is the conjugate of p, i.e., $q = p/(p-1) \approx 1 + 1/(2 \ln d)$ if $p = 2 \ln d$. This ℓ^q-ball contains the ℓ^1-ball $B_1(U)$ but is only a slight overapproximation of it since $e^{-1} \|\cdot\|_1 \leqslant \|\cdot\|_q \leqslant \|\cdot\|_1$.

for some nonnegative quantity $k_T \leqslant (2e \ln d) U^2 X^2$,

$$\begin{aligned}\widehat{L}_T &\leqslant L_T^* + 8k_T + 8\sqrt{(k_T L_T^*)/2 + k_T^2} \\ &\leqslant L_T^* + 8UX\sqrt{(e \ln d)\, L_T^*} + (32e \ln d)\, U^2 X^2 \;,\end{aligned}$$

where the quantities $L_T^* \triangleq \min_{\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U\}} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$, $X \triangleq \max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty$, and $Y \triangleq \max_{1 \leqslant t \leqslant T} |y_t|$ are unknown to the forecaster. The fact that we got a similar bound is not surprising because the p-norm algorithm is known to share many properties with the EG$^\pm$ algorithm (in the limit $p \to +\infty$ with an appropriate initial weight vector, or for p of the order of $\ln d$ with a zero initial weight vector, cf. [Gen03]). The bound of Corollary 2.2 corroborates this similarity.

Proof (of Corollary 2.2): We apply Corollary 2.1 with the square loss $\ell_t(\boldsymbol{u}) = (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$:

$$\begin{aligned}&\sum_{t=1}^T \ell_t(\widehat{\boldsymbol{u}}_t) - \min_{\boldsymbol{u} : \|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T \ell_t(\boldsymbol{u}) \\ &\qquad \leqslant 4U \sqrt{\left(\sum_{t=1}^T \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty^2 \right) \ln(2d)} + U \left(8 \ln(2d) + 12\right) \max_{1 \leqslant t \leqslant T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty \;.\end{aligned}$$

Using the equality $\nabla \ell_t(\boldsymbol{u}) = -2(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)\, \boldsymbol{x}_t$ for all $\boldsymbol{u} \in \mathbb{R}^d$, we get that, on the one hand, by the upper bound $\|\boldsymbol{x}_t\|_\infty \leqslant X$,

$$\|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty^2 \leqslant 4X^2 \ell_t(\widehat{\boldsymbol{u}}_t) \;, \tag{2.35}$$

and, on the other hand, $\max_{1 \leqslant t \leqslant T} \|\nabla \ell_t(\widehat{\boldsymbol{u}}_t)\|_\infty \leqslant 2(Y + UX)X$ (indeed, by Hölder's inequality, $\left|\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t\right| \leqslant \|\widehat{\boldsymbol{u}}_t\|_1 \|\boldsymbol{x}_t\|_\infty \leqslant UX$). Substituting the last two inequalities in the bound of Corollary 2.1, setting $\widehat{L}_T \triangleq \sum_{t=1}^T \ell_t(\widehat{\boldsymbol{u}}_t)$, and recalling that $L_T^* \triangleq \min_{\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U\}} \sum_{t=1}^T \ell_t(\boldsymbol{u})$, we get that

$$\widehat{L}_T \leqslant L_T^* + 8UX\sqrt{\widehat{L}_T \ln(2d)} + \underbrace{\left(16 \ln(2d) + 24\right)\left(UXY + U^2X^2\right)}_{\triangleq C} \;.$$

Solving for $\widehat{L}_T$ via Lemma A.2 in Appendix A.4, we get that

$$\begin{aligned}\widehat{L}_T &\leqslant L_T^* + C + \left(8UX\sqrt{\ln(2d)}\right)\sqrt{L_T^* + C} + \left(8UX\sqrt{\ln(2d)}\right)^2 \\ &\leqslant L_T^* + 8UX\sqrt{L_T^* \ln(2d)} + 8UX\sqrt{C \ln(2d)} + 64U^2X^2 \ln(2d) + C \;.\end{aligned}$$

Using that

$$\begin{aligned}UX\sqrt{C \ln(2d)} &= UX \ln(2d) \sqrt{\left(16 + 24/\ln(2d)\right)\left(UXY + U^2X^2\right)} \\ &\leqslant \sqrt{U^2X^2 + UXY}\, \ln(2d) \sqrt{\left(16 + 24/\ln(2)\right)\left(UXY + U^2X^2\right)} \\ &= \sqrt{16 + 24/\ln(2)}\, \left(UXY + U^2X^2\right) \ln(2d)\end{aligned}$$

and performing some simple upper bounds concludes the proof. □

In the last corollary we used the adaptive EG$^{\pm}$ algorithm with the square loss functions $\ell_t : \boldsymbol{u} \mapsto (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$. In Chapter 4 we use yet another instance of the adaptive EG$^{\pm}$ algorithm that we call the Lipschitzifying Exponentiated Gradient (LEG) algorithm. It corresponds to the adaptive EG$^{\pm}$ algorithm applied not to the square loss but to a Lipschitz continuous modification $\widetilde{\ell}_t : \mathbb{R}^d \to \mathbb{R}$ of the square loss. Both the modified loss function $\widetilde{\ell}_t$ and the threshold used to perform an additional clipping are updated as a function of the available data only. Applying Corollary 2.1 again, we show in Theorem 4.3 of Chapter 4 that the cumulative loss $\widehat{L}_T$ of the LEG algorithm is upper bounded by

$$\widehat{L}_T \leqslant \widetilde{L}_T^* + 8UX\sqrt{\widetilde{L}_T^* \ln(2d)} + \big(153\ln(2d) + 58\big)\big(UXY + U^2X^2\big) + 12Y^2 \,,$$

where $\widetilde{L}_T^* \triangleq \min_{\{\boldsymbol{u}\in\mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U\}} \sum_{t=1}^T \widetilde{\ell}_t(\boldsymbol{u})$ is the optimal cumulative *Lipschitzified* loss within $B_1(U)$. The main two terms of the last bound slightly improve on those of Corollary 2.2 since, by Figure 4.2 of Chapter 4, we always have

$$\widetilde{L}_T^* \leqslant \min_{\boldsymbol{u}\in\mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \,.$$

As explained therein, the improvement brought about by the Lipschitzification step is more significant for loss functions with higher curvature than the square loss, e.g., loss functions of the form $\boldsymbol{u} \mapsto \big|y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\big|^\alpha$ with $\alpha > 2$.

2.5 From online to batch bounds

In this section we explain how to convert an online algorithm into a method suitable for a probabilistic batch setting. We then point out that, contrary to a common belief, online methods can be used in the regression model with random design even if the outputs are unbounded.

2.5.1 The online-to-batch conversion

Let $\mathcal{D}$ be a convex decision space, $\mathcal{Z}$ be an outcome space[17], and $\ell : \mathcal{D} \times \mathcal{Z} \to \mathbb{R}$ be a loss function convex in its first argument. In the sequel we consider the following batch stochastic setting. The forecaster is given at the beginning of the game T independent random copies $Z_1, \ldots, Z_T$ of $Z \in \mathcal{Z}$, whose common distribution is unknown. The goal of the forecaster is to predict the next outcome $Z_{T+1} \sim Z$ almost as well as does any fixed element in a non-empty subset $\Theta \subset \mathcal{D}$. More precisely, its goal is to output a decision $\widehat{a}_T \in \mathcal{D}$ based on the sample $(Z_1, \ldots, Z_T)$ so as to minimize its excess expected risk

$$\mathbb{E}\big[\ell\big(\widehat{a}_T, Z\big)\big] - \inf_{a\in\Theta} \mathbb{E}\big[\ell\big(a, Z\big)\big] \,, \tag{2.36}$$

where the expectation on the left is taken with respect to all sources of randomness (i.e., with respect to $(Z_1, \ldots, Z_T)$ and Z).

[17] We use the notation $\mathcal{Z}$ instead of $\mathcal{Y}$ to avoid any ambiguity with the regression model with random design where Y_t only denotes the output while we observe the whole pair $Z_t = (X_t, Y_t) \in \mathcal{X} \times \mathbb{R}$. In this setting, $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$.

We illustrate in the sequel the links between this batch[18] setting and the online framework of prediction with expert advice studied in the previous sections. More precisely, consider the online protocol of Figure 2.1 when the outcome space is $\mathcal{Z}$ and when the expert advice are constant and given by $a_{\theta,t} = \theta$ for all $\theta \in \Theta$ and all $t \geqslant 1$. In this particular setting, the online protocol of Figure 2.1 can be rewritten as follows: a forecaster repeatedly outputs a decision $\widetilde{a}_t \in \mathcal{D}$, observes the new outcome $z_t \in \mathcal{Z}$, and incurs the loss $\ell(\widetilde{a}_t, z_t)$; after T time steps, its regret against the set of experts Θ reads:

$$\sum_{t=1}^{T} \ell(\widetilde{a}_t, z_t) - \inf_{a \in \Theta} \sum_{t=1}^{T} \ell(a, z_t) \; . \tag{2.37}$$

(The above online protocol is known as *online convex optimization* when $\Theta = \mathcal{D}$; see, e.g., [Zin03, SSSSS09].)

Next we show that any online algorithm—i.e., any strategy of the forecaster— $(\widetilde{a}_t)_{t \geqslant 1}$ that has a small regret (2.37) in the above online protocol can be converted into a method $\widehat{a}_T$ that has a small excess expected risk (2.36) in the batch stochastic setting.

The following *online-to-batch conversion* is a standard trick in the machine learning community that can be traced back to around [Lit89] (see also the earlier references given in [DS06]). High-probability data-dependent risk bounds for arbitrary convex decision spaces and convex and bounded loss functions were derived by [CBCG04]. The latter paper also addresses the case when either the decision space or the (bounded) loss function is not convex via a more sophisticated online-to-batch conversion. Several improved high-probability risk bounds were then obtained by [CBG08] in the possibly non-convex setting and by [Zha05, KT09] for convex decision spaces and "Bernstein-friendly" loss functions (e.g., strongly convex losses).

The conversion consists in treating the sample $Z_{1:T} \triangleq (Z_1, \dots, Z_T)$ in a sequential fashion: even if all the Z_t are known at the beginning of the game, they are only used one at a time from round 1 to round T, that is, the online algorithm $(\widetilde{a}_t)_{t \geqslant 1}$ sequentially outputs its decisions $\widetilde{a}_t(Z_{1:t-1}) \in \mathcal{D}$ based on the past data $Z_{1:t-1} \triangleq (Z_1, \dots, Z_{t-1})$, $t = 1, \dots, T$ ($\widetilde{a}_1$ is deterministic). Finally, the simplest way to define $\widehat{a}_T(Z_{1:T})$ when $\mathcal{D}$ is convex is to consider the average:

$$\widehat{a}_T(Z_{1:T}) = \frac{1}{T} \sum_{t=1}^{T} \widetilde{a}_t(Z_{1:t-1}) \; .$$

Proposition 2.5. *Let $\mathcal{D}$ be a convex decision space, $\mathcal{Z}$ be an outcome space, and $\ell : \mathcal{D} \times \mathcal{Z} \to \mathbb{R}$ be a loss function convex in its first argument. Let $(\widetilde{a}_t)_{t \geqslant 1}$ be any online algorithm and $(R_T)_{T \geqslant 1}$ be any real-valued sequence such that, for all $T \geqslant 1$, uniformly over all $z_1, \dots, z_T \in \mathcal{Z}$,*

$$\sum_{t=1}^{T} \ell(\widetilde{a}_t, z_t) - \inf_{a \in \Theta} \sum_{t=1}^{T} \ell(a, z_t) \leqslant R_T \; . \tag{2.38}$$

Then the above online to batch conversion applied to $(\widetilde{a}_t)_{t \geqslant 1}$ yields a procedure $\widehat{a}_T$ such that, for

[18] "Batch" means that all outcomes $Z_1, \dots, Z_T$ are available at the beginning of the game—as opposed to the online setting.

all i.i.d. samples $(Z_1, \dots, Z_T) \in \mathcal{Z}^T$,

$$\mathbb{E}\big[\ell\big(\widehat{a}_T, Z\big)\big] - \inf_{a\in\Theta} \mathbb{E}\big[\ell\big(a, Z\big)\big] \leqslant \frac{R_T}{T} \,,$$

where the expectations are taken with respect to both the sample $(Z_1, \dots, Z_T)$ *and a random variable* $Z \in \mathcal{Z}$ *independent of* $(Z_1, \dots, Z_T)$ *and distributed as* Z_1.

Proof: In the sequel we explicitely write all the dependencies $\widetilde{a}_t(Z_{1:t-1})$ and $\widehat{a}_T(Z_{1:T})$. By assumption, the regret bound (2.38) holds uniformly over all individual sequences $(z_1, \dots, z_T) \in \mathcal{Z}^T$. Therefore, *almost surely*,

$$\sum_{t=1}^{T} \ell\big(\widetilde{a}_t(Z_{1:t-1}), Z_t\big) \leqslant \inf_{a\in\Theta} \sum_{t=1}^{T} \ell(a, Z_t) + R_T \,. \tag{2.39}$$

In particular the last inequality holds in expectation, so that, dividing by T and using the fact that $\mathbb{E}\big[\inf_{a\in\Theta} \sum_{t=1}^{T} \ell(a, Z_t)\big] \leqslant \inf_{a\in\Theta} \mathbb{E}\big[\sum_{t=1}^{T} \ell(a, Z_t)\big]$, we get

$$\begin{aligned}\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} \ell\big(\widetilde{a}_t(Z_{1:t-1}), Z_t\big)\right] &\leqslant \frac{1}{T} \inf_{a\in\Theta} \mathbb{E}\left[\sum_{t=1}^{T} \ell(a, Z_t)\right] + \frac{R_T}{T} \\ &= \inf_{a\in\Theta} \mathbb{E}\big[\ell(a, Z)\big] + \frac{R_T}{T} \,, \end{aligned} \tag{2.40}$$

where the last equality follows from the fact that Z_t and Z are identically distributed for all $t = 1, \dots, T$. We conclude the proof via Jensen's inequality: by definition of $\widehat{a}_T(Z_{1:T})$ and by convexity of ℓ in its first argument, we have

$$\begin{aligned}\mathbb{E}\Big[\ell\big(\widehat{a}_T(Z_{1:T}), Z\big)\Big] = \mathbb{E}\left[\ell\left(\frac{1}{T}\sum_{t=1}^{T} \widetilde{a}_t(Z_{1:t-1}), Z\right)\right] &\leqslant \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\Big[\ell\big(\widetilde{a}_t(Z_{1:t-1}), Z\big)\Big] \\ &= \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\Big[\ell\big(\widetilde{a}_t(Z_{1:t-1}), Z_t\big)\Big] \,, \end{aligned}$$

where we used the fact that $(Z_{1:t-1}, Z)$ and $(Z_{1:t-1}, Z_t)$ are identically distributed (since Z and Z_t are identically distributed and both independent of $Z_{1:t-1}$). Combining the last inequality with (2.40) concludes the proof. □

Note that Z_T was not used to construct the procedure $\widehat{a}_T(Z_{1:T})$. Taking instead the average $\widehat{a}_T(Z_{1:T}) \triangleq (T+1)^{-1} \sum_{t=1}^{T+1} \widetilde{a}_t(Z_{1:t-1})$ up to time $T+1$, we can see from the above analysis that it has an excess risk upper bounded by $R_{T+1}/(T+1)$, which is usually smaller than R_T/T. (This was carried out, e.g., in [Aud09, Bar11].)

More importantly, the following improvements or extensions are possible:

(a) We assumed that R_T was a uniform upper bound on the regret of $(\widetilde{a}_t)_{t\geqslant 1}$ (cf. (2.38)). The above analysis also works if $R_T = R_T(z_1, \dots, z_T)$ is data-dependent. The risk bound becomes:

$$\mathbb{E}\big[\ell\big(\widehat{a}_T, Z\big)\big] - \inf_{a\in\Theta} \mathbb{E}\big[\ell\big(a, Z\big)\big] \leqslant \frac{\mathbb{E}\big[R_T(Z_{1:T})\big]}{T} \,.$$

(b) Proposition 2.5 tranforms the almost-sure regret bound (2.39) into a risk bound in expectation. It is also possible to transform (2.39) into a high-probability data-dependent risk bound via standard martingale concentration tools (e.g., the Hoeffding-Azuma inequality or Bernstein's inequality for martingales). See [CBG08] for a different conversion suited for general decision spaces and bounded loss functions (not necessarily convex) and [Zha05, KT09] for the same conversion as the one studied above but in the particular case of "Bernstein-friendly" losses (e.g., strongly convex).

(c) As showed in [Bar11], the online-to-batch conversion described above can be slightly modified to handle sequences $(Z_1, \dots, Z_T, Z)$ that are generated by a stationary process — independence is no longer required. This can be achieved by defining

$$\widehat{a}_T(Z_{1:T}) = \frac{1}{T+1} \sum_{t=0}^{T} \widetilde{a}_{T-t+1}(Z_{t+1:T}) \, .$$

In this case the online algorithm is repeatedly re-initialized: for each $t = 0, \dots, T$, the algorithm is restarted and run on the sub-sample $Z_{t+1:T} \triangleq (Z_{t+1}, \dots, Z_T)$ of length $T - t$ (so that the corresponding decision function is $\widetilde{a}_{T-t+1}$). By similar arguments, this procedure is seen to satisfy the bound

$$\mathbb{E}\big[\ell(\widehat{a}_T, Z)\big] - \inf_{a \in \Theta} \mathbb{E}\big[\ell(a, Z)\big] \leqslant \frac{R_{T+1}}{T+1} \, .$$

2.5.2 Application: regression model with random design and unbounded outputs

In this section, we focus on the online-to-batch conversion from the online linear regression setting to the regression model with random design. The case of a fixed design is not addressed here, but it can to some extent be dealt with via similar techniques (see Section 3.4.2 in Chapter 3). In the sequel, we first make some preliminary comments and introduce the regression model with random design together with some related aggregation problems. Afterward we study the online-to-batch conversion in the easy case of bounded outputs. We then discuss a trick of [BN08] to deal with unbounded outputs when the forecaster has access to some prior knowledge on the regression function and on the dictionary at hand. We finally explain how to overcome the last limitation, i.e., how to design a fully automatic online algorithm whose batch conversion satisfies adaptivity properties.

Preliminary comments

In the last section we showed that, unsurprisingly, worst-case regret bounds imply risk bounds in expectation. In this respect, individual sequence prediction methods can be thought of as being more robust than standard batch methods as they can be lead to controls under almost no assumption on the data at hand. The only assumptions are on the loss function (e.g., boundedness and convexity, exp-concavity) and on the set of experts' indices Θ (e.g., finiteness, convexity).

A major criticism that has however sometimes been issued as far as the square loss is concerned is that only bounded outputs can be dealt with in online (deterministic) linear regression, while this restriction fails even in a stochastic setting as simple as the regression model with random design and Gaussian noise. It actually turns out that, with additional care, individual sequence methods

can still be used in this setting. This was illustrated by [BN08], and we make further progress in this direction.

Note that most methods studied in the batch stochastic setting already handle unbounded outputs (at least in the Gaussian case, or more generally, when the regression function is uniformly bounded and when the deviation of the output from the regression function has a bounded exponential moment). See, e.g., [Nem00, Cat99, Tsy03, Aud04b] and [Cat04, Chapters 3 and 4]; see also [AC11] for PAC-Bayesian methods under even weaker assumptions. It also turns out that some of these methods have an online nature: e.g., the *progressive mixture rule* is nothing else than the exponentially weighted average forecaster with constant parameter η combined with the online-to-batch conversion. Its theoretical properties with unbounded outputs were derived in [Cat04, Chapter 3] using arguments different from that of the individual sequence framework (see also [BN08, Theorem 1-(a)]). The corresponding risk bounds were proved when the parameter η is tuned as a function of a known uniform bound on the regression function and on the base regressors and of a known bound related to the moment-generating function of the noise. It is not clear whether the latter tuning ensures non-trivial regret bounds for individual sequences. Next we focus on online algorithms (e.g., properly tuned variants of the exponentially weighted average forecaster) that have provable guarantees for individual sequences and for which the standard online-to-batch conversion yields interesting risk bounds in the stochastic setting. This individual sequence approach not only ensures that the resulting method is robust in some sense, but also provides adaptivity results to unknown quantities in the stochastic setting.

Regression model with random design

Next we introduce the regression model with random design and related aggregation problems. In this batch setting the forecaster is given at the beginning of the game T independent random copies $(X_1, Y_1), \dots, (X_T, Y_T)$ of $(X, Y) \in \mathcal{X} \times \mathbb{R}$ whose common distribution is unknown. The random variables Y_t, $1 \leqslant t \leqslant T$, are called *outputs* or (somewhat abusively) *observations*. We assume thereafter that $\mathbb{E}[Y^2] < \infty$; the goal of the forecaster is to estimate the regression function $f : \mathcal{X} \to \mathbb{R}$ defined by $f(x) \triangleq \mathbb{E}[Y|X = x]$ for all $x \in \mathcal{X}$. The quality of a regressor $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ based on the sample $(X_1, Y_1), \dots, (X_T, Y_T)$ is measured by its L^2-risk $\left\| f - \widehat{f}_T \right\|_{L^2}^2$, where, denoting the distribution of X by P^X, we set, for all measurable functions $h : \mathcal{X} \to \mathbb{R}$,

$$\|h\|_{L^2} \triangleq \left(\int_{\mathcal{X}} h(x)^2 P^X(\mathrm{d}x) \right)^{1/2} = \left(\mathbb{E}\big[h(X)^2\big] \right)^{1/2} .$$

In the sequel we focus on the expected L^2-risk $\mathbb{E}\big[\left\| f - \widehat{f}_T \right\|_{L^2}^2 \big]$.

> Nota: Setting $\varepsilon_t \triangleq Y_t - f(X_t)$ for all $t = 1, \dots, T$, the regression model can be rewritten as
>
> $$Y_t = f(X_t) + \varepsilon_t \,, \quad 1 \leqslant t \leqslant T \,,$$
>
> where the pairs $(X_1, \varepsilon_1), \dots, (X_T, \varepsilon_T)$ are i.i.d. with $\mathbb{E}[\varepsilon_1^2] < \infty$, $\mathbb{E}[f^2(X_1)] < \infty$, and $\mathbb{E}[\varepsilon_1 | X_1] = 0$ almost surely (note that ε_1 and X_1 are not necessarily independent). This equivalent description is sometimes chosen to state the regression model with random design.

We consider the following aggregation problem. The forecaster is given a dictionnary $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_d)$ of base regressors $\varphi_j : \mathcal{X} \to \mathbb{R}$, $1 \leqslant j \leqslant d$ (the φ_j can be, e.g., elements of a suitably chosen functional basis or estimators computed on an independent sample). The goal of the forecaster is to output a regressor $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ based on the sample $(X_1, Y_1), \dots, (X_T, Y_T)$ and whose expected L^2-risk is almost as small as that of the best linear regressor $\boldsymbol{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^d u_j \varphi_j$ in a given reference class $\{\boldsymbol{u} \cdot \boldsymbol{\varphi} : \boldsymbol{u} \in \mathcal{U}\}$, where $\mathcal{U} \subset \mathbb{R}^d$. Namely, its goal is to satisfy a risk bound of the form

$$\mathbb{E}\left[\left\| f - \widehat{f} \right\|_{L^2}^2\right] \leqslant \inf_{\boldsymbol{u} \in \mathcal{U}} \left\{ \|f - \boldsymbol{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 + \psi_{T,d,\mathcal{U}}(\boldsymbol{u}) \right\} \tag{2.41}$$

for a remainder term $\psi_{T,d,\mathcal{U}}(\boldsymbol{u})$ that should be as small as possible. Note that $\psi_{T,d,\mathcal{U}}(\boldsymbol{u})$ depends on the sample size T, the ambient dimension d, and the comparison set $\mathcal{U}$. To be more rigorous, it may actually also depend on the joint distribution $\mathbb{P}$ of (X, Y) (through, e.g., the noise level $\mathbb{E}[(Y - f(X))^2]$) and on the dictionary $\boldsymbol{\varphi}$ (through, e.g., some norm $\|\boldsymbol{\varphi}\|$). Following [Nem00, Chapter 5] and [Tsy03] (see also [Lou07]), the comparison set $\mathcal{U} \subset \mathbb{R}^d$ can be taken as, e.g., the set of the vertices of the simplex in $\mathbb{R}^d$ (which corresponds to the problem of *model-selection* aggregation), the whole simplex (which corresponds to *convex* aggregation), or the whole $\mathbb{R}^d$ space (which corresponds to *linear* aggregation).

From online to batch: straightforward bounds for bounded outputs

Next we discuss the applicability of algorithms designed for online linear regression (cf. Section 2.4) to the regression model with random design. For the three aggregation problems mentioned above, there are online algorithms[19] $(\widetilde{f}_t)_{t \geqslant 1}$ that satisfy regret bounds of the following form: uniformly over all sequences $(x_1, y_1), \dots, (x_T, y_T) \in \mathcal{X} \times [-B_y, B_y]$,

$$\sum_{t=1}^T \big(y_t - \widetilde{f}_t(x_t)\big)^2 \leqslant \inf_{\boldsymbol{u} \in \mathcal{U}} \left\{ \sum_{t=1}^T \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + \Delta_{T,d,B_y}(\boldsymbol{u}) \right\} , \tag{2.42}$$

where B_y is a bound on the observations $|y_1|, \dots, |y_T|$ and where the regret term[20] $\Delta_{T,d,B_y}(\boldsymbol{u})$ is:

- of the order of $B_y^2 \ln d$ for the problem of model-selection type aggregation (see, e.g., Lemma 4.5 in Chapter 4, Appendix 4.B);
- of the order of $B_y \|\boldsymbol{\varphi}\|_\infty \sqrt{T \ln d}$ for convex aggregation in high dimension d (cf. Chapter 4, Section 4.2.1);
- of the order of $B_y^2 d \ln\big[T \|\boldsymbol{u}\|_2^2 \|\boldsymbol{\varphi}\|_\infty^2 / (d B_y^2)\big]$ for linear aggregation (cf. (2.27)), where we set $\|\boldsymbol{\varphi}\|_\infty \triangleq \max_{1 \leqslant j \leqslant d} \sup_{x \in \mathcal{X}} \big|\varphi_j(x)\big|$.

In their basic forms, most online algorithms satisfying (2.42) are tuned as a function of the bound B_y (e.g., the exponentially weighted average forecaster for model-selection aggregation,

[19] By *online algorithm*, we mean here any sequence $(\widetilde{f}_t)_{t \geqslant 1}$ of functions such that $\widetilde{f}_t : \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R})^{t-1} \to \mathbb{R}$ maps at time t the new input $\boldsymbol{x}_t$ and the past data $(\boldsymbol{x}_s, y_s)_{1 \leqslant s \leqslant t-1}$ to a prediction $\widetilde{f}_t\big(\boldsymbol{x}_t ; (\boldsymbol{x}_s, y_s)_{1 \leqslant s \leqslant t-1}\big)$, also denoted by $\widetilde{f}_t(\boldsymbol{x}_t)$ or by $\widehat{y}_t$ for notational convenience.

[20] As mentioned above, we recall that, for the sake of clarity, we only write the more important dependencies in T, d, and B_y, but the regret term $\Delta_{T,d,B_y}(\boldsymbol{u})$ usually also depends on the dictionary $\boldsymbol{\varphi}$ (through, e.g., some norm $\|\boldsymbol{\varphi}\|$).

the Exponentiated Gradient algorithm for convex aggregation, or the sequential ridge regression forecaster with the ideal tuning (2.27) of the end of Section 2.4.2 for linear aggregation).

We can then use the online to batch conversion of Section 2.5.1 with $\mathcal{Z} = \mathcal{X} \times [-B_y, B_y]$, $\mathcal{D}$ being the set of all measurable functions from $\mathcal{X}$ to $[-B_y, B_y]$, $\ell : \mathcal{D} \times \mathcal{Z} \to \mathbb{R}$ defined by $\ell\big(f, (x,y)\big) \triangleq \big(y - f(x)\big)^2$, and $\Theta = \{\boldsymbol{u} \cdot \boldsymbol{\varphi} : \boldsymbol{u} \in \mathcal{U}\}$. If the output Y lies almost surely in $[-B_y, B_y]$, this online to batch conversion transforms the aforementioned online algorithms $(\widetilde{f}_t)_{t \geqslant 1}$ into data-based regressors $\widehat{f}_T \triangleq \frac{1}{T}\sum_{t=1}^T \widetilde{f}_t$ that satisfy the risk bound

$$\mathbb{E}\Big[\big(Y - \widehat{f}_T(X)\big)^2\Big] \leqslant \inf_{\boldsymbol{u} \in \mathcal{U}} \left\{ \mathbb{E}\Big[\big(Y - \boldsymbol{u} \cdot \boldsymbol{\varphi}(X)\big)^2\Big] + \frac{\Delta_{T,d,B_y}(\boldsymbol{u})}{T} \right\}.$$

Elementary manipulations then yield the desired risk bound (see the proof of Theorem 3.2 in Chapter 3, Appendix 3.A.3 for more details):

$$\mathbb{E}\left[\left\| f - \widehat{f}_T \right\|_{L^2}^2\right] \leqslant \inf_{\boldsymbol{u} \in \mathcal{U}} \left\{ \|f - \boldsymbol{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 + \frac{\Delta_{T,d,B_y}(\boldsymbol{u})}{T} \right\}.$$

What about unbounded outputs Y_t? — The method of [BN08] under some prior knowledge on the regression function f and the dictionary $\boldsymbol{\varphi}$.

In the previous paragraphs, we assumed that the output Y lied in some bounded interval $[-B_y, B_y]$. Assume now that Y is unbounded in the sense that

$$\forall B > 0, \quad \mathbb{P}\big(|Y| > B\big) > 0\,. \tag{2.43}$$

In this case, one method suggested by [BN08] is to truncate the outputs Y_t to some threshold $\gamma = b$ or $\gamma = b \ln T$ (up to constant factors) for some known bound $b > 0$ on the infinity norms of the regression function f and the base regressors φ_j, $1 \leqslant j \leqslant d$, i.e.,

$$b \geqslant \max\big\{\|f\|_\infty, \|\varphi_1\|_\infty, \ldots, \|\varphi_d\|_\infty\big\}\,.$$

The authors then apply an exponentially weighted average forecaster to the truncated outputs

$$\widetilde{Y}_t = [Y_t]_\gamma\,, \quad 1 \leqslant t \leqslant T\,,$$

where $[x]_\gamma \triangleq \min\big\{\gamma, \max\{-\gamma, x\}\big\}$ denotes the truncation (or *clipping*) of $x \in \mathbb{R}$ to the threshold level γ. They then prove a risk bound for $\widehat{f}_T$ in terms of the truncated output $\widetilde{Y} = [Y]_\gamma$, from which they derive by an approximation argument a risk bound in terms of the non-truncated (and possibly unbounded) output Y.

A drawback of the previous approach is that the bound b is assumed to be known in advance. Besides, the base forecasters $\boldsymbol{u} \cdot \boldsymbol{\varphi}$, $\boldsymbol{u} \in \mathcal{U}$, are not uniformly bounded if $\boldsymbol{\varphi} \neq \mathbf{0}$ and $\mathcal{U} = \mathbb{R}^d$, so that the approach followed in [BN08] for model-selection or convex aggregation does not readily apply to linear aggregation. For these two reasons, we truncate the base forecasters $\boldsymbol{u} \cdot \boldsymbol{\varphi}$ instead of the outputs Y_t (see below); truncation is carried out in an automatic way in the sense that no a priori knowledge is required.

The idea of truncating the base forecasts was used many times in the past; see, e.g., [Vov01]

for the online linear regression setting, [GKKW02, Chapter 10] for the regression problem with random design, and [GO07, BBGO10] for sequential prediction of unbounded time series under the square loss. A key ingredient in our work (i.e., in Chapters 3 and 4) is to perform truncation with respect to a data-driven threshold.

Online adaptation to the almost-sure bound $\max_{1\leqslant t\leqslant T}|Y_t|$ — Dealing with all previous issues.

Another way to handle the case where the output Y is unbounded is to note that, almost surely, the finite sequence $Y_1,\dots,Y_T$ lies in the bounded interval $[-B_y,B_y]$ where $B_y \triangleq \max_{1\leqslant t\leqslant T}|Y_t|$ ($B_y<+\infty$ a.s. since $\mathbb{E}\big[|Y|\big]<+\infty$ by assumption). The almost-sure boundedness of the Y_t should not be confused with the boundedness of Y in the sense of (2.43).

The above remark suggests to design algorithms $(\widetilde{f}_t)_{t\geqslant 1}$that satisfy regret bounds of the form (2.42) without knowing the random bound $B_y=\max_{1\leqslant t\leqslant T}|Y_t|$ in advance. Indeed, by Remark (a) at the end of Section 2.5.1 and by elementary manipulations carried out in the proof of Theorem 3.2 (Chapter 3), the resulting batch method $\widehat{f}_T \triangleq \frac{1}{T}\sum_{t=1}^T \widetilde{f}_t$ satisfies

$$\mathbb{E}\left[\left\|f-\widehat{f}_T\right\|_{L^2}^2\right] \leqslant \inf_{\boldsymbol{u}\in\mathcal{U}}\left\{\|f-\boldsymbol{u}\cdot\boldsymbol{\varphi}\|_{L^2}^2+\frac{\mathbb{E}\big[\Delta_{T,d,B_y}(\boldsymbol{u})\big]}{T}\right\}.$$

As can be seen from the examples of $\Delta_{T,d,B_y}(\boldsymbol{u})$ following Equation (2.42), the regret term $\Delta_{T,d,B_y}(\boldsymbol{u})$ usually (roughly) scales as B_y^2 or B_y. Hence the term $\mathbb{E}\big[\Delta_{T,d,B_y}(\boldsymbol{u})\big]$ scales as $\mathbb{E}[B_y^2]=\mathbb{E}\big[\max_{1\leqslant t\leqslant T}Y_t^2\big]$ or as $\mathbb{E}[B_y]=\mathbb{E}\big[\max_{1\leqslant t\leqslant T}|Y_t|\big]$. The last two quantities can be both upper bounded under general assumptions on the distribution of Y, e.g., when $Y-\mathbb{E}[Y]$ satisfies a tail assumption such as boundedness, a subgaussian tail, or a bounded exponential moment (see Corollary 3.5 in Chapter 3). In a word, an online adaptation to the unknown bound B_y is key to handle unbounded outputs with individual sequence techniques.

In Chapter 3, Section 3.4, we design such an algorithm $(\widetilde{f}_t)_{t\geqslant 1}$ when $\mathcal{U}=\mathbb{R}^d$, i.e., it satisfies a regret bound of the form (2.42) without knowing the random bound B_y. This algorithm is a properly tuned exponentially weighted average forecaster (with continuous weights on $\mathbb{R}^d$) applied to truncated base forecasts $\big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(X_t)\big]_{B_t}$ — note that, contrary to [BN08, Corollary 1 and Theorem 1-(b)], we truncate the base forecasts instead of truncating the outputs Y_t. Ideally we would like to choose the threshold B_t equal to the unknown random bound $\max_{1\leqslant t\leqslant T}|Y_t|$, since this can only improve prediction. The actual truncation (and the tuning of η_t) is thus performed with respect to a time-varying threshold B_t that adapts to $\max_{1\leqslant t\leqslant T}|Y_t|$ via parameter-tuning techniques provided by [ACBG02, CBMS07].

Adaptation means here that we are able to prove regret bounds within constant factors of (quasi-optimal) bounds that could be proved if $\max_{1\leqslant t\leqslant T}|Y_t|$ was known in advance by the forecaster. By a remark above on the upper bounding of $\mathbb{E}\big[\max_{1\leqslant t\leqslant T}Y_t^2\big]$, this adaptivity property in the online deterministic setting leads to sparsity oracle inequalities in the stochastic setting that are adaptive to the unknown variance of the noise at least whenever the latter is Gaussian — see Chapter 3, page 113 .

2.6 Sparsity oracle inequalities in the stochastic setting

Sparsity has been extensively studied in the stochastic setting over the past decade. Among the tools introduced for this purpose the notion of *sparsity oracle inequality* plays a fundamental role. In high-dimensional linear regression, such inequalities indicate that the task consisting in predicting almost as well as an unknown target vector is still statistically feasible if the target vector has only few non-zero coordinates. Such theoretical guarantees and the associated statistical methods have proved useful in many contemporary applications such as computational biology (e.g., analysis of DNA sequences), collaborative filtering (e.g., Netflix, Amazon), satellite and hyperspectral imaging, and high-dimensional econometrics (e.g., cross-country growth regression problems).

In this section we recall the basic ideas underlying the notion of sparsity oracle inequality in the stochastic batch setting. In Chapter 3 we use similar ideas in the framework of individual sequences to introduce a new type of (deterministic) regret bounds under a sparsity scenario.

Framework

We first consider the (generalized) linear regression model with fixed or random design. The forecaster observes independent random pairs $(X_1, Y_1), \dots, (X_T, Y_T) \in \mathcal{X} \times \mathbb{R}$ given by

$$Y_t = \boldsymbol{u}^* \cdot \boldsymbol{\varphi}(X_t) + \varepsilon_t \,, \quad 1 \leqslant t \leqslant T \,, \tag{2.44}$$

where the $X_t \in \mathcal{X}$ are either i.i.d random variables (random design) or fixed elements (fixed design), denoted in both cases by capital letters in this section, where $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_d)$ is a dictionary of base regressors $\varphi_j : \mathcal{X} \to \mathbb{R}$, $1 \leqslant j \leqslant d$, where $\boldsymbol{u}^* \in \mathbb{R}^d$ is the unknown linear combination (recall that $\boldsymbol{u}^* \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^d u_j^* \varphi_j$), and where the ε_t are i.i.d. square-integrable real random variables with zero mean (conditionally on the X_t if the design is random).

Three main statistical problems arise in this linear regression framework:

- prediction: estimating $\big(\boldsymbol{u}^* \cdot \boldsymbol{\varphi}(X_t)\big)_{1 \leqslant t \leqslant T}$ (fixed design) or $\boldsymbol{u}^* \cdot \boldsymbol{\varphi}$ (random design);
- estimation: estimating $\boldsymbol{u}^*$;
- support estimation: estimating the set of the non-zero coordinates of $\boldsymbol{u}^*$.

These three tasks have been extensively studied over the past decade in the high-dimensional setting under a *sparsity scenario*, namely, when

$$\|\boldsymbol{u}^*\|_0 \ll T \ll d \,, \tag{2.45}$$

where $\|\boldsymbol{u}^*\|_0 \triangleq \big|\{j : u_j^* \neq 0\}\big|$ denotes the number of non-zero coordinates of $\boldsymbol{u}^*$. In this thesis, we focus on the prediction problem. As we will see later, this problem can be addressed in rather general regression models both in the stochastic batch setting (see (2.46) below) and in the deterministic online setting (see Chapter 3).

In the stochastic setting, most risk bounds for the prediction problem under a sparsity scenario take the form of sparsity oracle inequalities. We explain below the basic idea that underlies such risk bounds, as well as their main consequences.

2.6.1 An ideal ordinary least-squares estimator

In the (generalized) linear regression model with fixed or random design (2.44), the ordinary least squares estimator

$$\widehat{\boldsymbol{u}}_T \in \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^d} \frac{1}{T}\sum_{t=1}^{T}\big(Y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(X_t)\big)^2$$

has an expected risk $\mathbb{E}\big[R\big(\widehat{\boldsymbol{u}}_T\big)\big]$ at most of the order of d/T (see [GKKW02] and the references therein for the fixed design, and the more recent advances in [AC11] for the random design under weak assumptions on the output distribution). Here, we defined the risk $R(\boldsymbol{u})$ of any linear combination $\boldsymbol{u}\in\mathbb{R}^d$ by

$$R(\boldsymbol{u}) \triangleq \begin{cases} \|\boldsymbol{u}^*\cdot\boldsymbol{\varphi} - \boldsymbol{u}\cdot\boldsymbol{\varphi}\|_{L^2}^2 & \text{(random design)},\\ \dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\big(\boldsymbol{u}^*\cdot\boldsymbol{\varphi}(X_t) - \boldsymbol{u}\cdot\boldsymbol{\varphi}(X_t)\big)^2 & \text{(fixed design)}, \end{cases}$$

where, in the random design case, we set $\|h\|_{L^2} \triangleq \big(\mathbb{E}[h(X_1)^2]\big)^{1/2}$ for all measurable functions $h:\mathcal{X}\to\mathbb{R}$ such that $\mathbb{E}[h(X_1)^2]<\infty$.

When the ambient dimension d is much larger than the sample size T, a direct minimization of the least-squares criterion on $\mathbb{R}^d$ can lead to overfitting, which is reflected in the non-vanishing upper bound d/T. However, as suggested by the following remark, it is still possible to achieve a small risk under the additional assumption $\|\boldsymbol{u}^*\|_0 = s \ll T$. Indeed, if the support

$$J(\boldsymbol{u}^*) \triangleq \big\{j\in\{1,\dots,d\} : u_j^*\neq 0\big\}$$

of $\boldsymbol{u}^*$ was known in advance, the oracle applying the ordinary least squares estimator to the linear subspace $\big\{\boldsymbol{u}\in\mathbb{R}^d : \forall j\notin J(\boldsymbol{u}^*), u_j = 0\big\}$ would have a risk at most of the order of $s/T \ll 1$. This suggests that the prediction task in high dimension is still feasible under a sparsity scenario. However this ordinary least-squares is ideal since the support $J(\boldsymbol{u}^*)$ is unknown in practice; it is closely related to the notion of *oracle* in the terminology of model selection (see Chapter 6 for further details).

2.6.2 Adaptivity to the unknown sparsity by model selection

The rate s/T of the above ideal least-squares estimator can actually be achieved up to a $\ln d$ factor without the prior knowledge of the set $J(\boldsymbol{u}^*)$ nor even of its cardinality $s = \|\boldsymbol{u}^*\|_0$. This was first done by adding a ℓ^0-complexity penalty to the least-squares criterion:

$$\widehat{\boldsymbol{u}}_T^{\mathrm{pen}} \in \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^d} \left\{\frac{1}{T}\sum_{t=1}^{T}\big(Y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(X_t)\big)^2 + \mathrm{pen}(\boldsymbol{u})\right\},$$

where the penalty $\mathrm{pen}(\boldsymbol{u})$ is proportional to the number $\|\boldsymbol{u}\|_0$ of non-zero coefficients of $\boldsymbol{u}$ (see [Aka71] for the AIC criterion, [Mal73] for Mallows' C_p, and [Sch78, FG94] for the BIC criterion). In the fixed design case, [BM01a] proved via model-selection arguments that for a penalty $\mathrm{pen}(\boldsymbol{u})$ of the order of $\|\boldsymbol{u}\|_0\big[1+\ln(d/\|\boldsymbol{u}\|_0)\big]$, the penalized least-squares estimator $\widehat{\boldsymbol{u}}_T^{\mathrm{pen}}$ "mimics" the

ℓ^0-oracle in the sense that its risk is at most of the order of

$$\frac{s\ln(d/s)}{T}\,.$$

For a detailed proof of this fact with a Bayesian variant of the estimator $\widehat{\boldsymbol{u}}_T^{\mathrm{pen}}$, see Chapter 6, Section 6.4.1. The above rate follows from (6.41) therein.

Numerous works addressed this model-selection-type problem: see, e.g., [BM07a, ABDJ06, BTW07a] for the fixed design setting and [BTW04] for the random design setting. Further references can be found, e.g., in [ABDJ06] and in [AGS11, Chapter 4].

The above rate holds without any assumption on the dictionary $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_d)$ and without any prior knowledge on the support $J(\boldsymbol{u}^*)$. Note that the rate contains an additional multiplicative factor $\ln(d/s)$ compared to the upper bound s/T satisfied by the ideal least-squares estimator of the previous section. This logarithmic factor is the price to pay for not knowing $J(\boldsymbol{u}^*)$ in advance. Indeed, [RWY11, Ver10] proved that the rate $s\ln(d/s)/T$ is minimax optimal on ℓ^0-balls for fixed or Gaussian random designs (see also [BM01b] in the infinite-dimensional Gaussian sequence model). In this respect, methods with a risk at most of order $s\ln(d/s)/T$ are termed *adaptive* to the unknown sparsity s of $\boldsymbol{u}^*$.

The risk bounds proved in [BM01a, BTW04, BTW07a] are actually stronger, since they are stated in a more general regression model already encountered in the previous sections:

$$Y_t = f(X_t) + \varepsilon_t\,, \quad 1 \leqslant t \leqslant T\,, \tag{2.46}$$

where the X_t are either i.i.d. random variables (random design) or deterministic elements (fixed design). Since f is not necessarily assumed to be of the form $f = \boldsymbol{u}^* \cdot \boldsymbol{\varphi}$, the risk bounds mentioned earlier

$$\mathbb{E}\big[R\big(\widehat{\boldsymbol{u}}_T\big)\big] = \mathcal{O}\left(\frac{s\ln(d/s)}{T}\right)$$

are replaced with bounds on the differences $\mathbb{E}\big[R\big(\widehat{\boldsymbol{u}}_T\big)\big] - R(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathbb{R}^d$, where the risk $R(\boldsymbol{u})$ is now defined by

$$R(\boldsymbol{u}) \triangleq \begin{cases} \|f - \boldsymbol{u}\cdot\boldsymbol{\varphi}\|_{L^2}^2 & \text{(random design)} \\ \dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\big(f(X_t) - \boldsymbol{u}\cdot\boldsymbol{\varphi}(X_t)\big)^2 & \text{(fixed design).} \end{cases}$$

The risk bounds proved in [BTW04, BTW07a] are indeed of the form

$$\mathbb{E}\big[R\big(\widehat{\boldsymbol{u}}_T\big)\big] \leqslant (1+a)\inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{R(\boldsymbol{u}) + C(a)\,\frac{\|\boldsymbol{u}\|_0}{T}\ln\left(\frac{e\,d}{\max\big\{\|\boldsymbol{u}\|_0\,,1\big\}}\right)\right\}\,, \tag{2.47}$$

where $C(a) \sim a^{-1} \to +\infty$ as $a \to 0$. The above upper bound is a typical example of what is called a *sparsity oracle inequality*, i.e., in the prediction problem, a risk bound involving a trade-off between the risk $R(\boldsymbol{u})$ and the number of non-zero coordinates $\|\boldsymbol{u}\|_0$ of all $\boldsymbol{u} \in \mathbb{R}^d$.

2.6.3 Other methods: ℓ^1-regularization and exponential weighting

We end this section with a brief computational efficiency-oriented overview of alternatives to ℓ^0-regularization. Indeed, a major drawback of ℓ^0-regularization is that the corresponding non-convex minimization problems are not computationally tractable. This complexity issue has been handled by replacing the ℓ^0-penalty with a ℓ^1-penalty (proportional to the sum of the absolute values of the coefficients). ℓ^1-regularization can indeed be seen as a 'convex relaxation' of ℓ^0-regularization, i.e., from a geometrical viewpoint, it behaves similarly to the ℓ^0-penalty but leads to convex and thus computationally tractable minimization problems. It was first proposed by [Tib96] for the so-called Lasso estimator and by [DJ94a] for a soft thresholding-based estimator in the context of wavelet regression. In its dual form, the simplest version of the Lasso is defined by

$$\widehat{\boldsymbol{u}}_T^{\mathrm{Lasso}} \in \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \frac{1}{T}\sum_{t=1}^{T} \big(Y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(X_t)\big)^2 + \lambda \left\|\boldsymbol{u}\right\|_1 \right\} ,$$

for some tuning parameter λ usually taken of the order of $\sigma\sqrt{\ln(d)/T}$ if $\varepsilon_t \sim \mathcal{N}(0,\sigma^2)$. The ℓ^0-oracle properties of the Lasso (and variants of the Lasso), i.e., risk bounds typically of the form $s\ln(d)/T$, have then been extensively studied over the past decade. A list of few references — but far from being comprehensive — includes [BTW07b, CT07, vdG08, BRT09, Kol09a, Kol09b, HvdG11, KLT11, LPvdGT11]. Until very recently all sparsity oracle inequalities proved for the Lasso had a leading constant strictly larger than 1. This apparent drawback (as compared to exponential weighting algorithms mentioned below) was overcome in [KLT11], who derived a *sharp* sparsity oracle inequality for the Lasso, i.e., a sparsity oracle inequality with leading constant equal to 1.

We also mention that [MM11] recently addressed the ℓ^1-oracle properties of the Lasso estimator from a different viewpoint. They analyze the Lasso not as a variable selector but as a model selector among a countably infinite collection of ℓ^1-balls. Their oracle-type inequalities follow from the general model selection theorem for nonlinear models of [Mas07, Theorem 4.18]. For further details, see Chapter 6 where we analyze a Bayesian extension of this model-selection procedure based on exponential weighting.

Despite their computational efficiency, the aforementioned ℓ^1-regularized methods still suffer from a drawback: their ℓ^0-oracle properties hold under rather restrictive assumptions on the (fixed or random) design; namely, that the covariates should be nearly orthogonal. We refer the reader to [vdGB09] for a detailed discussion on these assumptions.

Recently an attempt has thus been made to reach a compromise between strong theoretical guarantees (that hold under very weak assumptions on the design) and computational efficiency. In this respect [DT07, DT08, DT11] proposed an aggregation algorithm which is based on exponential weighting, which satisfies sharp sparsity oracle inequalities on a fixed or random design under almost no assumption on the dictionary, and which can be approximated numerically at a reasonable computational cost for large values of the ambient dimension d (cf. [DT09] who use Langevin Monte-Carlo methods). This is the algorithm from which our online forecaster SeqSEW of Chapter 3 in the deterministic setting is inspired.

More recently [RT11, AL11] designed aggregation algorithms that achieve optimal rates of

sparse aggregation in the regression model with fixed design (in the sense of [RT11]). [AL11] also addressed the regression model with random design but the corresponding risk bounds depend as in [DT11] on the logarithms of $\|\boldsymbol{u}^*\|_1$ and T (but their bound holds with large probability). In both papers [RT11, AL11] the corresponding algorithms were shown to be well approximated by MCMC methods with conclusive experimental results.

2.6.4 Some interesting consequences of sparsity oracle inequalities

As detailed in [BTW06, BTW07a, DT08], sparsity oracle inequalities have interesting consequences. They indeed imply that:

- In the high-dimensional linear regression model (2.44), prediction is still statistically feasible under a sparsity scenario (this is the main motivation we chose to introduce the notion of sparsity oracle inequality).

- Statistical procedures satisfying such sparsity oracle inequalities can be used to perform adaptive nonparametric regression (i.e., for an appropriately well chosen basis, these procedures are adaptive to the unknown smoothness of the regression function f). See also [BM01a, Mas07].

- Statistical procedures satisfying sharp sparsity oracle inequalities achieve (quasi-)optimal rates of model-selection, convex, and linear aggregation in the sense of [Nem00, Tsy03]. Namely, up to some small remainder terms, these procedures predict at least as well as the best among the base predictors $\boldsymbol{\varphi}_j$ (model-selection aggregation), the best convex combination of the $\boldsymbol{\varphi}_j$ (convex aggregation), and the best linear combination of the $\boldsymbol{\varphi}_j$ (linear aggregation); the corresponding remainder terms are the smallest possible ones. Further details can be found, e.g., in [RT11, Section 6] (note that there are also other types of aggregation than the three ones mentioned above, such as D-convex aggregation [Lou07]).

2.A Proofs

Proof (of Lemma 2.2): Note that the lower bound is trivial[21] if $K = 1$. Therefore, we assume in the sequel that $K \geqslant 2$. In the sequel $\mathrm{Ber}(q)$ denotes the Bernoulli distribution with parameter $q \in [0,1]$.

The proof technique is due to [CBLS05, Sto10b] and relies on arguments of [ACBFS02]. Consider the space $\Omega = (\{0,1\}^K)^T$ endowed with its discrete σ-algebra. For all $1 \leqslant t \leqslant T$, define the random variable $\boldsymbol{Y}_t : (\{0,1\}^K)^T \to \{0,1\}^K$ as the t-th coordinate mapping on $(\{0,1\}^K)^T$. We equip Ω with a family of probability distributions $\big(Q_j^{\otimes T}\big)_{1\leqslant j\leqslant K}$, where we set $Q_j = \bigotimes_{i=1}^K \mathrm{Ber}\big(1/2 - \varepsilon\,\mathbb{I}_{\{i=j\}}\big)$ for some $\varepsilon \in (0,1/2)$ to be determined by the analysis. The proof is dedicated to show that

$$\inf_S \max_{1\leqslant j\leqslant K} \mathbb{E}_{Q_j^{\otimes T}}\left[\sum_{t=1}^T \widehat{\boldsymbol{a}}_t \cdot \boldsymbol{Y}_t - \min_{1\leqslant i\leqslant K} \sum_{t=1}^T Y_{i,t}\right] \geqslant c_2\sqrt{\frac{T}{2}\ln K}\,. \tag{2.48}$$

[21] Indeed, the expected regret is nonnegative for the i.i.d. sequence $\boldsymbol{Y}_1 = \ldots = \boldsymbol{Y}_T = 0$, and $\sqrt{(T/2)\ln 1} = 0$.

This will then conclude the proof since $\ell(\widehat{\boldsymbol{a}}_t, \boldsymbol{Y}_t) = \widehat{\boldsymbol{a}}_t \cdot \boldsymbol{Y}_t$ and $\ell(\boldsymbol{\delta}_i, \boldsymbol{Y}_t) = \boldsymbol{\delta}_i \cdot \boldsymbol{Y}_t = Y_{i,t}$ for all $i \in \{1, \dots, K\}$ and $t \in \{1, \dots, T\}$ almost surely.

Note that, by construction, the random vectors $\boldsymbol{Y}_t = (Y_{1,t}, \dots, Y_{K,t})$, $1 \leqslant t \leqslant T$, are such that for all $j = 1, \dots, K$, under $Q_j^{\otimes T}$,

- the real random variables $Y_{i,t}$, $1 \leqslant i \leqslant K$, $1 \leqslant t \leqslant T$, are independent;
- for all $i \in \{1, \dots, K\}$, the random sequence $\big(Y_{i,t}\big)_{1 \leqslant t \leqslant T}$ associated with the i-th action is an i.i.d. Bernoulli sequence with parameter $1/2$ (if $i \neq j$) or $1/2 - \varepsilon$ (if $i = j$).

By Fano's lemma, we show next that if ε is small enough, then the forecaster cannot identify the best action j too quickly uniformly over all distributions $Q_j^{\otimes T}$, and therefore incurs a regret at least of the order of $\sqrt{T \ln K}$ for at least one distribution $Q_j^{\otimes T}$.

Let $S = \big(\widehat{\boldsymbol{a}}_t\big)_{t \geqslant 1}$ be any strategy of the forecaster. We split below the expected regret into two parts. On the one hand, denoting by $\boldsymbol{Y}_{1:t-1} \triangleq \big(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_{t-1}\big)$ the whole[22] information available to the forecaster before making its prediction at time t, and noting that $\widehat{\boldsymbol{a}}_t \cdot \boldsymbol{Y}_t = \sum_{i=1}^K \widehat{a}_{i,t} Y_{i,t}$, we get by the tower rule that, for all $j \in \{1, \dots, K\}$,

$$\begin{aligned}
\mathbb{E}_{Q_j^{\otimes T}}\left[\sum_{t=1}^T \widehat{\boldsymbol{a}}_t \cdot \boldsymbol{Y}_t\right] &= \sum_{t=1}^T \sum_{i=1}^K \mathbb{E}_{Q_j^{\otimes T}}\Big[\mathbb{E}_{Q_j^{\otimes T}}\big[\widehat{a}_{i,t} Y_{i,t} \,\big|\, \boldsymbol{Y}_{1:t-1}\big]\Big] \\
&= \sum_{t=1}^T \sum_{i=1}^K \mathbb{E}_{Q_j^{\otimes T}}\big[\widehat{a}_{i,t}\big]\left(\frac{1}{2} - \varepsilon \mathbb{I}_{\{i=j\}}\right) \qquad (2.49) \\
&= \frac{T}{2} - \varepsilon \sum_{t=1}^T \mathbb{E}_{Q_j^{\otimes T}}\big[\widehat{a}_{j,t}\big]\,, \qquad (2.50)
\end{aligned}$$

where (2.49) follows from the fact that $\widehat{a}_{i,t}$ is $\boldsymbol{Y}_{1:t-1}$-measurable (recall that the experts' advice $\boldsymbol{a}_{i,t} = \boldsymbol{\delta}_i$ are deterministic) and from the fact that, under $Q_j^{\otimes T}$, $Y_{i,t} \sim \mathrm{Ber}\big(1/2 - \varepsilon\, \mathbb{I}_{\{i=j\}}\big)$ is independent of $\boldsymbol{Y}_{1:t-1}$. As for (2.50), it follows from the almost sure equality $\sum_{i=1}^K \widehat{a}_{i,t} = 1$ (since $\widehat{\boldsymbol{a}}_t \in \mathcal{X}_K$).

Next we introduce an external randomization (as in [CBLS05, Sto10b]). Let $(\Omega_{\mathrm{ext}}, \mathcal{B}_{\mathrm{ext}}, \mathbb{Q}_{\mathrm{ext}})$ be a probability space, and let $I_1, \dots, I_T \in \{1, \dots, K\}$ be random variables defined on the augmented space $(\{0,1\}^K)^T \times \Omega_{\mathrm{ext}}$ such that[23] I_t is measurable with respect to the σ-field $\sigma(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_{t-1}) \otimes \mathcal{B}_{\mathrm{ext}}$, and, for all $j \in \{1, \dots, K\}$,

$$\forall t \in \{1, \dots, T\}, \quad \forall i \in \{1, \dots, K\}, \quad Q_j^{\otimes T} \otimes \mathbb{Q}_{\mathrm{ext}}\Big[I_t = i \,\Big|\, (\boldsymbol{Y}_1, I_1), \dots, (\boldsymbol{Y}_{t-1}, I_{t-1})\Big] = \widehat{a}_{i,t}\,.$$

(Recall that $\sum_{i=1}^K \widehat{a}_{i,t} = 1$ almost surely.) By the property above, (2.50) can be rewritten for all

[22] Since the expert advice are constant and known to the forecaster (they are given by $\boldsymbol{a}_{i,t} = \boldsymbol{\delta}_i$), the only useful information at time t is $\boldsymbol{Y}_{1:t-1} = \big(\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_{t-1}\big)$.

[23] The random variables I_t can be constructed as follows: at each time $t = 1, \dots, T$, pick $I_t \in \{1, \dots, K\}$ at random such that $I_t = i$ with probability $\widehat{a}_{i,t}$ (conditionally on the past data $(\boldsymbol{Y}_1, I_1), \dots, (\boldsymbol{Y}_{t-1}, I_{t-1})$).

$j \in \{1, \dots, K\}$ as

$$\mathbb{E}_{Q_j^{\otimes T}}\left[\sum_{t=1}^{T} \widehat{\boldsymbol{a}}_t \cdot \boldsymbol{Y}_t\right] = \frac{T}{2} - \varepsilon \sum_{t=1}^{T} Q_j^{\otimes T} \otimes \mathbb{Q}_{\mathrm{ext}}\big[I_t = j\big] \,. \tag{2.51}$$

On the other hand, by Jensen's inequality and by definition of Q_j, we get, for all $j \in \{1, \dots, K\}$,

$$\mathbb{E}_{Q_j^{\otimes T}}\left[\min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} Y_{i,t}\right] \leqslant \min_{1 \leqslant i \leqslant K} \mathbb{E}_{Q_j^{\otimes T}}\left[\sum_{t=1}^{T} Y_{i,t}\right] = \mathbb{E}_{Q_j^{\otimes T}}\left[\sum_{t=1}^{T} Y_{j,t}\right] = \frac{T}{2} - T\varepsilon \,.$$

Combining the last inequality with (2.51), we can lower bound the expected regret under each probability distribution $Q_j^{\otimes T}$ and get that

$$\max_{1 \leqslant j \leqslant K} \mathbb{E}_{Q_j^{\otimes T}}\left[\sum_{t=1}^{T} \widehat{\boldsymbol{a}}_t \cdot \boldsymbol{Y}_t - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} Y_{i,t}\right] \geqslant T\varepsilon \left(1 - \min_{1 \leqslant j \leqslant K} \frac{1}{T} \sum_{t=1}^{T} Q_j^{\otimes T} \otimes \mathbb{Q}_{\mathrm{ext}}\big[I_t = j\big]\right) \,. \tag{2.52}$$

To conclude the proof of (2.48), it suffices to lower bound the minimum in the parentheses by a positive absolute constant for ε of the order of $\sqrt{(\ln K)/T}$. But, by the extension of Fano's lemma to convex combinations due to [CBLS05] (see Lemma A.10 in Appendix A.7) and by the fact that $K \geqslant 2$, we get

$$\min_{1 \leqslant j \leqslant K} \frac{1}{T} \sum_{t=1}^{T} Q_j^{\otimes T} \otimes \mathbb{Q}_{\mathrm{ext}}\big[I_t = j\big] \leqslant \max\left\{\frac{2e}{2e+1}, \frac{\bar{\mathcal{K}}}{\ln K}\right\} \,, \tag{2.53}$$

where

$$\bar{\mathcal{K}} \triangleq \frac{1}{K-1} \sum_{j=2}^{K} \sum_{t=1}^{T} \frac{1}{T} \mathcal{K}\Big(Q_j^{\otimes T} \otimes \mathbb{Q}_{\mathrm{ext}}, Q_1^{\otimes T} \otimes \mathbb{Q}_{\mathrm{ext}}\Big) = \frac{T}{K-1} \sum_{j=2}^{K} \mathcal{K}\big(Q_j, Q_1\big) \,. \tag{2.54}$$

In the last equality, we used the fact that $\mathcal{K}\big(Q_j^{\otimes T} \otimes \mathbb{Q}_{\mathrm{ext}}, Q_1^{\otimes T} \otimes \mathbb{Q}_{\mathrm{ext}}\big) = T\mathcal{K}\big(Q_j, Q_1\big)$ by the chain rule for the Kullback-Leibler divergence. But, noting that for all $j = 2, \dots, K$, the probability distributions $Q_j = \otimes_{i=1}^{K} \mathrm{Ber}\big(1/2 - \varepsilon\, \mathbb{I}_{\{i=j\}}\big)$ and $Q_1 = \otimes_{i=1}^{K} \mathrm{Ber}\big(1/2 - \varepsilon\, \mathbb{I}_{\{i=1\}}\big)$ only differ on the two actions 1 and j by ε, we have, using again the chain rule for the Kullback-Leibler divergence,

$$\mathcal{K}\big(Q_j, Q_1\big) = \mathcal{K}\Big(\mathrm{Ber}(1/2), \mathrm{Ber}(1/2 - \varepsilon)\Big) + \mathcal{K}\Big(\mathrm{Ber}(1/2 - \varepsilon), \mathrm{Ber}(1/2)\Big) \leqslant 5\varepsilon^5 \,, \tag{2.55}$$

where the last inequality is proved in [Sto05, Lemma A.5] for all $0 \leqslant \varepsilon \leqslant 1/10$. Putting (2.53), (2.54), and (2.55) together, we get that, for all $0 < \varepsilon \leqslant 1/10$,

$$\min_{1 \leqslant j \leqslant K} \frac{1}{T} \sum_{t=1}^{T} Q_j^{\otimes T} \otimes \mathbb{Q}_{\mathrm{ext}}\big[I_t = j\big] \leqslant \max\left\{\frac{2e}{2e+1}, \frac{5T\varepsilon^2}{\ln K}\right\} = \frac{2e}{2e+1} \,, \tag{2.56}$$

where the last equality follows from the choice of

$$\varepsilon = \sqrt{\frac{2e}{2e+1} \frac{\ln K}{5T}} \,,$$

which is indeed smaller than $1/10$ (as required) if $T \geqslant [40e/(2e+1)] \ln K$. Substituting (2.56) in (2.52), we get (2.48) by setting $c_1 \triangleq 40e/(2e+1)$ and $c_2 \triangleq [2/(2e+1)]\sqrt{e/[5(2e+1)]}$. This concludes the proof. $\square$

Remark 2.5. *In the proof above, we showed via a version of Fano's lemma that there exists a probability distribution under which the random vectors $\boldsymbol{Y}_t$, $1 \leqslant t \leqslant T$, are i.i.d. and such that the expected regret is at least of the order of $\sqrt{T \ln K}$ (see (2.48)). This probability distribution is of the form $Q_{j^*}^{\otimes K}$, where $j^* \in \{1, \ldots, K\}$ minimizes the left-hand side of (2.53). Therefore, it depends on the strategy of the forecaster $(\widehat{\boldsymbol{a}}_t)_{t \geqslant 1}$ through $(I_t)_{t \geqslant 1}$.*

Note that we could have used another variant of Fano's lemma[24] for $K \geqslant 3$ or Pinsker's inequality (see Appendix A.7) for $K = 2$ to prove that the expected regret under $(1/K) \sum_{i=1}^K Q_j^{\otimes T}$ is also at least of the order of $\sqrt{T \ln K}$ for any *strategy of the forecaster. Interestingly, the probability distribution $(1/K) \sum_{i=1}^K Q_j^{\otimes T}$ is now independent of the forecaster. Besides, the aforementioned $\sqrt{T \ln K}$ lower bound on the expected regret under $(1/K) \sum_{i=1}^K Q_j^{\otimes T}$ yields a lower bound similar to (2.48), at the price of worst constants though.*

[24] See, e.g., [Bir01] and the references therein.

Chapter 3

Sparsity regret bounds for individual sequences in online linear regression

We consider the problem of online linear regression on arbitrary deterministic sequences when the ambient dimension d can be much larger than the number of time rounds T. We introduce the notion of sparsity regret bound, which is a deterministic online counterpart of the so-called sparsity oracle inequalities from the stochastic setting. We prove such regret bounds for an online-learning algorithm called SeqSEW and based on exponential weighting and data-driven truncation. In a second part we apply a parameter-free version of this algorithm to the regression model with random design (i.i.d. data) and derive risk bounds of the same flavor as in [DT11] but which solve two questions left open therein. In particular our risk bounds are adaptive (up to a logarithmic factor) to the unknown variance of the noise if the latter is Gaussian. We also address the regression model with fixed design as in [DT08].

NOTA: This chapter is the full version (with extensive proofs) of a conference paper [Ger11a] that appeared in the proceedings of COLT 2011. Corollary 3.5 and Section 3.4.2 are published here for the first time.

Contents

3.1 Introduction

Sparsity has been extensively studied in the stochastic setting over the past decade. Among the theoretical tools introduced for this purpose, the notion of *sparsity oracle inequality* plays a fundamental role. In high-dimensional linear regression, such inequalities indicate that the task consisting in predicting almost as well as an unknown target vector is still statistically feasible if the target vector has only few non-zero coordinates. A detailed motivation of such risk bounds and some bibliographic references are provided in Section 2.6 of Chapter 2.

In this chapter, we bring the notion of sparsity oracle inequality into the framework of prediction of individual sequences (of deterministic nature). The corresponding deterministic inequalities are called *sparsity regret bounds*. We prove such bounds for an online-learning algorithm called *SeqSEW* which is inspired from the Sparse Exponential Weighting algorithm introduced in the stochastic setting by [DT07]. Thanks to individual sequences techniques (e.g., online truncation and online tuning), the most sophisticated version of our algorithm is fully automatic in the sense that no a priori knowledge is needed for the choice of the tuning parameters.

The second contribution of this chapter deals with fruitful connections between the framework of individual sequences and the stochastic setting. More precisely, we show that, via the standard online to batch trick, the online truncation and parameter tuning performed by the algorithm SeqSEW for deterministic purposes yield, in the regression model with random or fixed design, sparsity oracle inequalities with leading constant 1 which are of the same flavor as in [DT08, DT11]. In addition our bounds are adaptive to the unknown variance σ^2 of the noise (up to a logarithmic factor) at least whenever the latter is Gaussian; weaker bounds are also proved under weaker assumptions. Therefore, in the batch stochastic setting, individual sequence techniques appear to be useful for adaptation purposes.

In the next paragraphs, we introduce our main setting and motivate the notion of sparsity regret bound from an online learning viewpoint (this motivation can be paralleled to that of Section 2.6.1 in Chapter 2). We then detail our main contributions with respect to the statistical literature and the machine learning literature.

Introduction of a deterministic counterpart of sparsity oracle inequalities

We consider the problem of online linear regression on arbitrary deterministic sequences. A forecaster has to predict in a sequential fashion the values $y_t \in \mathbb{R}$ of an unknown sequence of observations given some input data $x_t \in \mathcal{X}$ and some base forecasters $\varphi_j : \mathcal{X} \to \mathbb{R}$, $1 \leqslant j \leqslant d$, on the basis of which he outputs a prediction $\widehat{y}_t \in \mathbb{R}$. The quality of the predictions is assessed by the square loss. The goal of the forecaster is to predict almost as well as the best linear forecaster $\boldsymbol{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^{d} u_j \varphi_j$, where $\boldsymbol{u} \in \mathbb{R}^d$, i.e., to satisfy, uniformly over all individual sequences $(x_t, y_t)_{1 \leqslant t \leqslant T}$, a regret bound of the form

$$\sum_{t=1}^{T} \big(y_t - \widehat{y}_t\big)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + \Delta_{T,d}(\boldsymbol{u}) \right\}$$

for some regret term $\Delta_{T,d}(\boldsymbol{u})$ that should be as small as possible and, in particular, sublinear in T. (For the sake of introduction, we omit the dependencies of $\Delta_{T,d}(\boldsymbol{u})$ on the amplitudes $\max_{1\leqslant t\leqslant T}\|\boldsymbol{x}_t\|_\infty$ and $\max_{1\leqslant t\leqslant T}|y_t|$.)

In this setting the version of the sequential ridge regression forecaster[1] studied by [AW01] and [Vov01] and tuned with the illegal optimal tuning of Section 2.4.2 has a regret $\Delta_{T,d}(\boldsymbol{u})$ of order at most $d\ln\bigl(T\|\boldsymbol{u}\|_2^2\bigr)$; see (2.27) in the aforementioned section. When the ambient dimension d is much larger than the number of time rounds T, the bound $d\ln T$ may unfortunately be larger than T and is thus somehow trivial. Since the regret bound $d\ln T$ is optimal in a certain sense (see [Vov01, Theorem 2]), additional assumptions are needed to get interesting theoretical guarantees.

A natural assumption, which has already been extensively studied in the stochastic setting, is that there is a sparse linear combination $\boldsymbol{u}^*$ (i.e., with $s\ll T/(\ln T)$ non-zero coefficients) which has a small cumulative square loss. If the forecaster knew in advance the support $J(\boldsymbol{u}^*)\triangleq\{j:u_j^*\neq 0\}$ of $\boldsymbol{u}^*$, he could apply the same forecaster as above but only to the s-dimensional linear subspace $\bigl\{\boldsymbol{u}\in\mathbb{R}^d:\forall j\notin J(\boldsymbol{u}^*),u_j=0\bigr\}$. The regret bound of this "oracle" would be roughly of order $s\ln T$ and thus sublinear in T. Under this sparsity scenario, a sublinear regret thus seems possible, though, of course, the aforementioned regret bound $s\ln T$ can only be used as an ideal benchmark (since the support of $\boldsymbol{u}^*$ is unknown).

In this chapter, we prove that a regret bound proportional to s is achievable (up to logarithmic factors). In Corollary 3.1 and its refinements (Corollary 3.2 and Theorem 3.1), we indeed derive regret bounds of the form

$$\sum_{t=1}^T(y_t-\widehat{y}_t)^2\leqslant\inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^T\bigl(y_t-\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\bigr)^2+\bigl(\|\boldsymbol{u}\|_0+1\bigr)\,g_{T,d}\bigl(\|\boldsymbol{u}\|_1,\|\boldsymbol{\varphi}\|_\infty\bigr)\right\}, \tag{3.1}$$

where $\|\boldsymbol{u}\|_0$ denotes the number of non-zero coordinates of $\boldsymbol{u}$ and where g is increasing but grows at most logarithmically in T, d, $\|\boldsymbol{u}\|_1\triangleq\sum_{j=1}^d|u_j|$, and $\|\boldsymbol{\varphi}\|_\infty\triangleq\sup_{x\in\mathcal{X}}\max_{1\leqslant j\leqslant d}|\varphi_j(x)|$. We call regret bounds of the above form *sparsity regret bounds*.

This work is in connection with several papers that belong either to the statistical or to the machine learning literature. Next we discuss these papers and some related references.

Related works in the stochastic setting

The above regret bound (3.1) can be seen as a deterministic online counterpart of the so-called *sparsity oracle inequalities* introduced in the stochastic setting in the past decade. The latter are risk bounds expressed in terms of the number of non-zero coefficients of the oracle vector. Such inequalities were derived by [BM01a] through model selection arguments and later developed by, e.g., [BM07a, BTW07a] in the regression model with fixed design and by [BTW04] for the case of a random design. An introduction to the notion of sparsity oracle inequality can be found in Section 2.6 (Chapter 2); we refer the reader to this section for further references.

We only mention that, recently, sparsity oracle inequalities with leading constant equal to 1

[1] This forecaster is recalled in Chapter 2; see (2.26) in Section 2.4.2.

were proved for procedures based on exponential weighting; see [DT07, DT08, RT11, AL11] for the regression model with fixed design and [DT11, AL11] for the regression model with random design. These papers show that a trade-off can be reached between strong theoretical guarantees (as with ℓ^0-regularization) and computational efficiency (as with ℓ^1-regularization). They indeed propose aggregation algorithms which satisfy sparsity oracle inequalities under almost no assumption on the base forecasters $(\varphi_j)_j$, and which can be approximated numerically at a reasonable computational cost for large values of the ambient dimension d.

Our online-learning algorithm SeqSEW is inspired from [DT08, DT11]. Following the same lines as in [DT09], it is possible to slightly adapt its statement to make it computationally tractable by means of Langevin Monte-Carlo approximation while not affecting its statistical properties. The technical details are however omitted in this chapter, which only focuses on the theoretical guarantees of the algorithm SeqSEW.

Previous works on sparsity in the framework of individual sequences

To the best of our knowledge, Corollary 3.1 and its refinements (Corollary 3.2 and Theorem 3.1) provide the first examples of sparsity regret bounds in the sense of (3.1). To comment on the optimality of such regret bounds and compare them to related results in the framework of individual sequences, note that (3.1) can be rewritten in the equivalent form:

For all $s \in \mathbb{N}$ and all $U > 0$,

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 - \inf_{\substack{\|\boldsymbol{u}\|_0 \leqslant s \\ \|\boldsymbol{u}\|_1 \leqslant U}} \sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 \leqslant \big(s+1\big)\, g_{T,d}\big(U, \|\boldsymbol{\varphi}\|_\infty\big)\,,$$

where g grows at most logarithmically in T, d, U, and $\|\boldsymbol{\varphi}\|_\infty$. When $s \ll T$, this upper bound matches (up to logarithmic factors) the lower bound of order $s \ln T$ that follows in a straightforward manner from [Vov01, Theorem 2] or [CBL06, Chapter 11]. Indeed, if $s \ll T$, $\mathcal{X} = \mathbb{R}^d$, and $\varphi_j(x) = x_j$, then for any forecaster, there is an individual sequence $(x_t, y_t)_{1\leqslant t\leqslant T}$ such that the regret of this forecaster on $\big\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_0 \leqslant s \text{ and } \|\boldsymbol{u}\|_1 \leqslant d\big\}$ is bounded from below by a quantity of order $s \ln T$. Therefore, up to logarithmic factors, any algorithm satisfying a sparsity regret bound of the form (3.1) is minimax optimal on intersections of ℓ^0-balls (of radii $s \ll T$) and ℓ^1-balls. This is in particular the case for our algorithm SeqSEW, but this contrasts with related works discussed below.

Recent works in the field of online convex optimization addressed the sparsity issue in the online deterministic setting, but from a quite different angle. They focus on algorithms which output sparse linear combinations, while we are interested in algorithms whose regret is small under a sparsity scenario, i.e., on ℓ^0-balls of small radii. See, e.g., [LLZ09, SST09, Xia10, DSSST10] and the references therein. All these articles focus on convex regularization. In the particular case of ℓ^1-regularization under the square loss, the aforementioned works propose algorithms which predict as a sparse linear combination $\widehat{y}_t = \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{\varphi}(x_t)$ of the base forecasts (i.e., $\|\widehat{\boldsymbol{u}}_t\|_0$ is small), while no such guarantee can be proved for our algorithm SeqSEW. However they prove bounds on

the ℓ^1-regularized regret of the form

$$\sum_{t=1}^{T}\Big((y_t-\widehat{\boldsymbol{u}}_t\cdot\boldsymbol{x}_t)^2+\lambda\left\|\widehat{\boldsymbol{u}}_t\right\|_1\Big)\leqslant\inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\Big((y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t)^2+\lambda\left\|\boldsymbol{u}\right\|_1\Big)+\widetilde{\Delta}_{T,d}(\boldsymbol{u})\right\}\ , \quad (3.2)$$

for some regret term $\widetilde{\Delta}_{T,d}(\boldsymbol{u})$ which is suboptimal on intersections of ℓ^0- and ℓ^1-balls as explained below. The truncated gradient algorithm of [LLZ09, Corollary 4.1] satisfies[2] such a regret bound with $\widetilde{\Delta}_{T,d}(\boldsymbol{u})$ at least of order $\left\|\boldsymbol{\varphi}\right\|_\infty\sqrt{dT}$ when the base forecasts $\varphi_j(x_t)$ are dense in the sense that $\max_{1\leqslant t\leqslant T}\sum_{j=1}^d\varphi_j^2(x_t)\approx d\,\left\|\boldsymbol{\varphi}\right\|_\infty^2$. This regret bound grows as a power of and not logarithmically in d as is expected for sparsity regret bounds (recall that we are interested in the case when $d\gg T$).

The three other papers mentioned above do prove (some) regret bounds with a logarithmic dependence in d, but these bounds do not have the dependence in $\left\|\boldsymbol{u}\right\|_1$ and T we are looking for. For $p-1\approx 1/(\ln d)$, the p-norm RDA method of [Xia10] and the algorithm SMIDAS of [SST09] – the latter being a particular case of the algorithm COMID of [DSSST10] specialized to the p-norm divergence – satisfy regret bounds of the above form (3.2) with $\widetilde{\Delta}_{T,d}(\boldsymbol{u})\approx\mu\left\|\boldsymbol{u}\right\|_1\sqrt{T\ln d}$, for some gradient-based constant μ. Therefore, in all three cases, the function $\widetilde{\Delta}$ grows at least linearly in $\left\|\boldsymbol{u}\right\|_1$ and as $\sqrt{T}$. This is in contrast with the logarithmic dependence in $\left\|\boldsymbol{u}\right\|_1$ and the fast rate $\mathcal{O}(\ln T)$ we are looking for and prove, e.g., in Corollary 3.1.

Note that the suboptimality of the aforementioned algorithms is specific to the goal we are pursuing, i.e., prediction on ℓ^0-balls (intersected with ℓ^1-balls). On the contrary the rate $\left\|\boldsymbol{u}\right\|_1\sqrt{T\ln d}$ is more suited and actually nearly optimal for learning on ℓ^1-balls (see Chapter 4). Moreover, the predictions output by our algorithm SeqSEW are not necessarily sparse linear combinations of the base forecasts. A question left open is thus whether it is possible to design an algorithm which both ouputs sparse linear combinations (which is statistically useful and sometimes essential for computational issues) and satisfies a sparsity regret bound of the form (3.1).

PAC-Bayesian analysis in the framework of individual sequences

To derive our sparsity regret bounds, we follow a PAC-Bayesian approach combined with the choice of a sparsity-favoring prior. We do not have the space to review the PAC-Bayesian literature in the stochastic setting and only refer the reader to [Cat04] for a thorough introduction to the subject. As for the online deterministic setting, PAC-Bayesian-type inequalities were proved in the framework of prediction with expert advice, e.g., in [FSSW97] and [KW99], or in the same setting as ours with a Gaussian prior in [Vov01]. More recently, [Aud09] proved a PAC-Bayesian result on individual sequences for general losses and prediction sets. The latter result relies on a unifying assumption called the online variance inequality, which holds true, e.g., when the loss function is exp-concave. In the present chapter, we only focus on the particular case of the square loss. We first use Theorem 4.6 of [Aud09] to derive a non-adaptive sparsity regret bound. We then

[2]The bound stated in [LLZ09, Corollary 4.1] differs from (3.2) in that the constant before the infimum is equal to $C=1/(1-2c_d^2\eta)$, where $c_d^2\approx\max_{1\leqslant t\leqslant T}\sum_{j=1}^d\varphi_j^2(x_t)\leqslant d\left\|\boldsymbol{\varphi}\right\|_\infty^2$, and where a reasonable choice for η can easily be seen to be $\eta\approx 1/\sqrt{2c_d^2T}$. If the base forecasts $\varphi_j(x_t)$ are dense in the sense that $c_d^2\approx d\left\|\boldsymbol{\varphi}\right\|_\infty^2$, then we have $C\approx 1+\sqrt{2c_d^2/T}$, which yields a regret bound with leading constant 1 as in (3.2) and with $\widetilde{\Delta}_{T,d}(\boldsymbol{u})$ at least of order $\sqrt{c_d^2T}\approx\left\|\boldsymbol{\varphi}\right\|_\infty\sqrt{dT}$.

provide an adaptive online PAC-Bayesian inequality to automatically adapt to the unknown range of the observations $\max_{1\leqslant t\leqslant T}|y_t|$.

Open questions by Dalalyan and Tsybakov

In Section 3.4.1 we apply a parameter-free version of our algorithm SeqSEW on i.i.d. data and derive a risk bound of the same flavor as in [DT11]. However, our risk bound holds on the whole $\mathbb{R}^d$ space instead of ℓ^1-balls of finite radii, which solves one question left open by [DT11, Section 4.2]. Besides, our algorithm does not need the a priori knowledge of the variance factor of the noise when the latter is subgaussian, which solves a second question raised in [DT11, Section 5.1, Remark 6].

Outline of the chapter

This chapter is organized as follows. In Section 3.2 we describe our main (deterministic) setting as well as our main notations. In Section 3.3 we prove the aforementioned sparsity regret bounds for our algorithm SeqSEW, first when the forecaster has access to some a priori knowledge on the observations (Sections 3.3.1 and 3.3.2), and then when no a priori information is available (Section 3.3.3), which yields a fully automatic algorithm. In Section 3.4 we apply the algorithm SeqSEW to the regression model with random design (Section 3.4.1) and to the regression model with fixed design (Section 3.4.2). Some technical tools are finally given in appendix.

3.2 Setting and notations

The main setting considered in this chapter is an instance of the game of prediction with expert advice called *prediction with side information (under the square loss)* or, more simply, *online linear regression*. This online protocol is described in Figure 3.1. An introduction to this setting is provided in Section 2.4 of Chapter 2.

Note that our online protocol is described as if the environment were oblivious to the forecaster's predictions. Actually, since we only consider deterministic forecasters, all regret bounds of this chapter also hold when $(x_t)_{t\geqslant 1}$ and $(y_t)_{t\geqslant 1}$ are chosen by an adversarial environment. See Section 2.3.1 of Chapter 2 for further details.

Two stochastic batch settings are also considered later in this chapter. See Section 3.4.1 for the regression model with random design, and Section 3.4.2 for the regression model with fixed design.

Some notations

We now define some notations. Vectors in $\mathbb{R}^d$ will be denoted by bold letters. For all $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^d$, the standard inner product in $\mathbb{R}^d$ between $\boldsymbol{u}=(u_1,\dots,u_d)$ and $\boldsymbol{v}=(v_1,\dots,v_d)$ will be denoted

Parameters: input data set $\mathcal{X}$, base forecasters $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_d)$ with $\varphi_j : \mathcal{X} \to \mathbb{R}$, $1 \leqslant j \leqslant d$.

Initial step: the environment chooses a sequence of observations $(y_t)_{t \geqslant 1}$ in $\mathbb{R}$ and a sequence of input data $(x_t)_{t \geqslant 1}$ in $\mathcal{X}$ but the forecaster has not access to them.

At each time round $t \in \mathbb{N}^*$,

1. The environment reveals the input data $x_t \in \mathcal{X}$.
2. The forecaster chooses a prediction $\widehat{y}_t \in \mathbb{R}$
   (possibly as a linear combination of the $\varphi_j(x_t)$, but this is not necessary).
3. The environment reveals the observation $y_t \in \mathbb{R}$.
4. Each linear forecaster $\boldsymbol{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^d u_j \varphi_j$, $\boldsymbol{u} \in \mathbb{R}^d$, incurs the loss $\big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2$ and the forecaster incurs the loss $(y_t - \widehat{y}_t)^2$.

Figure 3.1: The online linear regression setting.

by $\boldsymbol{u} \cdot \boldsymbol{v} = \sum_{i=j}^d u_j \, v_j$; the ℓ^0-, ℓ^1-, and ℓ^2-norms of $\boldsymbol{u} = (u_1, \dots, u_d)$ are respectively defined by

$$\|\boldsymbol{u}\|_0 \triangleq \sum_{j=1}^d \mathbb{I}_{\{u_j \neq 0\}} = \big|\{j : u_j \neq 0\}\big|, \qquad \|\boldsymbol{u}\|_1 \triangleq \sum_{j=1}^d |u_j|, \qquad \text{and} \qquad \|\boldsymbol{u}\|_2 \triangleq \left(\sum_{j=1}^d u_j^2\right)^{1/2}.$$

The set of all probability distributions on a set Θ (endowed with some σ-algebra, e.g., the Borel σ-algebra when $\Theta = \mathbb{R}^d$) will be denoted by $\mathcal{M}_1^+(\Theta)$. For all $\rho, \pi \in \mathcal{M}_1^+(\Theta)$, the Kullback-Leibler divergence between ρ and π is defined by

$$\mathcal{K}(\rho, \pi) \triangleq \begin{cases} \displaystyle\int_{\mathbb{R}^d} \ln\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right) \mathrm{d}\rho & \text{if } \rho \text{ is absolutely continuous with respect to } \pi; \\ +\infty & \text{otherwise,} \end{cases}$$

where $\frac{\mathrm{d}\rho}{\mathrm{d}\pi}$ denotes the Radon-Nikodym derivative of ρ with respect to π.

For all $x \in \mathbb{R}$ and $B > 0$, we denote by $\lceil x \rceil$ the smallest integer larger than or equal to x, and by $[x]_B$ its thresholded (or clipped) value:

$$[x]_B \triangleq \begin{cases} -B & \text{if } x < -B; \\ x & \text{if } -B \leqslant x \leqslant B; \\ B & \text{if } x > B. \end{cases}$$

Finally, we will use the (natural) conventions $1/0 = +\infty$, $(+\infty) \times 0 = 0$, and $0 \ln(1 + U/0) = 0$ for all $U \geqslant 0$. Any sum $\sum_{s=1}^0 a_s$ indexed from 1 up to 0 is by convention equal to 0.

3.3 Sparsity regret bounds for individual sequences

In this section we prove sparsity regret bounds for different variants of our algorithm SeqSEW. We first assume in Section 3.3.1 that the forecaster has access in advance to a bound B_y on the observations $|y_t|$ and a bound B_Φ on the trace of the empirical Gram matrix. We then remove these requirements one by one in Sections 3.3.2 and 3.3.3.

3.3.1 Known bounds B_y on the observations and B_Φ on the trace of the empirical Gram matrix

To simplify the analysis, we first assume that, at the beginning of the game, the number of rounds T is known to the forecaster and that he has access to a bound B_y on all the observations $y_1, \dots, y_T$ and to a bound B_Φ on the trace of the empirical Gram matrix, i.e.,

$$y_1, \dots, y_T \in [-B_y, B_y] \qquad \text{and} \qquad \sum_{j=1}^{d} \sum_{t=1}^{T} \varphi_j^2(x_t) \leqslant B_\Phi \,.$$

The first version of the algorithm studied in this chapter is defined in Figure 3.2 (adaptive variants will be introduced later). We name it *SeqSEW* for it is a variant of the Sparse Exponential Weighting algorithm introduced in the stochastic setting by [DT07, DT08] which is tailored for the prediction of individual sequences.

The choice of the heavy-tailed prior π_τ is due to [DT07]. The role of heavy-tailed priors to tackle the sparsity issue was already pointed out earlier; see, e.g., the discussion in [See08, Section 2.1]. In high dimension, such heavy-tailed priors favor sparsity: sampling from these prior distributions (or posterior distributions based on them) typically results in approximately sparse vectors, i.e., vectors having most coordinates almost equal to zero and the few remaining ones with quite large values.

Proposition 3.1. *Assume that, for a known constant $B_y > 0$, the $(x_1, y_1), \dots, (x_T, y_T)$ are such that $y_1, \dots, y_T \in [-B_y, B_y]$. Then, for all $B \geqslant B_y$, all $\eta \leqslant 1/(8B^2)$, and all $\tau > 0$, the algorithm $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$ satisfies*

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + \frac{4}{\eta} \|\boldsymbol{u}\|_0 \ln\!\left(1 + \frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0 \, \tau}\right) \right\} + \tau^2 \sum_{j=1}^{d} \sum_{t=1}^{T} \varphi_j^2(x_t) \,. \tag{3.3}$$

Corollary 3.1. *Assume that, for some known constants $B_y > 0$ and $B_\Phi > 0$, the $(x_1, y_1), \dots, (x_T, y_T)$ are such that $y_1, \dots, y_T \in [-B_y, B_y]$ and $\sum_{j=1}^{d} \sum_{t=1}^{T} \varphi_j^2(x_t) \leqslant B_\Phi$. Then, when used with $B = B_y$, $\eta = \dfrac{1}{8B_y^2}$, and $\tau = \sqrt{\dfrac{16 B_y^2}{B_\Phi}}$, the algorithm $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$ satisfies*

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + 32\, B_y^2 \, \|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{B_\Phi}\, \|\boldsymbol{u}\|_1}{4\, B_y \, \|\boldsymbol{u}\|_0}\right) \right\} + 16 B_y^2 \,. \tag{3.4}$$

Parameters: threshold $B > 0$, inverse temperature $\eta > 0$, and prior scale $\tau > 0$ with which we associate the *sparsity prior* $\pi_\tau \in \mathcal{M}_1^+(\mathbb{R}^d)$ defined by

$$\pi_\tau(\mathrm{d}\boldsymbol{u}) \triangleq \prod_{j=1}^{d} \frac{(3/\tau)\,\mathrm{d}u_j}{2\big(1+|u_j|/\tau\big)^4} \,.$$

Initialization: $p_1 \triangleq \pi_\tau$.

At each time round $t \geqslant 1$,

1. Get the input data x_t and predict[a] as $\widehat{y}_t \triangleq \displaystyle\int_{\mathbb{R}^d} \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_B \, p_t(\mathrm{d}\boldsymbol{u})$;

2. Get the observation y_t and compute the posterior distribution $p_{t+1} \in \mathcal{M}_1^+(\mathbb{R}^d)$ as

$$p_{t+1}(\mathrm{d}\boldsymbol{u}) \triangleq \frac{\exp\left(-\eta \displaystyle\sum_{s=1}^{t}\Big(y_s - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_s)\big]_B\Big)^2\right)}{W_{t+1}} \, \pi_\tau(\mathrm{d}\boldsymbol{u}) \,,$$

where

$$W_{t+1} \triangleq \int_{\mathbb{R}^d} \exp\left(-\eta \sum_{s=1}^{t}\Big(y_s - \big[\boldsymbol{v}\cdot\boldsymbol{\varphi}(x_s)\big]_B\Big)^2\right) \pi_\tau(\mathrm{d}\boldsymbol{v}) \,.$$

[a] The clipping operator $[\cdot]_B$ is defined in Section 3.2.

Figure 3.2: The algorithm $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$.

Note that, if $\|\boldsymbol{\varphi}\|_\infty \triangleq \sup_{x\in\mathcal{X}} \max_{1\leqslant j\leqslant d} |\varphi_j(x)|$ is finite, then the last corollary provides a *sparsity regret bound* in the sense of (3.1). Indeed, in this case, we can take $B_\Phi = dT\,\|\boldsymbol{\varphi}\|_\infty^2$, which yields a regret bound proportional to $\|\boldsymbol{u}\|_0$ and that grows logarithmically in d, T, $\|\boldsymbol{u}\|_1$, and $\|\boldsymbol{\varphi}\|_\infty$.

To prove Proposition 3.1, we first need the following deterministic PAC-Bayesian inequality which is at the core of our analysis. It is a straightforward consequence of Theorem 4.6 of [Aud09] when applied to the square loss (see also Appendix 3.A.1 for a self-contained proof). An adaptive variant of this inequality will be provided in Section 3.3.2.

Lemma 3.1. *Assume that for some known constant $B_y > 0$, we have $y_1, \dots, y_T \in [-B_y, B_y]$. For all $\tau > 0$, if the algorithm $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$ is used with $B \geqslant B_y$ and $\eta \leqslant 1/(8B^2)$, then*

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\rho\in\mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^{T}\Big(y_t - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_B\Big)^2 \rho(\mathrm{d}\boldsymbol{u}) \;+\; \frac{\mathcal{K}(\rho,\pi_\tau)}{\eta} \right\} \tag{3.5}$$

$$\leqslant \inf_{\rho\in\mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 \rho(\mathrm{d}\boldsymbol{u}) \;+\; \frac{\mathcal{K}(\rho,\pi_\tau)}{\eta} \right\} . \tag{3.6}$$

Proof (of Lemma 3.1): Inequality (3.5) is a straightforward consequence of Theorem 4.6 of [Aud09] when applied to the square loss, the set of prediction functions $\mathcal{G} \triangleq \big\{x \mapsto \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x)\big]_B : \boldsymbol{u} \in \mathbb{R}^d\big\}$, and the prior[3] π on $\mathcal{G}$ induced by the prior π_τ on $\mathbb{R}^d$ via the mapping $\boldsymbol{u} \in \mathbb{R}^d \mapsto \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(\cdot)\big]_B \in \mathcal{G}$.

To apply the aforementioned theorem, recall from Appendix A.2 that the square loss is $1/(8B^2)$-exp-concave on $[-B, B]$ and thus η-exp-concave[4] (since $\eta \leqslant 1/(8B^2)$ by assumption). Therefore, by Theorem 4.6 of [Aud09] with the variance function $\delta_\eta \equiv 0$ (see the comments following Remark 4.1 therein), we get

$$\begin{aligned}\sum_{t=1}^T (y_t - \widehat{y}_t)^2 &\leqslant \inf_{\mu \in \mathcal{M}_1^+(\mathcal{G})} \left\{ \int_{\mathcal{G}} \sum_{t=1}^T \big(y_t - g(x_t)\big)^2 \mu(\mathrm{d}g) \, + \, \frac{\mathcal{K}(\mu, \pi)}{\eta} \right\} \\ &\leqslant \inf_{\rho \in \mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \Big(y_t - \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big]_B\Big)^2 \rho(\mathrm{d}\boldsymbol{u}) \, + \, \frac{\mathcal{K}(\widetilde{\rho}, \pi)}{\eta} \right\} ,\end{aligned}$$

where the last inequality follows by restricting the infimum over $\mathcal{M}_1^+(\mathcal{G})$ to the subset $\big\{\widetilde{\rho} : \rho \in \mathcal{M}_1^+(\mathbb{R}^d)\big\} \subset \mathcal{M}_1^+(\mathcal{G})$, where $\widetilde{\rho} \in \mathcal{M}_1^+(\mathcal{G})$ denotes the probability distribution induced by $\rho \in \mathcal{M}_1^+(\mathbb{R}^d)$ via the mapping $\boldsymbol{u} \in \mathbb{R}^d \mapsto \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(\cdot)\big]_B \in \mathcal{G}$. Inequality (3.5) then follows from the fact that for all $\rho \in \mathcal{M}_1^+(\mathbb{R}^d)$, we have $\mathcal{K}(\widetilde{\rho}, \pi) \leqslant \mathcal{K}(\rho, \pi_\tau)$ by joint convexity of $\mathcal{K}(\cdot, \cdot)$.

As for Inequality (3.6), it follows from (3.5) by noting that

$$\forall y \in [-B, B], \quad \forall x \in \mathbb{R}, \qquad \big|y - [x]_B\big| \leqslant |y - x| \, .$$

Therefore, truncation to $[-B, B]$ can only improve prediction under the square loss if the observations are $[-B, B]$-valued, which is the case here since by assumption $y_t \in [-B_y, B_y] \subset [-B, B]$ for all $t = 1, \ldots, T$. □

Remark 3.1. *As can be seen from the previous proof, Lemma 3.1 still holds when π_τ is replaced with* any *prior $\pi \in \mathcal{M}_1^+(\mathbb{R}^d)$ (both in the statement of the lemma and in the definition of the algorithm SeqSEW). This fact is standard in the PAC-Bayesian approach; see, e.g., [Cat04] and [DT08]. As a consequence, any algorithm satisfying (3.6) will also satisfy Proposition 3.1 and Corollary 3.1.*

Proof (of Proposition 3.1): Our proof mimics the proof of Theorem 5 in [DT08]. We thus only write the outline of the proof and stress the minor changes that are needed to derive Inequality (3.3). The key technical tools provided in [DT08] are reproduced in Appendix 3.B.1 for the convenience of the reader.

[3] The set $\mathcal{G}$ is endowed with the σ-algebra generated by all the coordinate mappings $g \in \mathcal{G} \mapsto g(x) \in \mathbb{R}$, $x \in \mathcal{X}$ (where $\mathbb{R}$ is endowed with its Borel σ-algebra).

[4] This means that for all $y \in [-B, B]$, the function $x \mapsto \exp\big(-\eta(y - x)^2\big)$ is concave on $[-B, B]$.

Let $\boldsymbol{u}^* \in \mathbb{R}^d$. Since $B \geqslant B_y$ and $\eta \leqslant 1/(8B^2)$, we can apply Lemma 3.1 and get

$$\begin{aligned}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 &\leqslant \inf_{\rho\in\mathcal{M}_1^+(\mathbb{R}^d)}\left\{\int_{\mathbb{R}^d}\sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2\rho(\mathrm{d}\boldsymbol{u}) \;+\; \frac{\mathcal{K}(\rho,\pi_\tau)}{\eta}\right\}\\ &\leqslant \underbrace{\int_{\mathbb{R}^d}\sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2\rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u})}_{(1)} + \underbrace{\frac{\mathcal{K}(\rho_{\boldsymbol{u}^*,\tau},\pi_\tau)}{\eta}}_{(2)}\,. \end{aligned} \tag{3.7}$$

In the last inequality, $\rho_{\boldsymbol{u}^*,\tau}$ is taken as the translated of π_τ at $\boldsymbol{u}^*$, namely,

$$\rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u}) \triangleq \frac{\mathrm{d}\pi_\tau}{\mathrm{d}\boldsymbol{u}}(\boldsymbol{u}-\boldsymbol{u}^*)\,\mathrm{d}\boldsymbol{u} = \prod_{j=1}^{d}\frac{(3/\tau)\,\mathrm{d}u_j}{2\big(1+|u_j-u_j^*|/\tau\big)^4}\,.$$

The two terms (1) and (2) can be upper bounded as in the proof of Theorem 5 in [DT08]. By a symmetry argument recalled in Lemma 3.3, the first term (1) can be rewritten as

$$\int_{\mathbb{R}^d}\sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2\rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u}) = \sum_{t=1}^{T}\big(y_t - \boldsymbol{u}^*\cdot\boldsymbol{\varphi}(x_t)\big)^2 + \tau^2\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t)\,. \tag{3.8}$$

As for the term (2), we have, as is recalled in Lemma 3.4,

$$\frac{\mathcal{K}(\rho_{\boldsymbol{u}^*,\tau},\pi_\tau)}{\eta} \leqslant \frac{4}{\eta}\,\|\boldsymbol{u}^*\|_0\ln\left(1+\frac{\|\boldsymbol{u}^*\|_1}{\|\boldsymbol{u}^*\|_0\,\tau}\right)\,. \tag{3.9}$$

Combining (3.7), (3.8), and (3.9), which all hold for all $\boldsymbol{u}^* \in \mathbb{R}^d$, we get Inequality (3.3). □

Proof (of Corollary 3.1): Applying Proposition 3.1, we have, since $B \geqslant B_y$ and $\eta \leqslant 1/(8B^2)$,

$$\begin{aligned}\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 &\leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + \frac{4}{\eta}\,\|\boldsymbol{u}\|_0\ln\left(1+\frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0\,\tau}\right)\right\} + \tau^2\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t)\\ &\leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + \frac{4}{\eta}\,\|\boldsymbol{u}\|_0\ln\left(1+\frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0\,\tau}\right)\right\} + \tau^2 B_\Phi\,, \end{aligned} \tag{3.10}$$

since $\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t) \leqslant B_\Phi$ by assumption. The particular choices for η and τ given in the statement of the corollary then yield the desired inequality (3.4). □

We end this subsection with a remark on the choices of B, η, and τ suggested in Corollary 3.1. The best choice of (B,η) that satisfies the assumptions of Proposition 3.1 is $B = B_y$ and $\eta = 1/(8B_y^2)$. As for the choice of τ, it approximately minimizes the upper bound given in (3.10). Indeed, for all $C_1, C_2, C_3 > 0$, the function $f : (0,+\infty) \to \mathbb{R}$ defined by

$$f(\tau) \triangleq C_1\ln\left(\frac{C_2}{\tau}\right) + C_3\tau^2$$

has a derivative equal to $f'(\tau) = -C_1/\tau + 2C_3\tau = \tau^{-1}(2C_3\tau^2 - C_1)$, which is negative on

$\big(0, \sqrt{C_1/(2C_3)}\,\big)$ and positive on $\big(\sqrt{C_1/(2C_3)}, +\infty\big)$. The function f thus admits a global minimum in $\tau = \sqrt{C_1/(2C_3)}$. Since the sum of the last two terms of (3.10) is approximately of the form of $f(\tau)$ with[5] $C_1 = 4/\eta = 32B_y^2$ and $C_3 = B_\Phi$, a reasonable choice for τ is given by

$$\tau = \sqrt{\frac{32B_y^2}{2B_\Phi}} = \sqrt{\frac{16B_y^2}{B_\Phi}}\,.$$

3.3.2 Unknown bound B_y on the observations but known bound B_Φ on the trace of the empirical Gram matrix

In the previous section, to prove the upper bounds stated in Lemma 3.1 and Proposition 3.1, we assumed that the forecaster had access to a bound B_y on the observations $|y_t|$ and to a bound B_Φ on the trace of the empirical Gram matrix. In this section, we remove the first requirement and prove a sparsity regret bound for a variant of the algorithm $\mathrm{SeqSEW}_\tau^{\mathrm{B},\eta}$ which is adaptive to the unknown bound $B_y = \max_{1\leqslant t\leqslant T} |y_t|$; see Proposition 3.2 and Remark 3.2 below.

Parameter: prior scale $\tau > 0$ with which we associate the *sparsity prior* $\pi_\tau \in \mathcal{M}_1^+(\mathbb{R}^d)$ defined by

$$\pi_\tau(\mathrm{d}\boldsymbol{u}) \triangleq \prod_{j=1}^d \frac{(3/\tau)\,\mathrm{d}u_j}{2\big(1+|u_j|/\tau\big)^4}\,.$$

Initialization: $B_1 \triangleq 0$, $\eta_1 \triangleq +\infty$, and $p_1 \triangleq \pi_\tau$.

At each time round $t \geqslant 1$,

1. Get the input data x_t and predict[a] as $\widehat{y}_t \triangleq \displaystyle\int_{\mathbb{R}^d} \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t} p_t(\mathrm{d}\boldsymbol{u})$;
2. Get the observation y_t and update:
   - the threshold $B_{t+1} \triangleq \Big(2^{\lceil \log_2 \max_{1\leqslant s\leqslant t} y_s^2\rceil}\Big)^{1/2}$,
   - the inverse temperature $\eta_{t+1} \triangleq 1/\big(8B_{t+1}^2\big)$,
   - and the posterior distribution $p_{t+1} \in \mathcal{M}_1^+(\mathbb{R}^d)$ as

$$p_{t+1}(\mathrm{d}\boldsymbol{u}) \triangleq \frac{\exp\left(-\eta_{t+1}\displaystyle\sum_{s=1}^t \Big(y_s - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_s)\big]_{B_s}\Big)^2\right)}{W_{t+1}}\,\pi_\tau(\mathrm{d}\boldsymbol{u})\,,$$

   where

$$W_{t+1} \triangleq \int_{\mathbb{R}^d} \exp\left(-\eta_{t+1}\sum_{s=1}^t \Big(y_s - \big[\boldsymbol{v}\cdot\boldsymbol{\varphi}(x_s)\big]_{B_s}\Big)^2\right)\pi_\tau(\mathrm{d}\boldsymbol{v})\,.$$

[a] The clipping operator $[\cdot]_B$ is defined in Section 3.2.

Figure 3.3: The algorithm SeqSEW_τ^*.

[5] We omit the factor $\|\boldsymbol{u}\|_0$ in C_1, since the ℓ^0-norm of the minimizer $\boldsymbol{u}$ of (3.10) is unknown and "small" under a sparsity scenario. This approximation leads to a reasonable tuning as can be seen from Corollary 3.1.

For this purpose we consider the algorithm of Figure 3.3, which we call SeqSEW^*_τ thereafter. It differs from $\mathrm{SeqSEW}^{\mathrm{B},\eta}_\tau$ defined in the previous section in that the threshold B and the inverse temperature η are now allowed to vary over time and are chosen at each time round as a function of the data available to the forecaster.

The idea of truncating the base forecasts was used many times in the past; see, e.g., [Vov01] for the online linear regression setting, [GKKW02, Chapter 10] for the regression problem with random design, and [GO07, BBGO10] for sequential prediction of unbounded time series under the square loss. A key ingredient in the present chapter is to perform truncation with respect to a data-driven threshold. The online tuning of this threshold is based on a pseudo-doubling-trick technique provided in [CBMS07]. (We use the prefix *pseudo* since the algorithm does not restart at the beginning of each new regime.)

Proposition 3.2. *For all $\tau > 0$, the algorithm SeqSEW^*_τ satisfies*

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + 32 B_{T+1}^2 \,\|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0\,\tau}\right) \right\} \qquad (3.11)$$
$$+\ \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \ + 16 B_{T+1}^2 \,,$$

where $B_{T+1}^2 \triangleq 2^{\lceil \log_2 \max_{1\leqslant t\leqslant T} y_t^2 \rceil} \leqslant 2 \max_{1\leqslant t\leqslant T} y_t^2$.

Remark 3.2. In view of Proposition 3.1, the algorithm SeqSEW^*_τ satisfies a sparsity regret bound which is adaptive to the unknown bound $B_y = \max_{1\leqslant t\leqslant T} |y_t|$. The price for the automatic tuning with respect to B_y consists only of a multiplicative factor smaller than 2 and the additive factor $16 B_{T+1}^2$ which is smaller than $32 B_y^2$.

As in the previous section, several corollaries can be derived from Proposition 3.2. If the forecaster has access beforehand to a quantity $B_\Phi > 0$ such that $\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leqslant B_\Phi$, then a suboptimal but reasonable choice of τ is given by $\tau = 1/\sqrt{B_\Phi}$; see Corollary 3.2 below. The simpler tuning[6] $\tau = 1/\sqrt{dT}$ of Corollary 3.3 will be useful in the stochastic batch setting (cf. Section 3.4). The proofs of the next corollaries are immediate.

Corollary 3.2. *Assume that, for a known constant $B_\Phi > 0$, the $(x_1, y_1), \ldots, (x_T, y_T)$ are such that $\sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) \leqslant B_\Phi$. Then, when used with $\tau = 1/\sqrt{B_\Phi}$, the algorithm SeqSEW^*_τ satisfies*

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + 32 B_{T+1}^2 \,\|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{B_\Phi}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right) \right\} \qquad (3.12)$$
$$+ 16 B_{T+1}^2 + 1 \,,$$

where $B_{T+1}^2 \triangleq 2^{\lceil \log_2 \max_{1\leqslant t\leqslant T} y_t^2 \rceil} \leqslant 2 \max_{1\leqslant t\leqslant T} y_t^2$.

[6] The tuning $\tau = 1/\sqrt{dT}$ only uses the knowledge of T, which is known by the forecaster in the stochastic batch setting. In that framework, another simple and easy-to-analyse tuning is given by $\tau = 1/(\|\boldsymbol{\varphi}\|_\infty \sqrt{dT})$ — which corresponds to $B_\Phi = dT\,\|\boldsymbol{\varphi}\|_\infty^2$ — but it requires that $\|\boldsymbol{\varphi}\|_\infty \triangleq \sup_{x\in\mathcal{X}} \max_{1\leqslant j\leqslant d} |\varphi_j(x)|$ be finite. Note that the last tuning satisfies the scale-invariant property pointed out in [DT11, Remark 4].

Corollary 3.3. *Assume that T is known to the forecaster at the beginning of the prediction game. Then, when used with $\tau = 1/\sqrt{dT}$, the algorithm* SeqSEW^*_τ *satisfies*

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + 32 B_{T+1}^2 \, \|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right) \right\} \quad (3.13)$$
$$+ \frac{1}{dT} \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) + 16 B_{T+1}^2 \,,$$

where $B_{T+1}^2 \triangleq 2^{\lceil \log_2 \max_{1\leqslant t\leqslant T} y_t^2 \rceil} \leqslant 2 \max_{1\leqslant t\leqslant T} y_t^2$.

As in the previous section, to prove Proposition 3.2, we first need a key PAC-Bayesian inequality. The next lemma is an adaptive variant of Lemma 3.1.

Lemma 3.2. *For all $\tau > 0$, the algorithm* SeqSEW^*_τ *satisfies*

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 \leqslant \inf_{\rho\in\mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \Big(y_t - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t}\Big)^2 \rho(\mathrm{d}\boldsymbol{u}) + 8 B_{T+1}^2 \, \mathcal{K}(\rho,\pi_\tau) \right\} + 8 B_{T+1}^2 \quad (3.14)$$

$$\leqslant \inf_{\rho\in\mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 \rho(\mathrm{d}\boldsymbol{u}) + 8 B_{T+1}^2 \, \mathcal{K}(\rho,\pi_\tau) \right\} + 16 B_{T+1}^2 \,, \quad (3.15)$$

where $B_{T+1}^2 \triangleq 2^{\lceil \log_2 \max_{1\leqslant t\leqslant T} y_t^2 \rceil} \leqslant 2 \max_{1\leqslant t\leqslant T} y_t^2$.

Proof (of Lemma 3.2): The proof is based on arguments that are similar to those underlying Lemma 3.1, except that we now need to deal with B and η changing over time. In the same spirit as in [ACBG02, CBMS07, GO07], our analysis relies on the control of $(\ln W_{t+1})/\eta_{t+1} - (\ln W_t)/\eta_t$ where $W_1 \triangleq 1$ and, for all $t \geqslant 2$,

$$W_t \triangleq \int_{\mathbb{R}^d} \exp\left(-\eta_t \sum_{s=1}^{t-1} \Big(y_s - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_s)\big]_{B_s}\Big)^2 \right) \pi_\tau(\mathrm{d}\boldsymbol{u}) \,.$$

On the one hand, we have

$$\frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} = \frac{1}{\eta_{T+1}} \ln \int_{\mathbb{R}^d} \exp\left(-\eta_{T+1} \sum_{t=1}^T \Big(y_t - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t}\Big)^2 \right) \pi_\tau(\mathrm{d}\boldsymbol{u}) - \frac{1}{\eta_1} \ln 1$$
$$= - \inf_{\rho\in\mathcal{M}_1^+(\mathbb{R}^d)} \left\{ \int_{\mathbb{R}^d} \sum_{t=1}^T \Big(y_t - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t}\Big)^2 \rho(\mathrm{d}\boldsymbol{u}) + \frac{\mathcal{K}(\rho,\pi_\tau)}{\eta_{T+1}} \right\} , \quad (3.16)$$

where the last equality follows from a convex duality argument for the Kullback-Leibler divergence (cf., e.g., [Cat04, p. 159]) which we recall in Proposition A.1 in Appendix A.1.

On the other hand, we can rewrite $(\ln W_{T+1})/\eta_{T+1} - (\ln W_1)/\eta_1$ as a telescopic sum and get

$$\frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} = \sum_{t=1}^{T}\left(\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W_t}{\eta_t}\right) = \sum_{t=1}^{T}\Bigg(\underbrace{\frac{\ln W_{t+1}}{\eta_{t+1}} - \frac{\ln W'_{t+1}}{\eta_t}}_{(1)} + \underbrace{\frac{1}{\eta_t}\ln\frac{W'_{t+1}}{W_t}}_{(2)}\Bigg)\,, \tag{3.17}$$

where W'_{t+1} is obtained from W_{t+1} by replacing η_{t+1} with η_t; namely,

$$W'_{t+1} \triangleq \int_{\mathbb{R}^d} \exp\left(-\eta_t \sum_{s=1}^{t}\Big(y_s - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_s)\big]_{B_s}\Big)^2\right)\pi_\tau(\mathrm{d}\boldsymbol{u})\,.$$

Let $t \in \{1,\dots,T\}$. The first term (1) is non-positive by Jensen's inequality (note that $x \mapsto x^{\eta_{t+1}/\eta_t}$ is concave on $\mathbb{R}_+^*$ since $\eta_{t+1} \leqslant \eta_t$ by construction). As for the second term (2), by definition of W'_{t+1},

$$\begin{aligned}
&\frac{1}{\eta_t}\ln\frac{W'_{t+1}}{W_t}\\
&\quad= \frac{1}{\eta_t}\ln\int_{\mathbb{R}^d}\frac{\exp\Big(-\eta_t\Big(y_t - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t}\Big)^2\Big)\exp\Big(-\eta_t\sum_{s=1}^{t-1}\Big(y_s - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_s)\big]_{B_s}\Big)^2\Big)}{W_t}\pi_\tau(\mathrm{d}\boldsymbol{u})\\
&\quad= \frac{1}{\eta_t}\ln\int_{\mathbb{R}^d}\exp\left(-\eta_t\Big(y_t - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t}\Big)^2\right)p_t(\mathrm{d}\boldsymbol{u}) \qquad (3.18)\\
&\quad\leqslant \begin{cases} -(y_t-\widehat{y}_t)^2 & \text{if } B_{t+1} = B_t;\\ -(y_t-\widehat{y}_t)^2 + (2B_{t+1})^2 & \text{if } B_{t+1} > B_t; \end{cases} \qquad (3.19)
\end{aligned}$$

where (3.18) follows by definition of p_t. To get Inequality (3.19) when $B_{t+1} = B_t$, or, equivalently, $|y_t| \leqslant B_t$, we used the fact that the square loss is $1/(8B_t^2)$-exp-concave on $[-B_t, B_t]$ (as in Lemma 3.1). Indeed, by definition of $\eta_t \triangleq 1/(8B_t^2)$ and by Jensen's inequality, we get

$$\int_{\mathbb{R}^d} e^{-\eta_t\left(y_t - \left[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\right]_{B_t}\right)^2} p_t(\mathrm{d}\boldsymbol{u}) \leqslant \exp\left(-\eta_t\left(y_t - \int_{\mathbb{R}^d}\big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t}p_t(\mathrm{d}\boldsymbol{u})\right)^2\right) = e^{-\eta_t(y_t-\widehat{y}_t)^2},$$

where the last equality follows by definition of $\widehat{y}_t$. Taking the logarithms of both sides of the last inequality and dividing by η_t, we get (3.19) when $B_{t+1} = B_t$.

As for the rounds t such that $B_{t+1} > B_t$, the square loss $x \mapsto (y_t - x)^2$ is no longer $1/(8B_t^2)$-exp-concave on $[-B_t, B_t]$. In this case (3.19) follows from the cruder upper bound $(1/\eta_t)\ln(W'_{t+1}/W_t) \leqslant 0 \leqslant -(y_t-\widehat{y}_t)^2 + (2B_{t+1})^2$ (since $|y_t|, |\widehat{y}_t| \leqslant B_{t+1}$). Summing (3.19) over $t = 1,\dots,T$, Equation (3.17) yields

$$\frac{\ln W_{T+1}}{\eta_{T+1}} - \frac{\ln W_1}{\eta_1} \leqslant -\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 + 4\sum_{\substack{t=1\\ t:B_{t+1}>B_t}}^{T} B_{t+1}^2 \leqslant -\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 + 8B_{T+1}^2\,, \tag{3.20}$$

where, setting $K \triangleq \lceil \log_2 \max_{1\leqslant t\leqslant T} y_t^2\rceil$, we bounded the geometric sum $\sum_{t:B_{t+1}>B_t}^{T} B_{t+1}^2$ from above by $\sum_{k=-\infty}^{K} 2^k = 2^{K+1} \triangleq 2B_{T+1}^2$ in the same way as in Theorem 6 of [CBMS07].

Putting Equations (3.16) and (3.20) together, we get the PAC-Bayesian inequality

$$\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 \leqslant \inf_{\rho\in\mathcal{M}_1^+(\mathbb{R}^d)}\left\{\int_{\mathbb{R}^d}\sum_{t=1}^{T}\Big(y_t-\big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_{B_t}\Big)^2\rho(\mathrm{d}\boldsymbol{u})\;+\;\frac{\mathcal{K}(\rho,\pi_\tau)}{\eta_{T+1}}\right\}+8B_{T+1}^2\,,$$

which yields (3.14) by definition of $\eta_{T+1}\triangleq 1/(8B_{T+1}^2)$. The other PAC-Bayesian inequality (3.15), which is stated for non-truncated base forecasts, follows from (3.14) by the fact that truncation to B_t can only improve prediction if $|y_t|\leqslant B_t$. The remaining t's such that $|y_t|>B_t$ then just account for an overall additional term at most equal to $\sum_{t:B_{t+1}>B_t}^{T}\big(2B_{t+1}\big)^2\leqslant 8B_{T+1}^2$, which concludes the proof. □

Proof (of Proposition 3.2): The proof follows the exact same lines as in Proposition 3.1 except that we apply Lemma 3.2 instead of Lemma 3.1. Indeed, using Lemma 3.2 and restricting the infimum to the $\rho_{\boldsymbol{u}^*,\tau}$, $\boldsymbol{u}^*\in\mathbb{R}^d$ (cf. (3.43)), we get that

$$\begin{aligned}\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 &\leqslant \inf_{\boldsymbol{u}^*\in\mathbb{R}^d}\left\{\int_{\mathbb{R}^d}\sum_{t=1}^{T}\big(y_t-\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2\rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u})+8B_{T+1}^2\mathcal{K}(\rho_{\boldsymbol{u}^*,\tau},\pi_\tau)\right\}+16B_{T+1}^2\\ &\leqslant \inf_{\boldsymbol{u}^*\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t-\boldsymbol{u}^*\cdot\boldsymbol{\varphi}(x_t)\big)^2+32B_{T+1}^2\,\|\boldsymbol{u}^*\|_0\ln\left(1+\frac{\|\boldsymbol{u}^*\|_1}{\|\boldsymbol{u}^*\|_0\,\tau}\right)\right\}\\ &\quad+\tau^2\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t)+16B_{T+1}^2\,,\end{aligned}$$

where the last inequality follows from Lemmas 3.3 and 3.4. □

3.3.3 A fully automatic algorithm

In the previous section, we proved that adaptation to B_y was possible. If we also no longer assume that a bound B_Φ on the trace of the empirical Gram matrix is available to the forecaster, then we can use a doubling trick on the nondecreasing quantity

$$\gamma_t\triangleq\ln\left(1+\sqrt{\sum_{s=1}^{t}\sum_{j=1}^{d}\varphi_j^2(x_s)}\right)$$

and repeatedly run the algorithm SeqSEW_τ^* of the previous section for rapidly-decreasing values of τ. This yields a sparsity regret bound with extra logarithmic multiplicative factors as compared to Proposition 3.2, but which holds for a fully automatic algorithm; see Theorem 3.1 below.

More formally, our algorithm SeqSEW_*^* is defined as follows. The set of all time rounds $t=1,2,\ldots$ is partitioned into regimes $r=0,1,\ldots$ whose final time instances t_r are data-driven. Let $t_{-1}\triangleq 0$ by convention. We call *regime* r, $r=0,1,\ldots$, the sequence of time rounds $(t_{r-1}+1,\ldots,t_r)$ where t_r is the first date $t\geqslant t_{r-1}+1$ such that $\gamma_t>2^r$. At the beginning of regime r, we restart the algorithm SeqSEW_τ^* defined in Figure 3.3 with the parameter $\tau=\tau_r$, where τ_r is the solution of the equation $2^r=\ln(1+1/\tau)$, i.e., $\tau_r\triangleq 1/\big(\exp(2^r)-1\big)$.

Theorem 3.1. *Without requiring any preliminary knowledge at the beginning of the prediction game,* SeqSEW^*_* *satisfies, for all* $T \geqslant 1$ *and all* $(x_1, y_1), \dots, (x_T, y_T) \in \mathcal{X} \times \mathbb{R}$,

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \Bigg\{ \sum_{t=1}^{T} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + 256 \Big(\max_{1 \leqslant t \leqslant T} y_t^2\Big) \|\boldsymbol{u}\|_0 \ln \left(e + \sqrt{\sum_{t=1}^{T} \sum_{j=1}^{d} \varphi_j^2(x_t)} \right) + 64 \Big(\max_{1 \leqslant t \leqslant T} y_t^2\Big) A_T \|\boldsymbol{u}\|_0 \ln \left(1 + \frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0} \right) \Bigg\} + \Big(1 + 38 \max_{1 \leqslant t \leqslant T} y_t^2\Big) A_T \,,$$

where $A_T \triangleq 2 + \log_2 \ln \Big(e + \sqrt{\sum_{t=1}^{T} \sum_{j=1}^{d} \varphi_j^2(x_t)} \Big)$.

On each regime r, the current instance of the algorithm $\mathrm{SeqSEW}^*_{\tau_r}$ only uses the past observations y_s, $s \in \{t_{r-1} + 1, \dots, t-1\}$, to perform the online trunction and to tune the inverse temperature parameter. Therefore, the algorithm SeqSEW^*_* is fully automatic.

Note however that two possible improvements could be addressed in the future. From a theoretical viewpoint, can we contruct a fully automatic algorithm with a bound similar to Theorem 3.1 but without the extra logarithmic factor A_T? From a practical viewpoint, is it possible to perform the adaptation to B_Φ without restarting the algorithm repeatedly (just like we did for B_y)? A smoother time-varying tuning $(\tau_t)_{t \geqslant 2}$ might enable to answer both questions. This would be very probably at the price of a more involved analysis (e.g., if we adapt the PAC-Bayesian bound of Lemma 3.2, then a third approximation term would appear in (3.17) since π_{τ_t} changes over time).

Proof sketch (of Theorem 3.1): The proof relies on the application of Proposition 3.2 with $\tau = \tau_r$ on all regimes r visited up to time T. Summing the corresponding inequalities over r then concludes the proof. See Appendix 3.A.2 for a detailed proof. □

Theorem 3.1 yields the following corollary. It upper bounds the regret of the algorithm SeqSEW^*_* uniformly over all $\boldsymbol{u} \in \mathbb{R}^d$ such that $\|\boldsymbol{u}\|_0 \leqslant s$ and $\|\boldsymbol{u}\|_1 \leqslant U$, where the sparsity level $s \in \mathbb{N}$ and the ℓ^1-diameter $U > 0$ are both unknown to the forecaster. The proof is postponed to Appendix 3.A.2.

Corollary 3.4. *Fix* $s \in \mathbb{N}$ *and* $U > 0$. *Then, for all* $T \geqslant 1$ *and all* $(x_1, y_1), \dots, (x_T, y_T) \in \mathcal{X} \times \mathbb{R}$, *the regret of the algorithm* SeqSEW^*_* *on* $\big\{\boldsymbol{u} : \|\boldsymbol{u}\|_0 \leqslant s\big\} \bigcap \big\{\boldsymbol{u} : \|\boldsymbol{u}\|_1 \leqslant U\big\}$ *is bounded by*

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 - \inf_{\substack{\|\boldsymbol{u}\|_0 \leqslant s \\ \|\boldsymbol{u}\|_1 \leqslant U}} \sum_{t=1}^{T} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 \leqslant 256 \Big(\max_{1 \leqslant t \leqslant T} y_t^2\Big) s \ln \left(e + \sqrt{\sum_{t=1}^{T} \sum_{j=1}^{d} \varphi_j^2(x_t)} \right) + 64 \Big(\max_{1 \leqslant t \leqslant T} y_t^2\Big) A_T \, s \ln \Big(1 + \frac{U}{s} \Big) + \Big(1 + 38 \max_{1 \leqslant t \leqslant T} y_t^2\Big) A_T \,,$$

where $A_T \triangleq 2 + \log_2 \ln \Big(e + \sqrt{\sum_{t=1}^{T} \sum_{j=1}^{d} \varphi_j^2(x_t)} \Big)$.

3.4 Adaptivity to the unknown variance in the stochastic setting

In this section, we apply the online algorithm SeqSEW^*_τ of Section 3.3.2 to two related stochastic settings: the regression model with random design (Section 3.4.1) and the regression model with fixed design (Section 3.4.2). The sparsity regret bounds proved for this algorithm on individual sequences imply in both settings sparsity oracle inequalities with leading constant 1. These risk bounds are of the same flavor as in [DT08, DT11] but they are adaptive (up to a logarithmic factor) to the unknown variance σ^2 of the noise if the latter is Gaussian. In particular, we solve two questions left open in [DT11] in the random design case.

In the sequel, just like in the online deterministic setting, we assume that the forecaster has access to a dictionary $\boldsymbol{\varphi} = (\varphi_1, \dots, \varphi_d)$ of measurable base regressors $\varphi_j : \mathcal{X} \to \mathbb{R}$, $j = 1, \dots, d$.

3.4.1 Regression model with random design

In this section we apply the algorithm SeqSEW^*_τ to the regression model with random design. In this batch setting the forecaster is given at the beginning of the game T independent random copies $(X_1, Y_1), \dots, (X_T, Y_T)$ of $(X, Y) \in \mathcal{X} \times \mathbb{R}$ whose common distribution is unknown. We assume thereafter that $\mathbb{E}[Y^2] < \infty$; the goal of the forecaster is to estimate the regression function $f : \mathcal{X} \to \mathbb{R}$ defined by $f(x) \triangleq \mathbb{E}[Y|X = x]$ for all $x \in \mathcal{X}$. Setting $\varepsilon_t \triangleq Y_t - f(X_t)$ for all $t = 1, \dots, T$, note that

$$Y_t = f(X_t) + \varepsilon_t \,, \quad 1 \leqslant t \leqslant T \,,$$

and that the pairs $(X_1, \varepsilon_1), \dots, (X_T, \varepsilon_T)$ are i.i.d. and such that $\mathbb{E}[\varepsilon_1^2] < \infty$ and $\mathbb{E}[\varepsilon_1|X_1] = 0$ almost surely. In the sequel, we denote the distribution of X by P^X and we set, for all measurable functions $h : \mathcal{X} \to \mathbb{R}$,

$$\|h\|_{L^2} \triangleq \left(\int_{\mathcal{X}} h(x)^2 P^X(\mathrm{d}x) \right)^{1/2} = \left(\mathbb{E}\big[h(X)^2\big] \right)^{1/2} .$$

Next we construct a regressor $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ based on the sample $(X_1, Y_1), \dots, (X_T, Y_T)$ that satisfies a sparsity oracle inequality, i.e., its expected L^2-risk $\mathbb{E}\big[\big\| f - \widehat{f}_T \big\|_{L^2}^2 \big]$ is almost as small as the smallest L^2-risk $\|f - \boldsymbol{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2$, $\boldsymbol{u} \in \mathbb{R}^d$, up to some additive term proportional to $\|\boldsymbol{u}\|_0$.

Algorithm and main result

Even if the whole sample $(X_1, Y_1), \dots, (X_T, Y_T)$ is available at the beginning of the prediction game, we treat it in a sequential fashion. We run the algorithm SeqSEW^*_τ of Section 3.3.2 from time 1 to time T with $\tau = 1/\sqrt{dT}$ (note that T is known in this setting). Using the standard online to batch conversion (cf. Section 2.5 in Chapter 2), we define our data-based regressor $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ as the uniform average

$$\widehat{f}_T \triangleq \frac{1}{T} \sum_{t=1}^{T} \widetilde{f}_t \tag{3.21}$$

of the regressors $\widetilde{f}_t : \mathcal{X} \to \mathbb{R}$ sequentially built by the algorithm SeqSEW^*_τ as

$$\widetilde{f}_t(x) \triangleq \int_{\mathbb{R}^d} \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x)\big]_{B_t} p_t(\mathrm{d}\boldsymbol{u}) \,. \tag{3.22}$$

Note that, contrary to much prior work from the statistics community such as [Cat04, BN08, DT11], the regressors $\widetilde{f}_t : \mathcal{X} \to \mathbb{R}$ are tuned online. Therefore, $\widehat{f}_T$ does not depend on any prior knowledge on the unknown distribution of the (X_t, Y_t), $1 \leqslant t \leqslant T$, such as the unknown variance $\mathbb{E}\big[(Y - f(X))^2\big]$ of the noise, the $\|\varphi_j\|_\infty$, or the $\|f - \varphi_j\|_\infty$ (actually, the φ_j and the $f - \varphi_j$ do not even need to be bounded in ℓ^∞-norm).

In this respect, as explained in Section 2.5.2 (Chapter 2), this work improves on [BN08] who tune their online forecasters as a function of $\max_{1 \leqslant j \leqslant d} \|\varphi_j\|_\infty$. The major technique difference is that we truncate the base forecasts $\boldsymbol{u} \cdot \boldsymbol{\varphi}(X_t)$ instead of truncating the observations Y_t. In particular, this enables to aggregate the base regressors $\boldsymbol{u} \cdot \boldsymbol{\varphi}$ for all $\boldsymbol{u} \in \mathbb{R}^d$, i.e., in the whole $\mathbb{R}^d$ space.

The next sparsity oracle inequality is the main result of this section. It follows from the deterministic regret bound of Corollory 3.3 and from Jensen's inequality. Two corollaries are to be derived later.

Theorem 3.2. *Assume that $(X_1, Y_1), \dots, (X_T, Y_T) \in \mathcal{X} \times \mathbb{R}$ are independent random copies of $(X, Y) \in \mathcal{X} \times \mathbb{R}$, where $\mathbb{E}[Y^2] < +\infty$ and $\|\varphi_j\|_{L^2}^2 \triangleq \mathbb{E}[\varphi_j(X)^2] < +\infty$ for all $j = 1, \dots, d$. Then, the data-based regressor $\widehat{f}_T$ defined in (3.21)-(3.22) satisfies*

$$\mathbb{E}\left[\left\|f - \widehat{f}_T\right\|_{L^2}^2\right] \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \|f - \boldsymbol{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 + 64 \frac{\mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2\right]}{T} \|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right) \right\}$$
$$+ \frac{1}{dT} \sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + 32 \frac{\mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2\right]}{T} .$$

Note that our risk bounds are stated in expectation (which already improves on existing results in the stochastic setting, see the next section). However, by convexity (and closedness) of all sets of the form $\big\{\boldsymbol{u} \cdot \boldsymbol{\varphi} : J(\boldsymbol{u}) \subset J_0, \|\boldsymbol{u}\|_1 \leqslant U\big\}$, where $U \geqslant 0$ and $J_0 \subset \{1, \dots, d\}$, and where $J(\boldsymbol{u}) \triangleq \{j : u_j \neq 0\}$, it is possible to use [Zha05, Theorem 8] to transform our results into risk bounds with high probability (at least when the output Y is bounded, but similar results should hold true under reasonable assumptions on the output distribution).

Proof sketch (of Theorem 3.2): By Corollary 3.3 and by definition of $\widetilde{f}_t$ above and $\widehat{y}_t \triangleq \widetilde{f}_t(X_t)$ in Figure 3.3, we have, *almost surely*,

$$\sum_{t=1}^T (Y_t - \widetilde{f}_t(X_t))^2 \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^T \big(Y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(X_t)\big)^2 + 64 \left(\max_{1 \leqslant t \leqslant T} Y_t^2\right) \|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right) \right\}$$
$$+ \frac{1}{dT} \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(X_t) + 32 \max_{1 \leqslant t \leqslant T} Y_t^2 .$$

Taking the expectations of both sides and applying Jensen's inequality yields the desired result. For a detailed proof, see Appendix 3.A.3. □

Theorem 3.2 above can be used under several assumptions on the distribution of the output Y. In all cases, it suffices to upper bound the amplitude $\mathbb{E}\big[\max_{1 \leqslant t \leqslant T} Y_t^2\big]$. We present below a general corollary and explain later why our fully automatic procedure $\widehat{f}_T$ solves two questions left open by [DT11] (see Corollary 3.6).

A general corollary

Using Lemmas 3.5, 3.6, and 3.7 in Appendix 3.B to upper bound the two terms $\mathbb{E}\big[\max_{1\leqslant t\leqslant T} Y_t^2\big]$ of Theorem 3.2, we get the following sparsity oracle inequality. The proof is postponed to Appendix 3.A.3.

Corollary 3.5. *Assume that $(X_1, Y_1), \dots, (X_T, Y_T) \in \mathcal{X} \times \mathbb{R}$ are independent random copies of $(X, Y) \in \mathcal{X} \times \mathbb{R}$, that $\sup_{1\leqslant j\leqslant d} \|\varphi_j\|_{L^2}^2 < +\infty$, that $\mathbb{E}|Y| < +\infty$, and that one of the following assumptions holds on the distribution of $\Delta Y \triangleq Y - \mathbb{E}[Y]$.*

- $\big(\mathrm{BD(B)}\big)$: *$|\Delta Y| \leqslant B$ almost surely for a given constant $B > 0$;*
- $\big(\mathrm{SG}(\sigma^2)\big)$: *ΔY is subgaussian with variance factor $\sigma^2 > 0$, that is, $\mathbb{E}\left[e^{\lambda \Delta Y}\right] \leqslant e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$;*
- $\big(\mathrm{BEM}(\alpha, \mathrm{M})\big)$: *ΔY has a bounded exponential moment, that is, $\mathbb{E}\left[e^{\alpha|\Delta Y|}\right] \leqslant M$ for some given constants $\alpha > 0$ and $M > 0$;*
- $\big(\mathrm{BM}(\alpha, \mathrm{M})\big)$: *ΔY has a bounded moment, that is, $\mathbb{E}\big[|\Delta Y|^\alpha\big] \leqslant M$ for some given constants $\alpha > 2$ and $M > 0$.*

Then, the data-based regressor $\widehat{f}_T$ defined above satisfies

$$\mathbb{E}\left[\left\|f - \widehat{f}_T\right\|_{L^2}^2\right] \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d} \left\{ \|f - \boldsymbol{u}\cdot\boldsymbol{\varphi}\|_{L^2}^2 + 128\left(\frac{\mathbb{E}[Y]^2}{T} + \psi_T\right)\|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\right\} + \frac{1}{dT}\sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + 64\left(\frac{\mathbb{E}[Y]^2}{T} + \psi_T\right),$$

where

$$\psi_T \triangleq \frac{1}{T}\mathbb{E}\left[\max_{1\leqslant t\leqslant T}\big(Y_t - \mathbb{E}[Y_t]\big)^2\right] \leqslant \begin{cases} \dfrac{B^2}{T} & \text{under Assumption } \big(\mathrm{BD(B)}\big), \\ \dfrac{2\sigma^2\ln(2eT)}{T} & \text{under Assumption } \big(\mathrm{SG}(\sigma^2)\big), \\ \dfrac{\ln^2\big((M+e)T\big)}{\alpha^2\,T} & \text{under Assumption } \big(\mathrm{BEM}(\alpha,\mathrm{M})\big), \\ \dfrac{M^{2/\alpha}}{T^{(\alpha-2)/\alpha}} & \text{under Assumption } \big(\mathrm{BM}(\alpha,\mathrm{M})\big). \end{cases}$$

Several comments can be made about Corollary 3.5. We first stress that, if $T \geqslant 2$, then the two "bias" terms $\mathbb{E}[Y]^2/T$ above can be avoided, at least at the price of a multiplicative factor of $2T/(T-1) \leqslant 4$. This can be achieved via a slightly more sophisticated online clipping — see Remark 3.4 in Appendix 3.A.3.

Second, under the assumptions $\big(\mathrm{BD(B)}\big)$, $\big(\mathrm{SG}(\sigma^2)\big)$, or $\big(\mathrm{BEM}(\alpha,\mathrm{M})\big)$, the key quantity ψ_T is respectively of the order of $1/T$, $\ln(T)/T$ and $\ln^2(T)/T$. Up to a logarithmic factor, this corresponds to the classical fast rate of convergence $1/T$ obtained in the random design setting for different aggregation problems (see, e.g., [Cat99, JRT08, Aud09] for model-selection-type

aggregation and [DT11] for linear aggregation). However, the rate $T^{-(\alpha-2)/\alpha}$ we proved under the bounded moment assumption $\big(\mathrm{BM}(\alpha,\mathrm{M})\big)$ does not match the faster rate $T^{-\alpha/(\alpha+2)}$ obtained in [JRT08, Aud09] under a similar assumption. In particular the bound of Corollary 3.5 goes to $M>0$ as $\alpha\to 2$, while the optimal rate for $\alpha=2$ in similar situations is $T^{-1/2}$ [Aud09].

The minor logarithmic difference under Assumptions $\big(\mathrm{SG}(\sigma^2)\big)$ or $\big(\mathrm{BEM}(\alpha,\mathrm{M})\big)$ and the clear difference in the rates under Assumption $\big(\mathrm{BM}(\alpha,\mathrm{M})\big)$ come from the fact that our online algorithm SeqSEW^*_τ was primarily designed for bounded individual sequences with an unknown bound. As remarked in Section 2.5.2 (Chapter 2), the finite i.i.d. sequence $Y_1,\dots,Y_T$ is almost surely uniformly bounded by the random bound $\max_{1\leqslant t\leqslant T}|Y_t|$. Our individual sequence techniques adapt sequentially to this random bound, yielding a regret bound that scales as $\max_{1\leqslant t\leqslant T}Y_t^2$. As a result, the risk bounds obtained after the online to batch conversion scale as $\mathbb{E}\big[\max_{1\leqslant t\leqslant T}Y_t^2\big]/T$. If the distribution of the output Y is bounded or lightly-tailed, then we can (almost) recover for free the fast rate of convergence $1/T$ (the extra logarithmic factor coming from the "slight" non-boundedness of Y). But if the distribution of Y is heavy-tailed, then a different tuning of η_t or a more sophisticated online truncation seem necessary. Doing so without requiring any prior knowledge on the output distribution such as the quantities σ^2, α, M... — as is the case for our fully automatic procedure $\widehat{f}_T$ — is a challenging task.

Third, several variations on the assumptions are possible. First note that several classical assumptions on Y expressed in terms of $f(X)$ and $\varepsilon\triangleq Y-f(X)$ are either particular cases of the above corollary or can be treated similarly. Indeed, each of the four assumptions above on $\Delta Y\triangleq Y-\mathbb{E}[Y]=f(X)-\mathbb{E}[f(X)]+\varepsilon$ is satisfied as soon as both the distribution of $f(X)-\mathbb{E}[f(X)]$ and the conditional distribution of ε (conditionally on X) satisfy the same type of assumption. For example, if $f(X)-\mathbb{E}[f(X)]$ is subgaussian with variance factor σ_X^2 and if ε is subgaussian conditionally on X with a variance factor uniformly bounded by a constant σ_ε^2, then ΔY is subgaussian with variance factor $\sigma_X^2+\sigma_\varepsilon^2$ (see also Remark 3.5 in Appendix 3.A.3 to avoid conditioning).

The assumptions on $f(X)-\mathbb{E}[f(X)]$ and ε can also be mixed together. For instance, as explained in Remark 3.5 in Appendix 3.A.3, under the classical assumptions

$$\|f\|_\infty<+\infty \qquad \text{and} \qquad \mathbb{E}\Big[e^{\alpha|\varepsilon|}\,\Big|\,X\Big]\leqslant M \quad \text{a.s.} \tag{3.23}$$

or

$$\|f\|_\infty<+\infty \qquad \text{and} \qquad \mathbb{E}\Big[e^{\lambda\varepsilon}\,\Big|\,X\Big]\leqslant e^{\lambda^2\sigma^2/2} \quad \text{a.s.}, \quad \forall\lambda\in\mathbb{R}\,, \tag{3.24}$$

the key quantity ψ_T in the corollary can be bounded from above by

$$\psi_T\leqslant\begin{cases}\dfrac{8\,\|f\|_\infty^2}{T}+\dfrac{2\ln^2\big((M+e)T\big)}{\alpha^2\,T} & \text{under the set of assumptions (3.23),}\\[2ex] \dfrac{8\,\|f\|_\infty^2}{T}+\dfrac{4\sigma^2\ln(2eT)}{T} & \text{under the set of assumptions (3.24).}\end{cases}$$

In particular, under the set of assumptions (3.24), our procedure $\widehat{f}_T$ solves two questions left open in [DT11]. We discuss below our contributions in this particular case.

Questions left open by Dalalyan and Tsybakov

In this subsection we focus on the case when the regression function f is bounded (by an unknown constant) and when the noise $\varepsilon \triangleq Y - f(X)$ is subgaussian conditionally on X in the sense that, for some (unknown) constant $\sigma^2 > 0$,

$$\|f\|_\infty < +\infty \qquad \text{and} \qquad \mathbb{E}\Big[e^{\lambda\varepsilon} \,\Big|\, X\Big] \leqslant e^{\lambda^2\sigma^2/2} \quad \text{a.s.}, \quad \forall \lambda \in \mathbb{R}\,. \tag{3.25}$$

In this case, the two terms $\mathbb{E}\big[\max_{1\leqslant t\leqslant T} Y_t^2\big]$ of Theorem 3.2 can be upper bounded in a simpler and slightly tighter way as compared to the proof of Corollary 3.5 (we only use the inequality $(x+y)^2 \leqslant 2x^2 + 2y^2$ once, instead of twice). It yields the following sparsity oracle inequality.

Corollary 3.6. *Assume that $(X_1, Y_1), \dots, (X_T, Y_T) \in \mathcal{X} \times \mathbb{R}$ are independent random copies of $(X, Y) \in \mathcal{X} \times \mathbb{R}$ such that the set of assumptions (3.25) above holds true. Then, the data-based regressor $\widehat{f}_T$ defined in (3.21)-(3.22) satisfies*

$$\begin{aligned}
&\mathbb{E}\left[\left\|f - \widehat{f}_T\right\|_{L^2}^2\right] \\
&\quad\leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\|f - \boldsymbol{u}\cdot\boldsymbol{\varphi}\|_{L^2}^2 + 128\Big(\|f\|_\infty^2 + 2\sigma^2\ln(2eT)\Big)\,\frac{\|\boldsymbol{u}\|_0}{T}\ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\right\} \\
&\qquad + \frac{1}{dT}\sum_{j=1}^d \|\varphi_j\|_{L^2}^2 + \frac{64}{T}\Big(\|f\|_\infty^2 + 2\sigma^2\ln(2eT)\Big)\,.
\end{aligned}$$

Proof: We apply Theorem 3.2 and bound $\mathbb{E}\big[\max_{1\leqslant t\leqslant T} Y_t^2\big]$ from above. By the elementary inequality $(x+y)^2 \leqslant 2x^2 + 2y^2$ for all $x, y \in \mathbb{R}$, we get

$$\begin{aligned}
\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right] &= \mathbb{E}\left[\max_{1\leqslant t\leqslant T}\big(f(X_t) + \varepsilon_t\big)^2\right] \leqslant 2\left(\|f\|_\infty^2 + \mathbb{E}\left[\max_{1\leqslant t\leqslant T}\varepsilon_t^2\right]\right) \\
&\leqslant 2\Big(\|f\|_\infty^2 + 2\sigma^2\ln(2eT)\Big)\,,
\end{aligned}$$

where the last inequality follows from Lemma 3.5 in Appendix 3.B and from the fact that, for all $1 \leqslant t \leqslant T$ and all $\lambda \in \mathbb{R}$, we have $\mathbb{E}\big[e^{\lambda\varepsilon_t}\big] = \mathbb{E}\big[e^{\lambda\varepsilon}\big] = \mathbb{E}\big[\mathbb{E}\big[e^{\lambda\varepsilon} \,\big|\, X\big]\big] \leqslant e^{\lambda^2\sigma^2/2}$ by (3.25). (Note that the assumption of conditional subgaussianity in (3.25) is stronger than what we need, i.e., subgaussianity without conditioning.) This concludes the proof. □

The above bound is of the same order (up to a $\ln T$ factor) as the sparsity oracle inequality proved in Proposition 1 of [DT11]. For the sake of comparison we state below with our notations (e.g., β therein corresponds to $1/\eta$ in this chapter) a straightforward consequence of this proposition, which follows by Jensen's inequality and the particular[7] choice $\tau = 1/\sqrt{dT}$.

[7] Proposition 1 of [DT11] may seem more general than Theorem 3.2 at first sight since it holds for all $\tau > 0$, but this is actually also the case for Theorem 3.2. The proof of the latter would indeed have remained true had we replaced $\tau = 1/\sqrt{dT}$ with any value of $\tau > 0$. We however chose the reasonable value $\tau = 1/\sqrt{dT}$ to make our algorithm parameter-free. As noted earlier, if $\|\boldsymbol{\varphi}\|_\infty \triangleq \sup_{x\in\mathcal{X}} \max_{1\leqslant j\leqslant d} |\varphi_j(x)|$ is finite and known by the forecaster, another simple and easy-to-analyse tuning is given by $\tau = 1/(\|\boldsymbol{\varphi}\|_\infty\sqrt{dT})$.

Proposition 3.3 (A consequence of Prop. 1 of [DT11])**.**
Assume that $\sup_{1\leqslant j\leqslant d}\|\varphi_j\|_\infty < \infty$ *and that the set of assumptions (3.25) above hold true. Then, for all* $R > 2\sqrt{d/T}$ *and all* $\eta \leqslant \bar\eta(R) \triangleq \big(2\sigma^2 + 2\sup_{\|\boldsymbol{u}\|_1\leqslant R}\|\boldsymbol{u}\cdot\boldsymbol{\varphi} - f\|_\infty^2\big)^{-1}$*, the mirror averaging aggregate* $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ *defined in [DT11, Equations (1) and (3)] satisfies*

$$\mathbb{E}\left[\left\|f - \widehat{f}_T\right\|_{L^2}^2\right] \leqslant \inf_{\|\boldsymbol{u}\|_1\leqslant R-2\sqrt{d/T}}\left\{\|f - \boldsymbol{u}\cdot\boldsymbol{\varphi}\|_{L^2}^2 + \frac{4}{\eta}\,\frac{\|\boldsymbol{u}\|_0}{T+1}\ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\right\} + \frac{4}{dT}\sum_{j=1}^d\|\varphi_j\|_{L^2}^2 + \frac{1}{(T+1)\eta}\,.$$

We can now discuss the two questions left open by [DT11]. Despite the similarity of the two bounds, the sparsity oracle inequality stated in Proposition 3.3 above only holds for vectors $\boldsymbol{u}$ within an ℓ^1-ball of finite radius $R - 2\sqrt{d/T}$, while our bound holds over the whole $\mathbb{R}^d$ space. Moreover, the parameter R above has to be chosen in advance, but it cannot be chosen too large since $1/\eta \geqslant 1/\bar\eta(R)$, which grows as R^2 when $R \to +\infty$ (if $\boldsymbol{\varphi} \neq \mathbf{0}$). The authors asked in [DT11, Section 4.2] whether it was possible to get a bound with $1/\eta < +\infty$ such that the infimum in Proposition 3.3 extends to the whole $\mathbb{R}^d$ space. Our results show that, thanks to data-driven truncation, the answer is positive.

Note that it is still possible to transform the bound of Proposition 3.3 into a bound over the whole $\mathbb{R}^d$ space if the parameter R is chosen (illegally) as $R = \|\boldsymbol{u}^*\|_1 + 2\sqrt{d/T}$ (or as a tight upper bound of the last quantity), where $\boldsymbol{u}^* \in \mathbb{R}^d$ minimizes over $\mathbb{R}^d$ the regularized risk

$$\|f - \boldsymbol{u}\cdot\boldsymbol{\varphi}\|_{L^2}^2 + \frac{4}{\bar\eta(\|\boldsymbol{u}\|_1 + 2\sqrt{d/T})}\,\frac{\|\boldsymbol{u}\|_0}{T+1}\ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right) + \frac{4}{dT}\sum_{j=1}^d\|\varphi_j\|_{L^2}^2 + \frac{1}{(T+1)\bar\eta(\|\boldsymbol{u}\|_1 + 2\sqrt{d/T})}\,.$$

For instance, choosing $R = \|\boldsymbol{u}^*\|_1 + 2\sqrt{d/T}$ and $\eta = \bar\eta(R)$, we get from Proposition 3.3 that the expected L^2-risk $\mathbb{E}\big[\|f - \widehat{f}_T\|_{L^2}^2\big]$ of the corresponding procedure is upper bounded by the infimum of the above regularized risk over all $\boldsymbol{u} \in \mathbb{R}^d$. However, this parameter tuning is illegal since $\|\boldsymbol{u}^*\|_1$ is not known in practice. On the contrary, thanks to data-driven truncation, the prior knowledge of $\|\boldsymbol{u}^*\|_1$ is not required by our procedure.

The second open question, which was raised in [DT11, Section 5.1, Remark 6], deals with the prior knowledge of the variance factor σ^2 of the noise. The latter is indeed required by their algorithm for the choice of the inverse temperature parameter η. The authors thus asked whether adaptivity to σ^2 was possible. Corollary 3.6 above provides a positive answer (up to a $\ln T$ factor).

3.4.2 Regression model with fixed design

In this section, we consider the regression model with fixed design. In this batch setting the forecaster is given at the beginning of the game a T-sample $(x_1, Y_1), \dots, (x_T, Y_T) \in \mathcal{X} \times \mathbb{R}$,

where the x_t are deterministic elements in $\mathcal{X}$ and where

$$Y_t = f(x_t) + \varepsilon_t\,, \quad 1 \leqslant t \leqslant T, \tag{3.26}$$

for some i.i.d. sequence $\varepsilon_1, \dots, \varepsilon_T \in \mathbb{R}$ (with unknown distribution) and some unknown function $f : \mathcal{X} \to \mathbb{R}$.

In this setting, just like in Section 3.4.1, our algorithm and the corresponding analysis are a straightforward consequence of the general results on individual sequences developed in Section 3.3. As in the random design setting, the sample $(x_1, Y_1), \dots, (x_T, Y_T)$ is treated in a sequential fashion. We run the algorithm SeqSEW^*_τ defined in Figure 3.3 from time 1 to time T with the particular choice of $\tau = 1/\sqrt{dT}$. We then define our data-based regressor $\widehat{f}_T : \mathcal{X} \to \mathbb{R}$ by

$$\widehat{f}_T(x) \triangleq \begin{cases} \displaystyle\frac{1}{n_x} \sum_{\substack{1 \leqslant t \leqslant T \\ t : x_t = x}} \widetilde{f}_t(x) & \text{if } x \in \{x_1, \dots, x_T\}\,, \\ 0 & \text{if } x \notin \{x_1, \dots, x_T\}\,, \end{cases} \tag{3.27}$$

where $n_x \triangleq \big|\big\{t : x_t = x\big\}\big| = \sum_{t=1}^T \mathbb{I}_{\{x_t = x\}}$, and where the regressors $\widetilde{f}_t : \mathcal{X} \to \mathbb{R}$ sequentially built by the algorithm SeqSEW^*_τ are defined by

$$\widetilde{f}_t(x) \triangleq \int_{\mathbb{R}^d} \big[\boldsymbol{u} \cdot \boldsymbol{\varphi}(x)\big]_{B_t} p_t(\mathrm{d}\boldsymbol{u})\,. \tag{3.28}$$

In the particular case when the x_t are all distinct, $\widehat{f}_T$ is simply defined by $\widehat{f}_T(x) \triangleq \widetilde{f}_T(x)$ if $x \in \{x_1, \dots, x_T\}$ and by $\widehat{f}_T(x) = 0$ otherwise.

The next theorem is the main result of this subsection. It follows as in the random design setting from the deterministic regret bound of Corollory 3.3 and from Jensen's inequality. The proof is postponed to Appendix 3.A.4.

Theorem 3.3. *Consider the regression model with fixed design described in* (3.26)*. Then, the data-based regressor $\widehat{f}_T$ defined in* (3.27)–(3.28) *satisfies*

$$\begin{aligned} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \big(f(x_t) - \widehat{f}_T(x_t)\big)^2\right] \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \Bigg\{ & \frac{1}{T}\sum_{t=1}^T \big(f(x_t) - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 \\ & + 64\, \frac{\mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2\right]}{T}\, \|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{dT}\, \|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right) \Bigg\} \\ & + \frac{1}{dT^2} \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t) + 32\, \frac{\mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2\right]}{T}\,. \end{aligned}$$

As in Section 3.4.1, the amplitude $\mathbb{E}\big[\max_{1 \leqslant t \leqslant T} Y_t^2\big]$ can be upper bounded under various assumptions. The proof of the following corollary is postponed to Appendix 3.A.4.

Corollary 3.7. *Consider the regression model with fixed design described in* (3.26)*. Assume that one of the following assumptions holds on the distribution of* ε_1*.*

- $\big(\mathrm{BD}(\mathrm{B})\big)$ *:* $|\varepsilon_1| \leqslant B$ *almost surely for a given constant* $B > 0$*;*
- $\big(\mathrm{SG}(\sigma^2)\big)$ *:* ε_1 *is subgaussian with variance factor* $\sigma^2 > 0$*, that is,* $\mathbb{E}\left[e^{\lambda\varepsilon_1}\right] \leqslant e^{\lambda^2\sigma^2/2}$ *for all* $\lambda \in \mathbb{R}$*;*
- $\big(\mathrm{BEM}(\alpha, \mathrm{M})\big)$ *:* ε *has a bounded exponential moment, that is,* $\mathbb{E}\left[e^{\alpha|\varepsilon|}\right] \leqslant M$ *for some given constants* $\alpha > 0$ *and* $M > 0$*;*
- $\big(\mathrm{BM}(\alpha, \mathrm{M})\big)$ *:* ε *has a bounded moment, that is,* $\mathbb{E}\big[|\varepsilon|^\alpha\big] \leqslant M$ *for some given constants* $\alpha > 2$ *and* $M > 0$*.*

Then, the data-based regressor $\widehat{f}_T$ *defined in* (3.27)–(3.28) *satisfies*

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t) - \widehat{f}_T(x_t)\big)^2\right] \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\Bigg\{&\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t) - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 \\ &+ 128\left(\frac{\max_{1\leqslant t\leqslant T} f^2(x_t)}{T} + \psi_T\right)\|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\Bigg\} \\ &+ \frac{1}{dT^2}\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t) + 64\left(\frac{\max_{1\leqslant t\leqslant T} f^2(x_t)}{T} + \psi_T\right),\end{aligned}$$

where

$$\psi_T \triangleq \frac{1}{T}\,\mathbb{E}\left[\max_{1\leqslant t\leqslant T}\varepsilon_t^2\right] \leqslant \begin{cases} \dfrac{B^2}{T} & \text{if Assumption } \big(\mathrm{BD}(\mathrm{B})\big) \text{ holds,} \\ \dfrac{2\sigma^2\ln(2eT)}{T} & \text{if Assumption } \big(\mathrm{SG}(\sigma^2)\big) \text{ holds,} \\ \dfrac{\ln^2\left((M+e)T\right)}{\alpha^2\,T} & \text{if Assumption } \big(\mathrm{BEM}(\alpha,\mathrm{M})\big) \text{ holds,} \\ \dfrac{M^{2/\alpha}}{T^{(\alpha-2)/\alpha}} & \text{if Assumption } \big(\mathrm{BM}(\alpha,\mathrm{M})\big) \text{ holds.} \end{cases}$$

The above bound is of the same flavor as that of [DT08, Theorem 5]. It has one advantage and one drawback. On the one hand, we note two additional "bias" terms $\big(\max_{1\leqslant t\leqslant T} f^2(x_t)\big)/T$ as compared to the bound of [DT08, Theorem 5]. As of now, we have not been able to remove them using ideas similar to what we did in the random design case (see Remark 3.4 in Appendix 3.A.3). On the other hand, under Assumption $\big(\mathrm{SG}(\sigma^2)\big)$, contrary to [DT08], our algorithm does not require the prior knowledge of the variance factor σ^2 of the noise.

3.A Proofs

3.A.1 Another proof of Lemma 3.1 (Section 3.3.1)

In Section 3.3.1 we already provided a short proof of Lemma 3.1 via the use of [Aud09, Theorem 4.6]. Below is an alternative self-contained proof of Inequality (3.5) that we only provide for the convenience of the reader.

Proof (of Inequality (3.5) of Lemma 3.1): As is usually done in the online learning setting for the study of the exponentially weighted average forecaster, this proof relies on the control of $\sum_t \eta^{-1} \ln(W_{t+1}/W_t)$ where we recall that $W_1 \triangleq 1$ and, for all $t \geqslant 2$,

$$W_t \triangleq \int_{\mathbb{R}^d} \exp\left(-\eta \sum_{s=1}^{t-1} \Big(y_s - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_s)\big]_B\Big)^2\right) \pi_\tau(\mathrm{d}\boldsymbol{u}) \,.$$

On the one hand, we have

$$\begin{aligned}\frac{1}{\eta}\ln\frac{W_{T+1}}{W_1} &= \frac{1}{\eta}\ln\int_{\mathbb{R}^d} \exp\left(-\eta\sum_{t=1}^{T}\Big(y_s - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_s)\big]_B\Big)^2\right)\pi_\tau(\mathrm{d}\boldsymbol{u}) \;-\; \frac{1}{\eta}\ln 1 \\ &= -\inf_{\rho\in\mathcal{M}_1^+(\mathbb{R}^d)}\left\{\int_{\mathbb{R}^d}\sum_{t=1}^{T}\Big(y_t - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_B\Big)^2\rho(\mathrm{d}\boldsymbol{u}) \;+\; \frac{\mathcal{K}(\rho,\pi_\tau)}{\eta}\right\}, \qquad (3.29)\end{aligned}$$

where the last equality follows from a convex duality argument for the Kullback-Leibler divergence (cf., e.g., [Cat04, p. 159]) which we recall in Proposition A.1 in Appendix A.1.

On the other hand, we rewrite $\eta^{-1}\ln(W_{T+1}/W_1) = \sum_{t=1}^{T}\eta^{-1}\ln(W_{t+1}/W_t)$ as a telescopic sum and note that, for all $t = 1\ldots,T$,

$$\begin{aligned}&\frac{1}{\eta}\ln\frac{W_{t+1}}{W_t} \\ &\quad= \frac{1}{\eta}\ln\int_{\mathbb{R}^d}\frac{\exp\left(-\eta\Big(y_t - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_B\Big)^2\right)\exp\left(-\eta\sum_{s=1}^{t-1}\Big(y_s - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_s)\big]_B\Big)^2\right)}{W_t}\pi_\tau(\mathrm{d}\boldsymbol{u}) \\ &\quad= \frac{1}{\eta}\ln\int_{\mathbb{R}^d}\exp\left(-\eta\Big(y_t - \big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_B\Big)^2\right)p_t(\mathrm{d}\boldsymbol{u}) \,, \qquad (3.30)\end{aligned}$$

where (3.30) follows from the definition of p_t.

Let $t \in \{1,\ldots,T\}$. First note that by assumption $y_t \in [-B_y, B_y] \subset [-B, B]$ so that both y_t and $\big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_B$ are $[-B,B]$-valued for all $\boldsymbol{u}\in\mathbb{R}^d$. Moreover, from Proposition A.2 in Appendix A.2, the square loss is $1/(8B^2)$-exp-concave on $[-B,B]$ and thus η-exp-concave (since $\eta \leqslant 1/(8B^2)$ by assumption). Therefore, by Jensen's inequality,

$$\int_{\mathbb{R}^d} e^{-\eta\left(y_t - \left[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\right]_B\right)^2} p_t(\mathrm{d}\boldsymbol{u}) \leqslant \exp\left(-\eta\left(y_t - \int_{\mathbb{R}^d}\big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_B\, p_t(\mathrm{d}\boldsymbol{u})\right)^2\right) .$$

Taking the logarithms of both sides of the inequality yields

$$\begin{aligned}\ln\int_{\mathbb{R}^d} e^{-\eta\left(y_t - \left[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\right]_B\right)^2} p_t(\mathrm{d}\boldsymbol{u}) &\leqslant -\eta\left(y_t - \int_{\mathbb{R}^d}\big[\boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big]_B\, p_t(\mathrm{d}\boldsymbol{u})\right)^2 \\ &= -\eta(y_t - \widehat{y}_t)^2 \,. \qquad (3.31)\end{aligned}$$

Dividing the latter inequality by η, summing over $t \in \{1,\ldots,T\}$ and combining with Equa-

tion (3.30), we get

$$\frac{1}{\eta}\ln\frac{W_{T+1}}{W_1} \leqslant -\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2\,.$$

We conclude the proof of (3.5) of Lemma 3.1 by combining the last inequality with (3.29). □

3.A.2 Proofs of Theorem 3.1 and Corollary 3.4

Before proving Theorem 3.1, we first need the following comment. Since the algorithm SeqSEW^*_τ is restarted at the beginning of each regime, the threshold values B_t used on regime r by the algorithm SeqSEW^*_τ are not computed on the basis of all past observations $y_1,\dots,y_{t-1}$ but only on the basis of the past observations y_t, $t\in\{t_{r-1}+1,\dots,t-1\}$. To avoid any ambiguity, we set

$$B_{r,t} \triangleq \left(2^{\lceil \log_2 \max_{t_{r-1}+1\leqslant s\leqslant t-1} y_s^2\rceil}\right)^{1/2}\,, \quad t\in\{t_{r-1}+1,\dots,t_r\}\,. \tag{3.32}$$

Proof (of Theorem 3.1): We denote by $R\triangleq\min\{r\in\mathbb{N}: T\leqslant t_r\}$ the index of the last regime. For notational convenience, we re-define $t_R\triangleq T$ (even if $\gamma_T\leqslant 2^R$).

We upper bound the regret of the algorithm SeqSEW^*_* on $\{1,\dots,T\}$ by the sum of its regrets on each time interval. To do so, first note that

$$\begin{aligned}
\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 &= \sum_{r=0}^{R}\sum_{t=t_{r-1}+1}^{t_r}(y_t-\widehat{y}_t)^2 = \sum_{r=0}^{R}\left((y_{t_r}-\widehat{y}_{t_r})^2+\sum_{t=t_{r-1}+1}^{t_r-1}(y_t-\widehat{y}_t)^2\right)\\
&\leqslant \sum_{r=0}^{R}\left(2(y_{t_r}^2+B_{r,t_r}^2)+\sum_{t=t_{r-1}+1}^{t_r-1}(y_t-\widehat{y}_t)^2\right) \qquad (3.33)\\
&\leqslant \sum_{r=0}^{R}\left(\sum_{t=t_{r-1}+1}^{t_r-1}(y_t-\widehat{y}_t)^2\right)+6(R+1){y_T^*}^2\,, \qquad (3.34)
\end{aligned}$$

where we set $y_T^*\triangleq\max_{1\leqslant t\leqslant T}|y_t|$, where (3.33) follows from the upper bound $(y_{t_r}-\widehat{y}_{t_r})^2\leqslant 2(y_{t_r}^2+\widehat{y}_{t_r}^2)\leqslant 2(y_{t_r}^2+B_{r,t_r}^2)$ (since $|\widehat{y}_{t_r}|\leqslant B_{r,t_r}$ by construction), and where (3.34) follows from the inequality $y_{t_r}^2\leqslant {y_T^*}^2$ and the fact that

$$B_{r,t_r}^2\triangleq 2^{\lceil\log_2\max_{t_{r-1}+1\leqslant t\leqslant t_r-1}y_t^2\rceil}\leqslant 2\max_{t_{r-1}+1\leqslant t\leqslant t_r-1}y_t^2\leqslant 2\,{y_T^*}^2\,.$$

But, for every $r=0,\dots,R$, the trace of the empirical Gram matrix on $\{t_{r-1}+1,\dots,t_r-1\}$ is upper bounded by

$$\sum_{t=t_{r-1}+1}^{t_r-1}\sum_{j=1}^{d}\varphi_j^2(x_t)\leqslant\sum_{t=1}^{t_r-1}\sum_{j=1}^{d}\varphi_j^2(x_t)\leqslant(e^{2^r}-1)^2\,,$$

where the last inequality follows from the fact that $\gamma_{t_r-1}\leqslant 2^r$ (by definition of t_r). Since in addition $\tau_r\triangleq 1/\sqrt{(e^{2^r}-1)^2}$, we can apply Corollory 3.2 on each period $\{t_{r-1}+1,\dots,t_r-1\}$,

$r = 0, \dots, R$, with $B_\Phi = (e^{2^r} - 1)^2$ and get from (3.34) the upper bound

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \sum_{r=0}^{R} \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=t_{r-1}+1}^{t_r - 1} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + \Delta_r(\boldsymbol{u}) \right\} + 6(R+1) {y_T^*}^2 \,, \quad (3.35)$$

where

$$\Delta_r(\boldsymbol{u}) \triangleq 32 B_{t_r}^2 \, \|\boldsymbol{u}\|_0 \ln \left(1 + \frac{\left(e^{2^r} - 1\right) \|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0} \right) + 16 B_{r,t_r}^2 + 1 \,. \quad (3.36)$$

Since the infimum is superadditive and since $\big(y_{t_r} - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_{t_r})\big)^2 \geqslant 0$ for all $r = 0, \dots, R$, we get from (3.35) that

$$\begin{aligned} \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 &\leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \sum_{r=0}^{R} \left(\sum_{t=t_{r-1}+1}^{t_r} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + \Delta_r(\boldsymbol{u}) \right) + 6(R+1) {y_T^*}^2 \\ &= \inf_{\boldsymbol{u} \in \mathbb{R}^d} \left\{ \sum_{t=1}^{T} \big(y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 + \sum_{r=0}^{R} \Delta_r(\boldsymbol{u}) \right\} + 6(R+1) {y_T^*}^2 \,. \end{aligned} \quad (3.37)$$

Let $\boldsymbol{u} \in \mathbb{R}^d$. Next we bound $\sum_{r=0}^{R} \Delta_r(\boldsymbol{u})$ and $6(R+1){y_T^*}^2$ from above. First note that, by the upper bound $B_{r,t_r}^2 \leqslant 2{y_T^*}^2$ and by the elementary inequality $\ln(1+xy) \leqslant \ln\big((1+x)(1+y)\big) = \ln(1+x) + \ln(1+y)$ with $x = e^{2^r} - 1$ and $y = \|\boldsymbol{u}\|_1 / \|\boldsymbol{u}\|_0$, (3.36) yields

$$\Delta_r(\boldsymbol{u}) \leqslant 64 \, {y_T^*}^2 \, \|\boldsymbol{u}\|_0 \, 2^r + 64 \, {y_T^*}^2 \, \|\boldsymbol{u}\|_0 \ln \left(1 + \frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0} \right) + 32 {y_T^*}^2 + 1 \,.$$

Summing over $r = 0, \dots, R$, we get

$$\sum_{r=0}^{R} \Delta_r(\boldsymbol{u}) \leqslant 64 \left(2^{R+1} - 1\right) {y_T^*}^2 \, \|\boldsymbol{u}\|_0 + (R+1) \left(64 \, {y_T^*}^2 \, \|\boldsymbol{u}\|_0 \ln \left(1 + \frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0} \right) + 32 {y_T^*}^2 + 1 \right) . \quad (3.38)$$

First case: $R = 0$

Substituting (3.38) in (3.37), we conclude the proof by noting that $A_T \geqslant 2 + \log_2 1 \geqslant 1$ and that $\ln\Big(e + \sqrt{\sum_{t=1}^{T} \sum_{j=1}^{d} \varphi_j^2(x_t)}\Big) \geqslant 1$.

Second case: $R \geqslant 1$

Since $R \geqslant 1$, we have, by definition of t_{R-1},

$$2^{R-1} < \gamma_{t_{R-1}} \triangleq \ln \left(1 + \sqrt{\sum_{t=1}^{t_{R-1}} \sum_{j=1}^{d} \varphi_j^2(x_t)} \right) \leqslant \ln \left(e + \sqrt{\sum_{t=1}^{T} \sum_{j=1}^{d} \varphi_j^2(x_t)} \right) .$$

The last inequality entails that $2^{R+1} - 1 \leqslant 4 \cdot 2^{R-1} \leqslant 4 \ln\Big(e + \sqrt{\sum_{t=1}^{T} \sum_{j=1}^{d} \varphi_j^2(x_t)}\Big)$ and that

$R+1 \leqslant 2+\log_2 \ln\left(e+\sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)}\right) \triangleq A_T$. Therefore, one the one hand, via (3.38),

$$\sum_{r=0}^{R} \Delta_r(\boldsymbol{u}) \leqslant 256\, {y_T^*}^2\, \|\boldsymbol{u}\|_0 \ln\left(e+\sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)}\right) + 64\, {y_T^*}^2 A_T\, \|\boldsymbol{u}\|_0 \ln\left(1+\frac{\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right) + A_T\left(32{y_T^*}^2+1\right) ,$$

and, on the other hand,

$$6(R+1){y_T^*}^2 \leqslant 6A_T\, {y_T^*}^2 .$$

Substituting the last two inequalities in (3.37) and noting that ${y_T^*}^2 = \max_{1\leqslant t\leqslant T} y_t^2$ concludes the proof. □

Proof (of Corollary 3.4): The proof is straightforward. In view of Theorem 3.1, we just need to check that the quantity (continuously extended in $s=0$)

$$256\Big(\max_{1\leqslant t\leqslant T} y_t^2\Big)\, s\, \ln\left(e+\sqrt{\sum_{t=1}^T \sum_{j=1}^d \varphi_j^2(x_t)}\right) + 64\Big(\max_{1\leqslant t\leqslant T} y_t^2\Big) A_T\, s\, \ln\left(1+\frac{U}{s}\right)$$

is non-decreasing in $s \in \mathbb{R}_+$ and in $U \in \mathbb{R}_+$.

This is clear for U. The fact that it also non-decreasing in s comes from the following remark. For all $U \geqslant 0$, the function $s \in (0,+\infty) \mapsto s\, \ln(1+U/s)$ has a derivative equal to

$$\ln\left(1+\frac{U}{s}\right) - \frac{U/s}{1+U/s} \qquad \text{for all} \quad s>0 .$$

From the elementary inequality

$$\ln(1+u) = -\ln\left(\frac{1}{1+u}\right) \geqslant -\left(\frac{1}{1+u}-1\right) = \frac{u}{1+u} ,$$

which holds for all $u \in (-1,+\infty)$, the above derivative is nonnegative for all $s>0$ so that the continuous extension $s \in \mathbb{R}_+ \mapsto s\, \ln(1+U/s)$ is non-decreasing. □

3.A.3 Proofs of Theorem 3.2 and Corollary 3.5

In this subsection, we set $\varepsilon \triangleq Y - f(X)$, so that the pairs $(X_1,\varepsilon_1),\ldots,(X_T,\varepsilon_T)$ are independent copies of $(X,\varepsilon) \in \mathcal{X}\times\mathbb{R}$. We also define $\sigma \geqslant 0$ by

$$\sigma^2 \triangleq \mathbb{E}\big[\varepsilon^2\big] = \mathbb{E}\big[(Y-f(X))^2\big] .$$

Proof (of Theorem 3.2): By Corollory 3.3 and the definitions of $\widetilde{f}_t$ above and $\widehat{y}_t \triangleq \widetilde{f}_t(X_t)$ in Figure 3.3, we have, *almost surely*,

$$\sum_{t=1}^T (Y_t - \widetilde{f}_t(X_t))^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^T \big(Y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(X_t)\big)^2 + 64\Big(\max_{1\leqslant t\leqslant T} Y_t^2\Big)\, \|\boldsymbol{u}\|_0 \ln\left(1+\frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\right\}$$

$$+ \frac{1}{dT} \sum_{j=1}^{d} \sum_{t=1}^{T} \varphi_j^2(X_t) + 32 \max_{1 \leqslant t \leqslant T} Y_t^2 \, .$$

It remains to take the expectations of both sides with respect to $\big((X_1, Y_1), \dots, (X_T, Y_T)\big)$. First note that for all $t = 1, \dots, T$, since $\varepsilon_t \triangleq Y_t - f(X_t)$, we have

$$\begin{aligned} \mathbb{E}\left[\big(Y_t - \widetilde{f}_t(X_t)\big)^2\right] &= \mathbb{E}\left[\big(\varepsilon_t + f(X_t) - \widetilde{f}_t(X_t)\big)^2\right] \\ &= \sigma^2 + \mathbb{E}\left[\big(f(X_t) - \widetilde{f}_t(X_t)\big)^2\right] \, , \end{aligned}$$

since $\mathbb{E}\big[\varepsilon_t^2\big] = \mathbb{E}\big[\varepsilon^2\big] \triangleq \sigma^2$ one the one hand, and, on the other hand, $\widetilde{f}_t$ is a measurable function of $(X_s, Y_s)_{1 \leqslant s \leqslant t-1}$ and $\mathbb{E}\big[\varepsilon_t \big| (X_s, Y_s)_{1 \leqslant s \leqslant t-1}, X_t\big] = \mathbb{E}\big[\varepsilon_t \big| X_t\big] = 0$ (from the independence of $(X_s, Y_s)_{1 \leqslant s \leqslant t-1}$ and (X_t, Y_t) and by definition of f).

In the same way,

$$\mathbb{E}\left[\big(Y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(X_t)\big)^2\right] = \sigma^2 + \mathbb{E}\left[\big(f(X_t) - \boldsymbol{u} \cdot \boldsymbol{\varphi}(X_t)\big)^2\right] \, .$$

Therefore, by Jensen's inequality and the concavity of the infimum, the last inequality becomes, after taking the expectations of both sides,

$$\begin{aligned} T\sigma^2 + \sum_{t=1}^{T} \mathbb{E}\left[\big(f(X_t) - \widetilde{f}_t(X_t)\big)^2\right] \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \Bigg\{ & T\sigma^2 + \sum_{t=1}^{T} \mathbb{E}\left[\big(f(X_t) - \boldsymbol{u} \cdot \boldsymbol{\varphi}(X_t)\big)^2\right] \\ & + 64\, \mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2\right] \|\boldsymbol{u}\|_0 \, \ln\left(1 + \frac{\sqrt{dT}\, \|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right) \Bigg\} \\ & + \frac{1}{dT} \sum_{j=1}^{d} \sum_{t=1}^{T} \mathbb{E}\left[\varphi_j^2(X_t)\right] + 32\, \mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2\right] \, . \end{aligned}$$

Noting that the $T\sigma^2$ cancel out, dividing the two sides by T, and using the fact that $X_t \sim X$ in the right-hand side, we get

$$\begin{aligned} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\big(f(X_t) - \widetilde{f}_t(X_t)\big)^2\right] \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \Bigg\{ & \|f - \boldsymbol{u} \cdot \boldsymbol{\varphi}\|_{L^2}^2 \\ & + 64\, \frac{\mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2\right]}{T} \|\boldsymbol{u}\|_0 \, \ln\left(1 + \frac{\sqrt{dT}\, \|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right) \Bigg\} \\ & + \frac{1}{dT} \sum_{j=1}^{d} \|\varphi_j\|_{L^2}^2 + 32\, \frac{\mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2\right]}{T} \, . \end{aligned}$$

The right-hand side of the last inequality is exactly the upper bound stated in Theorem 3.2. To conclude the proof, we thus only need to check that $\|f - \widehat{f}_T\|_{L^2}^2$ is bounded from above by the left-hand side. But by definition of $\widehat{f}_T$ and by convexity of the square loss we have

$$\mathbb{E}\left[\left\| f - \widehat{f}_T \right\|_{L^2}^2\right] \triangleq \mathbb{E}\left[\left(f(X) - \frac{1}{T} \sum_{t=1}^{T} \widetilde{f}_t(X)\right)^2\right]$$

$$\leqslant \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(f(X)-\widetilde{f}_t(X)\right)^2\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(f(X_t)-\widetilde{f}_t(X_t)\right)^2\right] .$$

The last equality follows classically from the fact that, for all $t = 1, \dots, T$, $(X_s, Y_s)_{1\leqslant s\leqslant t-1}$ (on which $\widetilde{f}_t$ is constructed) is independent from both X_t and X and the fact that $X_t \sim X$. □

Remark 3.3. *The fact that the inequality stated in Corollary 3.3 has a leading constant equal to* 1 *on individual sequences is crucial to derive in the stochastic setting an oracle inequality in terms of the (excess) risks* $\mathbb{E}\left[\|f-\widehat{f}_T\|_{L^2}^2\right]$ *and* $\|f-\boldsymbol{u}\cdot\boldsymbol{\varphi}\|_{L^2}^2$. *Indeed, if the constant appearing in front of the infimum was equal to* $C > 1$, *then the* $T\sigma^2$ *would not cancel out in the previous proof, so that the resulting expected inequality would contain a non-vanishing additive term* $(C-1)\sigma^2$.

Proof (of Corollary 3.5): We can apply Theorem 3.2. Then, to prove the upper bound on $\mathbb{E}\left[\|f-\widehat{f}_T\|_{L^2}^2\right]$, it suffices to show that

$$\frac{\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right]}{T} \leqslant 2\left(\frac{\mathbb{E}[Y]^2}{T}+\psi_T\right) . \tag{3.39}$$

Recall that

$$\psi_T \triangleq \frac{1}{T}\mathbb{E}\left[\max_{1\leqslant t\leqslant T}\left(Y_t-\mathbb{E}[Y_t]\right)^2\right] = \frac{1}{T}\mathbb{E}\left[\max_{1\leqslant t\leqslant T}(\Delta Y)_t^2\right] ,$$

where we defined $(\Delta Y)_t \triangleq Y_t - \mathbb{E}[Y_t] = Y_t - \mathbb{E}[Y]$ for all $t = 1, \dots, T$.

From the elementary inequality $(x+y)^2 \leqslant 2x^2+2y^2$ for all $x, y \in \mathbb{R}$, we have

$$\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Y_t^2\right] \triangleq \mathbb{E}\left[\max_{1\leqslant t\leqslant T}\left(\mathbb{E}[Y]+(\Delta Y)_t\right)^2\right] \leqslant 2\,\mathbb{E}[Y]^2 + 2\,\mathbb{E}\left[\max_{1\leqslant t\leqslant T}(\Delta Y)_t^2\right] . \tag{3.40}$$

Dividing both sides by T, we get (3.39).

As for the upper bound on ψ_T, since the $(\Delta Y)_t$, $1 \leqslant t \leqslant T$, are distributed as ΔY, we can apply Lemmas 3.5, 3.6, and 3.7 in Appendix 3.B.2 to bound ψ_T from above under the assumptions $\big(\mathrm{SG}(\sigma^2)\big)$, $\big(\mathrm{BEM}(\alpha,\mathrm{M})\big)$, and $\big(\mathrm{BM}(\alpha,\mathrm{M})\big)$ respectively (the upper bound under $\big(\mathrm{BD}(\mathrm{B})\big)$ is straightforward):

$$\mathbb{E}\left[\max_{1\leqslant t\leqslant T}(\Delta Y)_t^2\right] \leqslant \begin{cases} B^2 & \text{if Assumption } \big(\mathrm{BD}(\mathrm{B})\big) \text{ holds,} \\ \sigma^2 + 2\sigma^2\ln(2eT) & \text{if Assumption } \big(\mathrm{SG}(\sigma^2)\big) \text{ holds,} \\ \dfrac{\ln^2\big((M+e)T\big)}{\alpha^2} & \text{if Assumption } \big(\mathrm{BEM}(\alpha,\mathrm{M})\big) \text{ holds,} \\ (MT)^{2/\alpha} & \text{if Assumption } \big(\mathrm{BM}(\alpha,\mathrm{M})\big) \text{ holds .} \end{cases}$$

□

Remark 3.4. *If* $T \geqslant 2$,*then the two "bias" terms* $\mathbb{E}[Y]^2/T$ *appearing in Corollary 3.5 can be avoided, at least at the price of a multiplicative factor of* $2T/(T-1) \leqslant 4$. *It suffices to use a slightly more sophisticated online clipping defined as follows. The first round* $t = 1$ *is only used*

to observe Y_1. Then, the algorithm SeqSEW^*_τ *is run with $\tau = 1/\sqrt{d(T-1)}$ from round 2 up to round T with the following important modification: instead of truncating the predictions to $[-B_t, B_t]$, which is best suited to the case $\mathbb{E}[Y] = 0$, we truncate them to the interval*

$$\big[Y_1 - B'_t, Y_1 + B'_t\big] \;, \quad \textit{where} \quad B'_t \triangleq \Big(2^{\lceil \log_2 \max_{2 \leqslant s \leqslant t-1} |Y_s - Y_1|^2 \rceil}\Big)^{1/2} \;.$$

If η_t is changed accordingly, i.e., if $\eta_t = 1/(8{B'_t}^2)$, then it easy to see that the resulting procedure $\widehat{f}_T \triangleq \frac{1}{T-1}\sum_{s=2}^T \widetilde{f}_s$ (where $\widetilde{f}_2, \dots, \widetilde{f}_T$ are the regressors output by SeqSEW^*_τ*) satisfies*

$$\mathbb{E}\left[\left\| f - \widehat{f}_T \right\|^2_{L^2}\right] \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{ \|f - \boldsymbol{u}\cdot\boldsymbol{\varphi}\|^2_{L^2} + 128\left(\frac{\mathit{Var}[Y]}{T-1} + \psi_{T-1}\right)\|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\right\}$$
$$+ \frac{1}{dT}\sum_{j=1}^d \|\varphi_j\|^2_{L^2} + 64\left(\frac{\mathit{Var}[Y]}{T-1} + \psi_{T-1}\right) \;,$$

where $\mathit{Var}[Y] \triangleq \mathbb{E}\big[(Y - \mathbb{E}[Y])^2\big]$. Comparing the last bound to that of Corollary 3.5, we note that the two terms $\mathbb{E}[Y]^2/T$ are absent, and that we loose a multiplicative factor at most of 4 since $\mathit{Var}[Y] \leqslant \mathbb{E}\big[\max_{2\leqslant t\leqslant T}(Y_t - \mathbb{E}[Y_t])^2\big] \triangleq (T-1)\psi_{T-1}$ so that

$$\frac{\mathit{Var}[Y]}{T-1} + \psi_{T-1} \leqslant 2\psi_{T-1} \leqslant 2\left(\frac{T}{T-1}\right)\psi_T \leqslant 4\psi_T \;.$$

Remark 3.5. *We mentioned after Corollary 3.5 that each of the four assumptions on ΔY is fulfilled as soon as both the distribution of $f(X) - \mathbb{E}[f(X)]$ and the conditional distribution of ε (conditionally on X) satisfy the same type of assumption. It actually extends to the more general case when the conditional distribution of ε given X is replaced with the distribution of ε itself (without conditioning). This relies on the elementary upper bound*

$$\mathbb{E}\left[\max_{1\leqslant t\leqslant T}(\Delta Y)^2_t\right] = \mathbb{E}\left[\max_{1\leqslant t\leqslant T}\big(f(X_t) - \mathbb{E}[f(X)] + \varepsilon_t\big)^2\right]$$
$$\leqslant 2\,\mathbb{E}\left[\max_{1\leqslant t\leqslant T}\big(f(X_t) - \mathbb{E}[f(X)]\big)^2\right] + 2\,\mathbb{E}\left[\max_{1\leqslant t\leqslant T}\varepsilon_t^2\right] \;.$$

From the last inequality, we can also see that assumptions of different nature can be made on $f(X) - \mathbb{E}[f(X)]$ and ε, such as the assumptions given in (3.23) or in (3.24).

3.A.4 Proofs of Theorem 3.3 and Corollary 3.7

Proof (of Theorem 3.3): The proof follows the sames lines as in the proof of Theorem 3.2. We thus only sketch the main arguments. In the sequel, we set $\sigma^2 \triangleq \mathbb{E}\big[\varepsilon_1^2\big]$.

Applying Corollory 3.3 we have, *almost surely*,

$$\sum_{t=1}^T \big(Y_t - \widetilde{f}_t(x_t)\big)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^T \big(Y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 + 64\left(\max_{1\leqslant t\leqslant T} Y_t^2\right)\|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{dT}\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\right\}$$

$$+ \frac{1}{dT} \sum_{j=1}^{d} \sum_{t=1}^{T} \varphi_j^2(x_t) + 32 \max_{1 \leqslant t \leqslant T} Y_t^2 \,.$$

Taking the expectations of both sides, expanding the squares $\big(Y_t - \widetilde{f}_t(x_t)\big)^2$ and $\big(Y_t - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2$, noting that two terms $T\sigma^2$ cancel out, and then dividing both sides by T, we get

$$\begin{aligned}\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t) - \widetilde{f}_t(x_t)\big)^2\right] \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d} \Bigg\{ & \frac{1}{T}\sum_{t=1}^{T}\big(f(x_t) - \boldsymbol{u} \cdot \boldsymbol{\varphi}(x_t)\big)^2 \\ & + 64\, \frac{\mathbb{E}\big[\max_{1 \leqslant t \leqslant T} Y_t^2\big]}{T}\, \|\boldsymbol{u}\|_0 \ln\left(1 + \frac{\sqrt{dT}\, \|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0}\right)\Bigg\} \\ & + \frac{1}{dT^2}\sum_{j=1}^{d}\sum_{t=1}^{T}\varphi_j^2(x_t) + 32\, \frac{\mathbb{E}\big[\max_{1 \leqslant t \leqslant T} Y_t^2\big]}{T} \,.\end{aligned}$$

The right-hand side is exactly the upper bound stated in Theorem 3.3. We thus only need to check that

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t) - \widehat{f}_T(x_t)\big)^2\right] \leqslant \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\big(f(x_t) - \widetilde{f}_t(x_t)\big)^2\right] \,. \tag{3.42}$$

This is an equality if the x_t are all distinct. In general we get an inequality which follows from the convexity of the square loss. Indeed, by definition of n_x, we have, almost surely,

$$\begin{aligned}\sum_{t=1}^{T}\big(f(x_t) - \widehat{f}_T(x_t)\big)^2 &= \sum_{x \in \{x_1,\dots,x_T\}} \sum_{\substack{1 \leqslant t \leqslant T \\ t : x_t = x}} \big(f(x_t) - \widehat{f}_T(x_t)\big)^2 = \sum_{x \in \{x_1,\dots,x_T\}} n_x \big(f(x) - \widehat{f}_T(x)\big)^2 \\ &= \sum_{x \in \{x_1,\dots,x_T\}} n_x \left(f(x) - \frac{1}{n_x} \sum_{\substack{1 \leqslant t \leqslant T \\ t : x_t = x}} \widetilde{f}_t(x)\right)^2 \\ &\leqslant \sum_{x \in \{x_1,\dots,x_T\}} n_x \frac{1}{n_x} \sum_{\substack{1 \leqslant t \leqslant T \\ t : x_t = x}} \big(f(x) - \widetilde{f}_t(x)\big)^2 = \sum_{t=1}^{T}\big(f(x_t) - \widetilde{f}_t(x_t)\big)^2 \,,\end{aligned}$$

where the second line is by definition of $\widehat{f}_T$ and where the last line follows from Jensen's inequality. Dividing both sides by T and taking their expectations, we get (3.42), which concludes the proof. □

Proof (of Corollary 3.7): First note that

$$\mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Y_t^2\right] \triangleq \mathbb{E}\left[\max_{1 \leqslant t \leqslant T}\big(f(x_t) + \varepsilon_t\big)^2\right] \leqslant 2\left(\max_{1 \leqslant t \leqslant T} f^2(x_t) + \mathbb{E}\left[\max_{1 \leqslant t \leqslant T} \varepsilon_t^2\right]\right) \,.$$

The proof then follows the exact same lines as for Corollary 3.5 with the sequence (ε_t) instead of the sequence $\big((\Delta Y)_t\big)$. □

3.B Tools

3.B.1 Some tools to exploit our PAC-Bayesian inequalities

In this section, we recall two results needed for the derivation of Proposition 3.1 and Proposition 3.2 from the PAC-Bayesian inequalities (3.6) and (3.15). The proofs are due to [DT07, DT08] and we only reproduce[8] them for the convenience of the reader.

For any $\boldsymbol{u}^* \in \mathbb{R}^d$ and $\tau > 0$, define $\rho_{\boldsymbol{u}^*,\tau}$ as the translated of π_τ at $\boldsymbol{u}^*$, namely,

$$\rho_{\boldsymbol{u}^*,\tau} \triangleq \frac{\mathrm{d}\pi_\tau}{\mathrm{d}\boldsymbol{u}}(\boldsymbol{u}-\boldsymbol{u}^*)\,\mathrm{d}\boldsymbol{u} = \prod_{j=1}^d \frac{(3/\tau)\,\mathrm{d}u_j}{2\big(1+|u_j-u_j^*|/\tau\big)^4}\,. \tag{3.43}$$

Lemma 3.3. *For all $\boldsymbol{u}^* \in \mathbb{R}^d$ and $\tau > 0$, the probability distribution $\rho_{\boldsymbol{u}^*,\tau}$ satisfies*

$$\int_{\mathbb{R}^d} \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 \rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u}) = \sum_{t=1}^T \big(y_t - \boldsymbol{u}^*\cdot\boldsymbol{\varphi}(x_t)\big)^2 + \tau^2 \sum_{j=1}^d \sum_{t=1}^T \varphi_j^2(x_t)\,.$$

Lemma 3.4. *For all $\boldsymbol{u}^* \in \mathbb{R}^d$ and $\tau > 0$, the probability distribution $\rho_{\boldsymbol{u}^*,\tau}$ satisfies*

$$\mathcal{K}(\rho_{\boldsymbol{u}^*,\tau}, \pi_\tau) \leqslant 4\,\|\boldsymbol{u}^*\|_0 \ln\left(1 + \frac{\|\boldsymbol{u}^*\|_1}{\|\boldsymbol{u}^*\|_0\,\tau}\right)\,.$$

Proof (of Lemma 3.3): For all $t \in \{1,\dots,T\}$ we expand the square $\big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 = \big(y_t - \boldsymbol{u}^*\cdot\boldsymbol{\varphi}(x_t) + (\boldsymbol{u}^*-\boldsymbol{u})\cdot\boldsymbol{\varphi}(x_t)\big)^2$ and use the linearity of the integral to get

$$\begin{aligned}
&\int_{\mathbb{R}^d} \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2 \rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u}) \\
&\quad= \sum_{t=1}^T \big(y_t - \boldsymbol{u}^*\cdot\boldsymbol{\varphi}(x_t)\big)^2 + \sum_{t=1}^T \int_{\mathbb{R}^d} \big((\boldsymbol{u}^*-\boldsymbol{u})\cdot\boldsymbol{\varphi}(x_t)\big)^2 \rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u}) \\
&\qquad+ \underbrace{\sum_{t=1}^T 2\big(y_t - \boldsymbol{u}^*\cdot\boldsymbol{\varphi}(x_t)\big) \int_{\mathbb{R}^d} (\boldsymbol{u}^*-\boldsymbol{u})\cdot\boldsymbol{\varphi}(x_t)\,\rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u})}_{=0}
\end{aligned} \tag{3.44}$$

The last sum equals zero by symmetry of $\rho_{\boldsymbol{u}^*,\tau}$ around $\boldsymbol{u}^*$, which entails that $\int_{\mathbb{R}} \boldsymbol{u}\,\rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u}) = \boldsymbol{u}^*$. As for the second sum of the right-hand side, it can be bounded from above similarly. Indeed, ex-

[8] The notations are however slightly modified because of the change in the statistical setting and goal. The target predictions $(f(x_1),\dots,f(x_T))$ are indeed replaced with the observations $(y_1,\dots,y_T)$ and the prediction loss $\|f - f_{\boldsymbol{u}}\|_n^2$ is replaced with the cumulative loss $\sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{\varphi}(x_t)\big)^2$. Moreover, the analysis of the present proof is slightly simpler since we just need to consider the case $L_0 = +\infty$ according to the notations of Theorem 5 in [DT08].

panding the inner product and then the square $\big((\boldsymbol{u}^*-\boldsymbol{u})\cdot\boldsymbol{\varphi}(x_t)\big)^2$ we have, for all $t=1,\dots,T$,

$$\big((\boldsymbol{u}^*-\boldsymbol{u})\cdot\boldsymbol{\varphi}(x_t)\big)^2=\sum_{j=1}^{d}(u_j^*-u_j)^2\varphi_j^2(x_t)+\sum_{1\leqslant j\neq k\leqslant d}(u_j^*-u_j)(u_k^*-u_k)\,\varphi_j(x_t)\,\varphi_k(x_t)\,.$$

By symmetry of $\rho_{\boldsymbol{u}^*,\tau}$ around $\boldsymbol{u}^*$ and the fact that $\rho_{\boldsymbol{u}^*,\tau}$ is a product-distribution, we get

$$\begin{aligned}\sum_{t=1}^{T}\int_{\mathbb{R}^d}\big((\boldsymbol{u}^*-\boldsymbol{u})\cdot\boldsymbol{\varphi}(x_t)\big)^2\rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u})&=\sum_{t=1}^{T}\sum_{j=1}^{d}\varphi_j^2(x_t)\int_{\mathbb{R}^d}(u_j^*-u_j)^2\rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u})\;+0\\&=\sum_{t=1}^{T}\sum_{j=1}^{d}\varphi_j^2(x_t)\int_{\mathbb{R}}(u_j^*-u_j)^2\frac{(3/\tau)\,\mathrm{d}u_j}{2\big(1+|u_j-u_j^*|/\tau\big)^4}\qquad(3.45)\\&=\tau^2\sum_{t=1}^{T}\sum_{j=1}^{d}\varphi_j^2(x_t)\int_{\mathbb{R}}\frac{3t^2\mathrm{d}t}{2(1+|t|)^4}\qquad(3.46)\\&=\tau^2\sum_{t=1}^{T}\sum_{j=1}^{d}\varphi_j^2(x_t)\,.\qquad(3.47)\end{aligned}$$

Equation (3.45) follows from the definition of $\rho_{\boldsymbol{u}^*,\tau}$. Equation (3.46) is obtained by the change of variables $t=(u_j-u_j^*)/\tau$. As for Equation (3.47), it follows from the equality $\displaystyle\int_{\mathbb{R}}\frac{3t^2\mathrm{d}t}{2\big(1+|t|\big)^4}=1$ that can be proved by integrating by parts.

Combined with (3.47), Equation (3.44) yields the desired equality. □

Proof (of Lemma 3.4): By definition of $\rho_{\boldsymbol{u}^*,\tau}$ and π_τ, we have

$$\begin{aligned}\mathcal{K}(\rho_{\boldsymbol{u}^*,\tau},\pi_\tau)\triangleq\int_{\mathbb{R}^d}\left(\ln\frac{\mathrm{d}\rho_{\boldsymbol{u}^*,\tau}}{\mathrm{d}\pi_\tau}(\boldsymbol{u})\right)\rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u})&=\int_{\mathbb{R}^d}\left(\ln\prod_{j=1}^{d}\frac{\big(1+|u_j|/\tau\big)^4}{\big(1+|u_j-u_j^*|/\tau\big)^4}\right)\rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u})\\&=4\int_{\mathbb{R}^d}\left(\sum_{j=1}^{d}\ln\frac{1+|u_j|/\tau}{1+|u_j-u_j^*|/\tau}\right)\rho_{\boldsymbol{u}^*,\tau}(\mathrm{d}\boldsymbol{u})\,.\qquad(3.48)\end{aligned}$$

But, for all $\boldsymbol{u}\in\mathbb{R}^d$, by the triangle inequality,

$$1+|u_j|/\tau\leqslant 1+|u_j^*|/\tau+|u_j-u_j^*|/\tau\leqslant\big(1+|u_j^*|/\tau\big)\big(1+|u_j-u_j^*|/\tau\big)\,,$$

so that Equation (3.48) yields the upper bound

$$\mathcal{K}(\rho_{\boldsymbol{u}^*,\tau},\pi_\tau)\leqslant 4\sum_{j=1}^{d}\ln\big(1+|u_j^*|/\tau\big)=4\sum_{j:u_j^*\neq 0}\ln\big(1+|u_j^*|/\tau\big)\,.$$

We now recall that $\|\boldsymbol{u}^*\|_0\triangleq\big|\{j:u_j^*\neq 0\}\big|$ and apply Jensen's inequality to the concave function

$x \in (-1, +\infty) \longmapsto \ln(1+x)$ to get

$$\sum_{j:u_j^*\neq 0} \ln\left(1+|u_j^*|/\tau\right) = \|\boldsymbol{u}^*\|_0 \, \frac{1}{\|\boldsymbol{u}^*\|_0} \sum_{j:u_j^*\neq 0} \ln\left(1+|u_j^*|/\tau\right) \leqslant \|\boldsymbol{u}^*\|_0 \ln\left(1+\frac{\sum_{j:u_j^*\neq 0}|u_j^*|}{\|\boldsymbol{u}^*\|_0\,\tau}\right)$$
$$\leqslant \|\boldsymbol{u}^*\|_0 \ln\left(1+\frac{\|\boldsymbol{u}^*\|_1}{\|\boldsymbol{u}^*\|_0\,\tau}\right) .$$

This concludes the proof. □

3.B.2 Some maximal inequalities

In this section, we prove three maximal inequalities needed for the derivation of Corollaries 3.5 and 3.7 from Theorems 3.2 and 3.3 respectively. Their proofs are quite standard but we provide them for the convenience of the reader.

Lemma 3.5. *Let $Z_1, \dots, Z_T$ be $T \geqslant 1$ (centered) real random variables such that, for a given constant $\nu \geqslant 0$, we have*

$$\forall t \in \{1,\dots,T\}, \quad \forall \lambda \in \mathbb{R}, \quad \mathbb{E}\Big[e^{\lambda Z_t}\Big] \leqslant e^{\lambda^2\nu/2} . \tag{3.49}$$

Then,

$$\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Z_t^2\right] \leqslant 2\nu \ln(2\mathrm{e}T) .$$

Lemma 3.6. *Let $Z_1, \dots, Z_T$ be $T \geqslant 1$ real random variables such that, for some given constants $\alpha > 0$ and $M > 0$, we have*

$$\forall t \in \{1,\dots,T\}, \quad \mathbb{E}\Big[e^{\alpha|Z_t|}\Big] \leqslant M .$$

Then,

$$\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Z_t^2\right] \leqslant \frac{\ln^2\big((M+\mathrm{e})T\big)}{\alpha^2} .$$

Lemma 3.7. *Let $Z_1, \dots, Z_T$ be $T \geqslant 1$ real random variables such that, for some given constants $\alpha > 2$ and $M > 0$, we have*

$$\forall t \in \{1,\dots,T\}, \quad \mathbb{E}\big[|Z_t|^\alpha\big] \leqslant M .$$

Then,

$$\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Z_t^2\right] \leqslant (MT)^{2/\alpha} .$$

Proof (of Lemma 3.5): Let $t \in \{1,\dots,T\}$. From the subgaussian assumption (3.49) it is well-known (see, e.g., [Mas07, Chapter 2]) that for all $x \geqslant 0$, we have

$$\forall t \in \{1,\dots,T\}\,, \quad \mathbb{P}\big(|Z_t| > x\big) \leqslant 2e^{-x^2/(2\nu)} .$$

Let $\delta \in (0,1)$. By the change of variables $x = \sqrt{2\nu \ln(2T/\delta)}$, the last inequality entails that, for all $t = 1, \dots, T$, we have $|Z_t| \leqslant \sqrt{2\nu \ln(2T/\delta)}$ with probability at least $1 - \delta/T$. Therefore, by a union bound, we get, with probability at least $1 - \delta$,

$$\forall t \in \{1, \dots, T\}\,, \quad |Z_t| \leqslant \sqrt{2\nu \ln(2T/\delta)}\,.$$

As a consequence, with probability at least $1 - \delta$,

$$\max_{1 \leqslant t \leqslant T} Z_t^2 \leqslant 2\nu \ln(2T/\delta) \leqslant 2\nu \ln(1/\delta) + 2\nu \ln(2T)\,.$$

Integrating the last high-probability bound via Lemma A.7 in Appendix A.6 (cf. Example A.1 with the change of variables $z = \ln(1/\delta)$), we get that $\mathbb{E}\big[\max_{1 \leqslant t \leqslant T} Z_t^2\big] \leqslant 2\nu + 2\nu \ln(2T)$, which concludes the proof. □

Proof (of Lemma 3.6): We first need the following definitions. Let $\psi_\alpha : \mathbb{R}_+ \to \mathbb{R}$ be a convex majorant of $x \mapsto e^{\alpha\sqrt{x}}$ on $\mathbb{R}_+$ defined by

$$\psi_\alpha(x) \triangleq \begin{cases} e & \text{if } x < 1/\alpha^2\,, \\ e^{\alpha\sqrt{x}} & \text{if } x \geqslant 1/\alpha^2\,. \end{cases}$$

We associate with ψ_α its generalized inverse $\psi_\alpha^{-1} : \mathbb{R} \to \mathbb{R}_+$ defined by

$$\psi_\alpha^{-1}(y) = \begin{cases} 1/\alpha^2 & \text{if } y < e\,, \\ (\ln y)^2/\alpha^2 & \text{if } y \geqslant e\,. \end{cases}$$

Elementary manipulations show that:

- ψ_α is nondecreasing and convex on $\mathbb{R}_+$;
- ψ_α^{-1} is nondecreasing on $\mathbb{R}$;
- $x \leqslant \psi_\alpha^{-1}\big(\psi_\alpha(x)\big)$ for all $x \in \mathbb{R}_+$.

The proof is based on a Pisier-type argument as is done, e.g., in [Mas07, Lemma 2.3] to prove the maximal inequality $\mathbb{E}[\max_{1 \leqslant t \leqslant T} \xi_t] \leqslant \sqrt{2\nu \ln T}$ for all subgaussian real random variables ξ_t, $1 \leqslant t \leqslant T$, with common variance factor $\nu \geqslant 0$ (see Lemma A.3 in Appendix A.5).

From the inequality $x \leqslant \psi_\alpha^{-1}\big(\psi_\alpha(x)\big)$ for all $x \in \mathbb{R}_+$ we have

$$\begin{aligned} \mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Z_t^2\right] &\leqslant \psi_\alpha^{-1}\left(\psi_\alpha\left(\mathbb{E}\Big[\max_{1 \leqslant t \leqslant T} Z_t^2\Big]\right)\right) \\ &\leqslant \psi_\alpha^{-1}\left(\mathbb{E}\left[\psi_\alpha\Big(\max_{1 \leqslant t \leqslant T} Z_t^2\Big)\right]\right) = \psi_\alpha^{-1}\left(\mathbb{E}\left[\max_{1 \leqslant t \leqslant T} \psi_\alpha\big(Z_t^2\big)\right]\right)\,, \end{aligned}$$

where the last two inequalities follow by Jensen's inequality (since ψ_α is convex) and the fact that both ψ_α^{-1} and ψ_α are nondecreasing.

Since $\psi_\alpha \geqslant 0$ and ψ_α^{-1} is nondecreasing we get

$$\begin{aligned}\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Z_t^2\right] &\leqslant \psi_\alpha^{-1}\left(\mathbb{E}\left[\sum_{t=1}^T \psi_\alpha\big(Z_t^2\big)\right]\right) = \psi_\alpha^{-1}\left(\sum_{t=1}^T \mathbb{E}\Big[\psi_\alpha\big(Z_t^2\big)\Big]\right) \\ &\leqslant \psi_\alpha^{-1}\left(\sum_{t=1}^T \mathbb{E}\Big[e^{\alpha|Z_t|} + e\Big]\right) \\ &\leqslant \psi_\alpha^{-1}\big(MT + eT\big) = \frac{\ln^2\big(MT+eT\big)}{\alpha^2}\,,\end{aligned}$$

where the second line follows from the inequality $\psi_\alpha(x) \leqslant e + e^{\alpha\sqrt{x}}$ for all $x \in \mathbb{R}_+$, and where the last line follows from the bounded exponential moment assumption and the definition of ψ_α^{-1}. It concludes the proof. □

Proof (of Lemma 3.7): As in the previous proof, we have, by Jensen's inequality and the fact that $x \mapsto x^{\alpha/2}$ is convex and nondecreasing on $\mathbb{R}_+$ (since $\alpha > 2$),

$$\begin{aligned}\mathbb{E}\left[\max_{1\leqslant t\leqslant T} Z_t^2\right] &\leqslant \mathbb{E}\left[\left(\max_{1\leqslant t\leqslant T} Z_t^2\right)^{\alpha/2}\right]^{2/\alpha} = \mathbb{E}\left[\max_{1\leqslant t\leqslant T} \big|Z_t\big|^\alpha\right]^{2/\alpha} \\ &\leqslant \mathbb{E}\left[\sum_{t=1}^T \big|Z_t\big|^\alpha\right]^{2/\alpha} \leqslant (MT)^{2/\alpha}\end{aligned}$$

by the bounded moment assumption, which concludes the proof. □

Chapter 4

Adaptive and optimal online linear regression on ℓ^1-balls

We consider the problem of online linear regression on individual sequences. The goal in this paper is for the forecaster to output sequential predictions which are, after T time rounds, almost as good as the ones output by the best linear predictor in a given ℓ^1-ball in $\mathbb{R}^d$. We consider both the cases where the dimension d is small and large relative to the time horizon T. We first present regret bounds with optimal dependencies on d, T, and on the sizes U, X and Y of the ℓ^1-ball, the input data and the observations. The minimax regret is shown to exhibit a regime transition around the point $d = \sqrt{T}UX/(2Y)$. Furthermore, we present efficient algorithms that are adaptive, i.e., that do not require the knowledge of U, X, Y, and T, but still achieve nearly optimal regret bounds.

NOTA: This chapter is the full version of a conference paper [GY11] to be presented at ALT 2011. Some improved bounds are published here for the first time (Theorem 4.3 and Remark 4.1).

Contents

4.1 Introduction

In this chapter, we consider the problem of online linear regression against arbitrary sequences of input data and observations, with the objective of being competitive with respect to the best linear predictor in an ℓ^1-ball of arbitrary radius. This extends the task of convex aggregation. We consider both low- and high-dimensional input data. Indeed, in a large number of contemporary problems, the available data can be high-dimensional—the dimension of each data point is larger than the number of data points. Examples include analysis of DNA sequences, collaborative filtering, astronomical data analysis, and cross-country growth regression. In such high-dimensional problems, performing linear regression on an ℓ^1-ball of small diameter may be helpful if the best linear predictor is sparse. Our goal is, in both low and high dimensions, to provide online linear regression algorithms along with bounds on ℓ^1-balls that characterize their robustness to worst-case scenarios.

4.1.1 Setting

We consider the online version of linear regression, which unfolds as follows (see also Section 2.4 for an introduction to this setting). First, the environment chooses a sequence of observations $(y_t)_{t\geqslant 1}$ in $\mathbb{R}$ and a sequence of input vectors $(\boldsymbol{x}_t)_{t\geqslant 1}$ in $\mathbb{R}^d$, both initially hidden from the forecaster. At each time instant $t \in \mathbb{N}^* = \{1, 2, \ldots\}$, the environment reveals the data $\boldsymbol{x}_t \in \mathbb{R}^d$; the forecaster then gives a prediction $\widehat{y}_t \in \mathbb{R}$; the environment in turn reveals the observation $y_t \in \mathbb{R}$; and finally, the forecaster incurs the square loss $(y_t - \widehat{y}_t)^2$. The dimension d can be either small or large relative to the number T of time steps: we consider both cases.

In the sequel, $\boldsymbol{u} \cdot \boldsymbol{v}$ denotes the standard inner product between $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, and we set $\|\boldsymbol{u}\|_\infty \triangleq \max_{1\leqslant j\leqslant d} |u_j|$ and $\|\boldsymbol{u}\|_1 \triangleq \sum_{j=1}^d |u_j|$. The ℓ^1-ball of radius $U > 0$ is the following bounded subset of $\mathbb{R}^d$:

$$B_1(U) \triangleq \left\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U\right\}.$$

Given a fixed radius $U > 0$ and a time horizon $T \geqslant 1$, the goal of the forecaster is to predict almost as well as the best linear forecaster in the reference set $\left\{\boldsymbol{x} \in \mathbb{R}^d \mapsto \boldsymbol{u} \cdot \boldsymbol{x} \in \mathbb{R} : \boldsymbol{u} \in B_1(U)\right\}$, i.e., to minimize the regret on $B_1(U)$ defined by

$$\sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \min_{\boldsymbol{u}\in B_1(U)} \left\{\sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2\right\}.$$

We shall present algorithms along with bounds on their regret that hold uniformly over all sequences[1] $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T}$ such that $\|\boldsymbol{x}_t\|_\infty \leqslant X$ and $|y_t| \leqslant Y$ for all $t = 1, \ldots, T$, where $X, Y > 0$. These regret bounds depend on four important quantities: U, X, Y, and T, which may be known or unknown to the forecaster.

[1] Actually our results hold whether $(\boldsymbol{x}_t, y_t)_{t\geqslant 1}$ is generated by an oblivious environment or a non-oblivious opponent since we consider deterministic forecasters. See Section 2.3.1 in Chapter 2 for further details.

4.1.2 Contributions and related works

Next we detail the main contributions of this chapter in view of related works in online linear regression.

Our first contribution consists of a minimax analysis of online linear regression on ℓ^1-balls in the arbitrary sequence setting. We first provide a refined regret bound expressed in terms of Y, d, and a quantity $\kappa = \sqrt{T}UX/(2dY)$. This quantity κ is used to distinguish two regimes: we show a distinctive regime transition[2] at $\kappa = 1$ or $d = \sqrt{T}UX/(2Y)$. Namely, for $\kappa < 1$, the regret is of the order of $\sqrt{T}$, whereas it is of the order of $\ln T$ for $\kappa > 1$.

The derivation of this regret bound partially relies on a Maurey-type argument used under various forms with i.i.d. data, e.g., in [Nem00, Tsy03, BN08, SSSZ10] (see also [Yan04]). We adapt it in a straightforward way to the deterministic setting. Therefore, this is yet another technique that can be applied to both the stochastic and individual sequence settings.

Unsurprisingly, the refined regret bound mentioned above matches the optimal risk bounds for stochastic settings[3] [BM01a, Tsy03] (see also [RWY11]). Hence, linear regression is just as hard in the stochastic setting as in the arbitrary sequence setting. Using the standard online to batch conversion, we make the latter statement more precise by establishing a lower bound for all κ at least of the order of $\sqrt{\ln d}/d$. This lower bound extends those of [CB99, KW97], which only hold for small κ of the order of $1/d$.

The algorithm achieving our minimax regret bound is both computationally inefficient and non-adaptive (i.e., it requires prior knowledge of the quantities U, X, Y, and T that may be unknown in practice). Those two issues were first overcome by [ACBG02] via an automatic tuning termed *self-confident* (since the forecaster somehow trusts himself in tuning its parameters). They indeed proved that the self-confident p-norm algorithm with $p = 2\ln d$ and tuned with U has a cumulative loss $\widehat{L}_T = \sum_{t=1}^T (y_t - \widehat{y}_t)^2$ bounded by

$$\begin{aligned}\widehat{L}_T &\leqslant L_T^* + 8UX\sqrt{(e\ln d)\, L_T^*} + (32e\ln d)\, U^2X^2 \\ &\leqslant 8UXY\sqrt{eT\ln d} + (32e\ln d)\, U^2X^2\,,\end{aligned}$$

where $L_T^* \triangleq \min_{\{\boldsymbol{u}\in\mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U\}} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \leqslant TY^2$. This algorithm is efficient, and our lower bound in terms of κ shows that it is optimal up to logarithmic factors in the regime $\kappa \leqslant 1$ without prior knowledge of X, Y, and T.

In Section 4.3, we study the adaptivity possibilities of a closely related forecaster due to [KW97] and called the EG$^{\pm}$ algorithm. A detailed presentation of this forecaster can be found in Section 2.4.3 (Chapter 2). As proved in [KW97, Theorem 5.11], when tuned as a function of U, X, and a known upper bound B on L_T^*, this algorithm has a regret bounded from above by

[2] In high dimensions (i.e., when $d > \omega T$, for some absolute constant $\omega > 0$), we do not observe this transition (cf. Figure 4.1).

[3] For example, $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T}$ may be i.i.d. , or $\boldsymbol{x}_t$ can be deterministic and $y_t = f(\boldsymbol{x}_t) + \varepsilon_t$ for an unknown function f and an i.i.d. sequence $(\varepsilon_t)_{1\leqslant t\leqslant T}$ of Gaussian noise.

$2UX\sqrt{2B\ln(2d)} + 2U^2X^2\ln(2d)$. Again, this algorithm is efficient and nearly optimal in the regime $\kappa \leqslant 1$. However, the EG$^\pm$ algorithm requires prior knowledge of U, X, and B — or, alternatively, U, X, Y, and T.

Our second contribution — already detailed in Chapter 2 — is a generic version of the EG$^\pm$ algorithm for general convex loss functions. When applied to the square loss and combined with the variance-based tuning of [CBMS07], the corresponding *adaptive EG$^\pm$ algorithm* satisfies a regret bound comparable to that of the self-confident p-norm algorithms (Corollary 2.2 in Section 2.4.3). In particular this algorithm adapts automatically to X, Y, and T when U is known.

Our third contribution is a generic technique called *loss Lipschitzification*. It transforms the loss functions $\boldsymbol{u} \mapsto (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$ (or $\boldsymbol{u} \mapsto \left|y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\right|^\alpha$ if the predictions are scored with the α-loss, $\alpha \geqslant 2$) into Lipschitz continuous functions. We illustrate this technique by applying the generic adaptive EG$^\pm$ algorithm to the modified loss functions. When the predictions are scored with the square loss, this yields an algorithm (the LEG algorithm) whose main regret term can only improve on that derived for the adaptive EG$^\pm$ algorithm without Lipschtizification. The benefits of this technique are clearer for loss functions with higher curvature: if $\alpha > 2$, the resulting regret bound roughly grows as U instead of a naive $U^{\alpha/2}$.

Finally, we provide a simple way to achieve minimax regret uniformly over all ℓ^1-balls $B_1(U)$ for $U > 0$. This method aggregates instances of an algorithm that require prior knowledge of U. For the sake of simplicity, we assume that X, Y, and T are known, but explain in the discussions how to extend the method to a fully adaptive algorithm that requires the knowledge neither of U, X, Y, nor T.

This chapter is organized as follows. In Section 4.2, we establish our refined upper and lower bounds in terms of the intrinsic quantity κ. In Section 4.3, we present an efficient and adaptive algorithm — the adaptive EG$^\pm$ algorithm with or without loss Lipschitzfication — that achieves the optimal regret on $B_1(U)$ when U is known. In Section 4.4, we use an aggregating strategy to achieve an optimal regret uniformly over all ℓ^1-balls $B_1(U)$, for $U > 0$, when X, Y, and T are known. Finally, in Section 4.5, we discuss as an extension a fully automatic algorithm that requires no prior knowledge of U, X, Y, or T. Some proofs and additional tools are postponed to the appendix.

4.2 Optimal rates

In this section, we first present a refined upper bound on the minimax regret on $B_1(U)$ for an arbitrary $U > 0$. In Corollary 4.1, we express this upper bound in terms of an intrinsic quantity $\kappa \triangleq \sqrt{T}UX/(2dY)$. The optimality of the latter bound is shown in Section 4.2.2.

We first consider the following definition to avoid any ambiguity. We call *online forecaster* any sequence $F = (\widetilde{f}_t)_{t\geqslant 1}$ of functions such that $\widetilde{f}_t : \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R})^{t-1} \to \mathbb{R}$ maps at time t the new input $\boldsymbol{x}_t$ and the past data $(\boldsymbol{x}_s, y_s)_{1\leqslant s\leqslant t-1}$ to a prediction $\widetilde{f}_t\big(\boldsymbol{x}_t; (\boldsymbol{x}_s, y_s)_{1\leqslant s\leqslant t-1}\big)$. Depending on the context, the latter prediction may be simply denoted by $\widetilde{f}_t(\boldsymbol{x}_t)$ or by $\widehat{y}_t$.

4.2.1 Upper bound

Theorem 4.1 (Upper bound)**.** *Let $d, T \in \mathbb{N}^*$, and $U, X, Y > 0$. The minimax regret on $B_1(U)$ for bounded base predictions and observations satisfies*

$$\inf_F \sup_{\|\boldsymbol{x}_t\|_\infty \leqslant X,\, |y_t| \leqslant Y} \left\{ \sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \right\}$$
$$\leqslant \begin{cases} 3UXY\sqrt{2T\ln(2d)} & \text{if} \quad U < \frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}}\,, \\ 26\,UXY\sqrt{T\ln\left(1+\frac{2dY}{\sqrt{T}UX}\right)} & \text{if} \quad \frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}} \leqslant U \leqslant \frac{2dY}{\sqrt{T}X}\,, \\ 32\,dY^2\ln\!\left(1+\frac{\sqrt{T}UX}{dY}\right) + dY^2 & \text{if} \quad U > \frac{2dY}{X\sqrt{T}}\,, \end{cases}$$

where the infimum is taken over all forecasters F and where the supremum extends over all sequences $(\boldsymbol{x}_t, y_t)_{1 \leqslant t \leqslant T} \in (\mathbb{R}^d \times \mathbb{R})^T$ such that $|y_1|, \ldots, |y_T| \leqslant Y$ and $\|\boldsymbol{x}_1\|_\infty, \ldots, \|\boldsymbol{x}_T\|_\infty \leqslant X$.

Theorem 4.1 improves the bound of [KW97, Theorem 5.11] for the EG$^\pm$ algorithm. First, our bound depends logarithmically—as opposed to linearly—on U for $U > 2dY/(\sqrt{T}X)$. Secondly, it is smaller by a factor ranging from 1 to $\sqrt{\ln d}$ when

$$\frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}} \leqslant U \leqslant \frac{2dY}{\sqrt{T}X}\,. \tag{4.1}$$

Hence, Theorem 4.1 provides a partial answer to a question[4] raised in [KW97] about the gap of $\sqrt{\ln(2d)}$ between the upper and lower bounds.

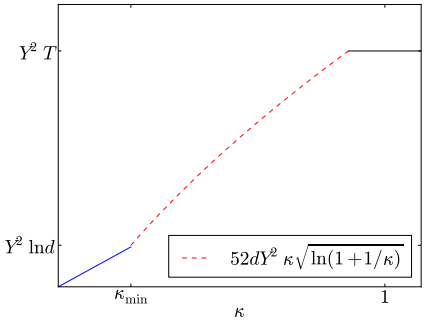
Before proving the theorem (see below), we state the following immediate corollary. It expresses the upper bound of Theorem 4.1 in terms of an intrinsic quantity $\kappa \triangleq \sqrt{T}UX/(2dY)$ that relates $\sqrt{T}UX/(2Y)$ to the ambient dimension d.

Corollary 4.1 (Upper bound in terms of an intrinsic quantity)**.** *Let $d, T \in \mathbb{N}^*$, and $U, X, Y > 0$. The upper bound of Theorem 4.1 expressed in terms of d, Y, and the intrinsic quantity $\kappa \triangleq \sqrt{T}UX/(2dY)$ reads:*

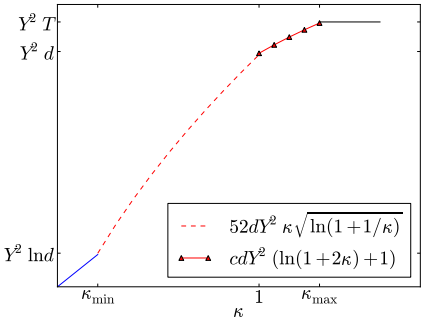
$$\inf_F \sup_{\|\boldsymbol{x}_t\|_\infty \leqslant X,\, |y_t| \leqslant Y} \left\{ \sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \right\}$$
$$\leqslant \begin{cases} 6\,dY^2\kappa\sqrt{2\ln(2d)} & \text{if} \quad \kappa < \frac{\sqrt{\ln(1+2d)}}{2d\sqrt{\ln 2}}\,, \\ 52\,dY^2\kappa\sqrt{\ln(1+1/\kappa)} & \text{if} \quad \frac{\sqrt{\ln(1+2d)}}{2d\sqrt{\ln 2}} \leqslant \kappa \leqslant 1\,, \\ 32\,dY^2\big(\ln(1+2\kappa)+1\big) & \text{if} \quad \kappa > 1\,. \end{cases}$$

The upper bound of Corollary 4.1 is shown in Figure 4.1. Observe that, in low dimension (Figure 4.1(b)), a clear transition from a regret of the order of $\sqrt{T}$ to one of $\ln T$ occurs at $\kappa = 1$. This transition is absent for high dimensions: for $d \geqslant \omega T$, where $\omega \triangleq \big(32(\ln(3)+1)\big)^{-1}$, the regret bound $32\,dY^2\big(\ln(1+2\kappa)+1\big)$ is worse than a trivial bound of TY^2 when $\kappa \geqslant 1$.

[4] The authors of [KW97] asked: "For large d there is a significant gap between the upper and lower bounds. We would like to know if it possible to improve the upper bounds by eliminating the $\ln d$ factors."



(a) High dimension $d \geqslant \omega T$.



(b) Low dimension $d < \omega T$.

Figure 4.1: The regret bound of Corollary 4.1 over $B_1(U)$ as a function of $\kappa = \sqrt{T}UX/(2dY)$. The constant c is chosen to ensure continuity at $\kappa = 1$, and $\omega \triangleq \big(32(\ln(3)+1)\big)^{-1}$. We define: $\kappa_{\min} = \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$ and $\kappa_{\max} = (e^{(T/d-1)/c}-1)/2$.

We now prove Theorem 4.1. The main part of the proof relies on a Maurey-type argument. Although this argument was used in the stochastic setting [Nem00, Tsy03, BN08, SSSZ10], we adapt it to the deterministic setting. This is yet another technique that can be applied to both the stochastic and individual sequence settings.

Proof (of Theorem 4.1): First note from Lemma 4.4 in Appendix 4.B that the minimax regret on $B_1(U)$ is upper bounded[5] by

$$\min\left\{3UXY\sqrt{2T\ln(2d)},\; 32\,dY^2\ln\left(1+\frac{\sqrt{T}UX}{dY}\right)+dY^2\right\}\,. \tag{4.2}$$

Therefore, the first case $U < \frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}}$ and the third case $U > \frac{dY}{X\sqrt{T}}$ are straightforward.

Therefore, we assume in the sequel that $\frac{Y}{X}\sqrt{\frac{\ln(1+2d)}{T\ln 2}} \leqslant U \leqslant \frac{2dY}{\sqrt{T}X}$.
We use a Maurey-type argument to refine the regret bound (4.2). This technique was used under various forms in the stochastic setting, e.g., in [Nem00, Tsy03, BN08, SSSZ10]. It consists of discretizing $B_1(U)$ and looking at a random point in this discretization to study its approximation properties. We also use clipping to get a regret bound growing as U instead of a naive U^2.

More precisely, we first use the fact that to be competitive against $B_1(U)$, it is sufficient to be competitive against its finite subset

$$\widetilde{B}_{U,m} \triangleq \left\{\left(\frac{k_1U}{m},\dots,\frac{k_dU}{m}\right) : (k_1,\dots,k_d)\in\mathbb{Z}^d, \sum_{j=1}^d |k_j| \leqslant m\right\} \subset B_1(U)\,,$$

[5] As proved in Lemma 4.4, the regret bound (4.2) is achieved either by the EG$^{\pm}$ algorithm, the algorithm SeqSEW$^{B,\eta}_{\tau}$ of Chapter 3 (we could also get a slightly worse bound with the sequential ridge regression forecaster), or the trivial null forecaster.

where $m \triangleq \lfloor \alpha \rfloor$ with $\alpha \triangleq \dfrac{UX}{Y}\sqrt{T(\ln 2)/\ln\left(1+\dfrac{2dY}{\sqrt{T}UX}\right)}$.

By Lemma 4.6 in appendix, and since $m > 0$ (see below), we indeed have

$$\begin{aligned}
&\inf_{\boldsymbol{u}\in\widetilde{B}_{U,m}} \sum_{t=1}^{T}(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \\
&\leqslant \inf_{\boldsymbol{u}\in B_1(U)} \sum_{t=1}^{T}(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 + \frac{TU^2X^2}{m} \\
&\leqslant \inf_{\boldsymbol{u}\in B_1(U)} \sum_{t=1}^{T}(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 + \frac{2}{\sqrt{\ln 2}}UXY\sqrt{T\ln\left(1+\frac{2dY}{\sqrt{T}UX}\right)}\,, \qquad (4.3)
\end{aligned}$$

where (4.3) follows from $m \triangleq \lfloor \alpha \rfloor \geqslant \alpha/2$ since $\alpha \geqslant 1$ (in particular, $m > 0$ as stated above).

To see why $\alpha \geqslant 1$, note that it suffices to show that $x\sqrt{\ln(1+x)} \leqslant 2d\sqrt{\ln 2}$ where we set $x \triangleq 2dY/(\sqrt{T}UX)$. But from the assumption $U \geqslant (Y/X)\sqrt{\ln(1+2d)/(T\ln 2)}$, we have $x \leqslant 2d\sqrt{\ln(2)/\ln(1+2d)} \triangleq y$, so that, by monotonicity, $x\sqrt{\ln(1+x)} \leqslant y\sqrt{\ln(1+y)} \leqslant y\sqrt{\ln(1+2d)} = 2d\sqrt{\ln 2}$.

Therefore it only remains to exhibit an algorithm which is competitive against $\widetilde{B}_{U,m}$ at an aggregation price of the same order as the last term in (4.3). This is the case for the standard exponentially weighted average forecaster applied to the clipped predictions

$$\big[\boldsymbol{u}\cdot\boldsymbol{x}_t\big]_Y \triangleq \min\Big\{Y, \max\big\{-Y, \boldsymbol{u}\cdot\boldsymbol{x}_t\big\}\Big\}\,, \quad \boldsymbol{u}\in\widetilde{B}_{U,m}\,,$$

and tuned with the inverse temperature parameter $\eta = 1/(8Y^2)$. More formally, this algorithm predicts at each time $t = 1, \dots, T$ as

$$\widehat{y}_t \triangleq \sum_{\boldsymbol{u}\in\widetilde{B}_{U,m}} p_t(\boldsymbol{u})\big[\boldsymbol{u}\cdot\boldsymbol{x}_t\big]_Y\,,$$

where $p_1(\boldsymbol{u}) \triangleq 1/\big|\widetilde{B}_{U,m}\big|$ (denoting by $\big|\widetilde{B}_{U,m}\big|$ the cardinality of the set $\widetilde{B}_{U,m}$), and where the weights $p_t(\boldsymbol{u})$ are defined for all $t = 2, \dots, T$ and $\boldsymbol{u}\in\widetilde{B}_{U,m}$ by

$$p_t(\boldsymbol{u}) \triangleq \frac{\exp\left(-\eta\sum_{s=1}^{t-1}\big(y_s - [\boldsymbol{u}\cdot\boldsymbol{x}_s]_Y\big)^2\right)}{\sum_{\boldsymbol{v}\in\widetilde{B}_{U,m}}\exp\left(-\eta\sum_{s=1}^{t-1}\big(y_s - [\boldsymbol{v}\cdot\boldsymbol{x}_s]_Y\big)^2\right)}\,.$$

By Lemma 4.5 in appendix, the above forecaster tuned with $\eta = 1/(8Y^2)$ satisfies

$$\begin{aligned}
&\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 - \inf_{\boldsymbol{u}\in\widetilde{B}_{U,m}}\sum_{t=1}^{T}(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \leqslant 8Y^2\ln\big|\widetilde{B}_{U,m}\big| \\
&\quad\leqslant 8Y^2\ln\left(\frac{\mathrm{e}(2d+m)}{m}\right)^m \qquad (4.4)\\
&\quad= 8Y^2 m\big(1+\ln(1+2d/m)\big) \leqslant 8Y^2\alpha\big(1+\ln(1+2d/\alpha)\big) \qquad (4.5)
\end{aligned}$$

$$= 8Y^2\alpha + 8Y^2\alpha \ln\left(1 + \frac{2dY}{\sqrt{T}UX}\sqrt{\frac{\ln\big(1+2dY/(\sqrt{T}UX)\big)}{\ln 2}}\right)$$

$$\leqslant 8Y^2\alpha + 16Y^2\alpha \ln\left(1 + \frac{2dY}{\sqrt{T}UX}\right) \tag{4.6}$$

$$\leqslant \left(\frac{8}{\sqrt{\ln 2}} + 16\sqrt{\ln 2}\right) UXY\sqrt{T\ln\left(1 + \frac{2dY}{\sqrt{T}UX}\right)}\,. \tag{4.7}$$

To get (4.4) we used Lemma 4.7 in appendix. Inequality (4.5) follows by definition of $m \leqslant \alpha$ and the fact that $x \mapsto x\big(1 + \ln(1 + A/x)\big)$ is nondecreasing on $\mathbb{R}_+^*$ for all $A > 0$. Inequality (4.6) follows from the assumption $U \leqslant 2dY/(\sqrt{T}X)$ and the elementary inequality $\ln\big(1 + x\sqrt{\ln(1+x)/\ln 2}\big) \leqslant 2\ln(1+x)$ which holds for all $x \geqslant 1$ and was used, e.g., at the end of [BN08, Theorem 2-a)]. Finally, elementary manipulations combined with the assumption that $2dY/(\sqrt{T}UX) \geqslant 1$ lead to (4.7).

Putting Equations (4.3) and (4.7) together, the previous algorithm has a regret on $B_1(U)$ which is bounded from above by

$$\left(\frac{10}{\sqrt{\ln 2}} + 16\sqrt{\ln 2}\right) UXY\sqrt{T\ln\left(1 + \frac{2dY}{\sqrt{T}UX}\right)}\,,$$

which concludes the proof since $10/\sqrt{\ln 2} + 16\sqrt{\ln 2} \leqslant 26$. □

4.2.2 Lower bound

Corollary 4.1 gives an upper bound on the regret in terms of the quantities d, Y, and $\kappa \triangleq \sqrt{T}UX/(2dY)$. We now show that for all $d \in \mathbb{N}^*$, $Y > 0$, and $\kappa \geqslant \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$, the upper bound can not be improved[6] up to logarithmic factors.

Theorem 4.2 (Lower bound)**.** *For all $d \in \mathbb{N}^*$, $Y > 0$, and $\kappa \geqslant \frac{\sqrt{\ln(1+2d)}}{2d\sqrt{\ln 2}}$, there exist $T \geqslant 1$, $U > 0$, and $X > 0$ such that $\sqrt{T}UX/(2dY) = \kappa$ and*

$$\inf_F \sup_{\|\boldsymbol{x}_t\|_\infty \leqslant X,\, |y_t| \leqslant Y} \left\{\sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2\right\}$$
$$\geqslant \begin{cases} \frac{c_1}{\ln\left(2+16d^2\right)} dY^2\kappa\sqrt{\ln\left(1 + 1/\kappa\right)} & \text{if} \quad \frac{\sqrt{\ln(1+2d)}}{2d\sqrt{\ln 2}} \leqslant \kappa \leqslant 1\,, \\ \frac{c_2}{\ln\left(2+16d^2\right)} dY^2 & \text{if} \quad \kappa > 1\,, \end{cases}$$

where $c_1, c_2 > 0$ are absolute constants. The infimum is taken over all forecasters F and the supremum extends over all sequences $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T} \in (\mathbb{R}^d \times \mathbb{R})^T$ such that $|y_1|, \ldots, |y_T| \leqslant Y$ and $\|\boldsymbol{x}_1\|_\infty, \ldots, \|\boldsymbol{x}_T\|_\infty \leqslant X$.

[6] For T sufficiently large, we may overlook the case $\kappa < \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$ or $\sqrt{T} < (Y/(UX))\sqrt{\ln(1+2d)/\ln 2}$. Observe that in this case, the minimax regret is already of the order of $Y^2\ln(1+d)$ (cf. Figure 4.1).

The above lower bound extends those of [CB99, KW97], which hold for small κ of the order of $1/d$. The proof is postponed to Appendix 4.A.1. We perform a reduction to the stochastic batch setting—via the standard online to batch conversion, and employ a version of a lower bound of [Tsy03].

4.3 Adaptation to unknown X, Y and T via exponential weights

Although the proof of Theorem 4.1 already gives an algorithm that achieves the minimax regret, the latter takes as inputs U, X, Y, and T, and it is inefficient in high dimensions. In this section, we present a new method that achieves the minimax regret both efficiently and without prior knowledge of X, Y, and T provided that U is known. Adaptation to an unknown U is considered in Section 4.4. Our method consists of modifying an underlying linear regression algorithm such as the EG$^{\pm}$ algorithm [KW97] or the sequential ridge regression [Vov01, AW01]. Next, we show that automatically tuned variants of the EG$^{\pm}$ algorithm – the first of which was introduced in Section 2.4.3 — nearly achieve the minimax regret for the regime $d \geqslant \sqrt{T}UX/(2Y)$. A similar modification could be applied to the ridge regression forecaster to achieve a nearly optimal regret bound of order $dY^2 \ln\Big(1 + d\Big(\frac{\sqrt{T}UX}{dY}\Big)^2\Big)$ in the regime $d < \sqrt{T}UX/(2Y)$. The latter analysis is more technical and hence is omitted.

4.3.1 An adaptive EG$^{\pm}$ algorithm

The second algorithm of the proof of Theorem 4.1 is computationally inefficient because it aggregates approximately $d^{\sqrt{T}}$ experts. In contrast, the EG$^{\pm}$ algorithm has a manageable computational complexity that is linear in d. In Section 2.4.3 of Chapter 2 we introduced a version of the EG$^{\pm}$ algorithm — called the adaptive EG$^{\pm}$ algorithm — that does not require prior knowledge of X, Y and T (as opposed to the original EG$^{\pm}$ algorithm of [KW97]). This version uses the automatic tuning of [CBMS07]. As proved in Corollary 2.2 of Chapter 2 — a consequence of [CBMS07, Corollary 1] — the adaptive EG$^{\pm}$ algorithm on $B_1(U)$ defined in Figure 2.5 with $\ell_t(\boldsymbol{u}) = (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$ satisfies, for all choices of $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_T, y_T) \in \mathbb{R}^d \times \mathbb{R}$,

$$\sum_{t=1}^{T} (y_t - \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t)^2 \leqslant L_T^* + 8UX\sqrt{L_T^* \ln(2d)} + \big(137\ln(2d) + 24\big)\big(UXY + U^2X^2\big)\,, \quad (4.8)$$

where the quantities $L_T^* \triangleq \min_{\{\boldsymbol{u} \in B_1(U)\}} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$, $X \triangleq \max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty$, and $Y \triangleq \max_{1 \leqslant t \leqslant T} |y_t|$ are unknown to the forecaster.

The above regret bound is an *improvement for small losses* (cf. (2.15) in Section 2.2.2). By the elementary inequality $L_T^* \leqslant TY^2$ (since $\mathbf{0} \in B_1(U)$), it yields the zero-order regret bound

$$\sum_{t=1}^{T} (y_t - \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t)^2 - \min_{\boldsymbol{u} \in B_1(U)} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \\ \leqslant 8UXY\sqrt{T\ln(2d)} + \big(137\ln(2d) + 24\big)\big(UXY + U^2X^2\big)\,.$$

Therefore, our version of the EG$^{\pm}$ algorithm is efficient and adaptive in X, Y, and T. It achieves approximately the regret bound of Theorem 4.1 in the regime $\kappa \leqslant 1$, i.e., $d \geqslant \sqrt{T}UX/(2Y)$.

Another way to perform the adaptation to X, Y, and T in an efficient way is provided by the self-confident p-norm algorithm of [ACBG02] with $p = 2\ln d$. As commented on after the statement of Corollary 2.2 in Chapter 2, this algorithm satisfies an improvement for small losses similar to (4.8). The fact that we got a similar bound is not surprising because the p-norm algorithms are known to share many properties with the EG$^{\pm}$ algorithm (in the limit $p \to +\infty$ with an appropriate initial weight vector, or for p of the order of $\ln d$ with a zero initial weight vector, cf. [Gen03]). The bound of Corollary 2.2 corroborates this similarity.

In the next subsections, we use yet another instance of the adaptive EG$^{\pm}$ algorithm that we call the Lipschitzifying Exponentiated Gradient (LEG) algorithm. It corresponds to the adaptive EG$^{\pm}$ algorithm applied not to the square loss but to a Lipschitz continuous modification $\widetilde{\ell}_t : \mathbb{R}^d \to \mathbb{R}$ of the square loss.

4.3.2 Lipschitzification of the loss function

Our key technique consists of transforming the loss functions $\boldsymbol{u} \mapsto (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$ into functions $\widetilde{\ell}_t$ that are Lipschitz continuous with respect to $\|\cdot\|_1$. Afterward, adaptation to the unknown Lipschitz constants $\|\nabla \widetilde{\ell}_t\|_\infty$ is carried out using the techniques of [CBMS07].

We point out that our Lipschitzification method can be applied to other convex loss functions with higher curvature, see Remark 4.1 later. Moreover, this technique is not specific to the pair of dual norms $(\|\cdot\|_1, \|\cdot\|_\infty)$ and to the EG$^{\pm}$ algorithm; it could be used with other pairs $(\|\cdot\|_q, \|\cdot\|_p)$ (with $1/p + 1/q = 1$) and other gradient-based algorithms, such as the p-norm algorithm [Gen03, ACBG02] and its regularized variants (SMIDAS and COMID) [SST09, DSSST10].

The Lipschitzification proceeds as follows. At each time $t \geqslant 1$, using adaptivity-oriented ideas from Chapter 3, we set

$$B_t \triangleq \left(2^{\lceil \log_2(\max_{1 \leqslant s \leqslant t-1} y_s^2) \rceil}\right)^{1/2} ,$$

so that B_t satisfies $|y_s| \leqslant B_t$ for all $s = 1, \dots, t-1$. The modified (or *Lipschitzified*) loss function $\widetilde{\ell}_t : \mathbb{R}^d \to \mathbb{R}$ is constructed as follows:

- if $|y_t| > B_t$, then
$$\widetilde{\ell}_t(\boldsymbol{u}) \triangleq 0 \quad \text{for all } \boldsymbol{u} \in \mathbb{R}^d ;$$
- if $|y_t| \leqslant B_t$, then $\widetilde{\ell}_t$ is the convex function that coincides with the square loss when $\left|\boldsymbol{u} \cdot \boldsymbol{x}_t\right| \leqslant B_t$ and is linear elsewhere. This function is shown in Figure 4.2 and can be formally defined as

$$\widetilde{\ell}_t(\boldsymbol{u}) \triangleq \begin{cases} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 & \text{if } \left|\boldsymbol{u} \cdot \boldsymbol{x}_t\right| \leqslant B_t, \\ (y_t - B_t)^2 + 2(B_t - y_t)(\boldsymbol{u} \cdot \boldsymbol{x}_t - B_t) & \text{if } \boldsymbol{u} \cdot \boldsymbol{x}_t > B_t, \\ (y_t + B_t)^2 + 2(-B_t - y_t)(\boldsymbol{u} \cdot \boldsymbol{x}_t + B_t) & \text{if } \boldsymbol{u} \cdot \boldsymbol{x}_t < -B_t. \end{cases}$$

Observe that in both cases $|y_t| > B_t$ and $|y_t| \leqslant B_t$, the function $\widetilde{\ell}_t$ is continuously differen-

tiable. Moreover, if $|y_t| \leqslant B_t$, then

$$\forall \boldsymbol{u} \in \mathbb{R}^d\,, \quad \nabla \widetilde{\ell}_t(\boldsymbol{u}) = -2\big(y_t - [\boldsymbol{u} \cdot \boldsymbol{x}_t]_{B_t}\big)\, \boldsymbol{x}_t\,, \tag{4.9}$$

where we define the clipping operator $[\cdot]_B$ by $[x]_B \triangleq \min\big\{B, \max\{-B, x\}\big\}$ for all $x \in \mathbb{R}$ and all $B > 0$.

Therefore, in both cases $|y_t| > B_t$ and $|y_t| \leqslant B_t$, the function $\widetilde{\ell}_t$ is Lipschitz continuous with respect to $\|\cdot\|_1$ with Lipschitz constant

$$\left\|\nabla \widetilde{\ell}_t\right\|_\infty \leqslant 2\left|y_t - [\boldsymbol{u} \cdot \boldsymbol{x}_t]_{B_t}\right| \|\boldsymbol{x}_t\|_\infty \tag{4.10}$$

$$\leqslant 2\big(|y_t| + B_t\big)\, \|\boldsymbol{x}_t\|_\infty \leqslant 2\big(1 + \sqrt{2}\big)\, \|\boldsymbol{x}_t\|_\infty \max_{1 \leqslant s \leqslant t} |y_s|\,, \tag{4.11}$$

where we used the fact that $B_t \leqslant \sqrt{2} \max_{1 \leqslant s \leqslant t-1} |y_s|$. We can also glean from Figure 4.2 that, when $|y_t| \leqslant B_t$, the modified loss function $\widetilde{\ell}_t : \mathbb{R}^d \to \mathbb{R}$ lies in between the square loss and its clipped version:

$$\forall \boldsymbol{u} \in \mathbb{R}^d, \quad \big(y_t - [\boldsymbol{u} \cdot \boldsymbol{x}_t]_{B_t}\big)^2 \leqslant \widetilde{\ell}_t(\boldsymbol{u}) \leqslant \big(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t\big)^2\,. \tag{4.12}$$

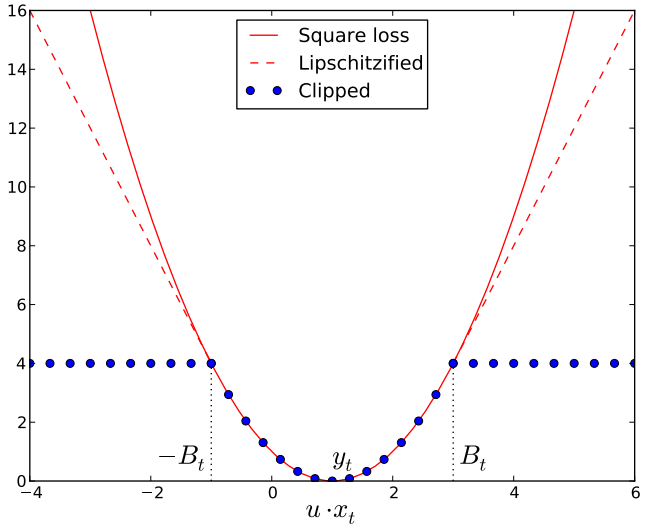


Figure 4.2: Example when $|y_t| \leqslant B_t$. The square loss $(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$, its clipped version $\big(y_t - [\boldsymbol{u} \cdot \boldsymbol{x}_t]_{B_t}\big)^2$ and its Lipschitzified version $\widetilde{\ell}_t(\boldsymbol{u})$ are plotted as a function of $\boldsymbol{u} \cdot \boldsymbol{x}_t$.

4.3.3 Lipschitzifying Exponentiated Gradient algorithm

In this section we illustrate the Lipschitzification technique described above with the adaptive EG$^\pm$ algorithm. We denote by $(\mathbf{e}_j)_{1 \leqslant j \leqslant d}$ the canonical basis of $\mathbb{R}^d$ and by ∇_j the j-th component of the gradient.

Parameter: radius $U > 0$.

Initialization: $B_1 \triangleq 0$, $\boldsymbol{p}_1 = (p^+_{1,1}, p^-_{1,1}, \dots, p^+_{d,1}, p^-_{d,1}) \triangleq \big(1/(2d), \dots, 1/(2d)\big) \in \mathbb{R}^{2d}$.

At each time round $t \geqslant 1$,

1. Compute the linear combination $\widehat{\boldsymbol{u}}_t \triangleq U \sum_{j=1}^{d} \big(p^+_{j,t} - p^-_{j,t}\big)\, \boldsymbol{e}_j \in B_1(U)$;
2. Get $\boldsymbol{x}_t \in \mathbb{R}^d$ and output the clipped prediction $\widehat{y}_t \triangleq \big[\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t\big]_{B_t}$;
3. Get $y_t \in \mathbb{R}$ and define the modified loss function $\widetilde{\ell}_t : \mathbb{R}^d \to \mathbb{R}$ as in Section 4.3.2;
4. Update the parameter η_{t+1} according to (4.13);
5. Update the weight vector $\boldsymbol{p}_{t+1} = (p^+_{1,t+1}, p^-_{1,t+1}, \dots, p^+_{d,t+1}, p^-_{d,t+1}) \in \mathcal{X}_{2d}$ defined for all $j = 1, \dots, d$ and $\gamma \in \{+, -\}$ by[a]

$$p^\gamma_{j,t+1} \triangleq \frac{\exp\left(-\eta_{t+1} \sum_{s=1}^{t} \gamma U \nabla_j \widetilde{\ell}_s(\widehat{\boldsymbol{u}}_s)\right)}{\displaystyle\sum_{\substack{1 \leqslant k \leqslant K \\ \mu \in \{+,-\}}} \exp\left(-\eta_{t+1} \sum_{s=1}^{t} \mu U \nabla_k \widetilde{\ell}_s(\widehat{\boldsymbol{u}}_s)\right)} \,.$$

6. Update the threshold $\quad B_{t+1} \triangleq \Big(2^{\lceil \log_2(\max_{1 \leqslant s \leqslant t} y_s^2) \rceil}\Big)^{1/2}$.

[a] For all $\gamma \in \{+, -\}$, by a slight abuse of notation, γU denotes U or $-U$ if $\gamma = +$ or $\gamma = -$ respectively.

Figure 4.3: The Lipschitzifying Exponentiated Gradient (LEG) algorithm.

Consider the Lipschitzifying Exponentiated Gradient (LEG) algorithm of Figure 4.3. It is yet another instance of the adaptive EG$^\pm$ algorithm on $B_1(U)$ (cf. Figure 2.5 of Section 2.4.3) applied not to the square loss but to the Lipschitzified loss functions $\widetilde{\ell}_t$, $t \geqslant 1$. In particular the LEG algorithm uses as a blackbox the exponentially weighted majority forecaster of [CBMS07] on $2d$ experts—namely, the vertices $\pm U \boldsymbol{e}_j$ of $B_1(U)$—as in [KW97]. It adapts to the unknown Lipschitz constants $\|\nabla \widetilde{\ell}_t\|_\infty$ by the particular choice of η_t due to [CBMS07] and defined for all $t \geqslant 2$ by

$$\eta_t = \min\left\{ \frac{1}{\widehat{E}_{t-1}}, C\sqrt{\frac{\ln K}{V_{t-1}}} \right\} , \tag{4.13}$$

where $C \triangleq \sqrt{2(\sqrt{2}-1)/(\mathrm{e}-2)}$ and where we set, for all $t = 1, \dots, T$,

$$z^+_{j,s} \triangleq U \nabla_j \widetilde{\ell}_s(\widehat{\boldsymbol{u}}_s) \quad \text{and} \quad z^-_{j,s} \triangleq -U \nabla_j \widetilde{\ell}_s(\widehat{\boldsymbol{u}}_s)\,, \quad j = 1, \dots, d, \quad s = 1, \dots, t\,,$$

$$\widehat{E}_t \triangleq \inf_{k \in \mathbb{Z}} \left\{ 2^k : 2^k \geqslant \max_{1 \leqslant s \leqslant t} \max_{\substack{1 \leqslant j,k \leqslant d \\ \gamma,\mu \in \{+,-\}}} \big| z^\gamma_{j,s} - z^\mu_{k,s} \big| \right\} ,$$

$$V_t \triangleq \sum_{s=1}^{t} \sum_{\substack{1 \leqslant j \leqslant d \\ \gamma \in \{+,-\}}} p_{j,s}^{\gamma} \left(z_{j,s}^{\gamma} - \sum_{\substack{1 \leqslant k \leqslant d \\ \mu \in \{+,-\}}} p_{k,s}^{\mu} z_{k,s}^{\mu} \right)^2 .$$

Note that $\widehat{E}_{t-1}$ approximates the range of the $z_{j,s}^{\gamma}$ up to time $t-1$, while V_{t-1} is the corresponding cumulative variance of the forecaster.

The next theorem bounds the regret of the LEG algorithm on $B_1(U)$. As with the square loss, this algorithm is efficient and adaptive in X, Y, and T; it achieves approximately the regret bound of Theorem 4.1 in the regime $\kappa \leqslant 1$, i.e., $d \geqslant \sqrt{T}UX/(2Y)$. The proof is postponed to Appendix 4.A.2. It follows from the bound on the adaptive EG$^{\pm}$ algorithm for general convex loss functions that we proved in Corollary 2.1 (Section 2.4.3).

Theorem 4.3. *Let $U > 0$. Then, the Lipschitzifying Exponentiated Gradient algorithm tuned with U satisfies, for all $T \geqslant 1$ and all individual sequences $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T) \in \mathbb{R}^d \times \mathbb{R}$,*

$$\sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 \leqslant \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \widetilde{\ell}_t(\boldsymbol{u}) + 8UX \sqrt{\left(\inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \widetilde{\ell}_t(\boldsymbol{u}) \right) \ln(2d)} + \left(153 \ln(2d) + 58\right) \left(UXY + U^2X^2\right) + 12Y^2 ,$$

where none of the three quantities $\inf_{\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U\}} \sum_{t=1}^{T} \widetilde{\ell}_t(\boldsymbol{u})$, $X \triangleq \max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_{\infty}$, and $Y \triangleq \max_{1 \leqslant t \leqslant T} |y_t|$ is known to the forecaster.

The first two terms of the bound of Theorem 4.3 slightly improve on those obtained without Lipschitzification (cf. (4.8)) since we always have

$$\inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \widetilde{\ell}_t(\boldsymbol{u}) \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 , \tag{4.14}$$

where we used the key property $\widetilde{\ell}_t(\boldsymbol{u}) \leqslant (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2$ for all $\boldsymbol{u} \in \mathbb{R}^d$ and all $t = 1, \ldots, T$ (by (4.12) if $|y_t| \leqslant B_t$, obvious otherwise). Though the improvement in the regret bound entailed by (4.14) is usually only of minor importance, Lipschitzification can be useful in at least two ways.

Remark 4.1 (Application to other convex loss functions with higher curvature)**.**
Lipschitzification can be used in a much more general setting than the one studied in this paper, i.e., when the loss functions are of the form $x \mapsto f\left(|y_t - x|\right)$ for an increasing twice differentiable function $f : \mathbb{R}_+ \to \mathbb{R}$ such that $f'' \geqslant 0$. Assume, e.g., that $f(x) = x^{\alpha}$ for all $x \geqslant 0$ and some $\alpha \geqslant 2$. As explained below, the benefits of Lipschitzification become clear when $\alpha > 2$: it yields a regret bound that depends linearly in U, instead of the rate $U^{\alpha/2}$ that would follow from a similar analysis for the adaptive EG$^{\pm}$ algorithm without loss Lipschitzification.

Next we assume that for some $\alpha \geqslant 2$, the predictions $\widehat{y}_t$ of the forecaster and the base predictions $\boldsymbol{u} \cdot \boldsymbol{x}_t$ are scored with the loss functions $x \mapsto |y_t - x|^{\alpha}$, $t \geqslant 1$. Correspondingly, we set

$\ell_t(\boldsymbol{u}) = |y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t|^\alpha$ for all $\boldsymbol{u}\in\mathbb{R}^d$. The Lipschtizification step of Section 4.3.2 can easily be extended to this case. The main two changes consist of the following:

- The adaptive clipping level is defined by $B_t \triangleq \left(2^{\lceil \log_2(\max_{1\leqslant s\leqslant t-1}|y_s|^\alpha)\rceil}\right)^{1/\alpha}$.
- At every round t such that $|y_t|\leqslant B_t$, the Lipschitzified loss function $\widetilde{\ell}_t:\mathbb{R}^d\to\mathbb{R}$ is defined by

$$\widetilde{\ell}_t(\boldsymbol{u}) \triangleq \begin{cases} \left|y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\right|^\alpha & \text{if } \left|\boldsymbol{u}\cdot\boldsymbol{x}_t\right|\leqslant B_t,\\ \left|y_t - B_t\right|^\alpha + \alpha\left|y_t - B_t\right|^{\alpha-1}(\boldsymbol{u}\cdot\boldsymbol{x}_t - B_t) & \text{if } \boldsymbol{u}\cdot\boldsymbol{x}_t > B_t,\\ \left|y_t + B_t\right|^\alpha - \alpha\left|y_t + B_t\right|^{\alpha-1}(\boldsymbol{u}\cdot\boldsymbol{x}_t + B_t) & \text{if } \boldsymbol{u}\cdot\boldsymbol{x}_t < -B_t. \end{cases}$$

Consider the adaptive EG$^\pm$ algorithm of Section 2.4.3 applied to the Lipschitzified loss functions $\widetilde{\ell}_t$. To analyse its performance, it suffices to follow the same lines as in the proof of Theorem 4.3. Again, a key property is that[7] $\nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) = -\alpha\,\mathrm{sgn}\big(y_t - [\widehat{\boldsymbol{u}}_t\cdot\boldsymbol{x}_t]_{B_t}\big)\left|y_t - [\widehat{\boldsymbol{u}}_t\cdot\boldsymbol{x}_t]_{B_t}\right|^{\alpha-1}\boldsymbol{x}_t = -\alpha\,\mathrm{sgn}\big(y_t-\widehat{y}_t\big)\left|y_t-\widehat{y}_t\right|^{\alpha-1}\boldsymbol{x}_t$. This entails that

$$\begin{aligned}\left\|\nabla\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)\right\|_\infty^2 &\leqslant \alpha^2X^2\left|y_t-\widehat{y}_t\right|^{2\alpha-2} = \alpha^2X^2\left|y_t-\widehat{y}_t\right|^{\alpha-2}\left|y_t-\widehat{y}_t\right|^\alpha\\ &\leqslant \alpha^2X^2(B_{t+1}+B_t)^{\alpha-2}\left|y_t-\widehat{y}_t\right|^\alpha \leqslant \alpha^2X^2\big((1+2^{1/\alpha})Y\big)^{\alpha-2}\left|y_t-\widehat{y}_t\right|^\alpha.\end{aligned}$$

Then, following the same lines as in the proof of Theorem 4.3, we can see that the adaptive EG$^\pm$ algorithm applied to $(\widetilde{\ell}_t)_{t\geqslant1}$ has a cumulative loss $\sum_{t=1}^T\left|y_t-\widehat{y}_t\right|^\alpha$ at most of

$$\begin{aligned}&\widetilde{L}_T^* + c_1(\alpha)UXY^{\frac{\alpha}{2}-1}\sqrt{\widetilde{L}_T^*\ln(2d)}\\ &\quad + c_2(\alpha)\left(UXY^{\alpha-1} + U^2X^2Y^{\alpha-2}\right)\ln(2d)\,,\end{aligned}$$

where $\widetilde{L}_T^* \triangleq \inf_{\{\boldsymbol{u}\in\mathbb{R}^d:\|\boldsymbol{u}\|_1\leqslant U\}}\sum_{t=1}^T\widetilde{\ell}_t(\boldsymbol{u}) \leqslant TY^\alpha$ and where $c_1(\alpha), c_2(\alpha) > 0$ are constants depending only on α (e.g., $c_1(\alpha) = 4\alpha\big(1+2^{1/\alpha}\big)^{\alpha/2-1}$). This bound improves on the bound we would have obtained via the same analysis for the adaptive EG$^\pm$ algorithm applied to the original losses $\ell_t(\boldsymbol{u}) = |y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t|^\alpha$:

$$\begin{aligned}&L_T^* + c_3(\alpha)UX(Y+UX)^{\frac{\alpha}{2}-1}\sqrt{L_T^*\ln(2d)}\\ &\quad + c_4(\alpha)\left(\alpha UX(Y+UX)^{\alpha-1} + \alpha^2U^2X^2(Y+UX)^{\alpha-2}\right)\ln(2d)\,,\end{aligned}$$

where we set $L_T^* \triangleq \inf_{\{\boldsymbol{u}\in\mathbb{R}^d:\|\boldsymbol{u}\|_1\leqslant U\}}\sum_{t=1}^T|y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t|^\alpha$, and where $c_3(\alpha), c_4(\alpha) > 0$ are constants depending only on α. The main difference between the two regret bounds above lies in the dependence in U: our main regret term scales as $UXY^{\alpha/2-1}$ while the one obtained without Lipschitzification scales as $UX(Y+UX)^{\alpha/2-1}$. The first term grows linearly in U while the second grows as $U^{\alpha/2}$, hence a clear improvement for $\alpha > 2$.

Remark 4.2 (A simpler analysis for the minimax regret)**.**
Another benefit of Lipschitzification is that all online convex optimization regret bounds expressed in terms of the maximal dual norm of the gradients — i.e., $\max_{1\leqslant t\leqslant T}\|\nabla\widetilde{\ell}_t\|_\infty$ *in our case — can*

[7] For all $x\in\mathbb{R}$, $\mathrm{sgn}(x)$ equals 1 (resp. -1, 0) if $x>0$ (resp. $x<0$, $x=0$).

be used fruitfully with the Lipschitzified loss functions $\widetilde{\ell}_t$. For instance, using the very simple bound (2.13) of Theorem 2.4 (a consequence of Corollary [CBMS07, Corollary 1], see Section 2.2.2), we can prove that

$$\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2-\inf_{\|\boldsymbol{u}\|_1\leqslant U}\sum_{t=1}^{T}(y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t)^2\leqslant c_1UXY\left(\sqrt{T\ln(2d)}+8\ln(2d)\right)+c_2Y^2\,,$$

where $c_1\triangleq 8\big(\sqrt{2}+1\big)$ and $c_2\triangleq 4\left(1+1/\sqrt{2}\right)^2$. The bound is no longer an improvement for small losses, but the analysis is even more straightforward (no need to solve a quadratic inequality); see below.

Proof (of Remark 4.2): By the key property (4.12) that holds for all rounds t such that $|y_t|\leqslant B_t$ (the other rounds accounting only for an additional total loss at most of c_2Y^2, see (4.48) in Appendix 4.A.2), we get that

$$\begin{aligned}
&\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2-\inf_{\|\boldsymbol{u}\|_1\leqslant U}\sum_{t=1}^{T}(y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t)^2\\
&\qquad\leqslant\sum_{t=1}^{T}\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)-\inf_{\|\boldsymbol{u}\|_1\leqslant U}\sum_{t=1}^{T}\widetilde{\ell}_t(\boldsymbol{u})+c_2Y^2\\
&\qquad\leqslant 4U\max_{1\leqslant t\leqslant T}\left\|\nabla\widetilde{\ell}_t\right\|_\infty\left(\sqrt{T\ln(2d)}+2\ln(2d)+3\right)+c_2Y^2 \qquad (4.15)\\
&\qquad\leqslant c_1UXY\left(\sqrt{T\ln(2d)}+8\ln(2d)\right)+c_2Y^2\,, \qquad (4.16)
\end{aligned}$$

where (4.15) follows from the last bound of Corollary 2.1 (i.e., with a uniform scaling factor $\max_{1\leqslant t\leqslant T}\|\nabla\widetilde{\ell}_t\|_\infty$), and where (4.16) follows from $\max_{1\leqslant t\leqslant T}\|\nabla\widetilde{\ell}_t\|_\infty\leqslant 2\big(1+\sqrt{2}\big)XY$ (by (4.11)) and from the elementary inequality $3\leqslant 6\ln(2d)$. □

4.4 Adaptation to unknown U

In the previous section, the forecaster is given a radius $U>0$ and asked to ensure a low worst-case regret on the ℓ^1-ball $B_1(U)$. In this section, U is no longer given: the forecaster is asked to be competitive against all balls $B_1(U)$, for $U>0$. Namely, its worst-case regret on each $B_1(U)$ should be almost as good as if U were known beforehand. For simplicity, we assume that X, Y, and T are known: we discuss in Section 4.5 how to simultaneously adapt to all parameters.

We define

$$R\triangleq\lceil\log_2(2T/c)\rceil_+\quad\text{and}\quad U_r\triangleq\frac{Y}{X}\frac{2^r}{\sqrt{T\ln(2d)}},\quad\text{for } r=0,\dots,R\,, \qquad (4.17)$$

where $c>0$ is a known absolute constant and

$$\lceil x\rceil_+\triangleq\min\big\{k\in\mathbb{N}:k\geqslant x\big\}\quad\text{for all } x\in\mathbb{R}\,.$$

The Scaling algorithm of Figure 4.4 works as follows. We have access to a sub-algorithm $\mathcal{A}(U)$

Parameters: $X, Y, \eta > 0$, $T \geqslant 1$, and $c > 0$ (a constant).
Initialization: $R = \lceil \log_2(2T/c) \rceil_+$, $\boldsymbol{w}_1 = \vec{1}/(R+1) \in \mathbb{R}^{R+1}$.
For time steps $t = 1, \ldots, T$:

1. For experts $r = 0, \ldots, R$:
   - Run the sub-algorithm $\mathcal{A}(U_r)$ on the ball $B_1(U_r)$ and obtain the prediction $\widehat{y}_t^{(r)}$.
2. Output the prediction $\widehat{y}_t = \sum_{r=0}^{R} \frac{w_t^{(r)}}{\sum_{r'=0}^{R} w_t^{(r')}} \big[\widehat{y}_t^{(r)}\big]_Y$.
3. Update $w_{t+1}^{(r)} = w_t^{(r)} \exp\Big(-\eta\big(y_t - \big[\widehat{y}_t^{(r)}\big]_Y\big)^2\Big)$ for $r = 0, \ldots, R$.

Figure 4.4: The Scaling algorithm.

which we run simultaneously for all $U = U_r$, $r = 0, \ldots, R$. Each instance of the sub-algorithm $\mathcal{A}(U_r)$ performs online linear regression on the ℓ^1-ball $B_1(U_r)$. We employ an exponentially weighted forecaster to aggregate these $R+1$ sub-algorithms to perform online linear regression simultaneously on the balls $B_1(U_0), \ldots, B_1(U_R)$. The following regret bound follows by exp-concavity of the square loss.

Theorem 4.4. *Suppose that $X, Y > 0$ are known. Let $c, c' > 0$ be two absolute constants. Suppose that for all $U > 0$, we have access to a sub-algorithm $\mathcal{A}(U)$ with regret against $B_1(U)$ of at most*

$$cUXY\sqrt{T\ln(2d)} + c'Y^2 \quad \text{for } T \geqslant T_0\,, \tag{4.18}$$

uniformly over all sequences $(\boldsymbol{x}_t)$ and (y_t) bounded by X and Y. Then, for a known $T \geqslant T_0$, the Scaling algorithm with $\eta = 1/(8Y^2)$ satisfies

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 + 2c\,\|\boldsymbol{u}\|_1\, XY\sqrt{T\ln(2d)}\right\} + 8Y^2\ln\big(\lceil\log_2(2T/c)\rceil_+ + 1\big) + (c+c')Y^2. \tag{4.19}$$

In particular, for every $U > 0$,

$$\sum_{t=1}^{T}(y_t - \widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in B_1(U)}\left\{\sum_{t=1}^{T}(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2\right\} + 2cUXY\sqrt{T\ln(2d)} + 8Y^2\ln\big(\lceil\log_2(2T/c)\rceil_+ + 1\big) + (c+c')Y^2.$$

Remark 4.3. *By Theorem 4.3 the LEG algorithm satisfies assumption (4.18) with $T_0 = \ln(2d)$, $c \triangleq 9c_1 = 72\big(\sqrt{2}+1\big)$, and $c' \triangleq c_2 = 4\left(1 + 1/\sqrt{2}\right)^2$.*

Proof: Since the Scaling algorithm is an exponentially weighted average forecaster (with clipping) applied to the $R+1$ experts $\mathcal{A}(U_r) = \big(\widehat{y}_t^{(r)}\big)_{t\geqslant 1}$, $r = 0, \ldots, R$, we have, by Lemma 4.5 in the

appendix,

$$\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 \leqslant \min_{r=0,\dots,R}\sum_{t=1}^{T}\left(\widehat{y}_t^{(r)}-\widehat{y}_t\right)^2 + 8Y^2\ln(R+1)$$
$$\leqslant \min_{r=0,\dots,R}\left\{\inf_{\boldsymbol{u}\in B_1(U_r)}\left\{\sum_{t=1}^{T}(y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t)^2\right\} + cU_rXY\sqrt{T\ln(2d)}\right\} + z\,, \tag{4.20}$$

where the last inequality follows by assumption (4.18), and where we set

$$z \triangleq 8Y^2\ln(R+1) + c'Y^2\,.$$

Let $\boldsymbol{u}_T^* \in \arg\min_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}(y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t)^2 + 2c\left\|\boldsymbol{u}\right\|_1 XY\sqrt{T\ln(2d)}\right\}$. Next, we proceed by considering three cases: $U_0 < \left\|\boldsymbol{u}_T^*\right\|_1 < U_R$, $\left\|\boldsymbol{u}_T^*\right\|_1 \leqslant U_0$, and $\left\|\boldsymbol{u}_T^*\right\|_1 \geqslant U_R$.

Case 1: $U_0 < \left\|\boldsymbol{u}_T^*\right\|_1 < U_R$. Let $r^* \triangleq \min\left\{r=0,\dots,R : U_r \geqslant \left\|\boldsymbol{u}_T^*\right\|_1\right\}$. Note that $r^* \geqslant 1$ since $\left\|\boldsymbol{u}_T^*\right\|_1 > U_0$. By (4.20) we have

$$\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 \leqslant \inf_{\boldsymbol{u}\in B_1(U_{r^*})}\left\{\sum_{t=1}^{T}(y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t)^2\right\} + cU_{r^*}XY\sqrt{T\ln(2d)} + z$$
$$\leqslant \sum_{t=1}^{T}(y_t-\boldsymbol{u}_T^*\cdot\boldsymbol{x}_t)^2 + 2c\left\|\boldsymbol{u}_T^*\right\|_1 XY\sqrt{T\ln(2d)} + z\,,$$

where the last inequality follows from $\boldsymbol{u}_T^* \in B_1(U_{r^*})$ and from the fact that $U_{r^*} \leqslant 2\left\|\boldsymbol{u}_T^*\right\|_1$ (since, by definition of r^*, $\left\|\boldsymbol{u}_T^*\right\|_1 > U_{r^*-1} = U_{r^*}/2$). Finally, we obtain (4.19) by definition of $\boldsymbol{u}_T^*$ and $z \triangleq 8Y^2\ln(R+1) + c'Y^2$.

Case 2: $\left\|\boldsymbol{u}_T^*\right\|_1 \leqslant U_0$. By (4.20) we have

$$\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 \leqslant \left\{\sum_{t=1}^{T}(y_t-\boldsymbol{u}_T^*\cdot\boldsymbol{x}_t)^2 + cU_0XY\sqrt{T\ln(2d)}\right\} + z\,, \tag{4.21}$$

which yields (4.19) by the equality $cU_0XY\sqrt{T\ln(2d)} = cY^2$ (by definition of U_0), by adding $2c\left\|\boldsymbol{u}_T^*\right\|_1 XY\sqrt{T\ln(2d)} \geqslant 0$, and by definition of $\boldsymbol{u}_T^*$ and z.

Case 3: $\left\|\boldsymbol{u}_T^*\right\|_1 \geqslant U_R$. By construction, we have $\widehat{y}_t \in [-Y,Y]$, and by assumption, we have $y_t \in [-Y,Y]$, so that

$$\sum_{t=1}^{T}(y_t-\widehat{y}_t)^2 \leqslant 4Y^2T \leqslant \sum_{t=1}^{T}(y_t-\boldsymbol{u}_T^*\cdot\boldsymbol{x}_t)^2 + 2cU_RXY\sqrt{T\ln(2d)}$$
$$\leqslant \sum_{t=1}^{T}(y_t-\boldsymbol{u}_T^*\cdot\boldsymbol{x}_t)^2 + 2c\left\|\boldsymbol{u}_T^*\right\|_1 XY\sqrt{T\ln(2d)}\,,$$

where the second inequality follows by $2cU_RXY\sqrt{T\ln(2d)} = 2cY^2 2^R \geqslant 4Y^2T$ (since $2^R \geqslant 2T/c$ by definition of R), and the last inequality uses the assumption $\left\|\boldsymbol{u}_T^*\right\|_1 \geqslant U_R$. We finally get

(4.19) by definition of $\boldsymbol{u}_T^*$.

This concludes the proof of the first claim (4.19). The second claim follows by bounding $\|\boldsymbol{u}\|_1 \leqslant U$. □

4.5 Extension to a fully adaptive algorithm and other discussions

The Scaling algorithm of Section 4.4 uses prior knowledge of Y, Y/X, and T. In order to obtain a fully automatic algorithm, we need to adapt efficiently to these quantities. Adaptation to Y is possible via a technique already used for the LEG algorithm, i.e., by updating the clipping range B_t based on the past observations $|y_s|$, $s \leqslant t-1$.

In parallel to adapting to Y, adaptation to Y/X can be carried out as follows. We replace the exponential sequence $\{U_0, \dots, U_R\}$ by another exponential sequence $\{U'_0, \dots, U'_{R'}\}$:

$$U'_r \triangleq \frac{1}{T^k} \frac{2^r}{\sqrt{T \ln(2d)}}, \quad r = 0, \dots, R', \tag{4.22}$$

where $R' \triangleq R + \lceil \log_2 T^{2k} \rceil = \lceil \log_2(2T/c) \rceil_+ + \lceil \log_2 T^{2k} \rceil$, and where $k > 1$ is a fixed constant. On the one hand, for $T \geqslant T_0 \triangleq \max\{(X/Y)^{1/k}, (Y/X)^{1/k}\}$, we have (cf. (4.17) and (4.22)),

$$[U_0, U_R] \subset [U'_0, U'_{R'}] .$$

Therefore, the analysis of Theorem 4.4 applied to the grid $\{U'_0, \dots, U_{R'}\}$ yields[8] a regret bound of the order of $UXY\sqrt{T \ln d} + Y^2 \ln(R'+1)$. On the other hand, clipping the predictions to $[-Y, Y]$ ensures the crude regret bound $4Y^2 T_0$ for small $T < T_0$. Hence, the overall regret for all $T \geqslant 1$ is of the order of

$$UXY\sqrt{T \ln d} + Y^2 \ln(k \ln T) + Y^2 \max\{(X/Y)^{1/k}, (Y/X)^{1/k}\} .$$

Adaptation to an unknown time horizon T can be carried out via a standard doubling trick on T. However, to avoid restarting the algorithm repeatedly, we can use a time-varying exponential sequence $\{U'_{-R'(t)}(t), \dots, U'_{R'(t)}(t)\}$ where $R'(t)$ grows at the rate of $k \ln(t)$. This gives[9] us an algorithm that is fully automatic in the parameters U, X, Y and T. In this case, we can show that the regret is of the order of

$$UXY\sqrt{T \ln d} + Y^2 k \ln(T) + Y^2 \max\left\{\left(\sqrt{T}X/Y\right)^{1/k}, \left(Y/(\sqrt{T}X)\right)^{1/k}\right\},$$

where the last two terms are negligible when $T \to +\infty$ (since $k > 1$).

Next we discuss another possible improvement. There is a logarithmic gap between the upper bound of Theorem 4.1 and the lower bound of Theorem 4.2. This gap comes from a concentration argument on a specific sequence of (unbounded) normal random variables in the proof of the lower bound. We think that in the interval $\kappa \geqslant cd$ (for some large enough absolute constant $c > 0$), we can recover the missing $\ln(1 + 2\kappa)$ in our lower bound by using the argument of [Vov01,

[8] The proof remains the same by replacing $8Y^2 \ln(R+1)$ with $8Y^2 \ln(R'+1)$.

[9] Each time the exponential sequence (U'_r) expands, the weights assigned to the existing points U'_r are appropriately reassigned to the whole new sequence.

Theorem 2] instead. As for the interval $\kappa \leqslant cd$, we could use a different sequence of random variables with bounded support, and, e.g., Assouad's Lemma.

4.A Proofs

4.A.1 Proof of Theorem 4.2

To prove Theorem 4.2, we perform a reduction to the stochastic batch setting (via the standard online to batch trick), and employ a version of the lower bound proved in [Tsy03] for convex aggregation.

We first need the following notations. Let $T \in \mathbb{N}^*$. Let (S, μ) be a probability space for which we can find an orthonormal family[10] $(\varphi_j)_{1 \leqslant j \leqslant d}$ with d elements in the space of square-integrable functions on S, which we denote by $\mathbb{L}^2(S, \mu)$ thereafter. For all $\boldsymbol{u} \in \mathbb{R}^d$ and $\gamma, \sigma > 0$, denote by $\mathbb{P}_{\boldsymbol{u}}^{\gamma,\sigma}$ the joint law of the i.i.d. sequence $(X_t, Y_t)_{1 \leqslant t \leqslant T}$ such that

$$Y_t = \gamma \varphi_{\boldsymbol{u}}(X_t) + \sigma \varepsilon_t \in \mathbb{R} \,, \tag{4.23}$$

where $\varphi_{\boldsymbol{u}} \triangleq \sum_{j=1}^d u_j \varphi_j$, where the X_t are i.i.d points in S drawn from μ, and where the ε_t are i.i.d standard Gaussian random variables such that $(X_t)_{1 \leqslant t \leqslant T}$ and $(\varepsilon_t)_{1 \leqslant t \leqslant T}$ are independent.

The next lemma is a direct adaptation of [Tsy03, Theorem 2], which we state with our notations in a slightly more precise form (we make clear how the lower bound depends on the noise level σ and the signal level γ).

Lemma 4.1 (An extension of Theorem 2 of [Tsy03])**.**
Let $d, T \in \mathbb{N}^*$ and $\gamma, \sigma > 0$. Let (S, μ) be a probability space for which we can find an orthonormal family $(\varphi_j)_{1 \leqslant j \leqslant d}$ in $\mathbb{L}^2(S, \mu)$, and consider the Gaussian linear model (4.23). Then there exist absolute constants $c_4, c_5, c_6, c_7 > 0$ such that

$$\inf_{\widehat{f}_T} \sup_{\substack{\boldsymbol{u} \in \mathbb{R}_+^d \\ \sum_j u_j \leqslant 1}} \left\{ \mathbb{E}_{\mathbb{P}_{\boldsymbol{u}}^{\gamma,\sigma}} \left\| \widehat{f}_T - \gamma \varphi_{\boldsymbol{u}} \right\|_\mu^2 \right\}$$
$$\geqslant \begin{cases} c_4 \frac{d\sigma^2}{T} & \text{if} \quad \frac{d}{\sqrt{T}} \leqslant c_5 \frac{\gamma}{\sigma} \,, \\ c_6 \gamma \sigma \sqrt{\frac{1}{T} \ln\left(1 + \frac{d\sigma}{\sqrt{T}\gamma}\right)} & \text{if} \quad c_5 \frac{\gamma}{\sigma} < \frac{d}{\sqrt{T}} \leqslant c_7 \frac{\gamma d}{\sigma \sqrt{\ln(1+d)}} \,, \end{cases}$$

where the infimum is taken over all estimators[11] $\widehat{f}_T : S \to \mathbb{R}$, where the supremum is taken over all nonnegative vectors with total mass at most 1, and where $\|f\|_\mu^2 \triangleq \int_S f(x)^2 \mu(\mathrm{d}x)$ for all measurable functions $f : S \to \mathbb{R}$.

Note that the lower bound we stated in Theorem 4.2 is very similar to T times the above lower bound with $\gamma \sim X$ and $\sigma \sim Y$ (recall that $\kappa \triangleq \sqrt{T}UX/(2dY)$). The main difference is that

[10] An example is given by $S = [-\pi, \pi]$, $\mu(\mathrm{d}x) = \mathrm{d}x/(2\pi)$, and $\varphi_j(x) = \sqrt{2}\sin(jx)$ for all $1 \leqslant j \leqslant d$ and $x \in [-\pi, \pi]$. We will use this particular case later.

[11] As usual, an estimator is a measurable function of the sample $(X_t, Y_t)_{1 \leqslant t \leqslant T}$, but the dependency on the sample is omitted.

the latter holds for unbounded observations, while we need bounded observations y_t, $1 \leqslant t \leqslant T$. A simple concentration argument will show that these observations lie in $[-Y, Y]$ with high probability, which will yield the desired lower bound. The proof of Theorem 4.2 thus consists of the following steps:

- step 1: reduction to the stochastic batch setting;
- step 2: application of Lemma 4.1;
- step 3: concentration argument.

Proof (of Theorem 4.2): We first assume that $\sqrt{\ln(1+2d)}/\big(2d\sqrt{\ln 2}\big) \leqslant \kappa \leqslant 1$. The case when $\kappa > 1$ will easily follow from the monotonicity of the minimax regret in κ (see the end of the proof). We set

$$T \triangleq 1 + \big\lceil (4d\kappa)^2 \big\rceil \,, \quad U \triangleq 1 \,, \quad \text{and} \quad X \triangleq \frac{2d\kappa Y}{\sqrt{T}} \,, \tag{4.24}$$

so that $T \geqslant 2$, $\sqrt{T}UX/(2dY) = \kappa$, and $X \leqslant Y/2$ (since $\sqrt{T} \geqslant 4d\kappa$).

Step 1: reduction to the stochastic batch setting.
First note that by clipping to $[-Y, Y]$, we have

$$\begin{aligned} &\inf_{(\widetilde{f}_t)_t} \sup_{\substack{\|\boldsymbol{x}_t\|_\infty \leqslant X \\ |y_t| \leqslant Y}} \left\{ \sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\} \\ &= \inf_{\substack{(\widetilde{f}_t)_t \\ |\widetilde{f}_t| \leqslant Y}} \sup_{\substack{\|\boldsymbol{x}_t\|_\infty \leqslant X \\ |y_t| \leqslant Y}} \left\{ \sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\}, \end{aligned} \tag{4.25}$$

where the first infimum is taken over all online forecasters[12] $(\widetilde{f}_t)_t$, where the second infimum is restricted to online forecasters $(\widetilde{f}_t)_t$ which output predictions in $[-Y, Y]$, and where both suprema are taken over all individual sequences $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T} \in (\mathbb{R}^d \times \mathbb{R})^T$ such that $|y_1|, \dots, |y_T| \leqslant Y$ and $\|\boldsymbol{x}_1\|_\infty, \dots, \|\boldsymbol{x}_T\|_\infty \leqslant X$.

Next we use the standard online to batch conversion (cf. Section 2.5.1) to bound from below the right-hand side of (4.25) by T times the lower bound of Lemma 4.1, which we apply to the particular case where $S = [-\pi, \pi]$, $\mu(\mathrm{d}x) = \mathrm{d}x/(2\pi)$, and $\varphi_j(x) = \sqrt{2}\sin(jx)$ for all $1 \leqslant j \leqslant d$ and $x \in [-\pi, \pi]$. Let

$$\gamma \triangleq c_8 X \quad \text{and} \quad \sigma \triangleq \frac{c_9 Y}{\sqrt{\ln T}} \,, \tag{4.26}$$

for some absolute constants $c_8, c_9 > 0$ to be chosen by the analysis.

Let $(\widetilde{f}_t)_{t\geqslant 1}$ be any online forecaster whose predictions lie in $[-Y, Y]$, and consider the estimator

[12] Recall that an online forecaster is a sequence of functions $(\widetilde{f}_t)_{t\geqslant 1}$, where $\widetilde{f}_t : \mathbb{R}^d \times (\mathbb{R}^d \times \mathbb{R})^{t-1} \to \mathbb{R}$ maps at time t the new input $\boldsymbol{x}_t$ and the past data $(\boldsymbol{x}_s, y_s)_{1\leqslant s\leqslant t-1}$ to a prediction $\widetilde{f}_t\big(\boldsymbol{x}_t; (\boldsymbol{x}_s, y_s)_{1\leqslant s\leqslant t-1}\big)$. However, unless mentioned otherwise, we omit the dependency in $(\boldsymbol{x}_s, y_s)_{1\leqslant s\leqslant t-1}$, and only write $\widetilde{f}_t(\boldsymbol{x}_t)$.

$\widehat{f}_T$ defined for each sample $(X_t, Y_t)_{1\leqslant t\leqslant T}$ and each new input X' by

$$\widehat{f}_T\Big(X'; (X_t, Y_t)_{1\leqslant t\leqslant T}\Big) \triangleq \frac{1}{T}\sum_{t=1}^{T} \widetilde{f}_t\Big(\gamma\boldsymbol{\varphi}(X'); (\gamma\boldsymbol{\varphi}(X_s), Y_s)_{1\leqslant s\leqslant t-1}\Big)\,, \tag{4.27}$$

where $\boldsymbol{\varphi} \triangleq (\varphi_1, \dots, \varphi_d)$, and where we explicitely wrote all the dependencies[12] of the $\widetilde{f}_t$, $t = 1, \dots, T$.
Take $\boldsymbol{u}^* \in \mathbb{R}^d_+$ achieving the supremum[13] in Lemma 4.1 for the estimator $\widehat{f}_T$. Note that $\|\boldsymbol{u}^*\|_1 \leqslant 1$. Besides, consider the i.i.d. random sequence $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T}$ in $\mathbb{R}^d \times \mathbb{R}$ defined for all $t = 1, \dots, T$ by

$$\boldsymbol{x}_t \triangleq \big(\gamma\varphi_1(X_t), \dots, \gamma\varphi_d(X_t)\big) \quad \text{and} \quad y_t \triangleq \gamma\varphi_{\boldsymbol{u}^*}(X_t) + \sigma\varepsilon_t\,, \tag{4.28}$$

where $\varphi_{\boldsymbol{u}^*} \triangleq \sum_{j=1}^d u_j^*\varphi_j$ (so that $y_t = \boldsymbol{u}^* \cdot \boldsymbol{x}_t + \sigma\varepsilon_t$ for all t), where the X_t are i.i.d points in $[-\pi, \pi]$ drawn from the uniform distribution $\mu(\mathrm{d}x) = \mathrm{d}x/(2\pi)$, and where the ε_t are i.i.d standard Gaussian random variables such that $(X_t)_t$ and $(\varepsilon_t)_t$ are independent. All the expectations below are thus taken with respect to the probability distribution $\mathbb{P}^{\gamma,\sigma}_{\boldsymbol{u}^*}$.

By standard manipulations (e.g., using the tower rule and Jensen's inequality), we get the following lower bound. A detailed proof can be found after the proof of the present theorem (page 154).

Lemma 4.2 (Reduction to the batch setting)**.**
With $(\widetilde{f}_t)_{1\leqslant t\leqslant T}$, $\widehat{f}_T$, and $\boldsymbol{u}^*$ defined above, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1\leqslant 1}\sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2\right] \geqslant T\,\mathbb{E}\left\|\widehat{f}_T - \gamma\varphi_{\boldsymbol{u}^*}\right\|_\mu^2\,.$$

Step 2: application of Lemma 4.1.
Next we use Lemma 4.1 to prove that, for some absolute constants $c_9, c_{11} > 0$,

$$T\,\mathbb{E}\left\|\widehat{f}_T - \gamma\varphi_{\boldsymbol{u}^*}\right\|_\mu^2 \geqslant \frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\,. \tag{4.29}$$

By Lemma 4.1 and by definition of $\boldsymbol{u}^*$, we have

$$\begin{aligned}
&\mathbb{E}\left\|\widehat{f}_T - \gamma\varphi_{\boldsymbol{u}^*}\right\|_\mu^2 \\
&\geqslant \begin{cases} c_4\frac{d\sigma^2}{T} & \text{if} \quad \frac{d}{\sqrt{T}} \leqslant c_5\frac{\gamma}{\sigma}\,, \\ c_6\gamma\sigma\sqrt{\frac{1}{T}\ln\left(1+\frac{d\sigma}{\sqrt{T}\gamma}\right)} & \text{if} \quad c_5\frac{\gamma}{\sigma} < \frac{d}{\sqrt{T}} \leqslant \frac{c_7\gamma d}{\sigma\sqrt{\ln(1+d)}}\,. \end{cases} \\
&\geqslant \begin{cases} \frac{c_4c_9^2}{T(\ln T)}dY^2 & \text{if} \quad \frac{d}{\sqrt{T}} \leqslant c_5\frac{\gamma}{\sigma}\,, \\ \frac{c_6c_8c_9}{\sqrt{\ln T}}UXY\sqrt{\frac{1}{T}\ln\left(1+\frac{c_9dY}{c_8\sqrt{T(\ln T)}UX}\right)} & \text{if} \quad c_5\frac{\gamma}{\sigma} < \frac{d}{\sqrt{T}} \leqslant \frac{c_7\gamma d}{\sigma\sqrt{\ln(1+d)}}\,, \end{cases}
\end{aligned} \tag{4.30}$$

[13] If the supremum in Lemma 4.1 is not achieved, then we can instead take an ε-almost-maximizer for any $\varepsilon > 0$. Letting $\varepsilon \to 0$ in the end will conclude the proof.

where the last inequality follows from (4.26) and from $U = 1$.

The above lower bound is only meaningful if the following condition holds true:

$$\frac{d}{\sqrt{T}} \leqslant \frac{c_7 \gamma d}{\sigma \sqrt{\ln(1+d)}} \,. \tag{4.31}$$

But, by definition of $T \triangleq 1 + \lceil (4d\kappa)^2 \rceil$ and by the assumption $\sqrt{\ln(1+2d)}/\big(2d\sqrt{\ln 2}\big) \leqslant \kappa$, elementary manipulations show that (4.31) actually holds true whenever[14] $c_9 \leqslant c_7 c_8 c_{10}$, where $c_{10} \triangleq \frac{1}{2} \inf_{x \geqslant 2\sqrt{\frac{\ln 3}{\ln 2}}} \left\{ \frac{x}{\sqrt{1+\lceil x^2 \rceil}} \right\}$ (note that $c_{10} > 0$).

Therefore, if $c_9 \leqslant c_7 c_8 c_{10}$, then (4.30) entails that

$$\begin{aligned} &\mathbb{E} \left\| \widehat{f}_T - \gamma \varphi_{\boldsymbol{u}^*} \right\|_\mu^2 \\ &\qquad \geqslant \min \left\{ \frac{c_4 c_9^2}{T(\ln T)} dY^2 \,,\; \frac{c_6 c_8 c_9}{\sqrt{\ln T}} UXY \sqrt{\frac{1}{T} \ln \left(1 + \frac{c_9 dY}{c_8 \sqrt{T(\ln T)} UX} \right)} \right\} . \end{aligned} \tag{4.32}$$

Moreover, note that if $c_9 \leqslant c_8 2\sqrt{\ln 2}$, then $c_8 \geqslant c_9/(2\sqrt{\ln 2}) \geqslant c_9/(2\sqrt{\ln T})$. In this case, since $x \mapsto x\sqrt{\ln(1 + A/x)}$ is nondecreasing on $\mathbb{R}_+^*$ for all $A > 0$, we can replace c_8 with $c_9/(2\sqrt{\ln T})$ in the next expression and get

$$\begin{aligned} &\frac{c_6 c_8 c_9}{\sqrt{\ln T}} UXY \sqrt{\frac{1}{T} \ln \left(1 + \frac{c_9 dY}{c_8 \sqrt{T(\ln T)} UX} \right)} \\ &\qquad \geqslant \frac{c_6 c_9^2}{2 \ln T} UXY \sqrt{\frac{1}{T} \ln \left(1 + \frac{2dY}{\sqrt{T} UX} \right)} = \frac{c_6 c_9^2}{T(\ln T)} dY^2 \kappa \sqrt{\ln(1 + 1/\kappa)} \,, \end{aligned}$$

where we used the definition of $\kappa \triangleq \sqrt{T} UX/(2dY)$.
In the sequel we will choose the absolute constants c_8 and c_9 such that

$$c_9 \leqslant c_7 c_8 c_{10} \quad \text{and} \quad c_9 \leqslant c_8 2\sqrt{\ln 2} \,. \tag{4.33}$$

Therefore, by the above remarks, by the fact that $\ln T \triangleq \ln\big(1 + \lceil (4d\kappa)^2 \rceil\big) \leqslant \ln\big(2 + 16d^2\big)$ (since $\kappa \leqslant 1$ by assumption), and multiplying both sides of (4.32) by T, we get

$$\begin{aligned} T\,\mathbb{E} \left\| \widehat{f}_T - \gamma \varphi_{\boldsymbol{u}^*} \right\|_\mu^2 &\geqslant \min \left\{ \frac{c_4 c_9^2}{\ln\big(2 + 16d^2\big)} dY^2 \,,\; \frac{c_6 c_9^2}{\ln\big(2 + 16d^2\big)} dY^2 \kappa \sqrt{\ln(1 + 1/\kappa)} \right\} \\ &\geqslant \frac{c_{11} c_9^2}{\ln\big(2 + 16d^2\big)} dY^2 \kappa \sqrt{\ln(1 + 1/\kappa)} \,, \end{aligned}$$

where we set $c_{11} \triangleq \min\big\{ c_4/\sqrt{\ln 2}, c_6 \big\}$, and where used the fact that $x \mapsto x\sqrt{\ln(1 + 1/x)}$ is nondecreasing on $\mathbb{R}_+^*$, so that its value at $x = \kappa \leqslant 1$ is smaller than $\sqrt{\ln 2}$. This concludes the

[14] By definition of γ and σ, (4.31) is equivalent to $T \ln T \geqslant c_9^2/(c_7^2 c_8^2)(Y/X)^2 \ln(1 + d)$. But by definition of X and by the assumption $\kappa \geqslant \sqrt{\ln(1 + 2d)}/(2d\sqrt{\ln 2})$, we have $Y/X \leqslant 1/c_{10}$. Therefore, (4.31) is implied by $T \ln T \geqslant c_9^2/(c_7^2 c_8^2 c_{10}^2) \ln(1 + d)$, which in turn is implied by the condition $c_9 \leqslant c_7 c_8 c_{10}$ (by definition of T).

proof of (4.29).

Combining Lemma 4.2 and (4.29), we get

$$\mathbb{E}\left[\sum_{t=1}^{T}\big(y_t-\widetilde{f}_t(\boldsymbol{x}_t)\big)^2-\inf_{\|\boldsymbol{u}\|_1\leqslant 1}\sum_{t=1}^{T}\big(y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2\right] \geqslant \frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)}dY^2\kappa\sqrt{\ln(1+1/\kappa)}\,. \tag{4.34}$$

Step 3: concentration argument.
At this stage it would be tempting to conclude by using (4.34) to assert that since the expectation is lower bounded, then there is at least one individual sequence with the same lower bound. However, we have no boundedness guarantee about such individual sequence since the random observations y_t lie outside of $[-Y,Y]$ with positive probability. Next we prove that the probability of the event

$$\mathcal{A}\triangleq\bigcap_{t=1}^{T}\big\{|y_t|\leqslant Y\big\}$$

is actually close to 1, and that

$$\mathbb{E}\left[\mathbb{I}_{\mathcal{A}}\left(\widehat{L}_T-\inf_{\|\boldsymbol{u}\|_1\leqslant 1}L_T(\boldsymbol{u})\right)\right]\geqslant\frac{1}{2}\frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)}dY^2\kappa\sqrt{\ln(1+1/\kappa)}\,. \tag{4.35}$$

(Note a missing factor of 2 between (4.34) and (4.35).) The last lower bound will then enable to conclude the proof of this theorem.

Set $\widehat{L}_T\triangleq\sum_{t=1}^{T}\big(y_t-\widetilde{f}_t(\boldsymbol{x}_t)\big)^2$ and $L_T(\boldsymbol{u})\triangleq\sum_{t=1}^{T}\big(y_t-\boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2$ for all $\boldsymbol{u}\in\mathbb{R}^d$. Denote by $\mathcal{A}^c$ the complement of $\mathcal{A}$, and by $\mathbb{I}_{\mathcal{A}}$ and $\mathbb{I}_{\mathcal{A}^c}$ the corresponding indicator functions. We have

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{I}_{\mathcal{A}}\left(\widehat{L}_T-\inf_{\|\boldsymbol{u}\|_1\leqslant 1}L_T(\boldsymbol{u})\right)\right]\\
&\quad=\mathbb{E}\left[\widehat{L}_T-\inf_{\|\boldsymbol{u}\|_1\leqslant 1}L_T(\boldsymbol{u})\right]-\mathbb{E}\left[\mathbb{I}_{\mathcal{A}^c}\left(\widehat{L}_T-\inf_{\|\boldsymbol{u}\|_1\leqslant 1}L_T(\boldsymbol{u})\right)\right]\\
&\quad\geqslant\frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)}dY^2\kappa\sqrt{\ln(1+1/\kappa)}-\mathbb{E}\Big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\Big]\;,
\end{aligned}\tag{4.36}$$

where the last inequality follows by (4.34) and by the fact that $L_T(\boldsymbol{u})\geqslant 0$ for all $\boldsymbol{u}\in\mathbb{R}^d$. The rest of the proof is dedicated to upper bounding the above quantity $\mathbb{E}\big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\big]$ by half the term on its left. This way, we will have proved (4.35).

First note that

$$\mathbb{E}\Big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\Big]\triangleq\mathbb{E}\left[\mathbb{I}_{\mathcal{A}^c}\sum_{t=1}^{T}\big(y_t-\widetilde{f}_t(\boldsymbol{x}_t)\big)^2\right]$$

$$\leqslant \mathbb{E}\left[\mathbb{I}_{\mathcal{A}^c}\sum_{t=1}^{T}\left(4Y^2\mathbb{I}_{\{|y_t|\leqslant Y\}} + \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2\mathbb{I}_{\{|y_t|>Y\}}\right)\right] \tag{4.37}$$

$$\leqslant 4TY^2\mathbb{P}\big(\mathcal{A}^c\big) + \sum_{t=1}^{T}\mathbb{E}\Big[\big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2\,\mathbb{I}_{\left\{|\varepsilon_t|>\frac{Y}{2\sigma}\right\}}\Big]\ , \tag{4.38}$$

where (4.37) follows from the fact that the online forecaster $(\widetilde{f}_t)_t$ outputs its predictions in $[-Y, Y]$. As for (4.38), note by definition of y_t that $|y_t| \leqslant \|\boldsymbol{u}^*\|_1\,\gamma\,\|\boldsymbol{\varphi}(X_t)\|_\infty + \sigma|\varepsilon_t| \leqslant \gamma\sqrt{2} + \sigma|\varepsilon_t|$ since $\|\boldsymbol{u}^*\|_1 \leqslant 1$ and $|\varphi_j(x)| \triangleq |\sqrt{2}\sin(jx)| \leqslant \sqrt{2}$ for all $j = 1, \dots, d$ and $x \in \mathbb{R}$. Therefore, by definition of $\gamma \triangleq c_9 X$, and since $X \leqslant Y/2$ (by definition of X), we get $|y_t| \leqslant c_9\sqrt{2}\,Y/2 + \sigma|\varepsilon_t| \leqslant Y/2 + \sigma|\varepsilon_t|$ provided that

$$c_9 \leqslant \frac{1}{\sqrt{2}}\ , \tag{4.39}$$

which we assume thereafter. The above remarks show that $\{|y_t| > Y\} \subset \{|\varepsilon_t| > Y/(2\sigma)\}$, which entails (4.38). By the same comments and since $|\widetilde{f}_t| \leqslant Y$, we have, for all $t = 1, \dots, T$,

$$\begin{aligned}
\mathbb{E}\Big[\big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2\mathbb{I}_{\left\{|\varepsilon_t|>\frac{Y}{2\sigma}\right\}}\Big] &\leqslant \mathbb{E}\Big[\big(Y/2 + \sigma|\varepsilon_t| + Y\big)^2\mathbb{I}_{\left\{|\varepsilon_t|>\frac{Y}{2\sigma}\right\}}\Big] \\
&\leqslant 2\left(\frac{3Y}{2}\right)^2\mathbb{P}\left(|\varepsilon_t| > \frac{Y}{2\sigma}\right) + 2\sigma^2\mathbb{E}\Big[\varepsilon_t^2\mathbb{I}_{\left\{|\varepsilon_t|>\frac{Y}{2\sigma}\right\}}\Big] && (4.40)\\
&\leqslant \frac{9Y^2}{2}\mathbb{P}\left(|\varepsilon_t| > \frac{Y}{2\sigma}\right) + 2\sigma^2\sqrt{3}\,\mathbb{P}^{1/2}\left(|\varepsilon_t| > \frac{Y}{2\sigma}\right) && (4.41)\\
&\leqslant 9Y^2T^{-1/(8c_9^2)} + 2\frac{c_9^2Y^2}{\ln 2}\sqrt{6}\,T^{-1/(16c_9^2)}\ , && (4.42)
\end{aligned}$$

where we used the following arguments. Inequality (4.40) follows by the elementary inequality $(a+b)^2 \leqslant 2(a^2+b^2)$ for all $a, b \in \mathbb{R}$. To get (4.41) we used the Cauchy-Schwarz inequality and the fact that $\mathbb{E}\big[\varepsilon_t^4\big] = 3$ (since ε_t is a standard Gaussian random variable). Finally, (4.42) follows by definition of $\sigma \triangleq c_9 Y/\sqrt{\ln T} \leqslant c_9 Y/\sqrt{\ln 2}$ and from the fact that, since ε_t is a standard Gaussian random variable,

$$\mathbb{P}\left(|\varepsilon_t| > \frac{Y}{2\sigma}\right) \leqslant 2e^{-\frac{1}{2}\left(\frac{Y}{2\sigma}\right)^2} = 2e^{-\frac{1}{2}\left(\frac{\sqrt{\ln T}}{2c_9}\right)^2} = 2T^{-1/(8c_9^2)}\ .$$

Using the fact that $\mathbb{P}\big(\mathcal{A}^c\big) \leqslant \sum_{t=1}^{T}\mathbb{P}\big(|y_t| > Y\big) \leqslant \sum_{t=1}^{T}\mathbb{P}\big(|\varepsilon_t| > Y/(2\sigma)\big) \leqslant 2T^{1-1/(8c_9^2)}$ by the inequality above and substituting (4.42) in (4.38), we get

$$\begin{aligned}
&\mathbb{E}\Big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\Big] \\
&\quad\leqslant 8Y^2T^{2-1/(8c_9^2)} + 9Y^2T^{1-1/(8c_9^2)} + \frac{2c_9^2\sqrt{6}}{\ln 2}Y^2T^{1-1/(16c_9^2)} \\
&\quad\leqslant 8Y^2 2^{2-1/(8c_9^2)} + 9Y^2 2^{1-1/(8c_9^2)} + \frac{2c_9^2\sqrt{6}}{\ln 2}Y^2 2^{1-1/(16c_9^2)}\ , && (4.43)
\end{aligned}$$

where the last inequality follows from the fact that $T^\alpha \leqslant 2^\alpha$ for all $\alpha < 0$ (since $T \geqslant 2$) and from a choice of c_9 such that $c_9 < 1/4$ (which we assume thereafter).

In order to further upper bound $\mathbb{E}\big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\big]$, we use the following technical lemma, which is proved after the proof of the present theorem (see page 155).

Lemma 4.3. *There exists an absolute constant $c_{13} > 0$ such that, for all $c_9 \in (0, c_{13})$,*

$$8Y^2 2^{2-1/(8c_9^2)} + 9Y^2 2^{1-1/(8c_9^2)} + \frac{2c_9^2\sqrt{6}}{\ln 2} Y^2 2^{1-1/(16c_9^2)} \leqslant \frac{1}{2}\frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\,.$$

We can now fix the values of the constants c_8 and c_9 and conclude the proof. Choosing c_9 and $c_8 \triangleq \max\big\{c_9/(2\sqrt{\ln 2}), c_9/(c_7c_{10})\big\}$ such that $c_9 < 1/\sqrt{2}$ (condition (4.39)), $c_9 < 1/4$, and $c_9 \leqslant c_{13}$, then the condition (4.33) also holds, and (4.43) combined with Lemma 4.3 entails that

$$\mathbb{E}\Big[\mathbb{I}_{\mathcal{A}^c}\widehat{L}_T\Big] \leqslant \frac{1}{2}\frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\,.$$

Substituting the last inequality in (4.36), we get that

$$\mathbb{E}\left[\mathbb{I}_{\mathcal{A}}\left(\widehat{L}_T - \inf_{\|\boldsymbol{u}\|_1\leqslant 1} L_T(\boldsymbol{u})\right)\right] \geqslant \frac{1}{2}\frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\,.$$

By the above lower bound and the fact that, $\mathbb{P}_{\boldsymbol{u}^*}^{\gamma,\sigma}$-almost surely, $\|\boldsymbol{x}_t\|_\infty \leqslant \gamma\sqrt{2} \leqslant X$ for all $t = 1, \dots, T$ (since $\gamma \triangleq c_9 X$ and $c_9 \leqslant 1/\sqrt{2}$), we get that

$$\sup_{\substack{\|\boldsymbol{x}_1\|_\infty,\dots,\|\boldsymbol{x}_T\|_\infty\leqslant X\\ y_1,\dots,y_T\in\mathbb{R}}} \left\{\mathbb{I}_{\mathcal{A}}\left(\widehat{L}_T - \inf_{\|\boldsymbol{u}\|_1\leqslant 1} L_T(\boldsymbol{u})\right)\right\} \geqslant \frac{1}{2}\frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\,.$$

Therefore, by definition of $\mathcal{A} \triangleq \bigcap_{t=1}^T \big\{|y_t| \leqslant Y\big\}$, of $\widehat{L}_T \triangleq \sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2$, and of $L_T(\boldsymbol{u}) \triangleq \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$, we get that, for all online forecasters $(\widetilde{f}_t)_{t\geqslant 1}$ whose predictions lie in $[-Y, Y]$,

$$\begin{aligned}\sup_{\substack{\|\boldsymbol{x}_1\|_\infty,\dots,\|\boldsymbol{x}_T\|_\infty\leqslant X\\ |y_1|,\dots,|y_T|\leqslant Y}} &\left\{\sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1\leqslant U}\sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2\right\}\\ &\geqslant \frac{1}{2}\frac{c_{11}c_9^2}{\ln\big(2+16d^2\big)} dY^2\kappa\sqrt{\ln(1+1/\kappa)}\,.\end{aligned}$$

Combining the last lower bound with (4.25) and setting $c_1 \triangleq c_{11}c_9^2/2$ concludes the proof under the assumption $\sqrt{\ln(1+2d)}/\big(2d\sqrt{\ln 2}\big) \leqslant \kappa \leqslant 1$.

Assume now that $\kappa > 1$.
The stated lower bound follows from the case when $\kappa = 1$ and by monotonicity of the minimax regret in κ (when d and Y are kept constant).

More formally, by the first part of this proof (when $\kappa = 1$), we can fix $T \geqslant 1$, $U_1 > 0$, and

$X > 0$ such that $\sqrt{T}U_1X/(2dY) = 1$ and

$$\inf_{(\widetilde{f}_t)_t} \sup_{\substack{\|\boldsymbol{x}_t\|_\infty \leqslant X \\ |y_t| \leqslant Y}} \left\{ \sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U_1} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\} \geqslant \frac{c_1}{\ln\big(2+16d^2\big)} dY^2 \sqrt{\ln 2}\,,$$

where the infimum is taken over all online forecasters $(\widetilde{f}_t)_{t\geqslant 1}$, and where the supremum is taken over all individual sequences bounded by X and Y.

Now take $\kappa > 1$, and set $U \triangleq \kappa U_1 > U_1$, so that $\sqrt{T}UX/(2dY) = \kappa$ (since $\sqrt{T}U_1X/(2dY) = 1$). Moreover, for all individual sequences bounded by X and Y, the regret on $B_1(U)$ is at least as large as the regret on $B_1(U_1)$ (since $U > U_1$). Combining the latter remark with the lower bound above and setting $c_2 \triangleq c_1\sqrt{\ln 2}$ concludes the proof. □

Proof (of Lemma 4.2): We use the same notations as in Step 1 of the proof of Theorem 4.2. Let (X', y') be a random copie of (X_1, y_1) independent of $(X_t, y_t)_{1\leqslant t\leqslant T}$, and define the random vector $\boldsymbol{x}' \triangleq \big(\gamma\varphi_1(X'), \dots, \gamma\varphi_d(X')\big)$. By the tower rule, we have

$$\mathbb{E}\big[(y_t - \widetilde{f}_t(\boldsymbol{x}_t)^2\big] = \mathbb{E}\Big[\mathbb{E}\big[(y_t - \widetilde{f}_t(\boldsymbol{x}_t))^2 \big| (\boldsymbol{x}_s, y_s)_{s\leqslant t-1}\big]\Big] = \mathbb{E}\big[(y' - \widetilde{f}_t(\boldsymbol{x}')^2\big]\,,$$

where we used the fact that $\widetilde{f}_t$ is built on the past data $(\boldsymbol{x}_s, y_s)_{s\leqslant t-1}$ and that $(\boldsymbol{x}', y')$ and $(\boldsymbol{x}_t, y_t)$ are both independent of $(\boldsymbol{x}_s, y_s)_{s\leqslant t-1}$ and are identically distributed. Similarly $\mathbb{E}\big[(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2\big] = \mathbb{E}\big[(y' - \boldsymbol{u}\cdot\boldsymbol{x}')^2\big]$. Using the last equalities and the fact that $\mathbb{E}\big[\inf\{\dots\}\big] \leqslant \inf \mathbb{E}\big[\{\dots\}\big]$, we get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^T \big(y_t - \widetilde{f}_t(\boldsymbol{x}_t)\big)^2 - \inf_{\|\boldsymbol{u}\|_1\leqslant 1} \sum_{t=1}^T \big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2\right] \\
&\quad\geqslant T\left(\frac{1}{T}\sum_{t=1}^T \mathbb{E}\Big[\big(y' - \widetilde{f}_t(\boldsymbol{x}')\big)^2\Big] - \inf_{\|\boldsymbol{u}\|_1\leqslant 1} \mathbb{E}\Big[\big(y' - \boldsymbol{u}\cdot\boldsymbol{x}'\big)^2\Big]\right) \\
&\quad\geqslant T\left(\mathbb{E}\Big[\big(y' - \widehat{f}_T(X')\big)^2\Big] - \inf_{\|\boldsymbol{u}\|_1\leqslant 1} \mathbb{E}\Big[\big(y' - \boldsymbol{u}\cdot\boldsymbol{x}'\big)^2\Big]\right) \qquad (4.44) \\
&\quad= T\,\mathbb{E}\Big[\big(\gamma\varphi_{\boldsymbol{u}^*}(X') - \widehat{f}_T(X')\big)^2\Big] \qquad (4.45) \\
&\quad= T\,\mathbb{E}\Big\|\widehat{f}_T - \gamma\varphi_{\boldsymbol{u}^*}\Big\|_\mu^2\,.
\end{aligned}$$

Inequality (4.44) follows by definition of $\widehat{f}_T \triangleq T^{-1}\sum_{t=1}^T \widetilde{f}_t$ (see (4.27)) and by Jensen's inequality. As for Inequality (4.45), it follows by expanding the square

$$\big(y' - \widehat{f}_t(X')\big)^2 = \big(\gamma\varphi_{\boldsymbol{u}^*}(X') - \widehat{f}_T(X') + y' - \gamma\varphi_{\boldsymbol{u}^*}(X')\big)^2\,,$$

by noting that $\mathbb{E}\big[y' - \gamma\varphi_{\boldsymbol{u}^*}(X')\big|X'\big] = 0$ (via (4.28)) and by the fact that

$$\inf_{\|\boldsymbol{u}\|_1\leqslant 1} \mathbb{E}\Big[\big(y' - \boldsymbol{u}\cdot\boldsymbol{x}'\big)^2\Big] = \mathbb{E}\Big[\big(y' - \gamma\varphi_{\boldsymbol{u}^*}(X')\big)^2\Big]\,,$$

where we used $\|\boldsymbol{u}^*\|_1 \leqslant 1$ (by definition of $\boldsymbol{u}^*$) and $\boldsymbol{u} \cdot \boldsymbol{x}' = \gamma\varphi_{\boldsymbol{u}}(X')$. This concludes the proof. □

Proof (of Lemma 4.3): We use the same notations and assumptions as in the proof of Theorem 4.2. Since the function $x \mapsto x\sqrt{\ln(1+1/x)}$ is nondecreasing on $\mathbb{R}_+^*$ and since $\kappa \geqslant \kappa_{\min} \triangleq \sqrt{\ln(1+2d)}/(2d\sqrt{\ln 2})$ by assumption, we have

$$\begin{aligned}
&\frac{c_{11}c_9^2}{\ln\bigl(2+16d^2\bigr)} dY^2 \kappa \sqrt{\ln(1+1/\kappa)} \\
&\quad \geqslant \frac{c_{11}c_9^2}{\ln\bigl(2+16d^2\bigr)} dY^2 \kappa_{\min} \sqrt{\ln(1+1/\kappa_{\min})} \\
&\quad = \frac{c_{11}c_9^2}{2\sqrt{\ln 2}} Y^2 \frac{\sqrt{\ln(1+2d)}\sqrt{\ln\Bigl[1+2d\sqrt{\ln 2}/\sqrt{\ln(1+2d)}\Bigr]}}{\ln\bigl(2+16d^2\bigr)} \qquad (4.46) \\
&\quad \geqslant \frac{c_{11}c_9^2}{2\sqrt{\ln 2}} Y^2 c_{12} \,, \qquad (4.47)
\end{aligned}$$

where c_{12} denotes the infimum of the last fraction of (4.46) over all $d \geqslant 1$; in particular, $c_{12} > 0$. It is now easy to see that by choosing the absolute constant $c_{13} > 0$ small enough (where c_{13} can be expressed in terms of c_{11} and c_{12}), we have, for all $c_9 \in (0, c_{13})$,

$$8 \cdot 2^{2-1/(8c_9^2)} + 9 \cdot 2^{1-1/(8c_9^2)} + \frac{2c_9^2\sqrt{6}}{\ln 2} 2^{1-1/(16c_9^2)} \leqslant \frac{c_{11}c_9^2}{2\sqrt{\ln 2}} c_{12} \,.$$

Multiplying both sides of the last inequality by Y^2 and combining it with (4.47) concludes the proof. □

4.A.2 Proof of Theorem 4.3

Proof (of Theorem 4.3): The proof follows directly from Corollary 2.2 of Chapter 2 and from the fact that the Lipschitzified losses are larger than their clipped versions. Indeed, first note that, by definition of $\widehat{y}_t$ and $B_{t+1} \geqslant |y_t|$, we have

$$\begin{aligned}
\sum_{t=1}^T \bigl(y_t - \widehat{y}_t\bigr)^2 &\leqslant \sum_{\substack{t=1 \\ t:|y_t|\leqslant B_t}}^T \Bigl(y_t - \bigl[\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t\bigr]_{B_t}\Bigr)^2 + \sum_{\substack{t=1 \\ t:|y_t|>B_t}}^T (B_{t+1}+B_t)^2 \\
&\leqslant \sum_{\substack{t=1 \\ t:|y_t|\leqslant B_t}}^T \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) + \left(1+\frac{1}{\sqrt{2}}\right)^2 \sum_{\substack{t=1 \\ t:B_{t+1}>B_t}}^T B_{t+1}^2 \\
&\leqslant \sum_{t=1}^T \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) + 4\left(1+\frac{1}{\sqrt{2}}\right)^2 Y^2 \,, \qquad (4.48)
\end{aligned}$$

where the second inequality follows from the fact that:

- if $|y_t| \leqslant B_t$ then $(y_t - [\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t]_{B_t})^2 \leqslant \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)$ by Equation (4.12);

- if $|y_t| > B_t$, which is equivalent to $B_{t+1} > B_t$ by definition of B_{t+1}, then $B_t \leqslant B_{t+1}/\sqrt{2}$, so that $B_{t+1} + B_t \leqslant \big(1 + 1/\sqrt{2}\big) B_{t+1}$.

As for the third inequality above, we used the non-negativity of $\widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)$ and upper bounded the geometric sum $\sum_{t:B_{t+1}>B_t}^{T} B_{t+1}^2$ in the same way as in [CBMS07, Theorem 6], i.e., setting $K \triangleq \lceil \log_2 \max_{1 \leqslant t \leqslant T} y_t^2 \rceil$,

$$\sum_{t:B_{t+1}>B_t}^{T} B_{t+1}^2 \leqslant \sum_{k=-\infty}^{K} 2^k = 2^{K+1} \leqslant 4Y^2 \,.$$

To bound (4.48) further from above, we now use the fact that, by construction, the LEG algorithm is the adaptive EG$^{\pm}$ algorithm applied to the modified loss functions $\widetilde{\ell}_t$. Therefore, we get from Corollary 2.1 (cf. Chapter 2) that

$$\begin{aligned}\sum_{t=1}^{T} \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \leqslant \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \widetilde{\ell}_t(\boldsymbol{u}) \\ + 4U \sqrt{\left(\sum_{t=1}^{T} \left\| \nabla \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \right\|_\infty^2 \right) \ln(2d)} + U \left(8 \ln(2d) + 12\right) \max_{1 \leqslant t \leqslant T} \left\| \nabla \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \right\|_\infty \,.\end{aligned} \tag{4.49}$$

We can now follow the same lines as in Corollary 2.2, except that we use the particular shape of the Lipschitzified losses. First note from (4.11) that $\max_{1 \leqslant t \leqslant T} \|\nabla \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t)\|_\infty \leqslant 2\big(1 + \sqrt{2}\big) XY$.

Moreover, using (4.9) and the definition of $\widehat{y}_t$ in Figure 4.3, we can see that the gradients satisfy $\nabla \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) = -2\big(y_t - [\widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t]_{B_t}\big)\, \boldsymbol{x}_t = -2(y_t - \widehat{y}_t)\, \boldsymbol{x}_t$. Combining the last equality with the upper bound $\|\boldsymbol{x}_t\|_\infty \leqslant X$, we get that

$$\left\| \nabla \widetilde{\ell}_t(\widehat{\boldsymbol{u}}_t) \right\|_\infty^2 \leqslant 4X^2 (y_t - \widehat{y}_t)^2 \,.$$

Substituting the last two inequalities in (4.49) and combining the resulting bound with (4.48), we get

$$\begin{aligned}\sum_{t=1}^{T} \big(y_t - \widehat{y}_t\big)^2 \leqslant \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T} \widetilde{\ell}_t(\boldsymbol{u}) + 8UX \sqrt{\left(\sum_{t=1}^{T} \big(y_t - \widehat{y}_t\big)^2 \right) \ln(2d)} \\ + \big(16 \ln(2d) + 24\big)\big(1 + \sqrt{2}\big) UXY + 4\big(1 + 1/\sqrt{2}\big)^2 Y^2 \,.\end{aligned}$$

Setting $C \triangleq \big(16 \ln(2d) + 24\big)\big(1 + \sqrt{2}\big) UXY + 4\big(1 + 1/\sqrt{2}\big)^2 Y^2$, $\widehat{L}_T \triangleq \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2$, and $\widetilde{L}_T^* \triangleq \min_{\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_1 \leqslant U\}} \sum_{t=1}^{T} \widetilde{\ell}_t(\boldsymbol{u})$, the previous inequality can be simply rewritten as

$$\widehat{L}_T \leqslant \widetilde{L}_T^* + C + 8UX \sqrt{\widehat{L}_T \ln(2d)} \,.$$

Solving for $\widehat{L}_T$ via Lemma A.2 in Appendix A.4, we get that

$$\widehat{L}_T \leqslant \widetilde{L}_T^* + C + \Big(8UX \sqrt{\ln(2d)}\Big) \sqrt{\widetilde{L}_T^* + C} + \Big(8UX \sqrt{\ln(2d)}\Big)^2$$

$$\leqslant \widetilde{L}_T^* + 8UX\sqrt{\widetilde{L}_T^* \ln(2d)} + 8UX\sqrt{C\ln(2d)} + 64U^2X^2\ln(2d) + C\,. \tag{4.50}$$

But, rewriting $C = C_1 + C_2$ with $C_1 \triangleq \big(16\ln(2d) + 24\big)\big(1+\sqrt{2}\big)UXY$ and $C_2 \triangleq 4\big(1 + 1/\sqrt{2}\big)^2 Y^2$, we get that

$$\begin{aligned} UX\sqrt{C\ln(2d)} &\leqslant UX\sqrt{C_1\ln(2d)} + UX\sqrt{C_2\ln(2d)} \\ &= UX\sqrt{C_1\ln(2d)} + 2\big(1+1/\sqrt{2}\big)UXY\sqrt{\ln(2d)}\,, \end{aligned} \tag{4.51}$$

where

$$\begin{aligned} UX\sqrt{C_1\ln(2d)} &= UX\ln(2d)\sqrt{\big(16 + 24/\ln(2d)\big)\big(1+\sqrt{2}\big)\,UXY} \\ &\leqslant \sqrt{U^2X^2 + UXY}\,\ln(2d)\,\sqrt{\big(16+24/\ln(2)\big)\big(1+\sqrt{2}\big)\big(UXY + U^2X^2\big)} \\ &= \sqrt{\big(16+24/\ln(2)\big)\big(1+\sqrt{2}\big)}\,\big(UXY + U^2X^2\big)\ln(2d)\,. \end{aligned}$$

Combining (4.50) with (4.51) and the last inequality and performing some simple upper bounds, we conclude the proof. □

4.B Lemmas

The next lemma is useful to prove Theorem 4.1. At the end of this section, we also provide an elementary lemma about the exponentially weighted average forecaster combined with clipping.

Lemma 4.4. *Let $d, T \in \mathbb{N}^*$, and $U, X, Y > 0$. The minimax regret on $B_1(U)$ for bounded base predictions and observations satisfies*

$$\begin{aligned} &\inf_F \sup_{\|\boldsymbol{x}_t\|_\infty \leqslant X,\, |y_t| \leqslant Y} \left\{ \sum_{t=1}^T (y_t - \widehat{y}_t)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \right\} \\ &\qquad \leqslant \min\left\{ 3UXY\sqrt{2T\ln(2d)},\, 32\,dY^2\ln\left(1 + \frac{\sqrt{T}UX}{dY}\right) + dY^2 \right\}\,, \end{aligned}$$

where the infimum is taken over all forecasters F and where the supremum extends over all sequences $(\boldsymbol{x}_t, y_t)_{1\leqslant t\leqslant T} \in (\mathbb{R}^d \times \mathbb{R})^T$ such that $|y_1|, \dots, |y_T| \leqslant Y$ and $\|\boldsymbol{x}_1\|_\infty, \dots, \|\boldsymbol{x}_T\|_\infty \leqslant X$.

Proof: We treat each of the two terms in the above minimum separately.

Step 1: We prove that their exists a forecaster F whose worst-case regret on $B_1(U)$ is upper bounded by $3UXY\sqrt{2T\ln(2d)}$.

First note that if $U \geqslant (Y/X)\sqrt{T/(2\ln(2d))}$, then the upper bound $3UXY\sqrt{2T\ln(2d)} \geqslant 3TY^2 \geqslant TY^2$ is trivial (by choosing the forecaster F which outputs $\widehat{y}_t = 0$ at each time t).

We can thus assume that $U < (Y/X)\sqrt{T/(2\ln(2d))}$. Consider the EG$^\pm$ algorithm as given in [KW97, Theorem 5.11], and denote by $\widehat{\boldsymbol{u}}_t \in B_1(U)$ the linear combination it outputs at each time

$t \geqslant 1$. Then, by the aforementioned theorem, this forecaster satisfies, uniformly over all individual sequences bounded by X and Y, that

$$\begin{aligned}
&\sum_{t=1}^{T}(y_t - \widehat{\boldsymbol{u}}_t \cdot \boldsymbol{x}_t)^2 - \inf_{\|\boldsymbol{u}\|_1 \leqslant U} \sum_{t=1}^{T}(y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \\
&\quad \leqslant 2UXY\sqrt{2T\ln(2d)} + 2U^2X^2\ln(2d) \\
&\quad \leqslant 2UXY\sqrt{2T\ln(2d)} + 2\left(Y\sqrt{\frac{T}{2\ln(2d)}}\right)UX\ln(2d) \qquad (4.52) \\
&\quad \leqslant 3UXY\sqrt{2T\ln(2d)}\,,
\end{aligned}$$

where (4.52) follows from the assumption $UX < Y\sqrt{T/(2\ln(2d))}$. This concludes the first step of this proof.

Step 2: We prove that their exists a forecaster F whose worst-case regret on $B_1(U)$ is upper bounded by $32\,dY^2\ln\!\left(1 + \frac{\sqrt{T}UX}{dY}\right) + dY^2$.

Such a forecaster is given by the algorithm $\mathrm{SeqSEW}_{\tau}^{B,\eta}$ of Section 3.3.1 (Chapter 3) tuned with $B = Y$, $\eta = 1/(8Y^2)$ and $\tau = Y/(\sqrt{T}X)$. Indeed, by Proposition 3.1 of Chapter 3, the cumulative square loss of this algorithm is upper bounded by

$$\begin{aligned}
&\inf_{\boldsymbol{u}\in\mathbb{R}^d}\left\{\sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2 + 32\,\|\boldsymbol{u}\|_0\,Y^2\ln\!\left(1 + \frac{\sqrt{T}X\,\|\boldsymbol{u}\|_1}{\|\boldsymbol{u}\|_0\,Y}\right)\right\} + dY^2 \\
&\leqslant \inf_{\|\boldsymbol{u}\|_1\leqslant U}\left\{\sum_{t=1}^{T}\big(y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t\big)^2\right\} + 32dY^2\ln\!\left(1 + \frac{\sqrt{T}XU}{dY}\right) + dY^2\,,
\end{aligned}$$

where the last inequality follows by monotonicity[15] in $\|\boldsymbol{u}\|_0$ and $\|\boldsymbol{u}\|_1$ of the second term of the left-hand side. This concludes the proof. □

Next we recall a regret bound satisfied by the standard exponentially weighted average forecaster applied to clipped base forecasts. Assume that at each time $t \geqslant 1$, the forecaster has access to $K \geqslant 1$ base forecasts $\widehat{y}_t^{(k)} \in \mathbb{R}$, $k = 1, \ldots, K$, and that for some known bound $Y > 0$ on the observations, the forecaster predicts at time t as

$$\widehat{y}_t \triangleq \sum_{k=1}^{K} p_{k,t}\big[\widehat{y}_t^{(k)}\big]_Y\,.$$

In the equation above, $[x]_Y \triangleq \min\{Y, \max\{-Y, x\}\}$ for all $x \in \mathbb{R}$, and the weight vectors

[15] Note that for all $A > 0$, the function $x \mapsto x\ln(1 + A/x)$ (continuously extended at $x = 0$) has a nonnegative first derivative and is thus nondecreasing on $\mathbb{R}_+$.

$\boldsymbol{p}_t \in \mathbb{R}^K$ are given by $\boldsymbol{p}_1 = (1/K, \dots, 1/K)$ and, for all $t = 2, \dots, T$, by

$$p_{k,t} \triangleq \frac{\exp\left(-\eta \sum_{s=1}^{t-1} \left(y_s - \left[\widehat{y}_s^{(k)}\right]_Y\right)^2\right)}{\sum_{j=1}^K \exp\left(-\eta \sum_{s=1}^{t-1} \left(y_s - \left[\widehat{y}_s^{(j)}\right]_Y\right)^2\right)}, \quad 1 \leqslant k \leqslant K,$$

for some inverse temperature parameter $\eta > 0$ to be chosen below. The next lemma is a straigthforward consequence of Theorem 2.2 in Chapter 2.

Lemma 4.5 (Exponential weighting with clipping)**.** *Assume that the forecaster knows beforehand a bound $Y > 0$ on the observations $|y_t|$, $t = 1, \dots, T$. Then, the exponentially weighted average forecaster tuned with $\eta \leqslant 1/(8Y^2)$ and with clipping $[\,\cdot\,]_Y$ satisfies*

$$\sum_{t=1}^T \left(y_t - \widehat{y}_t\right)^2 \leqslant \min_{1 \leqslant k \leqslant K} \sum_{t=1}^T \left(y_t - \widehat{y}_t^{(k)}\right)^2 + \frac{\ln K}{\eta}.$$

Proof (of Lemma 4.5): The proof follows straightforwardly from Theorem 2.2 in Chapter 2. To apply the latter result, recall from Appendix A.2 that the square loss is $1/(8Y^2)$-exp-concave on $[-Y, Y]$ and thus η-exp-concave[16] (since $\eta \leqslant 1/(8Y^2)$ by assumption). Therefore, by definition of our forecaster above, Theorem 2.2 yields

$$\sum_{t=1}^T \left(y_t - \widehat{y}_t\right)^2 \leqslant \min_{1 \leqslant k \leqslant K} \sum_{t=1}^T \left(y_t - \left[\widehat{y}_t^{(k)}\right]_Y\right)^2 + \frac{\ln K}{\eta}.$$

To conclude the proof, note for all $t = 1, \dots, T$ and $k = 1, \dots, K$ that $|y_t| \leqslant Y$ by assumption, so that clipping the base forecasts to $[-Y, Y]$ can only improve prediction, i.e., $\left(y_t - \left[\widehat{y}_t^{(k)}\right]_Y\right)^2 \leqslant \left(y_t - \widehat{y}_t^{(k)}\right)^2$. □

4.C Additional tools

The next approximation argument is originally due to Maurey, and was used under various forms, e.g., in [Nem00, Tsy03, BN08, SSSZ10].

Lemma 4.6 (Approximation argument)**.** *Let $U > 0$ and $m \in \mathbb{N}^*$. Define the following finite subset of $B_1(U)$:*

$$\widetilde{B}_{U,m} \triangleq \left\{ \left(\frac{k_1 U}{m}, \dots, \frac{k_d U}{m} \right) : (k_1, \dots, k_d) \in \mathbb{Z}^d, \sum_{j=1}^d |k_j| \leqslant m \right\} \subset B_1(U).$$

Then, for all $(\boldsymbol{x}_t, y_t)_{1 \leqslant t \leqslant T} \in \left(\mathbb{R}^d \times \mathbb{R}\right)^T$ such that $\max_{1 \leqslant t \leqslant T} \|\boldsymbol{x}_t\|_\infty \leqslant X$,

$$\inf_{\boldsymbol{u} \in \widetilde{B}_{U,m}} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 \leqslant \inf_{\boldsymbol{u} \in B_1(U)} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 + \frac{TU^2X^2}{m}.$$

[16] This means that for all $y \in [-Y, Y]$, the function $x \mapsto \exp\left(-\eta(y-x)^2\right)$ is concave on $[-Y, Y]$.

Proof: The proof is quite standard and follows the same lines as [Nem00, Proposition 5.2.2] or [BN08, Theorem 2] who addressed the aggregation task in the stochastic setting. We rewrite this argument below in our online deterministic setting.

Fix $\boldsymbol{u}^* \in \operatorname{argmin}_{\boldsymbol{u}\in B_1(U)} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2$. Define the probability distribution $\pi = (\pi_{-d}, \dots, \pi_d) \in \mathbb{R}_+^{2d+1}$ by

$$\pi_j \triangleq \begin{cases} \dfrac{(u_j^*)_+}{U} & \text{if } j \geqslant 1; \\ \dfrac{(u_j^*)_-}{U} & \text{if } j \leqslant -1; \\ 1 - \displaystyle\sum_{j=1}^d \frac{|u_j^*|}{U} & \text{if } j = 0\,. \end{cases}$$

Let $J_1, \dots, J_m \in \{-d, \dots, d\}$ be i.i.d. random integers drawn from π, and set

$$\widetilde{\boldsymbol{u}} \triangleq \frac{U}{m} \sum_{k=1}^m \boldsymbol{e}_{J_k}\,,$$

where $(\mathbf{e}_j)_{1\leqslant j\leqslant d}$ is the canonical basis of $\mathbb{R}^d$, where $\boldsymbol{e}_0 \triangleq \mathbf{0}$, and where $\boldsymbol{e}_{-j} \triangleq -\boldsymbol{e}_j$ for all $1 \leqslant j \leqslant d$. Note that $\widetilde{\boldsymbol{u}} \in \widetilde{B}_{U,m}$ by construction. Therefore,

$$\inf_{\boldsymbol{u}\in\widetilde{B}_{U,m}} \sum_{t=1}^T (y_t - \boldsymbol{u}\cdot\boldsymbol{x}_t)^2 \leqslant \mathbb{E}\left[\sum_{t=1}^T (y_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t)^2\right]\,. \tag{4.53}$$

The rest of the proof is dedicated to upper bounding the last expectation. Expanding all the squares $(y_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t)^2 = (y_t - \boldsymbol{u}^*\cdot\boldsymbol{x}_t + \boldsymbol{u}^*\cdot\boldsymbol{x}_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t)^2$, first note that

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T (y_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t)^2\right] &= \sum_{t=1}^T (y_t - \boldsymbol{u}^*\cdot\boldsymbol{x}_t)^2 + \sum_{t=1}^T \mathbb{E}\big[(\boldsymbol{u}^*\cdot\boldsymbol{x}_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t)^2\big] \\ &\quad + 2\sum_{t=1}^T (y_t - \boldsymbol{u}^*\cdot\boldsymbol{x}_t)\,\mathbb{E}\big[\boldsymbol{u}^*\cdot\boldsymbol{x}_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t\big]\,. \end{aligned}\tag{4.54}$$

But by definition of $\widetilde{\boldsymbol{u}}$ and π,

$$\begin{aligned}\mathbb{E}\big[\widetilde{\boldsymbol{u}}\big] &= U\,\mathbb{E}\big[\boldsymbol{e}_{J_1}\big] = U\sum_{j=-d}^d \pi_j \boldsymbol{e}_j \\ &= U\sum_{j=1}^d \left(\frac{(u_j^*)_+}{U}\boldsymbol{e}_j + \frac{(u_j^*)_-}{U}(-\boldsymbol{e}_j)\right) = U\sum_{j=1}^d \frac{u_j^*}{U}\boldsymbol{e}_j = \boldsymbol{u}^*\,,\end{aligned}$$

so that $\mathbb{E}\big[\widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t\big] = \boldsymbol{u}^*\cdot\boldsymbol{x}_t$ for all $1 \leqslant t \leqslant T$. Therefore, the last sum in (4.54) above equals zero, and

$$\mathbb{E}\Big[\big(\boldsymbol{u}^*\cdot\boldsymbol{x}_t - \widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t\big)^2\Big] = \operatorname{Var}\big(\widetilde{\boldsymbol{u}}\cdot\boldsymbol{x}_t\big) = \frac{U^2}{m^2}\sum_{k=1}^m \operatorname{Var}\big(\boldsymbol{e}_{J_k}\cdot\boldsymbol{x}_t\big) \leqslant \frac{U^2X^2}{m}\,,$$

where the second equality follows from $\widetilde{\boldsymbol{u}} \cdot \boldsymbol{x}_t = (U/m) \sum_{k=1}^m \boldsymbol{e}_{J_k} \cdot \boldsymbol{x}_t$ and from the independence of the J_k, $1 \leqslant k \leqslant m$, and where the last inequality follows from $|\boldsymbol{e}_{J_k} \cdot \boldsymbol{x}_t| \leqslant \|\boldsymbol{e}_{J_k}\|_1 \, \|\boldsymbol{x}_t\|_\infty \leqslant X$ for all $1 \leqslant k \leqslant m$.

Combining (4.54) with the remarks above, we get

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^T (y_t - \widetilde{\boldsymbol{u}} \cdot \boldsymbol{x}_t)^2\right] &\leqslant \sum_{t=1}^T (y_t - \boldsymbol{u}^* \cdot \boldsymbol{x}_t)^2 + \frac{TU^2X^2}{m} \\ &= \inf_{\boldsymbol{u} \in B_1(U)} \sum_{t=1}^T (y_t - \boldsymbol{u} \cdot \boldsymbol{x}_t)^2 + \frac{TU^2X^2}{m} \,,\end{aligned}$$

where the last line follows by definition of $\boldsymbol{u}^*$. Substituting the last inequality in (4.53) concludes the proof. □

The combinatorial result below (or variants of it) is well-known; see, e.g., [Tsy03, BN08]. We reproduce its proof for the convenience of the reader. We use the notation $\mathrm{e} \triangleq \exp(1)$.

Lemma 4.7 (An elementary combinatorial upper bound)**.**
Let $m, d \in \mathbb{N}^*$. Denoting by $|E|$ the cardinality of a set E, we have

$$\left|\left\{(k_1, \dots, k_d) \in \mathbb{Z}^d : \sum_{j=1}^d |k_j| \leqslant m\right\}\right| \leqslant \left(\frac{\mathrm{e}(2d+m)}{m}\right)^m .$$

Proof (of Lemma 4.7): Setting $(k'_{-j}, k'_j) \triangleq \big((k_j)_-, (k_j)_+\big)$ for all $1 \leqslant j \leqslant d$, and $k'_0 \triangleq m - \sum_{j=1}^d |k_j|$, we have

$$\begin{aligned}&\left|\left\{(k_1, \dots, k_d) \in \mathbb{Z}^d : \sum_{j=1}^d |k_j| \leqslant m\right\}\right| \\ &\quad\leqslant \left|\left\{(k'_{-d}, \dots, k'_d) \in \mathbb{N}^{2d+1} : \sum_{j=-d}^d k'_j = m\right\}\right| \\ &\quad= \binom{2d+m}{m} \qquad (4.55) \\ &\quad\leqslant \left(\frac{\mathrm{e}(2d+m)}{m}\right)^m . \qquad (4.56)\end{aligned}$$

To get inequality (4.55), we used the (elementary) fact that the number of $2d+1$ integer-valued tuples summing up to m is equal to the number of lattice paths from $(1,0)$ to $(2d+1, m)$ in $\mathbb{N}^2$, which is equal to $\binom{2d+1+m-1}{m}$. As for inequality (4.56), it follows straightforwardly from a classical combinatorial result stated, e.g., in [Mas07, Proposition 2.5]. □

Chapter 5

Minimax rates of internal and swap regrets

Within the framework of prediction with expert advice under linear losses, we study the minimax rates of two performance criteria related to game theory: internal regret and swap regret. We first prove the exact rates $\sqrt{T}$ and $\sqrt{T \ln K}$ respectively for internal and swap regrets when the loss vectors are i.i.d.. This shows that the missing $\sqrt{\ln K}$ factor between the known upper and lower bounds of [SL05] and [Sto05] on internal regret is unnecessary in the stochastic i.i.d. setting. Second, in the game with arbitrary deterministic loss vectors, we provide a lower bound of order $\sqrt{TK}$ on the swap regret; it improves on a lower bound of [BM07b]. Finally, we develop a generic technique that enables to reinterpret known deterministic regret bounds from a stochastic viewpoint, but also to derive a new regret bound in the problem of learning with global cost functions.

NOTA: A large part of this chapter was presented at the conferences 42èmes Journées de Statistique [Ger10b] and StatMathAppli 2010 [Ger10a]. Since then, [RST11] published an independent work that has significant overlaps with Section 5.5. However, some important questions remain open (see Section 5.1.2).

Contents

5.1 Introduction

In this chapter, we consider a decision-theoretic variant[1] of the framework of prediction with expert advice due to [FS97]. The problem is stated as a repeated game between a forecaster and an environment. At each time round $t \in \mathbb{N}^* = \{1, 2, \dots\}$, the forecaster chooses a weight vector $\boldsymbol{p}_t = (p_{1,t}, \dots, p_{K,t})$ over $K \geqslant 2$ different actions, i.e., $\boldsymbol{p}_t$ belongs to the simplex $\mathcal{X}_K \triangleq \{\boldsymbol{x} \in \mathbb{R}_+^K, \quad \sum_{i=1}^K x_i = 1\}$. The environment then reveals a loss vector $\boldsymbol{\ell}_t \triangleq (\ell_{i,t})_{1 \leqslant i \leqslant K}$ in $[0,1]^K$; each action $i \in \{1, \dots, K\}$ incurs the loss $\ell_{i,t}$ and the forecaster incurs the linear loss $\boldsymbol{p}_t \cdot \boldsymbol{\ell}_t = \sum_{i=1}^K p_{i,t}\, \ell_{i,t}$. After $T \geqslant 1$ time rounds, the cumulative loss of the forecaster equals $\sum_{t=1}^T \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t$, and his primary goal is to minimize it. In the sequel, we assume for simplicity that the loss sequence $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T$ is fixed in advance by the environment. However, by Section 2.3.1 (cf. Chapter 2), since we only consider deterministic forecasters, all upper bounds stated for *individual* sequences also hold true for *adversarial* environments.

The weight vectors $\boldsymbol{p}_t$ are chosen on the basis of the past loss vectors and can therefore be seen as values of functions $\boldsymbol{p}_t(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t-1})$. We call *strategy (of the forecaster)* any sequence $(\boldsymbol{p}_t)_{t \geqslant 1}$ of Borel functions $\boldsymbol{p}_t : [0,1]^{K(t-1)} \to \mathcal{X}_K$. For notational convenience, we often omit the dependency in $(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t-1})$ and write $\boldsymbol{p}_t$ instead of $\boldsymbol{p}_t(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t-1})$.

A classical way to assess the quality of a strategy $S = (\boldsymbol{p}_t)_{t \geqslant 1}$ on a finite loss sequence $\boldsymbol{\ell}_{1:T} \triangleq (\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T)$ is to compare its cumulative loss to that of the best action in hindsight. The difference between these two quantities, i.e.,

$$R_T^{\mathrm{ext}}(S, \boldsymbol{\ell}_{1:T}) \triangleq \sum_{t=1}^T \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{1 \leqslant j \leqslant K} \sum_{t=1}^T \ell_{j,t}\,, \tag{5.1}$$

is called the *external regret* of the forecaster and measures his difficulty to mimic the best action in hindsight while being compelled to output decisions in a sequential fashion. This performance criterion was introduced in Chapter 2 and studied in Chapters 3 and 4.

In this chapter, we study two stronger notions of regret called *internal regret* and *swap regret*, which play an important role in the theory of repeated games. Like for the external regret, the cumulative loss of the forecaster is compared to that of the best strategy (in hindsight) in a given reference class. However, the reference strategies are not external as for the external regret, but are instead given by consistent modifications of the forecaster's own strategy — hence the term *internal*.

The notion of internal regret was first studied by [FV97, FV98, FV99] (see also [FL99, HMC00, HMC01]). For any strategy $S = (\boldsymbol{p}_t)_{t \geqslant 1}$ and any finite loss sequence $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$, the

[1] This prediction protocol corresponds to the framework of prediction with expert advice described in Figure 2.1 (Chapter 2) with $\mathcal{D} = \mathcal{X}_K$, $\mathcal{Y} = [0,1]^K$, the linear loss $(\boldsymbol{p}, \boldsymbol{\ell}) \in \mathcal{X}_K \times [0,1]^K \mapsto \boldsymbol{p} \cdot \boldsymbol{\ell}$, and constant expert advice $\boldsymbol{a}_{i,t} = \boldsymbol{\delta}_i \triangleq (\mathbb{I}_{\{j=i\}})_{1 \leqslant j \leqslant K}$, $i = 1, \dots, K$, $t = 1, \dots, T$.

internal regret $R_T^{\mathrm{int}}(S, \boldsymbol{\ell}_{1:T})$ of the strategy S on the sequence $\boldsymbol{\ell}_{1:T} \triangleq (\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T)$ is defined by

$$R_T^{\mathrm{int}}(S, \boldsymbol{\ell}_{1:T}) \triangleq \sum_{t=1}^{T} \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^{T} \boldsymbol{p}_t^{i \to j} \cdot \boldsymbol{\ell}_t \,, \tag{5.2}$$

where the modified weight vector $\boldsymbol{p}_t^{i \to j} \in \mathcal{X}_K$ is obtained from $\boldsymbol{p}_t$ by replacing action i with action j. Namely, for all $k = 1, \dots, K$, the k-th component of $\boldsymbol{p}_t^{i \to j}$ equals

$$(p_t^{i \to j})_k = \begin{cases} 0 & \text{if } k = i, \\ p_{i,t} + p_{j,t} & \text{if } k = j, \\ p_{k,t} & \text{if } k \notin \{i, j\}. \end{cases} \tag{5.3}$$

Internal regret thus measures for all pairs (i, j), $i \neq j$, the regret the forecaster feels for not choosing action j each time he chose action i (all loss vectors being equal). Intuitively, if a forecaster has a small internal regret then he enjoys some stability properties. This has indeed been illustrated in game theory: [FV97, FV99] showed that if all players of a finite randomized game choose a strategy whose internal regret is almost surely sublinear in T, then the joint empirical frequencies of play converge almost surely to an equilibrium set called the set of correlated equilibria (see also [FL95, HMC00, SL07]). Internal regret also has some historical connections with another branch of game theory called calibration: the existence of strategies with sublinear internal regret implies the existence of calibrated forecasters (see [FV98] and the other references given in [CBL06, Chapter 4]).

The notion of swap regret was introduced by [BM07b] (see also [GJ03] for the broader notion of Φ-regret). The *swap regret* of a strategy is larger than its internal regret, since the pool of modified strategies $\{(\boldsymbol{p}_t^{i \to j})_{t \geqslant 1}, i \neq j\}$ to which the forecaster's strategy is compared is extended to all linear modifications $\{(\varphi(\boldsymbol{p}_t))_{t \geqslant 1}\}$, where φ extends over all linear mappings from the simplex $\mathcal{X}_K$ into itself. As is done in [Sto05, Chapter 3] via the Krein-Millman theorem, we can equivalently define the *swap regret* $R_T^{\mathrm{sw}}(S, \boldsymbol{\ell}_{1:T})$ of a strategy $S = (\boldsymbol{p}_t)_{t \geqslant 1}$ on a finite loss sequence $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0, 1]^K$ by

$$R_T^{\mathrm{sw}}(S, \boldsymbol{\ell}_{1:T}) \triangleq \sum_{t=1}^{T} \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F \in \mathcal{F}_K} \sum_{t=1}^{T} \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t \,, \tag{5.4}$$

where $\mathcal{F}_K$ denotes the set of all mappings $F : \{1 \dots, K\} \to \{1 \dots, K\}$ and where the modified weight vector $\boldsymbol{p}_t^F \in \mathcal{X}_K$ is obtained from $\boldsymbol{p}_t$ by replacing each action i with the action $F(i)$. Namely, for $1 \leqslant j \leqslant K$, the j-th component of $\boldsymbol{p}_t^F$ is defined by

$$(p_t^F)_j = \sum_{i : F(i) = j} p_{i,t} \,. \tag{5.5}$$

5.1.1 Known upper and lower bounds on internal and swap regrets

The existence of strategies with a small (sublinear in T) internal regret on individual sequences was first shown by [FV97]; see also [FL99, HMC00, HMC01]. More precisely, [FV97] designed an algorithm whose internal regret on all individual sequences $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0, 1]^K$ is upper bounded

by $\sqrt{2KT}$. The dependence in T is correct (see below), but the dependence in K is far from being optimal. Indeed, using an argument of [HMC01], [CBL03] proved the upper bound $2\sqrt{T \ln K}$ for a suitable exponentially weighted average forecaster. This bound was later lowered by [SL05] to $\sqrt{T \ln K}$ via a new analysis based on a fixed-point property. Therefore, up to now, the best[2] upper bound on the *minimax internal regret for individual sequences* reads:

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in [0,1]^K} R_T^{\mathrm{int}}(S, \boldsymbol{\ell}_{1:T}) \leqslant \sqrt{T \ln K}\,, \tag{5.6}$$

where the infimum is taken over all strategies $S = (\boldsymbol{p}_t)_{t \geqslant 1}$ of the forecaster. As shown by [Sto05], the last upper bound cannot be improved more than by a logarithmic factor $\sqrt{\ln K}$. Indeed, it is proved in [Sto05, Theorem 3.3] that, for all $K \geqslant 2$ and all $T \geqslant K^2/192$, the *minimax internal regret for i.i.d. loss vectors* is bounded from below by

$$\inf_S \sup_Q \mathbb{E}_{Q^{\otimes T}} \left[R_T^{\mathrm{int}}(S, \boldsymbol{\ell}_{1:T}) \right] \geqslant \sqrt{T}/\left(64\sqrt{3}\right)\,, \tag{5.7}$$

where the infimum is taken over all strategies $S = (\boldsymbol{p}_t)_{t \geqslant 1}$, where the supremum is taken over all probability distributions on $[0,1]^K$ (endowed with its Borel σ-algebra), and where the loss vectors $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$ are i.i.d. with common distribution Q. Since the minimax internal regret for individual sequences is larger than the minimax internal regret with i.i.d. loss vectors, the inequalities above show that the orders of magnitude in T and K of both minimax quantities lie between $\sqrt{T}$ and $\sqrt{T \ln K}$. We note a missing $\sqrt{\ln K}$ factor between the lower and upper bounds.

In the same spirit, [BM07b] proved that there exists an (efficient) algorithm whose swap regret is upper bounded by $\sqrt{(T/2)K \ln K}$ uniformly over all individual sequences; see also [SL05] for an alternative proof with a less efficient algorithm (of combinatorial nature). Therefore, as of now, the best upper bound on the *minimax swap regret for individual sequences* reads:

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in [0,1]^K} R_T^{\mathrm{sw}}(S, \boldsymbol{\ell}_{1:T}) \leqslant \sqrt{(T/2)K \ln K}\,.$$

The last upper bound was shown to be almost optimal by [BM07b], who exhibited a lower bound of the order of $\sqrt{TK}$. Their lower bound has however two limitations: first, it is proved in a randomized and adversarial setting for a quantity larger than the swap regret *stricto sensu*. Second, it is proved only for rounds T that are sub-exponential in K. See Section 5.4 for further details.

5.1.2 Main contributions

The main contributions of this chapter are the following. The first one is related to the stochastic protocol (i.i.d. loss vectors) while the other two ones are related to the deterministic protocol (arbitrary loss vectors).

- In the stochastic protocol, we derive the exact minimax rates of internal regret and swap regret for i.i.d. loss vectors. Using the same notations as above, we show that they are

[2] To be exact, the best upper bound known so far equals $\sqrt{\frac{T}{2} \ln[K(K-1)]}$ which is smaller but close to $\sqrt{T \ln K}$.

respectively of the order[3] of $\sqrt{T}$ and $\sqrt{T \ln K}$:

$$\inf_S \sup_Q \mathbb{E}_{Q^{\otimes T}}\Big[R_T^{\mathrm{int}}(S, \boldsymbol{\ell}_{1:T})\Big] \asymp \sqrt{T} \quad \text{and} \quad \inf_S \sup_Q \mathbb{E}_{Q^{\otimes T}}\Big[R_T^{\mathrm{sw}}(S, \boldsymbol{\ell}_{1:T})\Big] \asymp \sqrt{T \ln K}\,.$$

In particular, when the loss vectors are i.i.d., the optimal rate of internal regret is independent of the ambient dimension K.

- We prove a lower bound of the order of $\sqrt{TK}$ on the minimax swap regret for individual sequences: setting $c \triangleq 1/\big(16\sqrt{128\ln(4/3)}\,\big)$, we show that, for all $K \geqslant 2$ and all $T \geqslant \max\big\{128c^2K^5, K\big\}$,

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in [0,1]^K} R_T^{\mathrm{sw}}(S, \boldsymbol{\ell}_{1:T}) \geqslant c\sqrt{TK}\,.$$

This lower bound is stronger than that of [BM07b, Theorem 9] since it holds for the swap regret itself instead of a randomized variant of it (see Section 5.4). This solves a question left open in [BM07b, Section 9]. Besides, we do not need their assumption that T be sub-exponential in K.

Moreover, our lower bound of order $\sqrt{TK}$ highlights a major difference between external and swap regrets. On the one hand, as recalled in Chapter 2, the external regret behaves similarly on i.i.d. loss vectors and on individual sequences. Indeed, combining Theorem 2.1 and Lemma 2.2 therein (cf. pages 46 and 62), we get that, using the same notations as above, for all $K \geqslant 1$ and all $T \geqslant \big[40e/(2e+1)\big]\ln K$,

$$\begin{aligned}\frac{2}{2e+1}\sqrt{\frac{e\,T\ln K}{5(2e+1)}} &\leqslant \inf_S \sup_Q \mathbb{E}_{Q^{\otimes T}}\Big[R_T^{\mathrm{ext}}(S, \boldsymbol{\ell}_{1:T})\Big] \\ &\leqslant \inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in [0,1]^K} R_T^{\mathrm{ext}}(S, \boldsymbol{\ell}_{1:T}) \leqslant \sqrt{\frac{T}{2}\ln K}\,.\end{aligned}$$

On the other hand, contrary to external regret, swap regret is much harder to minimize with individual sequences than with i.i.d. losses (compare the rates $\sqrt{TK}$ and $\sqrt{T\ln K}$ above).

- We develop a stochastic technique to derive upper bounds on a generalized form of regret including external, internal, and swap regrets. This technique provides a new insight on the rates of these three types of regret and can be used to recover the best upper bounds known so far. We also derive — in a non-constructive way — a new upper bound of order $\sqrt{T\ln K}$ on the makespan regret, thus improving on the known bound of order $\ln(K)\sqrt{T}$ of [EDKMM09].

As is detailed in Section 5.5.1, a similar stochastic technique was independently studied in [RST11]. Since we work in a much more specific setting, we are able to get (sometimes tight) explicit constants. Our proof relies on arguments such as Bernoullization and an elementary maximal inequality for subgaussian random variables.

This work in progress raises some important questions. First, though the aforementioned stochastic technique is useful to better understand the problem at hand (since it provides an upper

[3] We write $a_{T,K} \asymp b_{T,K}$ if and only if there exist two absolute constants $c_1, c_2 > 0$ and a sequence $(t_K)_{K\geqslant 1}$ in $\mathbb{N}^*$ such that $c_1 b_{T,K} \leqslant a_{T,K} \leqslant c_2 b_{T,K}$ for all $K \geqslant 1$ and all $T \geqslant t_K$.

bound on the minimax regret), it is not constructive. Designing explicit algorithms that achieve the obtained upper bounds is an important task to be addressed in the future (e.g., is there any efficient algorithm with a makespan regret at most of order $\sqrt{T \ln K}$?). Note that the same issue arises in [RST11]. Second, the question of the missing logarithmic factor $\sqrt{\ln K}$ between the lower and upper bounds on the internal regret is still (partially) open. We did prove that the $\sqrt{\ln K}$ factor is unnecessary for i.i.d. loss vectors, but we still do not know whether this is also the case for arbitrary deterministic loss vectors.

Outline of the chapter

The rest of the chapter is organized as follows. In Section 5.2 we formally describe our setting and notations and state some basic facts. In Section 5.3 we derive the $\sqrt{T}$ minimax rate of internal regret in a stochastic environment. In Section 5.4 we prove a lower bound of the order of $\sqrt{TK}$ on the minimax swap regret with individual sequences, together with the optimal rate $\sqrt{T \ln K}$ for the minimax swap regret with i.i.d. loss vectors. The results in Sections 5.3 and 5.4 are obtained in a constructive way. In Section 5.5 we provide a general (non-constructive) stochastic technique to derive upper bounds with individual sequences on the quantities studied before and on other ones. Finally, some technical proofs can be found in Section 5.A while some elementary lemmas are provided in Section 5.B.

5.2 Setting, notations, and basic properties

5.2.1 Setting and notations

We give in Figure 5.1 a formal description of our repeated game. We consider two different assumptions on the way the loss vectors $\boldsymbol{\ell}_t$ are chosen before the beginning of the game: the $\boldsymbol{\ell}_t$ can either be drawn by a stochastic environment or they can form an individual sequence.

In the sequel, we denote by $\mathcal{X}_K \triangleq \big\{\boldsymbol{x} \in \mathbb{R}_+^K : \sum_{i=1}^K x_i = 1\big\}$ the simplex of order K and by $\mathcal{F}_K$ the set of all functions from $\{1, \dots, K\}$ to $\{1, \dots, K\}$. We also set, for all $x \in \mathbb{R}$,

$$\lfloor x \rfloor \triangleq \sup\big\{k \in \mathbb{Z} : k \leqslant x\big\} \qquad \text{and} \qquad \lceil x \rceil \triangleq \inf\big\{k \in \mathbb{Z} : k \geqslant x\big\} \,.$$

5.2.2 Basic properties

Next we recall some basic properties of external, internal, and swap regrets and well-known inequalities to compare them.

Equivalent definitions of external, internal, and swap regrets

In view of (5.1), the external regret can be rewritten as follows:

$$R_T^{\mathrm{ext}}(S, \boldsymbol{\ell}_{1:T}) = \max_{1 \leqslant j \leqslant K} \sum_{i=1}^K \sum_{t=1}^T p_{i,t}(\ell_{i,t} - \ell_{j,t}) \,. \tag{5.8}$$

Initial step: the environment chooses a sequence of loss vectors $(\boldsymbol{\ell}_t)_{t\in\mathbb{N}^*}$, where the $\boldsymbol{\ell}_t = (\ell_{i,t})_i \in [0,1]^K$ will be revealed round after round. Two different assumptions are considered:

- stochastic environment: $(\boldsymbol{\ell}_t)_{t\geqslant 1}$ is an i.i.d. sequence;
- individual sequence: $(\boldsymbol{\ell}_t)_{t\geqslant 1}$ is an arbitrary deterministic sequence.

At each time round $t \in \mathbb{N}^*$,

1. the forecaster chooses a convex combination $\boldsymbol{p}_t \in \mathcal{X}_K$;
2. the environment reveals the loss vector $\boldsymbol{\ell}_t \in [0,1]^K$;
3. each action i incurs the loss $\ell_{i,t}$ and the forecaster incurs the linear loss $\boldsymbol{p}_t \cdot \boldsymbol{\ell}_t = \sum_{i=1}^K p_{i,t}\,\ell_{i,t}$.

Figure 5.1: Description of the online protocol. The environment can be either stochastic (i.i.d. sequence) or deterministic (individual sequence).

As for the internal regret, note that since the weight vectors $\boldsymbol{p}_t^{i\to j}$ and $\boldsymbol{p}_t$ only differ in at most two coordinates, many terms cancel out in the difference (5.2). Therefore, we get that

$$R_T^{\mathrm{int}}(S, \boldsymbol{\ell}_{1:T}) = \max_{1\leqslant i\neq j\leqslant K} \sum_{t=1}^T p_{i,t}(\ell_{i,t} - \ell_{j,t})\,. \tag{5.9}$$

Finally, in view of (5.4) and (5.5), and noting that $\boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t = \sum_{i=1}^K p_{i,t}\ell_{F(i),t}$, the swap regret can be rewritten as follows:

$$\begin{aligned} R_T^{\mathrm{sw}}(S, \boldsymbol{\ell}_{1:T}) &= \max_{F\in\mathcal{F}_K} \sum_{i=1}^K \sum_{t=1}^T p_{i,t}\big(\ell_{i,t} - \ell_{F(i),t}\big) \\ &= \sum_{i=1}^K \max_{1\leqslant j\leqslant K} \sum_{t=1}^T p_{i,t}\big(\ell_{i,t} - \ell_{j,t}\big)\,. \end{aligned} \tag{5.10}$$

Comparison of external, internal, and swap regrets

The three notions of external, internal, and swap regret are closely related. Equations (5.1), (5.2), and (5.4) show that

$$R_T^{\mathrm{ext}} \leqslant R_T^{\mathrm{sw}} \qquad \text{and} \qquad R_T^{\mathrm{int}} \leqslant R_T^{\mathrm{sw}}\,. \tag{5.11}$$

Internal and swap regrets are of the same order of magnitude in T since $R_T^{\mathrm{sw}} \leqslant K\,(R_T^{\mathrm{int}})_+$ by (5.9) and (5.10), so that

$$R_T^{\mathrm{int}} \leqslant R_T^{\mathrm{sw}} \leqslant K\,(R_T^{\mathrm{int}})_+\,.$$

On the contrary, external and internal regrets are not necessarily of the same order of magnitude in T. On the one hand, we can see from (5.8) and (5.9) that

$$R_T^{\mathrm{ext}} \leqslant (K-1)\,(R_T^{\mathrm{int}})_+\,,$$

so that minimizing internal regret is a more difficult task than minimizing external regret (up to the constant factor $K-1$). On the other hand, we cannot upper bound the internal regret by a constant times the external regret in the general case. Indeed, as shown by [SL05], there exists an algorithm whose external regret is sublinear in T but whose internal regret grows linearly in T.

5.2.3 A new elementary upper bound on the internal regret

Next we design a strategy whose internal regret is almost bounded by $\min\{\sqrt{T\ln K}, T/K\}$ uniformly over all individual sequences $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$. The latter bound interpolates the $\sqrt{T\ln K}$ bound of [SL05] and the trivial T/K bound satisfied by the forecaster choosing constant weight vectors $\boldsymbol{p}_t = (1/K, \dots, 1/K)$. Such an interpolation improves on the bound $\sqrt{T\ln K}$ for large values of K since $T/K \leqslant \sqrt{T\ln K}$ when $K\sqrt{\ln K} \geqslant \sqrt{T}$.

If T is known in advance, the bound $\min\{\sqrt{T\ln K}, T/K\}$ can be easily achieved. Indeed, it suffices to predict either with uniform weight vectors if $K\sqrt{\ln K} \geqslant \sqrt{T}$ or with a strategy attaining the upper bound $\sqrt{T\ln K}$ if $K\sqrt{\ln K} < \sqrt{T}$ (e.g., the strategy described in [SL05, Theorem 3.1]). Next we design a simple trick which, up to a factor of $\sqrt{2}$ and a small remainder term, achieves the bound $\min\{\sqrt{T\ln K}, T/K\}$ without knowing T in advance.

Let S be any strategy whose internal regret after T time steps is upper bounded by $\sqrt{T\ln K}$. We denote its t-th weight vector by $\boldsymbol{p}_t^S(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t-1})$ when applied to the loss sequence $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2, \dots \in [0,1]^K$. Our meta-strategy $(\boldsymbol{p}_t)_{t\geqslant 1}$ is built on S as follows. We split the whole time interval $\mathbb{N}^*$ into two periods $\{1, \dots, T_0\}$ and $\{T_0+1, T_0+2, \dots\}$, where

$$T_0 \triangleq \lfloor K^2 \ln K \rfloor + 1 \,.$$

Note that T_0 approximately satisfies $K\sqrt{\ln K} \approx \sqrt{T_0}$, so that $K\sqrt{\ln K} \gtrsim \sqrt{T}$ for all $T \leqslant T_0$ while $K\sqrt{\ln K} \lesssim \sqrt{T}$ for all $T \geqslant T_0 + 1$. The last comment combined with the above remark about the case when T is known in advance suggests to define our meta-strategy as follows. On the first period, our meta-strategy outputs uniform weights, i.e.,

$$\boldsymbol{p}_t \triangleq (1/K, \dots, 1/K) \,, \qquad \text{for all} \quad t \in \{1, \dots, T_0\} \,. \tag{5.12}$$

On the second period, we start the strategy S at time T_0+1 from scratch (i.e., the past information $(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{T_0})$ is not used) and we output the same weight vectors as S, i.e.,

$$\boldsymbol{p}_t \triangleq \boldsymbol{p}_{t-T_0}^S(\boldsymbol{\ell}_{T_0+1}, \dots, \boldsymbol{\ell}_{t-1}) \,, \qquad \text{for all} \quad t \geqslant T_0 + 1 \,. \tag{5.13}$$

Proposition 5.1 (A new elementary upper bound on the internal regret)**.** *Let $K \geqslant 2$. Then, the internal regret of the strategy defined in (5.12) and (5.13) satisfies, for all $T \geqslant 1$ and all $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$,*

$$\max_{1\leqslant i\neq j\leqslant K} \sum_{t=1}^{T} p_{i,t}(\ell_{i,t} - \ell_{j,t}) \leqslant \sqrt{2}\min\left\{T/K, \sqrt{T\ln K}\right\} + 1/K \,.$$

Remark 5.1. *Up to a factor of $\sqrt{2}$ and a small remainder term, the above bound interpolates the two aforementioned bounds $\sqrt{T\ln K}$ and T/K. In particular, it improves on the $\sqrt{T\ln K}$ bound for large values of K. Note that it is achieved by a strategy that does not use prior knowledge of T.*

Proof: The case $T \leqslant T_0$ is straightforward: since the weight vectors $\boldsymbol{p}_t$ are uniform for all $t \leqslant T_0$,

$$\max_{1\leqslant i\neq j\leqslant K} \sum_{t=1}^{T} p_{i,t}(\ell_{i,t} - \ell_{j,t}) \leqslant T/K \leqslant \sqrt{2}\min\bigl\{T/K, \sqrt{T\ln K}\bigr\} + 1/K \,.$$

We can thus assume that $T > T_0$. Since the maximum is subadditive, we have

$$\max_{1\leqslant i\neq j\leqslant K} \sum_{t=1}^{T} p_{i,t}(\ell_{i,t} - \ell_{j,t}) \leqslant \max_{1\leqslant i\neq j\leqslant K} \sum_{t=1}^{T_0} p_{i,t}(\ell_{i,t} - \ell_{j,t}) + \max_{1\leqslant i\neq j\leqslant K} \sum_{t=T_0+1}^{T} p_{i,t}(\ell_{i,t} - \ell_{j,t}) \,.$$

In the last inequality, the first term of the right-hand side is bounded from above by T_0/K since the first T_0 weights are uniform. In view of the requirement imposed on the strategy S, the second term is bounded from above by $\sqrt{(T-T_0)\ln K}$. Therefore, we get that

$$\begin{aligned}
\max_{1\leqslant i\neq j\leqslant K} \sum_{t=1}^{T} p_{i,t}(\ell_{i,t} - \ell_{j,t}) &\leqslant \frac{T_0}{K} + \sqrt{(T-T_0)\ln K} \\
&\leqslant \frac{K^2\ln K + 1}{K} + \sqrt{(T - K^2\ln K)\ln K} \qquad (5.14)\\
&= 1/K + \sqrt{K^2\ln^2 K} + \sqrt{T\ln K - K^2\ln^2 K} \\
&\leqslant 1/K + \sqrt{2(T\ln K)} \,, \qquad (5.15)
\end{aligned}$$

where (5.14) follows from the fact that $T_0 - 1 \leqslant K^2\ln K \leqslant T_0$ (by definition of T_0), and where (5.15) follows from the elementary inequality $\sqrt{x} + \sqrt{y} \leqslant \sqrt{2(x+y)}$ that holds for all $x, y \geqslant 0$. Since $T \geqslant T_0 \geqslant K^2\ln K$, we have $\sqrt{T\ln K} \leqslant T/K$, so that

$$\sqrt{2(T\ln K)} \leqslant \sqrt{2}\min\bigl\{T/K, \sqrt{T\ln K}\bigr\} \,.$$

This concludes the proof. □

Since the bound T/K is easy to achieve, in the sequel, we focus on the most interesting regime (i.e., the second one, when T is large enough). For example, we show in the next section that when the loss vectors are i.i.d., the upper bound $\sqrt{T\ln K}$ can be lowered to $\sqrt{T}$.

5.3 Minimax rate of internal regret in a stochastic environment

In this section we derive the optimal rate of internal regret in a stochastic environment. Namely, we consider the repeated game of Figure 5.1 with i.i.d. loss vectors: the $\boldsymbol{\ell}_t \in [0,1]^K$, $t \geqslant 1$, are drawn independently at random from a common distribution $Q \in \mathcal{M}_1^+\bigl([0,1]^K\bigr)$.

In the sequel, all expectations are taken with respect to $Q^{\otimes T}$, where $T \geqslant 1$ is a fixed time horizon.

We set $m_i \triangleq \mathbb{E}[\ell_{i,1}] = \ldots = \mathbb{E}[\ell_{i,T}]$ for all $1 \leqslant i \leqslant K$ and define the gaps Δ_i by

$$\Delta_i \triangleq m_i - m_{i^*} \,, \quad \text{where} \quad i^* \in \operatorname*{argmin}_{1 \leqslant j \leqslant K} m_j \,.$$

Next we prove upper bounds for the internal regret that are of the order of $\sqrt{T}$ with high probability. This entails that the optimal rate of the expected internal regret against i.i.d. loss vectors is $\sqrt{T}$ and therefore does not depend on the ambient dimension K. For the sake of clarity, we first assume that the distribution Q of the loss vectors is known to the forecaster, and then extend the analysis to the case when it is unknown.

5.3.1 Known distribution

In this subsection we assume that the distribution Q of the loss vectors is known to the forecaster. We explain below why it is possible to achieve an internal regret independent of the ambient dimension K. A key remark is that, contrary to the external regret where the weights $p_{i,t}$ appear additively over $i \in \{1, \ldots, K\}$, the internal regret $\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^T p_{i,t}(\ell_{i,t} - \ell_{j,t})$ scales as a single $p_{i,t}$ (the sum over i is replaced by a maximum). Therefore, if several actions i are almost optimal (i.e., if they almost minimize m_i), then the probability mass of $\boldsymbol{p}_t$ should be well spread among those actions.

Let us illustrate the above remark with the toy situation where the losses $\ell_{i,t}$, $i = 1, \ldots, K$, $t = 1, \ldots, T$, are i.i.d. Bernoulli random variables with parameter $1/2$. In this case, the strategy that constantly outputs the Dirac probability distribution $\boldsymbol{p}_t = \boldsymbol{\delta}_{i^*}$ at some $i^* \in \operatorname{argmin}_{1 \leqslant i \leqslant K} m_i$ has an expected internal regret of

$$\mathbb{E}\left[\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^T p_{i,t}(\ell_{i,t} - \ell_{j,t})\right] = \mathbb{E}\left[\sum_{t=1}^T \ell_{i^*,t} - \min_{1 \leqslant j \leqslant K} \sum_{t=1}^T \ell_{j,t}\right] = \mathbb{E}\left[\max_{1 \leqslant j \leqslant K} \sum_{t=1}^T \left(\frac{1}{2} - \ell_{j,t}\right)\right] \,.$$

By central limit arguments[4], the last expectation is of the order of $\sqrt{T \ln K}$. On the contrary, the strategy that outputs uniform weight vectors $\boldsymbol{p}_t = (1/K, \ldots, 1/K)$ has an expected internal regret

$$\mathbb{E}\left[\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^T p_{i,t}(\ell_{i,t} - \ell_{j,t})\right] = \frac{1}{K}\,\mathbb{E}\left[\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^T (\ell_{i,t} - \ell_{j,t})\right]$$

of the order of $\sqrt{T \ln K}/K$. Therefore, though tempting at first sight, the naive strategy $(\boldsymbol{\delta}_{i^*})_{t \geqslant 1}$ is suboptimal against i.i.d. vectors, while averaging over the actions leads to a $\sqrt{T}$-internal regret.

Consider now the still ideal but slightly more difficult situation where the distribution Q of the loss vectors is such that, for some $K' \in \{1, \ldots, K\}$,

$$m_1 = \ldots = m_{K'} < m_{K'+1} = \ldots = m_K \qquad \text{where} \quad m_{K'+1} - m_{K'} \gg \sqrt{\frac{\ln K}{T}} \,.$$

In this case assigning uniform weights to all the actions is clearly a bad choice, and a suitable

[4] See ,e.g., [CBFH+97, Section 3.2] or [CBL06, Theorem 3.7] where asymptotic lower bounds of order $\sqrt{T \ln K}$ are derived on the minimax external regret.

trade-off between "selecting the good actions" and "spreading the probability mass sufficiently" is necessary. A simple and reasonable strategy consists in assigning zero weights to the suboptimal actions $i \in \{K'+1, \dots, K\}$ and uniform weights to the optimal actions $i \in \{1, \dots, K'\}$; more formally,

$$p_1 = \dots = p_{K'} = 1/K' \qquad \text{and} \qquad p_{K'+1} = \dots = p_{K'} = 0 \,.$$

Then, due to the averaging, the expected internal regret is not of order $\sqrt{T \ln(K')}$ (as would be the case with $(\boldsymbol{\delta}_{i^*})_{t \geqslant 1}$) but at most of order $\sqrt{T \ln(K')}/K'$, hence a $\sqrt{T}$-rate again. The latter statement is proved for a more general (smoothed) strategy in Theorem 5.1 and can be roughly explained as follows: decomposing $\sum_{t=1}^T p_i(\ell_{i,t} - \ell_{j,t})$ into a bias term and a deviation term,

$$\begin{aligned}
&\mathbb{E}\left[\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^T p_i(\ell_{i,t} - \ell_{j,t})\right] \\
&\quad = \mathbb{E}\left[\max_{1 \leqslant i \neq j \leqslant K} \left\{ p_i T(m_i - m_j) + p_i \sum_{t=1}^T \big(\ell_{i,t} - \ell_{j,t} - (m_i - m_j)\big) \right\}\right] \\
&\quad \approx \mathbb{E}\left[\max_{1 \leqslant i \neq j \leqslant K'} \left\{ p_i T(m_i - m_j) + p_i \sum_{t=1}^T \big(\ell_{i,t} - \ell_{j,t} - (m_i - m_j)\big) \right\}\right] \\
&\quad = \frac{1}{K'} \mathbb{E}\left[\max_{1 \leqslant i \neq j \leqslant K'} \sum_{t=1}^T \big(\ell_{i,t} - \ell_{j,t} - (m_i - m_j)\big)\right] \,, \qquad (5.16)
\end{aligned}$$

where the approximation above (i.e., the restriction of the minimum to $\{1, \dots, K'\}$) follows from the fact that $p_i = 0$ for all $i > K'$ and from the fact that, for all $i \leqslant K'$ and $j > K'$,

$$\underbrace{T(m_i - m_j)}_{\ll -\sqrt{T \ln K}} + \underbrace{\sum_{t=1}^T \big(\ell_{i,t} - \ell_{j,t} - (m_i - m_j)\big)}_{\lesssim \sqrt{T \ln K}} \ll 0 \,,$$

where we used the assumption $m_i - m_j \ll -\sqrt{(\ln K)/T}$ and the fact that, by Hoeffding-Azuma inequality (cf. Lemma A.6 in Appendix A.5), the deviations $\sum_{t=1}^T \big(\ell_{i,t} - \ell_{j,t} - (m_i - m_j)\big)$ are at most of the order of $\sqrt{T \ln K}$ with high-probability. But, by a well-known maximal inequality stated in [Mas07, Lemma 2.3] (see Lemma A.3 in Appendix A.5), the expectation in (5.16) is at most of the order of $\sqrt{T \ln(K')}/K'$.

Parameters: $Q \in \mathcal{M}_1^+\big([0,1]^K\big)$ and $T \geqslant 1$.

At each time round $t = 1, \dots, T$,
Output the same weight vector $\boldsymbol{p}^{\mathrm{int}}(Q) = \big(p_i^{\mathrm{int}}(Q)\big)_{1 \leqslant i \leqslant K}$ defined by

$$p_i^{\mathrm{int}}(Q) \triangleq \frac{e^{-\sqrt{T}\, m_i}}{\sum_{j=1}^K e^{-\sqrt{T}\, m_j}} \,, \qquad 1 \leqslant i \leqslant K \,, \qquad (5.17)$$

where $m_i \triangleq \mathbb{E}_Q[\ell_{i,1}]$ for all $i \in \{1, \dots, K\}$.

Figure 5.2: A simple internal-regret-minimizing strategy when the distribution Q of the loss vectors is known (cf. Theorem 5.1).

If the distribution Q of the loss vectors is arbitrary, then the trade-off between "selecting the good actions" and "spreading the probability mass sufficiently" can be carried out in a continuous way. The above explanation suggests to take constant weigth vectors $\boldsymbol{p}_t = \boldsymbol{p}_1$ such that $p_{i,1}$ decreases continuously with the gap $\Delta_i \triangleq m_i - m_{i^*}$. Such a choice is given by the exponential weights defined in Figure 5.2. Note that the corresponding weights $p_i^{\mathrm{int}}(Q)$ can be rewritten as

$$p_i^{\mathrm{int}}(Q) = \frac{e^{-\sqrt{T}\Delta_i}}{\sum_{j=1}^K e^{-\sqrt{T}\Delta_j}} = \frac{e^{-\sqrt{T}\Delta_i}}{K_{\mathrm{eff}}} \qquad \text{where} \quad K_{\mathrm{eff}} \triangleq \sum_{j=1}^K e^{-\sqrt{T}\Delta_j} \in [1, K] \, .$$

K_{eff} is a smooth generalization of the number K' considered in the example above; it can be thought of as the effective number of good actions.

The next theorem provides an (optimal) $\sqrt{T}$-high-probability upper bound on the internal regret of the simple strategy of Figure 5.2. In particular, it is independent of the ambient dimension K.

Theorem 5.1 (A $\sqrt{T}$-internal regret when the distribution Q of the loss vectors is known)**.** *Let $K \geqslant 2$ and $T \geqslant 1$. Assume that the loss vectors $\boldsymbol{\ell}_t \in [0,1]^K$, $1 \leqslant t \leqslant T$, are drawn independently at random from a common known distribution $Q \in \mathcal{M}_1^+([0,1]^K)$.*

Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$, the internal regret of the constant sequence $\big(\boldsymbol{p}^{\mathrm{int}}(Q)\big)_{t \geqslant 1}$ defined in (5.17) is upper bounded by

$$\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^T p_i^{\mathrm{int}}(Q)(\ell_{i,t} - \ell_{j,t}) \leqslant \frac{3}{K_{\mathrm{eff}}} \sqrt{T \ln\left(\frac{3K_{\mathrm{eff}}}{\delta}\right)} \leqslant 3\sqrt{T \ln\left(\frac{3}{\delta}\right)} \, ,$$

where $K_{\mathrm{eff}} \triangleq \sum_{i=1}^K e^{-\sqrt{T}\Delta_i} \in [1, K]$ can be interpreted as the effective number of good actions.

The proof is postponed to Appendix 5.A. As noted in Remark 5.6 therein, a weighted union-bound is key to derive an upper bound of the order of $\sqrt{T}$.

5.3.2 Unknown distribution

In this section the distribution Q of the loss vectors is no longer assumed to be known in advance by the forecaster. We adapt the simple strategy of Figure 5.2 to this setting by a plug-in method: the expectations $m_i \triangleq \mathbb{E}_Q[\ell_{i,1}]$ are sequentially estimated over exponentially growing epochs $\{1\}$ and $\big\{2^{r-1}+1, \dots, 2^r\big\}$, $r \in \mathbb{N}^*$. The resulting strategy is defined in Figure 5.3. We prove in Theorem 5.2 that it still achieves a $\sqrt{T}$-internal regret with high probability.

Theorem 5.2 (A $\sqrt{T}$-internal regret when the distribution Q of the loss vectors is unknown)**.** *There is an absolute constant $c_0 > 0$ such that the following holds true. Let $K \geqslant 2$ and $T \geqslant 1$. Assume that the loss vectors $\boldsymbol{\ell}_t \in [0,1]^K$, $1 \leqslant t \leqslant T$, are drawn independently at random from an unknown distribution $Q \in \mathcal{M}_1^+([0,1]^K)$.*

Then, for all $\delta \in (0,1)$, with probability at least $1-\delta$, the internal regret of the strategy defined in Figure 5.3 is upper bounded by

$$\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^T p_{i,t}(\ell_{i,t} - \ell_{j,t}) \leqslant c_0 \sqrt{T} \exp\Big(2\sqrt{2\ln(4/\delta)}\Big) + 1 \, .$$

Parameter: none.

Initialization: output the weight vector $\boldsymbol{p}_1 = (1/K, \dots, 1/K) \in \mathcal{X}_K$.

For each regime $r \in \mathbb{N}^*$,

1. Set $\widehat{m}_i^{(r)} = \dfrac{1}{2^{r-1}} \displaystyle\sum_{t=1}^{2^{r-1}} \ell_{i,t}$ for all $i \in \{1, \dots, K\}$;
2. At each time round $t \in \left\{2^{r-1}+1, \dots, 2^r\right\}$,
   output the same weight vector $\boldsymbol{p}_t \triangleq \boldsymbol{p}^{(r)} = \left(p_i^{(r)}\right)_{1 \leqslant i \leqslant K}$ defined by

$$p_i^{(r)} = \frac{\exp\left(-\sqrt{2^{r-1}}\, \widehat{m}_i^{(r)}\right)}{\displaystyle\sum_{j=1}^{K} \exp\left(-\sqrt{2^{r-1}}\, \widehat{m}_j^{(r)}\right)}, \quad 1 \leqslant i \leqslant K\,. \tag{5.18}$$

Figure 5.3: An internal-regret-minimizing strategy when the distribution Q of the loss vectors is unknown (cf. Theorem 5.2).

The proof of Theorem 5.2 is postponed to Appendix 5.A. It is a simple adaptation of that of Theorem 5.1. The only but important additional tool is a "backward weighted union-bound" carried out at the end of the proof.

Note that an explicit upper bound on the absolute constant c_0 can be computed at the end of the proof. However, since for the sake of clarity, we sometimes performed crude upper bounds, its value may be far from optimal.

Though Q is unknown, the bound of the above theorem is still independent of the ambient dimension K. Moreover, even if the deviation factor $\exp\bigl(2\sqrt{2\ln(4/\delta)}\bigr)$ above is much larger than the more standard factor $\sqrt{T\ln(3/\delta)}$ of Theorem 5.1, it is still small enough to yield a bound of order $\sqrt{T}$ in expectation. It suffices to integrate the above high-probability bound (see Section A.6) and to combine it with the lower bound of [Sto05] to get the following.

Corollary 5.1 (Minimax rate of internal regret with i.i.d. loss vectors)**.**
There exist absolute constants $c_1, c_2, c_3 > 0$ such that the following holds true. Let $K \geqslant 2$ and $T \geqslant c_1 K^2$. Then, the minimax internal regret with i.i.d. loss vectors satisfies

$$c_2\,\sqrt{T} \leqslant \inf_{S} \sup_{Q \in \mathcal{M}_1^+([0,1]^K)} \mathbb{E}_{Q^{\otimes T}} \left[\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^{T} p_{i,t}(\ell_{i,t} - \ell_{j,t}) \right] \leqslant c_3 \sqrt{T}\,,$$

where the infimum is taken over all strategies $S = (\boldsymbol{p}_t)_{t \geqslant 1}$ of the forecaster, and where, in the last expectation, the loss vectors $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T$ are i.i.d. with common distribution Q.

The proof of Corollary 5.1 is postponed to Appendix 5.A. Again, explicit bounds on the absolute constants c_1,c_2,c_3 can easily be computed at the end of the proof, but their values have not been optimized.

5.4 Lower bound on the swap regret with individual sequences

In this section we prove a lower bound of order $\sqrt{TK}$ on the minimax swap regret with individual sequences. This lower bound solves a question left open in [BM07b] — see below. It also highlights a major difference between external and swap regrets: contrary to external regret, swap regret is much harder to minimize with individual sequences than with i.i.d. losses — see Section 5.4.2.

5.4.1 Main result

The main result of this section is the following.

Theorem 5.3 (Lower bound on the minimax swap regret)**.**
There exists an absolute constant $c > 0$ such that the following holds true. Let $K \geqslant 2$ and $T \geqslant \max\{128c^2K^5, K\}$. Then the minimax swap regret with individual sequences satisfies

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \left\{ \sum_{t=1}^T \boldsymbol{p}_t\cdot\boldsymbol{\ell}_t - \min_{F\in\mathcal{F}_K} \sum_{t=1}^T \boldsymbol{p}_t^F\cdot\boldsymbol{\ell}_t \right\} \geqslant c\sqrt{TK}\,,$$

where the infimum is taken over all strategies $S = (\boldsymbol{p}_t)_{t\geqslant 1}$ of the forecaster, where $\mathcal{F}_K$ denotes the set of all functions from $\{1,\dots,K\}$ to $\{1,\dots,K\}$, and where the transformed weight vector $\boldsymbol{p}_t^F$ is defined in (5.5). In particular, we prove the theorem for $c = 1/\big(16\sqrt{128\ln(4/3)}\big)$.

The above theorem solves an open problem stated in [BM07b, Section 9]. The latter authors already proved a lower bound of order $\sqrt{TK}$ but only in a weak sense:

- Their lower bound was stated in a randomized and adversarial setting for a quantity larger than the swap regret *stricto sensu* (which makes the lower bound easier to prove). Their adversarial setting is defined recursively as follows. The environment – or *adversary* – has a strategy: it chooses a sequence $(\pi_t)_{t\geqslant 1}$ of conditional probability distributions on $[0,1]^K$ such that $\pi_t\big(\mathrm{d}\boldsymbol{\ell}_t \,\big|\, (\boldsymbol{\ell}_s,\boldsymbol{p}_s,I_s)_{s\leqslant t-1},\boldsymbol{p}_t\big)$ is the law of $\boldsymbol{\ell}_t$ conditionally on the available data $\big((\boldsymbol{\ell}_s,\boldsymbol{p}_s,I_s)_{s\leqslant t-1},\boldsymbol{p}_t\big)$. At each time t, the forecaster picks $I_t \in \{1,\dots,K\}$ at random such that, conditionally on the past data $(\boldsymbol{\ell}_s,I_s)_{1\leqslant s\leqslant t-1}$, the random variables I_t and $\boldsymbol{\ell}_t$ are independent and $I_t = i$ with probability $p_{i,t}$. The weight vectors $\boldsymbol{p}_t$ are now measurable functions of $(\boldsymbol{\ell}_s,I_s)_{1\leqslant s\leqslant t-1}$, and the corresponding sequence of functions $(\boldsymbol{p}_t)_{t\geqslant 1}$ is called a *randomized strategy*. Then, setting $\boldsymbol{\delta}_{I_t} \triangleq (\mathbb{I}_{\{I_t=i\}})_{1\leqslant i\leqslant K}$, Theorem 9 of [BM07b] provides a lower bound on the quantity

$$\sup_{(\pi_t)_{t\geqslant 1}} \inf_{S\text{ rand}} \mathbb{E}\left[\sum_{t=1}^T \boldsymbol{\delta}_{I_t}\cdot\boldsymbol{\ell}_t - \min_{F\in\mathcal{F}_K}\sum_{t=1}^T \boldsymbol{\delta}_{I_t}^F\cdot\boldsymbol{\ell}_t\right]\,,$$

  where the supremum is taken over all adversaries $(\pi_t)_{t\geqslant 1}$, where the infimum is taken over all randomized strategies S, and where the expectation is taken with respect to all sources of randomness (i.e., $(\boldsymbol{\ell}_t,I_t)_{1\leqslant t\leqslant T}$). By Jensen's inequality and by the fact that $\mathbb{E}\big[\boldsymbol{\delta}_{I_t} \,\big|\, (\boldsymbol{\ell}_s,I_s)_{s\leqslant t-1},\boldsymbol{\ell}_t\big] = \boldsymbol{p}_t$, the above quantity is larger than

$$\sup_{(\pi_t)_{t\geqslant 1}} \inf_{S\text{ rand}} \mathbb{E}\left[\sum_{t=1}^T \boldsymbol{p}_t\cdot\boldsymbol{\ell}_t - \min_{F\in\mathcal{F}_K}\sum_{t=1}^T \boldsymbol{p}_t^F\cdot\boldsymbol{\ell}_t\right]$$

$$\geqslant \sup_{\mathbb{Q}\in\mathcal{M}_1^+([0,1]^K)} \inf_S \mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^T \boldsymbol{p}_t\cdot\boldsymbol{\ell}_t - \min_{F\in\mathcal{F}_K}\sum_{t=1}^T \boldsymbol{p}_t^F\cdot\boldsymbol{\ell}_t\right] \tag{5.19}$$

$$= \inf_S \sup_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T\in[0,1]^K}\left\{\sum_{t=1}^T \boldsymbol{p}_t\cdot\boldsymbol{\ell}_t - \min_{F\in\mathcal{F}_K}\sum_{t=1}^T \boldsymbol{p}_t^F\cdot\boldsymbol{\ell}_t\right\}, \tag{5.20}$$

where the supremum in (5.19) is restricted[5] to the set of all probability distributions on $[0,1]^{KT}$ (the corresponding expectation is taken with respect to $(\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T)$ with joint distribution $\mathbb{Q}$), and where (5.20) follows from minimax duality (cf. Theorem 5.4 of the next section).

Therefore, our lower bound is stronger than the one of [BM07b, Theorem 9]. It solves a question left open in [BM07b, Section 9]: the authors showed that, in the randomized setting described above, it was not possible to ensure a worst-case swap regret of εT in a number of rounds T sublinear in K. They asked whether such impossiblity result remained true in the distributional setting (i.e., in our own non-randomized setting, with $\boldsymbol{p}_t$ instead of $\boldsymbol{\delta}_{I_t}^F$), where the task of the forecaster seems easier. Our lower bound provides a positive answer: a worst-case swap regret of εT is only possible for T at least of the order of K/ε^2.

- The lower bound $\sqrt{TK}/160 - 1$ of [BM07b, Theorem 9] is only stated for $K \leqslant T \leqslant \exp(K/288)/\sqrt{K}$, therefore, only for rounds T that are sub-exponential in K. On the contrary, our lower bound holds for all $T \geqslant \max\{128c^2K^5, K\}$.

The proof of Theorem 5.3 is postponed to Appendix 5.A.2. We use the key equality (5.10) of Section 5.2.2 to rewrite the swap regret as a sum of $K' \triangleq K/2$ internal regrets on time subintervals of length T/K'. The $\sqrt{T}$-lower bound on the internal regret of [Sto05, Theorem 3.3] then yields a lower bound on the swap regret of order $K'\sqrt{T/K'} = \sqrt{TK'}$. We make this statement more precise by using techniques borrowed from [Sto05, Theorem 3.3]. Namely, we use a reduction to stochastic losses for which, at each time t, only two of them are small, and then use Pinsker's inequality. However, due to the larger complexity of swap regret, our analysis is more involved than for internal regret — see the construction by induction in Appendix 5.A.2.

5.4.2 A major difference with classical works on external regret

In this section we point out a major difference between external and swap regrets: contrary to external regret, swap regret is much harder to minimize with individual sequences than with i.i.d. losses[6].

Indeed, on the one hand, all known lower bounds on the minimax external regret with individual sequences are proved with i.i.d. loss sequences (whose distribution may depend on the strategy of the forecaster). This is the case in the full information setting (see Section 2.3.2 in Chapter 2), but also in the bandit setting (cf. [ACBFS02, Theorem 5.1]), or in the label-efficient prediction setting (cf. [CBLS05, Theorem 13]).

[5] Since in (5.19) the environment is oblivious to the forecaster's past moves, the infimum $\inf_{\{S\text{ rand}\}}$ can be restricted to non-randomized strategies, i.e., such that $\boldsymbol{p}_t$ is a measurable function of $(\boldsymbol{\ell}_s)_{1\leqslant s\leqslant t-1}$ (by Jensen's inequality).

[6] As of now, we do not know if their is such difference for the internal regret. In any case, contrary to swap regret, such difference cannot be too large since the minimax internal regret for individual sequences is at most a factor of $\sqrt{\ln K}$ larger than the minimax internal regret for i.i.d. loss vectors — cf. (5.6) and (5.7).

On the other hand, there is a large gap between the minimax swap regret for individual sequences and the minimax swap regret for i.i.d. loss vectors. Indeed, as shown below, if the loss vectors are i.i.d. with common distribution $Q \in \mathcal{M}_1^+\big([0,1]^K\big)$, then the expected swap regret can be made as small as $\sqrt{T \ln K}$ (up to a constant factor) uniformly over all distributions $Q \in \mathcal{M}_1^+\big([0,1]^K\big)$. On the contrary, we proved in Theorem 5.3 that the minimax swap regret for individual sequences is at least of the order of $\sqrt{TK}$: setting $c \triangleq 1/\big(16\sqrt{128\ln(4/3)}\big)$, we showed that, for all $K \geqslant 2$ and all $T \geqslant \max\big\{128c^2K^5, K\big\}$,

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in [0,1]^K} \left\{ \sum_{t=1}^T \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F \in \mathcal{F}_K} \sum_{t=1}^T \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t \right\} \geqslant c\sqrt{TK}\,,$$

where the infimum is taken over all strategies $S = (\boldsymbol{p}_t)_{t \geqslant 1}$ of the forecaster. This lower bound was derived with *piecewise* i.i.d. loss vectors. Therefore, the lack of stationarity in the loss sequence deteriorates the ability of the forecaster to minimize his swap regret. This is in contrast with the external regret, for which arbitrary loss sequences are as easy to control as i.i.d. loss sequences.

Next we prove the aforementioned $\sqrt{T \ln K}$ bound: we design a simple strategy whose expected swap regret is at most of the order of $\sqrt{T \ln K}$ uniformly over all distributions $Q \in \mathcal{M}_1^+\big([0,1]^K\big)$. This upper bound is optimal (up to constant factors). Indeed, by the lower bound of order $\sqrt{T \ln K}$ on the external regret proved in Lemma 2.2 of Chapter 2 (which a fortiori implies a lower bound on the swap regret by (5.11)), we get that, for all $K \geqslant 1$ and all $T \geqslant \big[40e/(2e+1)\big] \ln K$,

$$\inf_S \sup_{Q \in \mathcal{M}_1^+([0,1]^K)} \mathbb{E}_{Q^{\otimes T}} \left[\sum_{t=1}^T \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F \in \mathcal{F}_K} \sum_{t=1}^T \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t \right] \geqslant \frac{2}{2e+1} \sqrt{\frac{e\,T \ln K}{5(2e+1)}}\,, \tag{5.21}$$

where the infimum is taken over all strategies $S = (\boldsymbol{p}_t)_{t \geqslant 1}$ of the forecaster and where in the last expectation, the loss vectors $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T$ are i.i.d. with common distribution Q.

For the sake of simplicity, we first assume that the distribution Q is known in advance by the forecaster. In this case, we set $m_i \triangleq \mathbb{E}_Q[\ell_{i,1}]$ for all $i = 1, \dots, K$ and consider the simple strategy that constantly outputs the Dirac probability distribution

$$\boldsymbol{p}_t = \boldsymbol{\delta}_{i^*}\,, \qquad 1 \leqslant t \leqslant T\,, \qquad \text{where} \qquad i^* \in \operatorname*{argmin}_{1 \leqslant i \leqslant K} m_i\,.$$

The following proposition indicates that this simple strategy, which we proved to be suboptimal for internal regret with i.i.d. loss vectors (cf. Section 5.3), is however sufficient to attain the optimal rate of swap regret.

Proposition 5.2. *Let $K \geqslant 2$ and $T \geqslant 1$. Assume that the loss vectors $\boldsymbol{\ell}_t \in [0,1]^K$, $1 \leqslant t \leqslant T$, are drawn independently at random from a common known distribution $Q \in \mathcal{M}_1^+([0,1]^K)$. Then the swap regret of the constant strategy $(\boldsymbol{\delta}_{i^*})_{t \geqslant 1}$ described above satisfies*

$$\mathbb{E}_{Q^{\otimes T}} \left[\sum_{t=1}^T \boldsymbol{\delta}_{i^*} \cdot \boldsymbol{\ell}_t - \min_{F \in \mathcal{F}_K} \sum_{t=1}^T \boldsymbol{\delta}_{i^*}^F \cdot \boldsymbol{\ell}_t \right] \leqslant \sqrt{\frac{T}{2} \ln K}\,.$$

Proof: The proof uses arguments that are similar to the ones of Theorem 5.1, but is even simpler due to the simpler form of the weight vector $\boldsymbol{\delta}_{i^*}$. Therefore, we only sketch below the main lines.

First note that since $\boldsymbol{\delta}_{i^*}^F = \boldsymbol{\delta}_{F(i^*)}$ for all $F \in \mathcal{F}_K$, we get that $\min_{F\in\mathcal{F}_K} \sum_{t=1}^T \boldsymbol{\delta}_{i^*}^F \cdot \boldsymbol{\ell}_t = \min_{1\leqslant j\leqslant K} \sum_{t=1}^T \ell_{j,t}$, so that the swap regret reduces to the external regret:

$$\begin{aligned}\mathbb{E}_{Q^{\otimes T}}\left[\sum_{t=1}^T \boldsymbol{\delta}_{i^*}\cdot\boldsymbol{\ell}_t - \min_{F\in\mathcal{F}_K}\sum_{t=1}^T \boldsymbol{\delta}_{i^*}^F\cdot\boldsymbol{\ell}_t\right] &= \mathbb{E}_{Q^{\otimes T}}\left[\sum_{t=1}^T \ell_{i^*,t} - \min_{1\leqslant j\leqslant K}\sum_{t=1}^T \ell_{j,t}\right] \\ &\leqslant \underbrace{Tm_{i^*} - \min_{1\leqslant j\leqslant K}\bigl\{Tm_j\bigr\}}_{=0} + \underbrace{\mathbb{E}_{Q^{\otimes T}}\left[\max_{1\leqslant j\leqslant K}\sum_{t=1}^T (m_j - \ell_{j,t})\right]}_{\leqslant\sqrt{(T/2)\ln K}},\end{aligned}$$

where the last inequality follows from the fact that $\min_i a_i - \min_i b_i \leqslant \max_i(a_i - b_i)$ for all $(a_i)_i, (b_i)_i \in \mathbb{R}^K$, and where the upper bound by $\sqrt{(T/2)\ln K}$ follows from Hoeffding's inequality combined with an elementary maximal inequality for subgaussian random variables (cf. Lemmas A.5 and A.3 respectively in Appendix A.5). This concludes the proof. □

We only stated a result in expectation. Note that a similar bound of the order of $\sqrt{T\ln(2K/\delta)}$ can be seen to hold true with probability at least $1-\delta$. Moreover, if the distribution Q of the loss vectors is unknown to the forecaster, then we can also derive a bound of the order of $\sqrt{T\ln(2K/\delta)}$ by adapting the above strategy via a plug-in method based on a doubling trick — in the same spirit as in Section 5.3.2, except that the weights are much simpler here. More precisely, set $\boldsymbol{p}_1 = (1/K,\dots,1/K) \in \mathcal{X}_K$ and set $\boldsymbol{p}_t = \boldsymbol{p}^{(r)}$ for all $t \in \{2^{r-1}+1,\dots,2^r\}$, $r \geqslant 1$, where

$$\boldsymbol{p}^{(r)} \triangleq \boldsymbol{\delta}_{\widehat{\imath}_r}\,, \qquad \text{with} \qquad \widehat{\imath}_r \in \operatorname*{argmin}_{1\leqslant j\leqslant K} \widehat{m}_j^{(r)} = \operatorname*{argmin}_{1\leqslant j\leqslant K}\left\{\frac{1}{2^{r-1}}\sum_{t=1}^{2^{r-1}}\ell_{j,t}\right\}.$$

Then, adapting the proof of Proposition 5.2 through the use a "backward weighted union-bound" as in the proof of Theorem 5.2, we could prove the following[7] (the proof is omitted for the sake of concision).

Proposition 5.3. *Let $K \geqslant 2$ and $T \geqslant 1$. Assume that the loss vectors $\boldsymbol{\ell}_t \in [0,1]^K$, $1 \leqslant t \leqslant T$, are drawn independently at random from a common unknown distribution $Q \in \mathcal{M}_1^+([0,1]^K)$. Then, for some absolute constant $c_4 > 0$, the swap regret of the strategy defined above satisfies, for all $\delta > 0$, with probability at least $1-\delta$,*

$$\sum_{t=1}^T \boldsymbol{p}_t\cdot\boldsymbol{\ell}_t - \min_{F\in\mathcal{F}_K}\sum_{t=1}^T \boldsymbol{p}_t^F\cdot\boldsymbol{\ell}_t \leqslant c_4\sqrt{T\ln\left(\frac{2K}{\delta}\right)}\,.$$

Integrating the last upper bound (via Lemma A.7 in Appendix A.6) and combining it with the lower bound (5.21), we get the next corollary.

[7] Contrary to Theorem 5.2, we are able to prove a bound that grows root-logarithmically in $1/\delta$. This is due to the simpler form of the weights (Dirac probability distributions) compared to those of Figure 5.3 (exponential weights).

Corollary 5.2 (Minimax rate of swap regret with i.i.d. loss vectors)**.**
There exist absolute constants $c_5, c_6, c_7 > 0$ such that the following holds true. Let $K \geqslant 2$ and $T \geqslant c_5 \ln K$. Then, the minimax swap regret with i.i.d. loss vectors satisfies

$$c_6 \sqrt{T \ln K} \leqslant \inf_{S} \sup_{Q \in \mathcal{M}_1^+([0,1]^K)} \mathbb{E}_{Q^{\otimes T}} \left[\sum_{t=1}^{T} \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F \in \mathcal{F}_K} \sum_{t=1}^{T} \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t \right] \leqslant c_7 \sqrt{T \ln K} \,,$$

where the infimum is taken over all strategies $S = (\boldsymbol{p}_t)_{t \geqslant 1}$ of the forecaster and where in the last expectation, the loss vectors $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T$ are i.i.d. with common distribution Q.

5.5 A stochastic technique for upper bounds with individual sequences

In this second part of the chapter, we develop a general stochastic technique to upper bound the minimax regret on arbitrary deterministic sequences. This technique is non-constructive but can be used for more general forms of regret than the ones studied before. It relies on a minimax duality theorem that enables to rewrite the minimax regret as a maximin regret where the loss vectors are random with a known joint distribution (see Sections 5.5.1 and 5.5.2 below). In Section 5.5.3 we then use this technique to recover known upper bounds on the external, internal and swap regrets. Finally, in Section 5.5.4, we derive a new upper bound of order $\sqrt{T \ln K}$ on the makespan regret, thus improving on the known bound of order $\ln(K)\sqrt{T}$ of [EDKMM09].

As is detailed in Section 5.5.1, page 183, a similar technique has been independently studied in [RST11]. Since we work in a much more specific setting, we are able to get (sometimes tight) explicit constants. Our proofs rely on less involved arguments (e.g., the Bernoullization technique of [Sch03] and an elementary maximal inequality for subgaussian random variables of [Mas07]).

We also stress that, though this stochastic technique is useful to better understand the problem at hand (since it provides an upper bound on the minimax regret), it is non-constructive. Designing explicit algorithms that achieve the obtained upper bounds is an important task to be addressed in the future (e.g., an efficient algorithm with a $\sqrt{T \ln K}$ makespan regret). Note that the same issue arises in [RST11].

5.5.1 Definitions and sketch of the stochastic technique

Next we introduce a generalized form of regret that includes as special cases the external, internal, and swap regrets, as well as the regret associated to global cost functions of [EDKMM09]. We then sketch the stochastic technique we use in the following sections to upper bound the minimax rate of such regrets on arbitrary deterministic sequences.

A generalized form of regret

Definition 5.1 ((ψ, φ)-regret)**.** *Let E be a real vector space, let $\psi = (\psi_t)_{t \geqslant 1}$ be a sequence of convex functions $\psi_t : E \to \mathbb{R}$, and let $\varphi : \mathbb{R}^K \times \mathbb{R}^K \to E$ be a bi-affine function, that is, $\varphi(\boldsymbol{u}, \cdot)$ and $\varphi(\cdot, \boldsymbol{v})$ are affine[8] for all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^K$.*

[8] A function $f : \mathbb{R}^K \to \mathbb{R}$ is affine if and only if $f - a$ is linear for some $a \in \mathbb{R}$.

Then, for any strategy $S = (\boldsymbol{p}_t)_{t\geqslant 1}$ of the forecaster, we define its (ψ,φ)-regret *after T time rounds on any loss sequence $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$ by*

$$\psi_T\left(\sum_{t=1}^{T} \varphi(\boldsymbol{p}_t, \boldsymbol{\ell}_t)\right) .$$

Example 5.1 (External regret)**.**
In view of (5.1), external regret corresponds to $E = \mathbb{R}^K$, $\psi_t : \mathbb{R}^K \to \mathbb{R}$, and $\varphi : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}^K$ defined by $\psi_t(\boldsymbol{x}) = \max_{1\leqslant i\leqslant K} x_i$ and by $\varphi(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{u}\cdot\boldsymbol{v} - v_i)_{1\leqslant i\leqslant K}$.

Example 5.2 (Internal regret)**.**
By (5.9), internal regret corresponds to $E = \mathbb{R}^{K\times K}$, $\psi_t : \mathbb{R}^{K\times K} \to \mathbb{R}$, and $\varphi : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}^{K\times K}$ defined by $\psi_t\big((x_{i,j})_{i,j}\big) = \max_{1\leqslant i\neq j\leqslant K} x_{i,j}$ and by $\varphi(\boldsymbol{u},\boldsymbol{v}) = \big(u_i\,(v_i - v_j)\big)_{1\leqslant i,j\leqslant K}$.

Example 5.3 (Swap regret)**.**
By (5.10), swap regret corresponds to $E = \mathbb{R}^{K\times K}$, $\psi_t : \mathbb{R}^{K\times K} \to \mathbb{R}$, and $\varphi : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}^{K\times K}$ defined by $\psi_t\big((x_{i,j})_{i,j}\big) = \sum_{i=1}^K \max_{1\leqslant j\leqslant K} x_{i,j}$ and by $\varphi(\boldsymbol{u},\boldsymbol{v}) = \big(u_i\,(v_i - v_j)\big)_{1\leqslant i,j\leqslant K}$.

Example 5.4 (Online learning with global cost functions)**.**
The framework of online learning with global cost functions recently introduced[9] by [EDKMM09] can also be cast into our generalized setting. More precisely, let $C : \mathbb{R}_+^K \to \mathbb{R}$ be a convex function such that $C^* : \mathbb{R}_+^K \to \mathbb{R}$ defined by

$$C^*(x_1,\dots,x_K) \triangleq \min_{\boldsymbol{\alpha}\in\mathcal{X}_K} C(\alpha_1 x_1, \dots, \alpha_K x_K)$$

is concave, and where $\mathcal{X}_K \triangleq \big\{\boldsymbol{x} \in \mathbb{R}_+^K : \sum_{i=1}^K x_i = 1\big\}$. Typical examples of C include the makespan ($C(x_1,\dots,x_K) = \max_i x_i$) and the d-norm cost ($C(x_1,\dots,x_K) = (\sum_i x_i^d)^{1/d}$). Then, for any strategy $S = (\boldsymbol{p}_t)_{t\geqslant 1}$ and any loss sequence $\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in [0,1]^K$, [EDKMM09] define the regret of S on $(\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T)$ with respect to the global cost function C by

$$C\left(\frac{1}{T}\sum_{t=1}^{T} p_{1,t}\ell_{1,t}, \dots, \frac{1}{T}\sum_{t=1}^{T} p_{K,t}\ell_{K,t}\right) - C^*\left(\frac{1}{T}\sum_{t=1}^{T}\ell_{1,t}, \dots, \frac{1}{T}\sum_{t=1}^{T}\ell_{K,t}\right) .$$

This regret corresponds to the (ψ,φ)-regret when $E = \mathbb{R}^K \times \mathbb{R}^K$ and when $\psi_t : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}$ and $\varphi : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{R}^K \times \mathbb{R}^K$ are defined by $\psi_t(\boldsymbol{u},\boldsymbol{v}) = C\big(t^{-1}\boldsymbol{u}\big) - C^*\big(t^{-1}\boldsymbol{v}\big)$ and by $\varphi(\boldsymbol{u},\boldsymbol{v}) = \big((u_i\,v_i)_{1\leqslant i\leqslant K}, \boldsymbol{v}\big)$.

Analysis from a stochastic viewpoint

In the sequel we derive upper bounds on the *minimax* (ψ,φ)*-regret* defined by

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \psi_T\left(\sum_{t=1}^{T}\varphi(\boldsymbol{p}_t,\boldsymbol{\ell}_t)\right)$$

for at least all (ψ,φ) corresponding to either external, internal, swap, or makespan regret (see the examples above). The infimum is taken over all strategies $S = (\boldsymbol{p}_t)_{t\geqslant 1}$ of the forecaster. We

[9] This setting was motivated by load balancing and job scheduling applications. See Section 5.5.4.

explain below how the minimax (ψ, φ)-regret – of deterministic nature – can be re-interpreted as a quantity involving random variables whose joint distribution is known and that are therefore easily manageable.

Step 1: Using minimax duality.
The first step consists in using a minimax duality theorem (see Theorem 5.4 below) to exchange the infimum and the supremum in the sense that

$$\inf_{S} \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in [0,1]^K} \psi_T\left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t, \boldsymbol{\ell}_t)\right) = \sup_{\mathbb{Q} \in \mathcal{M}_1^+([0,1]^{KT})} \inf_{S} \mathbb{E}_{\mathbb{Q}}\left[\psi_T\left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t, \boldsymbol{\ell}_t)\right)\right] ,$$

where the supremum in the right-hand-side is taken over all probability distributions on $[0,1]^{KT}$ and where its expectation is taken with respect to random variables $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$ with joint distribution $\mathbb{Q}$.

The left-hand-side quantity corresponds to a minimax game: the goal of the forecaster is to choose a strategy S whose worst-case regret is the smallest possible. In particular, the forecaster does not have any prior knowledge on the loss sequence to be dealt with. On the contrary, the right-hand-side quantity corresponds to a maximim game: the forecaster is first given the joint distribution $\mathbb{Q}$ of the future loss sequence $(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T)$ and then chooses a strategy S accordingly.

Step 2: Upper bounding the maximin regret.
By the equality above, we can see that the two aforementioned games are equally difficult. Therefore, the minimax regret can be upper bounded through its maximin counterpart. The last quantity is generally easier to control: at each time t, the forecaster knows the distributions of all future losses conditionnally on the past (since he knows $\mathbb{Q}$ and the past loss vectors). Therefore, the forecaster can minimize a "conditional variant" of the (ψ, φ)-regret given by

$$\psi_T\left(\sum_{t=1}^T \varphi\Big(\boldsymbol{p}_t, \mathbb{E}\big[\boldsymbol{\ell}_t \,\big|\, \boldsymbol{\ell}_{1:t-1}\big]\Big)\right) .$$

Deviations of the true (ψ, φ)-regret from this "conditional variant" are often small enough; in our proofs, we will control them via standard martingale concentration arguments.

One simple strategy for dealing with this "conditional variant" consists in putting at each round t a unit mass at the index I_t^* minimizing the next expected loss (conditionnally on the past). This corresponds to the strategy $S^*(\mathbb{Q}) = (\boldsymbol{p}_t^{\mathbb{Q}})_{t \geqslant 1}$ defined by

$$\boldsymbol{p}_t^{\mathbb{Q}} \triangleq \boldsymbol{\delta}_{I_t^*}, \qquad \text{with} \quad I_t^* \in \operatorname*{argmin}_{1 \leqslant i \leqslant K} \mathbb{E}_{\mathbb{Q}}\big[\ell_{i,t} \,\big|\, \boldsymbol{\ell}_{1:t-1}\big] , \tag{5.22}$$

where $\boldsymbol{\delta}_i \in \mathcal{X}_K$ denotes the Dirac distribution at $i \in \{1, \dots, K\}$ and where $\boldsymbol{\ell}_{1:T}$ is assumed to be drawn at random with joint distribution $\mathbb{Q} \in \mathcal{M}_1^+\big([0,1]^{KT}\big)$. Note that $\boldsymbol{p}_t^{\mathbb{Q}}$ depends on $\mathbb{Q}$ as suggested above. We will use $S^*(\mathbb{Q})$ later to revisit known upper bounds on external, internal, and swap regrets.

Comparison to the literature

The use of minimax duality to analyse the minimax rate of external regret was first exploited by [AABR09] and later by [RST10]. The analysis of the aforementioned papers is generic enough to cover various loss functions but relies in a somewhat crucial way on the fact that the weight vectors $\boldsymbol{p}_t$ appear additively in the external regret[10].

In the present chapter, we focus on the linear loss but extend the analysis of [AABR09] to other types of regret that do not satisfy this additivity property, e.g., the internal, swap, and makespan regrets. Such an extension has been independently carried out by [RST11] via the so-called "Triplex Inequality" and the control of "sequential Rademacher complexities". The setting considered in the last paper is much broader than ours: their analysis covers a wider sprectrum of cases and our notion of (ψ, φ)-regret resembles the regret of [RST11] in a particular situation that they called "when B is a function of the average" – see Section 3.2 therein.

However, in our simpler setting, our analysis relies on related but simpler tools such as Bernoullization and an elementary maximal inequality for subgaussian random variables (cf. Lemma 5.1 below and Lemma A.3 in Appendix A.5). The Bernoullization step allows to resort directly to a version of von Neumann's minimax theorem without the need to write the minimax regret as a cumbersome sequence of multiple infima and suprema — this is in contrast with [AABR09, RST10, RST11]. As for the aforementioned maximal inequality, it replaces a Dudley-entropy-type upper bound. The last tool is much more general, but it is unncessarily involved in our case since we only consider finite reference classes. Moreover, our approach yields explicit sharp constants that exactely recover the best constants known so far for the external, internal, and swap regrets.

5.5.2 A minimax theorem for the (ψ, φ)-regret

The next minimax duality theorem is the main result of this section.

Theorem 5.4 (A minimax theorem for the (ψ, φ)-regret)**.**
Let E be a real vector space, let $\psi = (\psi_t)_{t \geqslant 1}$ be a sequence of convex functions $\psi_t : E \to \mathbb{R}$, and let $\varphi : \mathbb{R}^K \times \mathbb{R}^K \to E$ be a bi-affine function (cf. Definition 5.1). Then the (ψ, φ)-regret satisfies the following duality formula:

$$\inf_S \sup_{\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K} \psi_T \left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t, \boldsymbol{\ell}_t) \right) = \sup_{\mathbb{Q} \in \mathcal{M}_1^+([0,1]^{KT})} \inf_S \mathbb{E}_{\mathbb{Q}} \left[\psi_T \left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t, \boldsymbol{\ell}_t) \right) \right] ,$$

where both infima are taken over all strategies $S = (\boldsymbol{p}_t)_{t \geqslant 1}$, where $\mathcal{M}_1^+\big([0,1]^{KT}\big)$ denotes the set of all probability distributions on $[0,1]^{KT}$, and where the expectation $\mathbb{E}_{\mathbb{Q}}[\,\cdot\,]$ is taken with respect to the random variables $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$ with joint distribution $\mathbb{Q}$.

The proof of Theorem 5.4 consists of a careful application of a version of von Neumann's minimax theorem (stated as Lemma A.1 in Appendix A.3). To use it in a convenient way, we will

[10] The fact that the $\boldsymbol{p}_t$ appear additively in the external regret has been used many other times. An example is in the water-filling technique used to derive the exact minimax external regret in the binary prediction problem under the absolute loss; see, e.g., [CBL06, Section 8.2] and the references therein.

first use the next technical lemma that enables a reduction to binary losses and relies on a technique due to [Sch03]. With this reduction, the compacity and continuity assumptions of Lemma A.1 are then satisfied.

Lemma 5.1 (Bernoullization)**.**
Let E be a real vector space, let $\psi = (\psi_t)_{t\geqslant 1}$ be a sequence of convex functions $\psi_t : E \to \mathbb{R}$, and let $\varphi : \mathbb{R}^K \times \mathbb{R}^K \to E$ be a bi-affine function (cf. Definition 5.1). Then the minimax (ψ, φ)-regret can be reduced to binary losses $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in \{0,1\}^K$ in the sense that

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \psi_T\left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t, \boldsymbol{\ell}_t)\right) = \inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in\{0,1\}^K} \psi_T\left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t, \boldsymbol{\ell}_t)\right) ,$$

where both infima are taken over all strategies $S = (\boldsymbol{p}_t)_{t\geqslant 1}$.

Note that the above lemma would be immediate if the considered strategies S were static, i.e., such that $\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}) = \boldsymbol{q}_t$ for all $t \geqslant 1$ and $\boldsymbol{\ell}_{1:t-1} \triangleq (\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t-1})$ and for some fixed sequence $(\boldsymbol{q}_t)_{t\geqslant 1}$ in $\mathcal{X}_K$. Indeed, the (ψ, φ)-regret of such strategies is convex on the polytope $[0,1]^{KT}$ and thus achieves its supremum on the hypercube $\{0,1\}^{KT}$. The above lemma shows that, even for non-static strategies, the hypercube is in some sense sufficient to assess the performance of any strategy (see also Remark 5.2 below).

Proof (of Lemma 5.1): Our proof is based on a Bernoullization argument of [Sch03] which we slightly simplify via the use of Jensen's inequality. Let $U_{i,t}$, $1 \leqslant i \leqslant K$, $1 \leqslant t \leqslant T$, be independent real random variables uniformly distributed on $[0,1]$. We set $\boldsymbol{U}_t \triangleq (U_{i,t})_{1\leqslant i\leqslant K}$ and $\boldsymbol{U}_{1:t} \triangleq (\boldsymbol{U}_s)_{1\leqslant s\leqslant t}$ for all $t \in \{1, \dots, T\}$. In this proof, we write $\mathbb{E}_{\boldsymbol{U}_{1:t}}$ or $\mathbb{E}_{\boldsymbol{U}_t}$ when the expectation is taken over $\boldsymbol{U}_{1:t}$ or $\boldsymbol{U}_t$ respectively. To avoid any ambiguity, we also explicitly see the weights as functions $\boldsymbol{p}_t : [0,1]^{K(t-1)} \to \mathcal{X}_K$ of the past loss vectors $\boldsymbol{\ell}_{1:t-1} \triangleq (\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t-1})$, and hence write $\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1})$.

We first introduce the following key definitions. We associate with any deterministic loss sequence $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$ its randomly thresholded version $\widehat{\boldsymbol{\ell}}_1, \dots, \widehat{\boldsymbol{\ell}}_T \in \{0,1\}^K$ defined for all t by $\widehat{\ell}_{i,t} \triangleq \mathbb{I}_{\{\ell_{i,t}\geqslant U_{i,t}\}}$, $1 \leqslant i \leqslant K$. Moreover, we associate with any strategy $S = (\boldsymbol{p}_t)_{t\geqslant 1}$ its bernoullized variant $\widetilde{S} = (\widetilde{\boldsymbol{p}}_t)_{t\geqslant 1}$ defined for all $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$ by

$$\widetilde{\boldsymbol{p}}_t(\boldsymbol{\ell}_{1:t-1}) \triangleq \mathbb{E}_{\boldsymbol{U}_{1:t-1}}\left[\boldsymbol{p}_t\left(\widehat{\boldsymbol{\ell}}_{1:t-1}\right)\right] = \mathbb{E}_{\boldsymbol{U}_{1:t-1}}\left[\boldsymbol{p}_t\left(\left(\mathbb{I}_{\{\ell_{i,s}\geqslant U_{i,s}\}}\right)_{\substack{1\leqslant i\leqslant K\\ 1\leqslant s\leqslant t-1}}\right)\right] . \tag{5.23}$$

Thus, at each round t, the Bernoullized strategy $\widetilde{S}$ first transforms the past losses $\ell_{i,s}$, $1 \leqslant i \leqslant K$, $1 \leqslant s \leqslant t-1$, into independent Bernoulli random variables $\widehat{\ell}_{i,s}$ with respective parameters $\ell_{i,s}$, then applies the function $\boldsymbol{p}_t$ to them, and finally averages the result out.

As noted in Remark 5.2 after the present proof, for any strategy S, the Bernoullized variant $\widetilde{S}$ has a lower worst-case regret than S. However, computing $\widetilde{\boldsymbol{p}}_t(\boldsymbol{\ell}_{1:t-1})$ in practice requires to evaluate the function $\boldsymbol{p}_t$ at the $2^{K(t-1)}$ vertices of the hypercube $\{0,1\}^{K(t-1)}$, so that $\widetilde{S}$ is only of theoretical interest.

To prove the lemma it suffices to show that

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \psi_T\left(\sum_{t=1}^T \varphi\Big(\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t\Big)\right) \leqslant \inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in\{0,1\}^K} \psi_T\left(\sum_{t=1}^T \varphi\Big(\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t\Big)\right). \tag{5.24}$$

First, restricting the infimum of the right-hand side to the set of all Bernoullized strategies $\widetilde{S} = (\widetilde{\boldsymbol{p}}_t)_{t\geqslant 1}$, we get that

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \psi_T\left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t)\right) \leqslant \inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \psi_T\left(\sum_{t=1}^T \varphi\Big(\widetilde{\boldsymbol{p}}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t\Big)\right). \tag{5.25}$$

Let $S = (\boldsymbol{p}_t)_{t\geqslant 1}$ and $\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in [0,1]^K$. The definition of $\widetilde{S}$ in (5.23) and the equality $\mathbb{E}_{\boldsymbol{U}_t}\big[\widehat{\ell}_{i,t}\big] = \mathbb{E}_{\boldsymbol{U}_t}\big[\mathbb{I}_{\{\ell_{i,t}\geqslant U_{i,t}\}}\big] = \ell_{i,t}$ for all $i = 1,\dots,K$ and $t = 1,\dots,T$ yield

$$\begin{aligned}
\psi_T\left(\sum_{t=1}^T \varphi\Big(\widetilde{\boldsymbol{p}}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t\Big)\right) &= \psi_T\left(\sum_{t=1}^T \varphi\Big(\mathbb{E}_{\boldsymbol{U}_{1:t-1}}\Big[\boldsymbol{p}_t\Big(\widehat{\boldsymbol{\ell}}_{1:t-1}\Big)\Big], \mathbb{E}_{\boldsymbol{U}_t}\Big[\widehat{\boldsymbol{\ell}}_t\Big]\Big)\right) \\
&= \psi_T\left(\sum_{t=1}^T \mathbb{E}_{\boldsymbol{U}_{1:t}}\Big[\varphi\Big(\boldsymbol{p}_t\Big(\widehat{\boldsymbol{\ell}}_{1:t-1}\Big),\widehat{\boldsymbol{\ell}}_t\Big)\Big]\right) \qquad (5.26) \\
&= \psi_T\left(\mathbb{E}_{\boldsymbol{U}_{1:T}}\left[\sum_{t=1}^T \varphi\Big(\boldsymbol{p}_t\Big(\widehat{\boldsymbol{\ell}}_{1:t-1}\Big),\widehat{\boldsymbol{\ell}}_t\Big)\right]\right) \\
&\leqslant \mathbb{E}_{\boldsymbol{U}_{1:T}}\left[\psi_T\left(\sum_{t=1}^T \varphi\Big(\boldsymbol{p}_t\Big(\widehat{\boldsymbol{\ell}}_{1:t-1}\Big),\widehat{\boldsymbol{\ell}}_t\Big)\right)\right], \qquad (5.27)
\end{aligned}$$

where (5.26) follows from the fact that $\varphi : \mathbb{R}^K \times \mathbb{R}^K \to \mathbb{E}$ is bi-affine (cf. Definition 5.1) and from Fubini's theorem (since $\boldsymbol{U}_{1:t-1}$ and $\boldsymbol{U}_t$ are independent), and where (5.27) follows from Jensen's inequality (since $\psi_T : E \to \mathbb{R}$ is convex). Note that all expectations above are actually taken over a finite number of points since $\widehat{\boldsymbol{\ell}}_{1:T} \in \{0,1\}^{KT}$ almost surely. Therefore no additional assumption on the real vector space E was needed to ensure that these expectations are well-defined or to apply Fubini's theorem and Jensen's inequality.

Since $\widehat{\boldsymbol{\ell}}_{1:T} \in \{0,1\}^{KT}$, the expectation in (5.27) is upper bounded by a supremum over $\{0,1\}^{KT}$. Therefore we have proved that for all $S = (\boldsymbol{p}_t)_{t\geqslant 1}$,

$$\sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \psi_T\left(\sum_{t=1}^T \varphi\Big(\widetilde{\boldsymbol{p}}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t\Big)\right) \leqslant \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in\{0,1\}^K} \psi_T\left(\sum_{t=1}^T \varphi\Big(\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t\Big)\right). \tag{5.28}$$

Combining the last inequality with (5.25) immediately yields (5.24). This concludes the proof. □

Remark 5.2 (Bernoullization can only help)**.**
Note from (5.28) above that the Bernoullized variant $\widetilde{S}$ of any strategy S performs always better than S in a worst-case sense. More precisely, since the weights $\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1})$ of S and the weights

$\widetilde{\boldsymbol{p}}_t(\boldsymbol{\ell}_{1:t-1})$ of $\widetilde{S}$ coincide for all $\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in \{0,1\}^{KT}$, Inequality (5.28) is actually an equality:

$$\sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \psi_T\left(\sum_{t=1}^T \varphi\Big(\widetilde{\boldsymbol{p}}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t\Big)\right) = \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in\{0,1\}^K} \psi_T\left(\sum_{t=1}^T \varphi\Big(\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t\Big)\right)$$
$$\leqslant \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \psi_T\left(\sum_{t=1}^T \varphi\Big(\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t\Big)\right) .$$

Therefore, the Bernoullization of any strategy can only improve its worst-case regret on $[0,1]^{KT}$. Even better, by the above equality, the worst-case regret of $\widetilde{S}$ on $[0,1]^{KT}$ equals that of S on the restricted set $\{0,1\}^{KT}$ and is therefore not influenced by the (potentially bad) performance of S outside of the hypercube $\{0,1\}^{KT}$. However, as mentioned earlier, the strategy $\widetilde{S}$ is unfortunately only of theoretical interest because of its exponential computational complexity.

Proof (of Theorem 5.4): The proof consists of a careful application of a version of von Neumann's minimax theorem. We first use a reduction to binary losses (via Lemma 5.1), then apply the aforementioned version of von Neumann's minimax theorem, and finally get back to $[0,1]$-valued losses. To avoid any ambiguity, we write all dependencies $\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1})$ explicitely. By Lemma 5.1, we have

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \psi_T\left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t)\right) = \inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in\{0,1\}^K} \psi_T\left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t)\right)$$
$$= \inf_S \sup_{\mathbb{Q}\in\mathcal{M}_1^+(\{0,1\}^{KT})} \underbrace{\mathbb{E}_{\mathbb{Q}}\left[\psi_T\left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t)\right)\right]}_{\triangleq F(\mathbb{Q},S)} \tag{5.29}$$
$$= \sup_{\mathbb{Q}\in\mathcal{M}_1^+(\{0,1\}^{KT})} \inf_S \mathbb{E}_{\mathbb{Q}}\left[\psi_T\left(\sum_{t=1}^T \varphi(\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t)\right)\right] , \tag{5.30}$$

where in the last two equalities, the expectations $\mathbb{E}_{\mathbb{Q}}[\,\cdot\,]$ are taken with respect to the random variables $\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T \in [0,1]^K$ with joint distribution $\mathbb{Q}$, and where we used the following arguments.

- (5.29) is elementary: the inequality "$\leqslant$" follows by considering the Dirac probability distributions $\mathbb{Q} = \delta_{(\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T)}$ for all $(\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T) \in \{0,1\}^{KT}$; the inequality "$\geqslant$" follows from the fact that any expectation is smaller than or equal to the supremum of its integrand.

- As for (5.30), the equality $\inf_S \sup_{\mathbb{Q}} F(\mathbb{Q},S) = \sup_{\mathbb{Q}} \inf_S F(\mathbb{Q},S)$ follows by applying a version of von Neumann's minimax theorem due to [Fan53, Theorem 2] to the function $F : \mathcal{M}_1^+(\{0,1\}^{KT}) \times \mathcal{S} \to \mathbb{R}$ defined by[11] ($\mathcal{S}$ denotes the set of all strategies[12])

$$F(\mathbb{Q},S) \triangleq \mathbb{E}_{\mathbb{Q}}\left[\psi_T\left(\sum_{t=1}^T \varphi\Big(\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}),\boldsymbol{\ell}_t\Big)\right)\right]$$

[11] We make a slight abuse of notation by denoting with the same symbols $\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T$ either random variables (with joint distribution $\mathbb{Q}$) or fixed elements of $\{0,1\}^K$.

[12] Recall that a strategy is a sequence $S = (\boldsymbol{p}_t)_{t\geqslant 1}$ of Borel functions $\boldsymbol{p}_t : [0,1]^{K(t-1)} \to \mathcal{X}_K$.

$$= \sum_{\boldsymbol{\ell}_{1:T} \in \{0,1\}^{KT}} \mathbb{Q}\big(\{\boldsymbol{\ell}_{1:T}\}\big)\, \psi_T\left(\sum_{t=1}^{T} \varphi\Big(\boldsymbol{p}_t(\boldsymbol{\ell}_{1:t-1}), \boldsymbol{\ell}_t\Big)\right).$$

The aforementioned version of von Neumann's minimax theorem is recalled in Lemma A.1 (Appendix A.3). Its assumptions are immediately satisfied in this finite-dimensional setting: $\mathcal{M}_1^+(\{0,1\}^{KT}) \equiv \mathcal{X}_{2^{KT}}$ is a convex, Haussdorff, and compact subset of $\mathbb{R}^{2^{KT}}$ (under the Euclidean topology), $\mathcal{S}$ is clearly convex, and F satisfies:

- for all $S \in \mathcal{S}$, $\mathbb{Q} \mapsto F(\mathbb{Q}, S)$ is linear on $\mathcal{M}_1^+(\{0,1\}^{KT}) \equiv \mathcal{X}_{2^{KT}}$ (and thus concave and continuous);
- for all $\mathbb{Q} \in \mathcal{M}_1^+(\{0,1\}^{KT})$, $S \mapsto F(\mathbb{Q}, S)$ is convex on $\mathcal{S}$ since ψ_T is convex and $\varphi(\cdot, \boldsymbol{v})$ is affine for all $\boldsymbol{v} \in \mathbb{R}^K$.

We can thus apply Lemma A.1, which yields (5.30).

To conclude the proof, we get back to $[0,1]$-valued losses by noting that

$$\begin{aligned}
\sup_{\mathbb{Q} \in \mathcal{M}_1^+(\{0,1\}^{KT})} \inf_S \mathbb{E}_{\mathbb{Q}}\left[\psi_T\left(\sum_{t=1}^{T} \varphi(\boldsymbol{p}_t, \boldsymbol{\ell}_t)\right)\right] &\leqslant \sup_{\mathbb{Q} \in \mathcal{M}_1^+([0,1]^{KT})} \inf_S \mathbb{E}_{\mathbb{Q}}\left[\psi_T\left(\sum_{t=1}^{T} \varphi(\boldsymbol{p}_t, \boldsymbol{\ell}_t)\right)\right] \\
&\leqslant \inf_S \sup_{\mathbb{Q} \in \mathcal{M}_1^+([0,1]^{KT})} \mathbb{E}_{\mathbb{Q}}\left[\psi_T\left(\sum_{t=1}^{T} \varphi(\boldsymbol{p}_t, \boldsymbol{\ell}_t)\right)\right] \\
&= \inf_S \sup_{\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T \in [0,1]^K} \psi_T\left(\sum_{t=1}^{T} \varphi(\boldsymbol{p}_t, \boldsymbol{\ell}_t)\right),
\end{aligned}$$

where the second line follows from the standard inequality $\sup \inf \leqslant \inf \sup$, and where the last equality follows from arguments similar to those used for (5.29). By (5.30), all the previous inequalities are equalities, which concludes the proof. □

Remark 5.3 (Why use Bernoullization?)**.**
Bernoullization enables a reduction to simple topological spaces — e.g., $\mathcal{M}_1^+(\{0,1\}^{KT})$ lies in a finite-dimensional space. In particular, there is no need here to use finer topological notions such as weak topology.

5.5.3 Rederivation of known bounds on external, internal, and swap regret

In this section we use the above minimax theorem to rederive known regret bounds on individual sequences from a stochastic viewpoint. A similar treatment will be carried out in the next section to derive a new bound on the makespan regret.

For all forms of regret considered below – e.g., external, internal, and swap regret – we use the distribution-dependent strategy $S^*(\mathbb{Q})$ defined in (5.22). Recall that this strategy assigns at each time t a unit mass at the random index $I_t^* \in \operatorname{argmin}_{1 \leqslant i \leqslant K} \mathbb{E}_{\mathbb{Q}}\big[\ell_{i,t} \big| \boldsymbol{\ell}_{1:t-1}\big]$.

Let $\Phi \subset \mathcal{F}_K$ be a set of functions from $\{1, \ldots, K\}$ to $\{1, \ldots, K\}$; we denote its cardinality

by $|\Phi|$. The next proposition provides an upper bound on the minimax Φ-regret defined by

$$\inf_{S} \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \left\{ \sum_{t=1}^T \boldsymbol{p}_t\cdot\boldsymbol{\ell}_t - \min_{F\in\Phi} \sum_{t=1}^T (\boldsymbol{p}_t)^F\cdot\boldsymbol{\ell}_t \right\},$$

where the infimum is taken over all strategies $S=(\boldsymbol{p}_t)_{t\geqslant 1}$ of the forecaster and where $(\boldsymbol{p}_t)^F$ is the weight vector induced by $\boldsymbol{p}_t$ via the mapping F (cf. (5.5)). The notion of Φ-regret was introduced by [GJ03]; it includes as special cases the external, internal, and swap regrets (see below). Though we are mainly interested in those three particular cases, our analysis is generic enough to cover the cases of all subsets $\Phi\subset\mathcal{F}_K$ (see also [RST11]).

Proposition 5.4 (Φ-regret from a stochastic viewpoint)**.**
Let $\mathbb{Q}\in\mathcal{M}_1^+\left([0,1]^{KT}\right)$. Then, the strategy $S^*(\mathbb{Q})$ defined in (5.22) satisfies

$$\mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^T \boldsymbol{p}_t\cdot\boldsymbol{\ell}_t - \min_{F\in\Phi}\sum_{t=1}^T (\boldsymbol{p}_t)^F\cdot\boldsymbol{\ell}_t\right] \leqslant \sqrt{\frac{T}{2}\ln|\Phi|}\,, \tag{5.31}$$

where $\boldsymbol{\ell}_{1:T}$ is drawn at random from the joint distribution $\mathbb{Q}\in\mathcal{M}_1^+\left([0,1]^{KT}\right)$. As a consequence, the minimax Φ-regret on individual sequences satisfies

$$\inf_{S} \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \left\{ \sum_{t=1}^T \boldsymbol{p}_t\cdot\boldsymbol{\ell}_t - \min_{F\in\Phi} \sum_{t=1}^T (\boldsymbol{p}_t)^F\cdot\boldsymbol{\ell}_t \right\} \leqslant \sqrt{\frac{T}{2}\ln|\Phi|}\,,$$

where the infimum is taken over all strategies of the forecaster $S=(\boldsymbol{p}_t)_{t\geqslant 1}$.

Before proving the last proposition, note that, as a corollary, we can recover the best upper bounds known so far for the external, internal, and swap regrets.

- External regret corresponds to the transformation set $\Phi=\{F_i : i=1,\dots,K\}$, where F_i is defined by $F_i(k)=i$ for all $k=1,\dots,K$. Since $|\Phi|=K$, the above proposition entails that the minimax external regret is upper bounded by $\sqrt{(T/2)\ln K}$.

- Internal regret corresponds to the transformation set $\Phi=\{F_{i,j} : 1\leqslant i\neq j\leqslant K\}$, where $F_{i,j}$ is defined by $F_{i,j}(k)=k$ for all $k\neq i$ and by $F_{i,j}(i)=j$. Since $|\Phi|=K(K-1)\leqslant K^2$, the above proposition entails that the minimax internal regret is upper bounded by $\sqrt{T\ln K}$.

- Swap regret corresponds to the whole transformation set $\Phi=\mathcal{F}_K$, whose cardinality equals K^K. Therefore, by the above proposition, the minimax swap regret is upper bounded by $\sqrt{(T/2)K\ln K}$.

Remark 5.4. *For the external regret, the stochastic viewpoint not only enables to get the optimal rate $\sqrt{T\ln K}$ but also the asymptotically optimal constant $1/\sqrt{2}$ (cf. Remark 2.3 in Chapter 2, Section 2.3.2). As for the internal and swap regrets, the bounds proved above are the best known so far, and we know from (5.7) and Theorem 5.3 that they are rate-optimal up to a factor at most of the order of $\sqrt{\ln K}$ (see also the next remark).*

Remark 5.5. *For the particular case of internal regret, we do not know yet whether the strategy $S^*(\mathbb{Q})$ is sufficient to get the optimal individual sequence rate (i.e., whether the missing $\sqrt{\ln K}$ factor is necessary), but we do know from Section 5.3 that $S^*(\mathbb{Q})$ is suboptimal if the loss vectors are i.i.d.. Indeed, in the last case, a less aggressive strategy that appropriately spreads the probability mass among the actions achieves a $\sqrt{T}$-upper bound (while $S^*(\mathbb{Q})$ does not). A natural question which should be addressed in the future is whether $S^*(\mathbb{Q})$ can be refined in the same spirit as in Section 5.3 to get an internal regret at most of the order of $\sqrt{T}$ on individual sequences (if such bound is possible). See Section 5.6 for some suggestions.*

Proof (of Proposition 5.4): In the sequel we set $\bar{\ell}_{i,t} \triangleq \mathbb{E}_{\mathbb{Q}}\big[\ell_{i,t} \,\big|\, \boldsymbol{\ell}_{1:t-1}\big]$ for all $i = 1, \dots, K$ and all $t = 1, \dots, T$ (the dependence in $\mathbb{Q}$ is omitted). Since for all $t = 1, \dots, T$, the weight vector $\boldsymbol{p}_t^{\mathbb{Q}}$ defined in (5.22) is measurable with respect to $\boldsymbol{\ell}_{1:t-1}$, we get that

$$\mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^{T} \boldsymbol{p}_t^{\mathbb{Q}} \cdot \boldsymbol{\ell}_t\right] = \sum_{t=1}^{T}\sum_{i=1}^{K} \mathbb{E}_{\mathbb{Q}}\Big[p_{i,t}^{\mathbb{Q}}\, \mathbb{E}_{\mathbb{Q}}\big[\ell_{i,t} \,\big|\, \boldsymbol{\ell}_{1:t-1}\big]\Big] = \mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^{T} \boldsymbol{p}_t^{\mathbb{Q}} \cdot \bar{\boldsymbol{\ell}}_t\right] ,$$

where we set $\bar{\boldsymbol{\ell}}_t = (\bar{\ell}_{1,t}, \dots, \bar{\ell}_{K,t})$. Using the last equality and the fact that $\min_i a_i - \min_i b_i \leqslant \max_i (a_i - b_i)$ for all $(a_i)_i, (b_i)_i \in \mathbb{R}^K$, we get that

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^{T} \boldsymbol{p}_t^{\mathbb{Q}} \cdot \boldsymbol{\ell}_t - \min_{F\in\Phi} \sum_{t=1}^{T} \big(\boldsymbol{p}_t^{\mathbb{Q}}\big)^F \cdot \boldsymbol{\ell}_t\right] \\
&\qquad \leqslant \mathbb{E}_{\mathbb{Q}}\Bigg[\underbrace{\sum_{t=1}^{T} \boldsymbol{p}_t^{\mathbb{Q}} \cdot \bar{\boldsymbol{\ell}}_t - \min_{F\in\Phi} \sum_{t=1}^{T} \big(\boldsymbol{p}_t^{\mathbb{Q}}\big)^F \cdot \bar{\boldsymbol{\ell}}_t}_{\leqslant 0 \quad \text{a.s.}}\Bigg] + \mathbb{E}_{\mathbb{Q}}\left[\max_{F\in\Phi} \sum_{t=1}^{T} \big(\boldsymbol{p}_t^{\mathbb{Q}}\big)^F \cdot \big(\bar{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t\big)\right] . \qquad (5.32)
\end{aligned}$$

The first expectation of the right-hand side corresponds to what we called the "conditional variant" of the regret in Section 5.5.1. It is non-positive since, by definition of the weight vector $\boldsymbol{p}_t^{\mathbb{Q}} \triangleq \boldsymbol{\delta}_{I_t^*}$ and of the index $I_t^* \in \operatorname{argmin}_{1\leqslant i\leqslant K} \mathbb{E}_{\mathbb{Q}}\big[\ell_{i,t} \,\big|\, \boldsymbol{\ell}_{1:t-1}\big]$, we have, almost surely,

$$\sum_{t=1}^{T} \boldsymbol{p}_t^{\mathbb{Q}} \cdot \boldsymbol{\ell}_t = \sum_{t=1}^{T} \min_{1\leqslant i\leqslant K} \bar{\ell}_{i,t} \leqslant \min_{F\in\Phi} \sum_{t=1}^{T} \big(\boldsymbol{p}_t^{\mathbb{Q}}\big)^F \cdot \bar{\boldsymbol{\ell}}_t \, .$$

The last expectation of (5.32), which is a deviation term, can be upper bounded via a classical maximal inequality for subgaussian random variables that can be found, e.g., in [Mas07, Lemma 2.3 and Section 6.1.1] and that we recall in Appendix A.5. Note indeed that for all $F \in \Phi$, the random sequence

$$\Big(\big(\boldsymbol{p}_t^{\mathbb{Q}}\big)^F \cdot \big(\bar{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t\big)\Big)_{t\geqslant 1}$$

is a martingale difference sequence with respect to the filtration generated by the $\boldsymbol{\ell}_t$. Moreover, it takes its values in the predictable intervals $[A_t, A_t + 1]$, where $A_t \triangleq \big(\boldsymbol{p}_t^{\mathbb{Q}}\big)^F \cdot \bar{\boldsymbol{\ell}}_t - 1$ (since the losses are $[0,1]$-valued). Therefore, by the Hoeffding-Azuma inequality (cf. Lemma A.6 in Appendix A.5), the random variables $\sum_{t=1}^{T} \big(\boldsymbol{p}_t^{\mathbb{Q}}\big)^F \cdot \big(\bar{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t\big)$, $F \in \Phi$, are subgaussian with common variance factor $v = T/4$. Hence, by Lemma A.3 in Appendix A.5,

$$\mathbb{E}_{\mathbb{Q}}\left[\max_{F\in\Phi} \sum_{t=1}^{T} \big(\boldsymbol{p}_t^{\mathbb{Q}}\big)^F \cdot \big(\bar{\boldsymbol{\ell}}_t - \boldsymbol{\ell}_t\big)\right] \leqslant \sqrt{2\,\frac{T}{4}\ln|\Phi|} = \sqrt{\frac{T}{2}\ln|\Phi|} \, .$$

Substituting the last inequality in (5.32), we get (5.31), which concludes the first part of the proposition. As a consequence, for all $\mathbb{Q} \in \mathcal{M}_1^+\left([0,1]^{KT}\right)$,

$$\inf_S \mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^T \boldsymbol{p}_t^{\mathbb{Q}} \cdot \boldsymbol{\ell}_t - \min_{F\in\Phi} \sum_{t=1}^T \left(\boldsymbol{p}_t^{\mathbb{Q}}\right)^F \cdot \boldsymbol{\ell}_t\right] \leqslant \sqrt{\frac{T}{2}\ln|\Phi|}\,,$$

where the infimum is taken over all strategies of the forecaster $S = (\boldsymbol{p}_t)_{t\geqslant 1}$. Therefore, to prove the second part of the proposition, it suffices to use the minimax duality result of Theorem 5.4 to get that

$$\begin{aligned}\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} &\left\{\sum_{t=1}^T \boldsymbol{p}_t^{\mathbb{Q}} \cdot \boldsymbol{\ell}_t - \min_{F\in\Phi} \sum_{t=1}^T \left(\boldsymbol{p}_t^{\mathbb{Q}}\right)^F \cdot \boldsymbol{\ell}_t\right\} \\ &= \sup_{\mathbb{Q}\in\mathcal{M}_1^+([0,1]^{KT})} \inf_S \mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^T \boldsymbol{p}_t^{\mathbb{Q}} \cdot \boldsymbol{\ell}_t - \min_{F\in\Phi} \sum_{t=1}^T \left(\boldsymbol{p}_t^{\mathbb{Q}}\right)^F \cdot \boldsymbol{\ell}_t\right] \leqslant \sqrt{\frac{T}{2}\ln|\Phi|}\,.\end{aligned}$$

This concludes the proof. □

5.5.4 A new bound on the makespan regret

In this section, we derive a new bound on the minimax makespan regret. Following[13] [EDKMM09] (see Example 5.4), we define the makespan regret of any strategy $S = (\boldsymbol{p}_t)_{t\geqslant 1}$ on any loss sequence $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$ by

$$\max_{1\leqslant i\leqslant K} \sum_{t=1}^T p_{i,t}\ell_{i,t} - \min_{\boldsymbol{q}\in\mathcal{X}_K} \left\{\max_{1\leqslant i\leqslant K} \sum_{t=1}^T q_i \ell_{i,t}\right\}\,.$$

This notion of regret is useful, e.g., to model job scheduling or load balancing problems. In such settings, a decision-maker repeatedly distributes a job to K machines; $p_{i,t}$ denotes the proportion of the t-th job assigned to the i-th machine and $\ell_{i,t}$ denotes the loss (or load) per job unit incurred by this machine at time t (so that the decision-maker incurs the weighted loss $p_{i,t}\ell_{i,t}$ on this machine). The goal of the decision-maker is to minimize the worst cumulative weighted loss $\max_{1\leqslant i\leqslant K} \sum_{t=1}^T p_{i,t}\ell_{i,t}$ over the K machines, and his performance is compared to that of the best static allocation $\boldsymbol{q} \in \mathcal{X}_K$.

Using the same non-constructive stochastic viewpoint as in the previous subsection, we prove next that the minimax makespan regret is upper bounded by $\sqrt{T} + \sqrt{T\ln(2K)/2}$. This improves on the bound of order $\ln(K)\sqrt{T}$ initially obtained by [EDKMM09] through an explicit algorithm. The design of an explicit algorithm with the better rate $\sqrt{T\ln K}$ should of course be addressed in the future.

We mention that a similar upper bound of the order of $\sqrt{T\ln K}$ was derived independently by [RST11]; see Section 5.5.1, page 183, for further details.

[13] Note that, contrary to [EDKMM09], we chose not to normalize the sums $\sum_{t=1}^T p_{i,t}\ell_{i,t}$ and $\sum_{t=1}^T q_i\ell_{i,t}$ by T. This definition is of course equivalent to that of [EDKMM09], but it is more consistent with the other definitions of regret considered in this chapter (none of which is normalized).

We first define a new distribution-dependent strategy. Let $\mathbb{Q} \in \mathcal{M}_1^+\left([0,1]^{KT}\right)$ and set $\bar{\ell}_{i,t} \triangleq \mathbb{E}_{\mathbb{Q}}\left[\ell_{i,t} \,\middle|\, \boldsymbol{\ell}_{1:t-1}\right]$ for all $i = 1, \dots, K$ and $t = 1, \dots, T$ (where $\boldsymbol{\ell}_{1:T}$ has joint distribution $\mathbb{Q}$). We associate with $\mathbb{Q}$ the strategy $S^{\mathrm{mk}}(\mathbb{Q})$ whose weight vectors $(\boldsymbol{p}_t)_{t\geqslant 1}$ are recursively defined by

$$\boldsymbol{p}_t \in \operatorname*{argmin}_{\boldsymbol{q}\in\mathcal{X}_K} \max_{1\leqslant i\leqslant K} \left\{ \sum_{s=1}^{t-1} p_{i,s}\bar{\ell}_{i,s} + q_i\bar{\ell}_{i,t} \right\}, \quad 1 \leqslant t \leqslant T\,. \tag{5.33}$$

Note that $S^{\mathrm{mk}}(\mathbb{Q})$ is a greedy-type strategy minimizing the makespan regret associated with the conditional losses $\bar{\ell}_{i,s}$. But, by an elementary induction and by Lemma 5.2 in Appendix 5.B, we can see that, for all $t = 1, \dots, T$ and all $i = 1, \dots, K$,

$$p_{i,t}\bar{\ell}_{i,t} = \frac{1}{\sum_{j=1}^K 1/\bar{\ell}_{j,t}} \qquad \text{and} \qquad p_{i,t} = \frac{1/\bar{\ell}_{i,t}}{\sum_{j=1}^K 1/\bar{\ell}_{j,t}} \quad \text{if } \bar{\ell}_{i,t} > 0\,, \tag{5.34}$$

with the convention that $1/0 = +\infty$, $1/(+\infty) = 0$, and $x + (+\infty) = +\infty$ for all $x \in \mathbb{R}_+$.

Proposition 5.5 (A new bound on the makespan regret)**.**
Let $\mathbb{Q} \in \mathcal{M}_1^+\left([0,1]^{KT}\right)$. Then, the makespan regret of the strategy $S^{mk}(\mathbb{Q})$ defined in (5.33) satisfies

$$\mathbb{E}_{\mathbb{Q}}\left[\max_{1\leqslant i\leqslant K} \sum_{t=1}^T p_{i,t}\ell_{i,t} - \min_{\boldsymbol{q}\in\mathcal{X}_K} \max_{1\leqslant i\leqslant K} \sum_{t=1}^T q_i\ell_{i,t} \right] \leqslant \sqrt{T} + \sqrt{T\ln(2K)/2}\,, \tag{5.35}$$

where $\boldsymbol{\ell}_{1:T}$ is drawn at random from the joint distribution $\mathbb{Q} \in \mathcal{M}_1^+\left([0,1]^{KT}\right)$. As a consequence, the minimax makespan regret on individual sequences satisfies

$$\inf_S \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K} \left\{ \max_{1\leqslant i\leqslant K} \sum_{t=1}^T p_{i,t}\ell_{i,t} - \min_{\boldsymbol{q}\in\mathcal{X}_K} \max_{1\leqslant i\leqslant K} \sum_{t=1}^T q_i\ell_{i,t} \right\} \leqslant \sqrt{T} + \sqrt{T\ln(2K)/2}\,.$$

Proof: First note that, by subadditivity of the maximum, we can upper bound the makespan regret by the sum of its "conditional variant" and two deviation terms, i.e., almost surely,

$$\begin{aligned}
\max_{1\leqslant i\leqslant K} \sum_{t=1}^T p_{i,t}\ell_{i,t} - \min_{\boldsymbol{q}\in\mathcal{X}_K} \max_{1\leqslant i\leqslant K} \sum_{t=1}^T q_i\ell_{i,t} \leqslant{}& \max_{1\leqslant i\leqslant K} \sum_{t=1}^T p_{i,t}\bar{\ell}_{i,t} - \min_{\boldsymbol{q}\in\mathcal{X}_K} \max_{1\leqslant i\leqslant K} \sum_{t=1}^T q_i\bar{\ell}_{i,t} \\
&+ \max_{1\leqslant i\leqslant K} \sum_{t=1}^T p_{i,t}\big(\ell_{i,t} - \bar{\ell}_{i,t}\big) \\
&+ \min_{\boldsymbol{q}\in\mathcal{X}_K} \max_{1\leqslant i\leqslant K} \sum_{t=1}^T q_i\bar{\ell}_{i,t} - \min_{\boldsymbol{q}\in\mathcal{X}_K} \max_{1\leqslant i\leqslant K} \sum_{t=1}^T q_i\ell_{i,t}\,.
\end{aligned} \tag{5.36}$$

Next we upper bound each of the three terms of the righ-hand side separately.

Term 1.
The first term is non-positive since, by (5.34),

$$\max_{1\leqslant i\leqslant K}\sum_{t=1}^{T} p_{i,t}\bar{\ell}_{i,t} = \sum_{t=1}^{T}\frac{1}{\sum_{j=1}^{K} 1/\bar{\ell}_{j,t}} \leqslant \frac{1}{\sum_{j=1}^{K} 1/\big(\sum_{t=1}^{T}\bar{\ell}_{j,t}\big)} = \min_{\boldsymbol{q}\in\mathcal{X}_K}\max_{1\leqslant j\leqslant K} q_j\sum_{t=1}^{T}\bar{\ell}_{j,t}\,, \tag{5.37}$$

where the inequality above follows from Lemma 5.3 in Appendix 5.B applied to the vectors $(\bar{\ell}_{j,t})_{1\leqslant j\leqslant K}\in\mathbb{R}_+^K$, $t=1,\dots,T$, and where the last equality follows from Lemma 5.2 in Appendix 5.B.

Term 2.
The second term is upper bounded by $\sqrt{T}$ in expectation. Indeed, by the elementary maximal inequality $\mathbb{E}[\max_i Z_i]\leqslant\big(\sum_i\mathbb{E}[Z_i^2]\big)^{1/2}$ that holds for all integrable random vectors $(Z_i)_{1\leqslant i\leqslant K}\in\mathbb{R}^K$, we get that

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}\left[\max_{1\leqslant i\leqslant K}\sum_{t=1}^{T} p_{i,t}\big(\ell_{i,t}-\bar{\ell}_{i,t}\big)\right] &\leqslant \left(\sum_{i=1}^{K}\mathbb{E}_{\mathbb{Q}}\left[\left(\sum_{t=1}^{T} p_{i,t}\big(\ell_{i,t}-\bar{\ell}_{i,t}\big)\right)^2\right]\right)^{1/2}\\
&= \left(\sum_{i=1}^{K}\sum_{t=1}^{T}\mathbb{E}_{\mathbb{Q}}\Big[p_{i,t}^2\big(\ell_{i,t}-\bar{\ell}_{i,t}\big)^2\Big]\right)^{1/2} \qquad (5.38)\\
&\leqslant \left(\sum_{t=1}^{T}\mathbb{E}_{\mathbb{Q}}\left[\sum_{i=1}^{K} p_{i,t}^2\right]\right)^{1/2} \leqslant \sqrt{T}\,, \qquad (5.39)
\end{aligned}$$

where (5.38) follows from the Pythagorean theorem since, for every $i=1,\dots,K$, the random sequence $\big(\sum_{t=1}^{T} p_{i,t}\big(\ell_{i,t}-\bar{\ell}_{i,t}\big)\big)_{1\leqslant t\leqslant T}$ is a square-integrable martingale and therefore has orthogonal increments. As for (5.39) it follows from the boundedness property $|\ell_{i,t}-\bar{\ell}_{i,t}|\leqslant 1$ and from the inequality $\sum_{i=1}^{K} p_{i,t}^2\leqslant\sum_{i=1}^{K} p_{i,t}=1$.

Term 3.
Set $\bar{\boldsymbol{L}}_T\triangleq\big(\sum_{t=1}^{T}\bar{\ell}_{i,t}\big)_{1\leqslant i\leqslant K}$ and $\boldsymbol{L}_T\triangleq\big(\sum_{t=1}^{T}\ell_{i,t}\big)_{1\leqslant i\leqslant K}$, and define the function $f:\mathbb{R}_+^K\to\mathbb{R}$ by

$$f(\boldsymbol{x})\triangleq\min_{\boldsymbol{q}\in\mathcal{X}_K}\max_{1\leqslant i\leqslant K} q_i x_i\,.$$

Since f is 1-Lipschitz continuous with respect to the infinity norm $\|\cdot\|_\infty$ (essentially because the minimum and maximum functions are also 1-Lipschitz continuous), the last term of (5.36) reads, almost surely,

$$\begin{aligned}
&\min_{\boldsymbol{q}\in\mathcal{X}_K}\max_{1\leqslant i\leqslant K}\sum_{t=1}^{T} q_i\bar{\ell}_{i,t} - \min_{\boldsymbol{q}\in\mathcal{X}_K}\max_{1\leqslant i\leqslant K}\sum_{t=1}^{T} q_i\ell_{i,t}\\
&\qquad = f(\bar{\boldsymbol{L}}_T)-f(\boldsymbol{L}_T)\leqslant\big\|\bar{\boldsymbol{L}}_T-\boldsymbol{L}_T\big\|_\infty = \max_{1\leqslant i\leqslant K}\left|\sum_{t=1}^{T}\big(\bar{\ell}_{i,t}-\ell_{i,t}\big)\right|\,.
\end{aligned}$$

Taking the expectations of both sides of the inequality above, we get, using again the Hoeffding-Azuma inequality and an elementary maximal inequality for subgaussian random variables (cf.

Lemmas A.6 and A.3 in Appendix A.5),

$$\mathbb{E}_{\mathbb{Q}}\left[\min_{\boldsymbol{q}\in\mathcal{X}_K}\max_{1\leqslant i\leqslant K}\sum_{t=1}^{T} q_i\bar{\ell}_{i,t} - \min_{\boldsymbol{q}\in\mathcal{X}_K}\max_{1\leqslant i\leqslant K}\sum_{t=1}^{T} q_i\ell_{i,t}\right] \leqslant \sqrt{\frac{T}{2}\ln(2K)}\,. \tag{5.40}$$

Putting everything together.
We conclude the proof of (5.35) by substituting the upper bounds (5.37), (5.39), and (5.40) in (5.36). As for the second part of the proposition, it follows again by using the minimax duality property of Theorem 5.4. □

5.6 Future works

We recall that the present chapter is a work in progress, which raises some important questions. First, though the stochastic technique of Section 5.5 is useful to better understand the problem at hand (since it provides an upper bound on the minimax regret), it is non-constructive. Designing explicit algorithms that achieve the obtained upper bounds is an important task to be addressed in the future. For instance, is there any efficient algorithm with a makespan regret at most of order $\sqrt{T\ln K}$?

Another fundamental question that remains open is related to the missing logarithmic factor between the known lower and upper bounds on internal regret (of the order of $\sqrt{T}$ and $\sqrt{T\ln K}$ respectively). Is this logarithmic factor necessary or not? We proved that it is unncessary for i.i.d. loss vectors, we also recovered the best known upper bound $\sqrt{T\ln K}$ on individual sequences (with the best known constant) through a new viewpoint, but we still do not know whether the $\sqrt{\ln K}$ factor is necessary for individual sequences. We briefly sketch below some ideas to tackle either the lower bound or the upper bound. (Note that both directions could be useful in case the order of magnitude of the minimax internal regret for individual sequences lies strictly between $\sqrt{T}$ and $\sqrt{T\ln K}$.)

Note that similar questions arise about the minimax swap regret for individual sequences (the rate of which lies between $\sqrt{TK}$ and $\sqrt{TK\ln K}$). The next suggestions are however more suited for the internal regret.

Refinement of the $\sqrt{T\ln K}$ upper bound on internal regret?

By minimax duality, to prove a $\sqrt{T}$-upper bound on the minimax internal regret for individual sequences, it is sufficient to prove a $\sqrt{T}$-upper bound in the maximin game of Section 5.5.3. For arbitrary joint distributions $\mathbb{Q}$ on $[0,1]^{KT}$, the strategy $S^*(\mathbb{Q})$ achieves a $\sqrt{T\ln K}$-upper bound. The results of Section 5.3 indicate that this strategy is suboptimal in the particular case of i.i.d. loss vectors (i.e., for $\mathbb{Q}$ of the form $\mathbb{Q}=Q^{\otimes T}$, $Q\in\mathcal{M}_1^+([0,1]^K)$). In that setting, it can indeed be refined through exponential weighting to yield a $\sqrt{T}$-upper bound. Is such an improvement also possible for all joint distributions $\mathbb{Q}$ on $[0,1]^{KT}$? We could study a smooth variant of $S^*(\mathbb{Q})$ given, e.g., by the exponential weights $\boldsymbol{p}_t=(p_{i,t})_{1\leqslant i\leqslant K}$ defined as

$$p_{i,t} = \frac{e^{-c\sqrt{T}\,\bar{\ell}_{i,t}}}{\sum_{j=1}^{K} e^{-c\sqrt{T}\,\bar{\ell}_{j,t}}}\,, \quad 1\leqslant i\leqslant K\,,$$

where $\bar{\ell}_{i,t} \triangleq \mathbb{E}_{\mathbb{Q}}\big[\ell_{i,t} \,\big|\, \boldsymbol{\ell}_{1:t-1}\big]$ for all $i \in \{1,\dots,K\}$ and $t \geqslant 1$, and where $c > 0$ is an absolute constant. The above strategy generalizes that of Section 5.3 for i.i.d. loss vectors when T is known in advance (which is the case here since T is a parameter of the minimax rate).

Note that the weight vectors $\boldsymbol{p}_t$ suggested above are no longer constant over time in general. This prevents from factorizing the internal regret by p_i as we did in Section 5.3. Therefore, in this more general setting, the use of Bernstein's inequality for martingales (see [Fre75]) seems more appropriate than the Hoeffding-Azuma inequality. (Indeed, the conditional variances of the random variables $p_{i,t}(\ell_{i,t} - \ell_{j,t})$ should be taken into account to see that the deviations of $\sum_{t=1}^T p_{i,t}(\ell_{i,t} - \ell_{j,t})$ from $\sum_{t=1}^T p_{i,t}(\bar{\ell}_{i,t} - \bar{\ell}_{j,t})$ scale as $\sqrt{\sum_{t=1}^T p_{i,t}^2}$ instead of $\sqrt{T}$.)

Refinement of the $\sqrt{T}$ lower bound on internal regret?

Another direction consists in proving a larger lower bound (if such improvement is possible). In view of all known lower bounds on the external regret, it could be tempting to try to construct a suitable i.i.d. sequence (possibly depending on the strategy of the forecaster) for which the internal regret of the forecaster is at least of the order of $\sqrt{Tf(K)}$ with $f(K) \to +\infty$ when $K \to +\infty$. The results of Section 5.3 provide a negative answer by indicating that this is not possible (note that mixtures of i.i.d. sequences are banned as well). More sophisticated stochastic sequences could thus be studied in the future, e.g., piecewise i.i.d. Bernoulli sequences.

5.A Proofs

In this section we provide the proofs of Theorem 5.1, Theorem 5.2, and Corollary 5.1 (internal regret in a stochastic environment), as well as the proof of Theorem 5.3 (swap regret with individual sequences).

5.A.1 Proofs related to internal regret in a stochastic environment

Proof (of Theorem 5.1):
In the sequel we write $p_i = p_i^{\mathrm{int}}(Q)$ for notational convenience. First note that

$$\begin{aligned}\max_{1\leqslant i\neq j\leqslant K} \sum_{t=1}^T p_i(\ell_{i,t} - \ell_{j,t}) &\leqslant \max_{1\leqslant i,j\leqslant K} \sum_{t=1}^T p_i(\ell_{i,t} - \ell_{j,t}) \\ &= \max_{1\leqslant i\leqslant K} \left\{ p_i \left(\sum_{t=1}^T \ell_{i,t} - \min_{1\leqslant j\leqslant K} \sum_{t=1}^T \ell_{j,t} \right) \right\} , \end{aligned} \tag{5.41}$$

where the last equality follows from the fact that the weights p_i are constant over time.

Next we bound with high probability $\sum_{t=1}^T \ell_{i,t}$ from above and $\min_{1\leqslant j\leqslant K} \sum_{t=1}^T \ell_{j,t}$ from below. Since for all $i \in \{1,\dots,K\}$, the $\ell_{i,t}$, $1 \leqslant t \leqslant T$, are independent, $[0,1]$-valued, and have common mean m_i, Hoeffding's inequality (see Lemma A.5 in Appendix A.5) entails that

$$\forall i \in \{1,\dots,K\}, \qquad \forall \delta_i \in (0,1)\,, \qquad \mathbb{P}\left[\left| \sum_{t=1}^T \ell_{i,t} - Tm_i \right| > \sqrt{\frac{T}{2} \ln\left(\frac{2}{\delta_i}\right)} \right] \leqslant \delta_i\,.$$

Now, let $\delta \in (0,1)$ and $\alpha_1, \dots, \alpha_K > 0$ such that $\sum_{i=1}^K \alpha_i = 1$ (the α_i will be determined by the analysis). Combining the above inequality with a union bound, we get that, with probability at least $1 - \sum_{i=1}^K \alpha_i \delta = 1 - \delta$,

$$\forall i \in \{1, \dots, K\}, \qquad Tm_i - \sqrt{\frac{T}{2} \ln\left(\frac{2}{\alpha_i \delta}\right)} \leqslant \sum_{t=1}^T \ell_{i,t} \leqslant Tm_i + \sqrt{\frac{T}{2} \ln\left(\frac{2}{\alpha_i \delta}\right)} \,. \tag{5.42}$$

Therefore, with probability at least $1 - \delta$, for all $1 \leqslant i \leqslant K$,

$$\begin{aligned}
\sum_{t=1}^T \ell_{i,t} - \min_{1 \leqslant j \leqslant K} \sum_{t=1}^T \ell_{j,t} &\leqslant Tm_i + \sqrt{\frac{T}{2} \ln\left(\frac{2}{\alpha_i \delta}\right)} - \min_{1 \leqslant j \leqslant K} \left\{ Tm_j - \sqrt{\frac{T}{2} \ln\left(\frac{2}{\alpha_j \delta}\right)} \right\} \\
&= 2Tm_i - Tm_i + \sqrt{\frac{T}{2} \ln\left(\frac{2}{\alpha_i \delta}\right)} + \max_{1 \leqslant j \leqslant K} \left\{ -Tm_j + \sqrt{\frac{T}{2} \ln\left(\frac{2}{\alpha_j \delta}\right)} \right\} \\
&\leqslant 2Tm_i + 2 \max_{1 \leqslant j \leqslant K} \left\{ -Tm_j + \sqrt{\frac{T}{2} \ln\left(\frac{2}{\alpha_j \delta}\right)} \right\} .
\end{aligned}$$

Substituting the last inequality in (5.41) and using the fact that $m_i = m_{i^*} + \Delta_i$ by definition, we get that, with probability at least $1 - \delta$,

$$\begin{aligned}
\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^T p_i (\ell_{i,t} - \ell_{j,t}) &\leqslant \max_{1 \leqslant i \leqslant K} p_i \left(2T\Delta_i + 2 \max_{1 \leqslant j \leqslant K} \left\{ -T\Delta_j + \sqrt{\frac{T}{2} \ln\left(\frac{2}{\alpha_j \delta}\right)} \right\} \right) \\
&\leqslant 2\sqrt{T} \max_{1 \leqslant i \leqslant K} \left\{ p_i \sqrt{T} \Delta_i \right\} + 2 \left(\max_{1 \leqslant i \leqslant K} p_i \right) \left(\max_{1 \leqslant j \leqslant K} \left\{ -T\Delta_j + \sqrt{\frac{T}{2} \ln\left(\frac{2}{\alpha_j \delta}\right)} \right\} \right) ,
\end{aligned} \tag{5.43}$$

where in the last inequality, we used the subadditivity of the maximum and the fact the last maximum is nonnegative (since $-T\Delta_{i^*} = 0$). But, multiplying the numerator and the denominator of (5.17) by $\exp\big(\sqrt{T} m_{i^*}\big)$, the weights p_i can be rewritten for all $i \in \{1, \dots, K\}$ as

$$p_i = \frac{e^{-\sqrt{T}\Delta_i}}{\sum_{j=1}^K e^{-\sqrt{T}\Delta_j}} = \frac{e^{-\sqrt{T}\Delta_i}}{K_{\text{eff}}} \qquad \text{where} \quad K_{\text{eff}} \triangleq \sum_{j=1}^K e^{-\sqrt{T}\Delta_j} \in [1, K] \,. \tag{5.44}$$

Next we combine (5.43) with (5.44). First note from (5.44) that $p_i \sqrt{T} \Delta_i = K_{\text{eff}}^{-1} e^{-\sqrt{T}\Delta_i} \sqrt{T} \Delta_i$ so that $p_i \sqrt{T} \Delta_i \leqslant K_{\text{eff}}^{-1} \sup_{x \geqslant 0} \{e^{-x} x\} = 1/(e\, K_{\text{eff}})$. Second, we get $\max_{1 \leqslant i \leqslant K} p_i = 1/K_{\text{eff}}$ from (5.44). Subtituting the last two upper bounds in (5.43), we get, with probability at least $1 - \delta$,

$$\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^T p_i (\ell_{i,t} - \ell_{j,t}) \leqslant \frac{2\sqrt{T}}{e\, K_{\text{eff}}} + \frac{2}{K_{\text{eff}}} \max_{1 \leqslant i \leqslant K} \left\{ -T\Delta_i + \sqrt{\frac{T}{2} \ln\left(\frac{2}{\alpha_i \delta}\right)} \right\} . \tag{5.45}$$

It turns out that a convenient choice of the α_i is enough to get the claimed bound. Note that it is such that the lower confidence bounds of (5.42) on the quantities $\sum_{t=1}^T \ell_{i,t}$, $1 \leqslant i \leqslant K$, are

approximately equalized; see (5.47) below. More precisely, we set

$$\alpha_i \triangleq \frac{e^{-2T\Delta_i^2}}{\sum_{j=1}^K e^{-2T\Delta_j^2}}, \qquad 1 \leqslant i \leqslant K\,. \tag{5.46}$$

Substituting the definition of the α_i in (5.45), we get, with probability at least $1-\delta$,

$$\begin{aligned}
&\max_{1\leqslant i\neq j\leqslant K} \sum_{t=1}^T p_i(\ell_{i,t}-\ell_{j,t}) \\
&\quad\leqslant \frac{2\sqrt{T}}{e\,K_{\mathrm{eff}}} + \frac{2}{K_{\mathrm{eff}}} \max_{1\leqslant i\leqslant K} \left\{ -T\Delta_i + \sqrt{\frac{T}{2}\ln\Big(e^{2T\Delta_i^2}\Big) + \frac{T}{2}\ln\left(\frac{2}{\delta}\sum_{j=1}^K e^{-2T\Delta_j^2}\right)} \right\} \\
&\quad\leqslant \frac{2\sqrt{T}}{e\,K_{\mathrm{eff}}} + \frac{1}{K_{\mathrm{eff}}} \sqrt{2T\ln\left(\frac{2}{\delta}\sum_{j=1}^K e^{-2T\Delta_j^2}\right)}\,,
\end{aligned} \tag{5.47}$$

where the last inequality follows from the elementary upper bound $\sqrt{x+y} \leqslant \sqrt{x}+\sqrt{y}$ for all $x, y \geqslant 0$ (so that $-T\Delta_i$ and $T\Delta_i$ cancel out). But note that for all $1 \leqslant j \leqslant K$, we have $e^{-2T\Delta_j^2} \leqslant e^{-\sqrt{T}\Delta_j} \sup_{x\geqslant 0}\{e^{-2x^2+x}\} = e^{-\sqrt{T}\Delta_j}\, e^{1/8}$, so that

$$\sum_{j=1}^K e^{-2T\Delta_j^2} \leqslant e^{1/8} \sum_{j=1}^K e^{-\sqrt{T}\Delta_j} = e^{1/8} K_{\mathrm{eff}}\,. \tag{5.48}$$

Substituting the last inequality in (5.47), we get, with probability at least $1-\delta$,

$$\begin{aligned}
\max_{1\leqslant i\neq j\leqslant K} \sum_{t=1}^T p_i(\ell_{i,t}-\ell_{j,t}) &\leqslant \frac{2\sqrt{T}}{e\,K_{\mathrm{eff}}} + \frac{1}{K_{\mathrm{eff}}}\sqrt{2T\ln\left(\frac{2\,e^{1/8}K_{\mathrm{eff}}}{\delta}\right)} \\
&\leqslant \left(\frac{2}{e}+\sqrt{2}\right)\frac{1}{K_{\mathrm{eff}}}\sqrt{T\ln\left(\frac{3K_{\mathrm{eff}}}{\delta}\right)} \leqslant \frac{3}{K_{\mathrm{eff}}}\sqrt{T\ln\left(\frac{3K_{\mathrm{eff}}}{\delta}\right)}\,,
\end{aligned}$$

where the second inequality follows from the fact that $2\,e^{1/8} \leqslant 3$ and that $\sqrt{\ln(3K_{\mathrm{eff}}/\delta)} \geqslant 1$ (since $K_{\mathrm{eff}} \geqslant 1$ and $\delta \leqslant 1$). As for the last inequality, it follows from the elementary upper bound $2/e + \sqrt{2} \leqslant 3$.

We have just proved the first inequality of the theorem. The second one follows from the fact that the function $x \mapsto x^{-1}\sqrt{\ln(3x/\delta)}$ is nonincreasing[14] on $\big[1,+\infty\big)$ and from the inequality $K_{\mathrm{eff}} \geqslant 1$. This concludes the proof. □

Remark 5.6. *The weighted union bound with $(\alpha_1,\dots,\alpha_K)$ in the previous proof is key to derive an upper bound of order $\sqrt{T}$. It will also be useful in the case when Q is unknown — see below.*

[14] Indeed, the first derivative of $x \mapsto x^{-2}\ln(x)$ is equal to $x^{-3}\big(1-2\ln(x)\big)$, which is non-positive on $\big[e^{1/2},+\infty\big)$. Therefore, the function $x \mapsto x^{-1}\sqrt{\ln(3x/\delta)}$ is non-increasing on $\big[e^{1/2}\delta/3,+\infty\big)$ and in particular on $[1,+\infty)$ (since $e^{1/2}\delta/3 \leqslant e^{1/2}/3 \leqslant 1$).

Proof (of Theorem 5.2): The proof uses the same key arguments as that of Theorem 5.1 — in particular, careful weighted union bounds are central to our analysis. Therefore, we omit some details already encountered in the previous proof and only stress the major changes (namely, we now need to control the deviations of the estimates $\widehat{m}_i^{(r)}$ from their expectations m_i and to deal with the change of regimes).

In the sequel, we set $R \triangleq \lceil \log_2 T \rceil$, $t_{-1} \triangleq 0$, $t_r \triangleq 2^r$ for all $r \in \{0, \dots, R-1\}$, and $t_R \triangleq T$. Therefore $\{t_{r-1}+1, \dots, t_r\} = \{2^{r-1}+1, \dots, 2^r\} \cap [1, T]$ for all $1 \leqslant r \leqslant R$ and we have the partition $\{1, \dots, T\} = \bigcup_{r=0}^{R} \{t_{r-1}+1, \dots, t_r\}$.

First, rewriting the sum $\sum_{t=1}^{T} = \sum_{r=0}^{R} \sum_{t=t_{r-1}+1}^{t_r}$ and using the subadditivity of the maximum, we get that, almost surely,

$$\begin{aligned}\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^{T} p_{i,t}(\ell_{i,t} - \ell_{j,t}) &\leqslant \sum_{r=0}^{R} \max_{1 \leqslant i,j \leqslant K} \sum_{t=t_{r-1}+1}^{t_r} p_{i,t}(\ell_{i,t} - \ell_{j,t}) \\ &\leqslant 1 + \sum_{r=1}^{R} \max_{1 \leqslant i \leqslant K} \left\{ p_i^{(r)} \left(\sum_{t=t_{r-1}+1}^{t_r} \ell_{i,t} - \min_{1 \leqslant j \leqslant K} \sum_{t=t_{r-1}+1}^{t_r} \ell_{j,t} \right) \right\} . \end{aligned} \tag{5.49}$$

To get the last equality, we upper bounded the summand at $r = 0$ by 1 (since the losses lie in $[0, 1]$ and since $t_0 - t_{-1} = 1$), and we used for $r \geqslant 1$ the fact that $p_{i,t} = p_i^{(r)}$ for all $t \in \{t_{r-1}+1, \dots, t_r\}$.

Next we control the deviations of the sums $\sum_{t=t_{r-1}+1}^{t_r} \ell_{i,t}$ and of the estimators $\widehat{m}_i^{(r)}$ around their expectations, uniformly over all $r = 1, \dots, R$. Let $\delta \in (0, 1)$, and fix $\beta_1, \dots, \beta_R > 0$ such that $\sum_{r=1}^{R} \beta_r \leqslant 1$. Fix also $\alpha_{1,r}, \dots, \alpha_{K,r} > 0$ and $\alpha'_{1,r}, \dots, \alpha'_{K,r} > 0$ such that $\sum_{i=1}^{K} \alpha_{i,r} = 1$ and $\sum_{i=1}^{K} \alpha'_{i,r} = 1$ for all $r = 1, \dots, R$ (the β_r, $\alpha_{i,r}$, and $\alpha'_{i,r}$ will be determined by the analysis).

Following the same lines that led to (5.42), we get that, by Hoeffding's inequality and by several union bounds: on some event Ω_δ of probability at least $1 - \delta$, for all $1 \leqslant r \leqslant R$ and all $1 \leqslant i \leqslant K$,

$$\tau_r m_i - \sqrt{\frac{\tau_r}{2} \ln\left(\frac{4}{\beta_r \alpha_{i,r} \delta}\right)} \leqslant \sum_{t=t_{r-1}+1}^{t_r} \ell_{i,t} \leqslant \tau_r m_i + \sqrt{\frac{\tau_r}{2} \ln\left(\frac{4}{\beta_r \alpha_{i,r} \delta}\right)} \tag{5.50}$$

and

$$2^{r-1} m_i - \sqrt{\frac{2^{r-1}}{2} \ln\left(\frac{4}{\beta_r \alpha'_{i,r} \delta}\right)} \leqslant \sum_{t=1}^{2^{r-1}} \ell_{i,t} \leqslant 2^{r-1} m_i + \sqrt{\frac{2^{r-1}}{2} \ln\left(\frac{4}{\beta_r \alpha'_{i,r} \delta}\right)} , \tag{5.51}$$

where we set $\tau_r \triangleq t_r - t_{r-1}$. Note that, by definition of the t_r, we have $\tau_0 = 1$, $\tau_r = 2^{r-1}$ for all $r \in \{1, \dots, R-1\}$, and $\tau_R = T - 2^{R-1}$. In particular, $\tau_r \leqslant 2^{r-1}$ for all $1 \leqslant r \leqslant R$.

By (5.50) and by the same upper bounding that led to (5.43), we get, on Ω_δ, for all $r \in \{1, \dots, R\}$,

$$\begin{aligned}
&\max_{1\leqslant i\leqslant K} \left\{ p_i^{(r)} \left(\sum_{t=t_{r-1}+1}^{t_r} \ell_{i,t} - \min_{1\leqslant j\leqslant K} \sum_{t=t_{r-1}+1}^{t_r} \ell_{j,t} \right) \right\} \\
&\quad \leqslant 2\sqrt{\tau_r} \max_{1\leqslant i\leqslant K} \left\{ p_i^{(r)} \sqrt{\tau_r} \Delta_i \right\} + 2 \left(\max_{1\leqslant i\leqslant K} p_i^{(r)} \right) \max_{1\leqslant i\leqslant K} \left\{ -\tau_r \Delta_i + \sqrt{\frac{\tau_r}{2} \ln\left(\frac{4}{\beta_r \alpha_{i,r} \delta} \right)} \right\} \\
&\quad \leqslant 2\sqrt{2^{r-1}} \max_{1\leqslant i\leqslant K} \left\{ p_i^{(r)} \sqrt{2^{r-1}} \Delta_i \right\} + 2 \left(\max_{1\leqslant i\leqslant K} p_i^{(r)} \right) \sqrt{\frac{\tau_r}{2} \ln\left(\frac{4}{\beta_r \delta} \sum_{j=1}^K e^{-2\tau_r \Delta_j^2} \right)} \,, \qquad (5.52)
\end{aligned}$$

where the last inequality follows from the fact that $\tau_r \leqslant 2^{r-1}$ for all $1 \leqslant r \leqslant R$ and from a choice of $\alpha_{i,r}$ similar to (5.46), i.e.,

$$\alpha_{i,r} \triangleq \frac{e^{-2\tau_r \Delta_i^2}}{\sum_{j=1}^K e^{-2\tau_r \Delta_j^2}} \,, \qquad 1 \leqslant i \leqslant K \,, \qquad 1 \leqslant r \leqslant R \,.$$

(For the moment, we do not upper bound τ_r by 2^{r-1} in the last term of (5.52); see the tighter bound in (5.56).)

Contrary to the proof of Theorem 5.1, the weights $p_i^{(r)}$ are not exactly of the form $e^{-\sqrt{2^{r-1}}\Delta_i}/K_{\text{eff}}^{(r)}$ (since the gaps Δ_i are now estimated by their empirical counterparts $\widehat{m}_i^{(r)}$). Next we show that $p_i^{(r)} \leqslant C_r(\delta) e^{-\sqrt{2^{r-1}}\Delta_i/2}/K_{\text{eff}}^{(r)}$ on Ω_δ, where $C_r(\delta)$ is small enough, and where

$$K_{\text{eff}}^{(r)} \triangleq \sum_{j=1}^K \exp\left(-\sqrt{2^{r-1}} \frac{3\Delta_j}{2} \right) \,, \qquad 1 \leqslant r \leqslant R \,. \qquad (5.53)$$

For this purpose, we choose the $\alpha'_{i,r}$ in a way similar to the $\alpha_{i,r}$:

$$\alpha'_{i,r} \triangleq \frac{e^{-2^{r-1}\Delta_i^2/2}}{\sum_{j=1}^K e^{-2^{r-1}\Delta_j^2/2}} \,, \qquad 1 \leqslant i \leqslant K \,, \qquad 1 \leqslant r \leqslant R \,.$$

By definition of $\widehat{m}_i^{(r)} = 2^{-(r-1)} \sum_{t=1}^{2^{r-1}} \ell_{i,t}$ and of $\Delta_i \triangleq m_i - m_{i^*}$, it is easy to see from (5.51) and from the choice of $\alpha'_{i,r}$ above that, on Ω_δ, for all $i \in \{1, \dots, K\}$,

$$m_{i^*} + \frac{\Delta_i}{2} - B_r(\delta) \leqslant \widehat{m}_i^{(r)} \leqslant m_{i^*} + \frac{3\Delta_i}{2} + B_r(\delta) \,,$$

where we set

$$B_r(\delta) \triangleq \sqrt{\frac{1}{2^r} \ln\left(\frac{4}{\beta_r \delta} \sum_{j=1}^K e^{-2^{r-1}\Delta_j^2/2} \right)} \,.$$

Subtituting the last inequalities in the definition of $p_i^{(r)}$ (cf. Figure 5.3) and using the definition of

$K_{\text{eff}}^{(r)}$ in (5.53), we get that, on Ω_δ, for all $i \in \{1, \dots, K\}$,

$$p_i^{(r)} \leqslant \exp\Bigl(2\sqrt{2^{r-1}} B_r(\delta)\Bigr) \, \frac{\exp\Bigl(-\sqrt{2^{r-1}}\,\Delta_i/2\Bigr)}{K_{\text{eff}}^{(r)}} \,. \tag{5.54}$$

Before combining (5.52) with the above inequalities, note that, following the same lines that led to (5.48) and using the elementary equality $\sup_{x \geqslant 0} \bigl\{e^{-x^2/2+3x/2}\bigr\} = e^{9/8}$, we get the upper bound $\sum_{j=1}^K e^{-2^{r-1}\Delta_j^2/2} \leqslant e^{9/8} K_{\text{eff}}^{(r)}$, so that, by definition of $B_r(\delta)$ above,

$$2\sqrt{2^{r-1}} B_r(\delta) \leqslant \sqrt{2 \ln\left(\frac{4\, e^{9/8} K_{\text{eff}}^{(r)}}{\beta_r \delta}\right)} \,. \tag{5.55}$$

In the same way, the elementary equality $\sup_{x \geqslant 0} \bigl\{e^{-2x^2+3x/2}\bigr\} = e^{9/32}$ yields the upper bound $\sum_{j=1}^K e^{-2\tau_r \Delta_j^2} \leqslant \bigl(e^{9/32}\bigr)^{2^{r-1}/\tau_r} K_{\text{eff}}^{(r)}$, so that

$$\tau_r \ln\left(\frac{4}{\beta_r \delta} \sum_{j=1}^K e^{-2\tau_r \Delta_j^2}\right) \leqslant \tau_r \ln\left(\frac{4}{\beta_r \delta} \left(e^{9/32}\right)^{2^{r-1}/\tau_r} K_{\text{eff}}^{(r)}\right) \leqslant 2^{r-1} \ln\left(\frac{4\, e^{9/32} K_{\text{eff}}^{(r)}}{\beta_r \delta}\right) \,, \tag{5.56}$$

where the last inequality follows from the bound $\tau_r \leqslant 2^{r-1}$ and from the fact that $x \mapsto x \ln\bigl(a\, b^{1/x}\bigr)$ is nondecreasing on $\mathbb{R}_+^*$ for all $a \geqslant 1$ and all $b > 0$ (note that $4K_{\text{eff}}^{(r)}/(\beta_r \delta) \geqslant 1$). Substituting the upper bounds (5.54), (5.55), and (5.56) in (5.52), we get that, on Ω_δ, for all $r \in \{1, \dots, R\}$,

$$\begin{aligned}
&\max_{1 \leqslant i \leqslant K} \left\{ p_i^{(r)} \left(\sum_{t=t_{r-1}+1}^{t_r} \ell_{i,t} - \min_{1 \leqslant j \leqslant K} \sum_{t=t_{r-1}+1}^{t_r} \ell_{j,t} \right) \right\} \\
&\leqslant \frac{2}{K_{\text{eff}}^{(r)}} \exp\left(\sqrt{2 \ln\left(\frac{4\, e^{9/8} K_{\text{eff}}^{(r)}}{\beta_r \delta}\right)}\right) \left(\sqrt{2^{r-1}} \sup_{x \geqslant 0} \bigl\{e^{-x/2}\, x\bigr\} + \sqrt{\frac{2^{r-1}}{2} \ln\left(\frac{4\ e^{9/32} K_{\text{eff}}^{(r)}}{\beta_r \delta}\right)}\right) \\
&\leqslant c \sqrt{2^{r-1}} \exp\left(\sqrt{2 \ln\left(\frac{4}{\beta_r \delta}\right)}\right) \sqrt{\ln\left(\frac{4}{\beta_r \delta}\right)} \,,
\end{aligned} \tag{5.57}$$

where

$$c \triangleq \sup_{K' \geqslant 1} \left\{ \frac{2}{K'} \exp\left(\sqrt{2 \ln\bigl(e^{9/8} K'\bigr)}\right) \left(\sup_{x \geqslant 0} \bigl\{e^{-x/2}\, x\bigr\} + \frac{1}{\sqrt{2}} + \sqrt{\frac{\ln\bigl(e^{9/32} K'\bigr)}{2}} \right) \right\} \,.$$

Elementary manipulations show that $c \leqslant 2\, e^{3/2} \bigl(2/e + 1/\sqrt{2} + 1\bigr) e^2 \leqslant 162$. Substituting (5.57) in (5.49) and using the fact that $e^x\, x \leqslant e^{2x}$ for all $x \in \mathbb{R}$, we get that, on Ω_δ,

$$\begin{aligned}
\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^T p_{i,t} (\ell_{i,t} - \ell_{j,t}) &\leqslant 1 + \sum_{r=1}^R \frac{c}{\sqrt{2}} \sqrt{2^{r-1}} \exp\left(2 \sqrt{2 \ln\left(\frac{4}{\beta_r \delta}\right)}\right) \\
&\leqslant 1 + \frac{c}{\sqrt{2}} \exp\left(2 \sqrt{2 \ln(4/\delta)}\right) \sum_{r=1}^R \sqrt{2^{r-1}} \exp\left(2 \sqrt{2 \ln\bigl(1/\beta_r\bigr)}\right) \,.
\end{aligned}$$

Recall that the β_r are parameters of the analysis and can therefore be chosen at our own convenience. To avoid any extra $\exp\bigl(\sqrt{\ln T}\bigr)$ factor, we carry out a "backward weighted union bound" by choosing $\beta_r \triangleq \dfrac{6/\pi^2}{(R-r+1)^2}$ for all $1 \leqslant r \leqslant R$, so that $\sum_{r=1}^R \beta_r \leqslant 1$. This yields the bound

$$\begin{aligned}\sum_{r=1}^R \sqrt{2^{r-1}} \exp\biggl(2\sqrt{2\ln\bigl(1/\beta_r\bigr)}\biggr) &\leqslant \exp\biggl(2\sqrt{2\ln\bigl(\pi^2/6\bigr)}\biggr) \sum_{r=1}^R \sqrt{2^{r-1}} \exp\Bigl(4\sqrt{\ln(R-r+1)}\Bigr) \\ &\leqslant 8\sqrt{2^R} \underbrace{\sum_{k=1}^R \sqrt{2^{-k}} \exp\Bigl(4\sqrt{\ln(k)}\Bigr)}_{\triangleq c_0' < \infty} \leqslant 8c_0'\sqrt{2T}\,,\end{aligned}$$

where the second inequality follows from the change of variables $k = R-r+1$, and where the last inequality follows from the fact that $R \triangleq \lceil \log_2(T) \rceil \leqslant \log_2(T) + 1$. Combining the inequalities above, we conclude the proof by setting $c_0 \triangleq 8\,c\,c_0' < \infty$. □

Proof (of Corollary 5.1): The lower bound with the constants $c_1 \triangleq 1/192$ and $c_2 \triangleq 1/\bigl(64\sqrt{3}\bigr)$ follows straightforwardly from the proof of [Sto05, Theorem 3.3]. As for the upper bound, it follows by integrating the high-probability bound of Theorem 5.2. More precisely, combining Theorem 5.2 and Example A.2 in Appendix A.6, we get that

$$\inf_S \sup_{Q \in \mathcal{M}_1^+([0,1]^K)} \mathbb{E}_{Q^{\otimes T}}\left[\max_{1 \leqslant i \neq j \leqslant K} \sum_{t=1}^T p_{i,t}(\ell_{i,t} - \ell_{j,t})\right] \leqslant c_3'\sqrt{T} + 1 \leqslant (c_3'+1)\sqrt{T}\,,$$

where, using the constant c_0 of Theorem 5.2, we set $c_3' \triangleq c_0 \exp\bigl[2\sqrt{2\ln(4)}\,\bigr]\bigl(e^{16}+1\bigr)$ (this constant can be improved). We conclude the proof by setting $c_3 \triangleq c_3' + 1$. □

5.A.2 Proof of the lower bound on the swap regret

Proof (of Theorem 5.3): In the sequel we first assume that T is a mutiple of $\lfloor K/2 \rfloor \triangleq \max\bigl\{k \in \mathbb{N} : k \leqslant K/2\bigr\}$ (the general case will follow by monotonicity of the minimax swap regret in T — see the end of the proof). Under this assumption, we prove in what follows that

$$\sup_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T \in [0,1]^K} \left\{\sum_{t=1}^T \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F \in \mathcal{F}_K} \sum_{t=1}^T \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t\right\} \geqslant c\sqrt{2TK}\,. \tag{5.58}$$

We use the standard reduction to stochastic losses. Let $S = (\boldsymbol{p}_t)_{t \geqslant 1}$ be *any* strategy of the forecaster. First note that

$$\begin{aligned}&\sup_{\boldsymbol{\ell}_1,\ldots,\boldsymbol{\ell}_T \in [0,1]^K} \left\{\sum_{t=1}^T \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F \in \mathcal{F}_K} \sum_{t=1}^T \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t\right\} \\ &\qquad \geqslant \sup_{\mathbb{Q} \in \mathcal{M}_1^+\left([0,1]^{KT}\right)} \mathbb{E}_{\mathbb{Q}}\left[\sum_{t=1}^T \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F \in \mathcal{F}_K} \sum_{t=1}^T \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t\right]\,,\end{aligned} \tag{5.59}$$

where $\mathcal{M}_1^+\bigl([0,1]^{KT}\bigr)$ denotes the set of all probability distributions on $[0,1]^{KT}$ (endowed with its

Borel σ-algebra) and where the loss vectors $\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T \in [0,1]^K$ appearing inside the expectation $\mathbb{E}_{\mathbb{Q}}[\,\cdot\,]$ are drawn at random with joint distribution $\mathbb{Q}$.

Next we consider a finite family $(\mathbb{Q}_{\boldsymbol{\gamma}})_{\boldsymbol{\gamma}}$ of probability distributions on $[0,1]^{KT}$ under which the random[15] vectors $\boldsymbol{\ell}_t \in [0,1]^K$ are piecewise i.i.d.. The time interval $\{1, \dots, T\}$ is divided into $\lfloor K/2 \rfloor$ sub-intervals $\{t_{r-1}+1, \dots, t_r\}$ such that $t_0 \triangleq 0 < t_1 < \dots < t_{\lfloor K/2 \rfloor} \triangleq T$ and $t_r - t_{r-1} = T/\lfloor K/2 \rfloor$ for all $r = 1, \dots, \lfloor K/2 \rfloor$. Then, for all $\boldsymbol{\gamma} = (\gamma_r)_{1 \leqslant r \leqslant \lfloor K/2 \rfloor} \in \{0,1\}^{\lfloor K/2 \rfloor}$, we define the probability distribution $\mathbb{Q}_{\boldsymbol{\gamma}}$ on $[0,1]^{KT}$ such that, under $\mathbb{Q}_{\boldsymbol{\gamma}}$:

- the (real-valued) losses $\ell_{i,t}$, $1 \leqslant i \leqslant K$, $1 \leqslant t \leqslant T$ are independent;
- on each sub-interval $\{t_{r-1}+1, \dots, t_r\}$ ($1 \leqslant r \leqslant \lfloor K/2 \rfloor$), the loss vectors $\boldsymbol{\ell}_t \in [0,1]^K$ are i.i.d. and

$$\begin{cases} \ell_{i,t} = 1 \quad \text{a.s.} & \text{if} \quad i \notin \{2r-1, 2r\}\,, \\ \ell_{i,t} \sim \mathrm{Ber}(1/2 - \gamma_r \varepsilon) & \text{if} \quad i = 2r-1\,, \\ \ell_{i,t} \sim \mathrm{Ber}(1/2 - (1-\gamma_r)\varepsilon) & \text{if} \quad i = 2r\,, \end{cases}$$

where $\mathrm{Ber}(q)$ denotes the Bernoulli distribution with parameter $q \in [0,1]$ and where $\varepsilon \in (0, 1/2)$ will be chosen by the analysis. We also set $i_r(\boldsymbol{\gamma}) \triangleq 2r - (1-\gamma_r)$ and $j_r(\boldsymbol{\gamma}) \triangleq 2r - \gamma_r$, so that $\big\{i_r(\boldsymbol{\gamma}), j_r(\boldsymbol{\gamma})\big\} = \{2r-1, 2r\}$ and, for all $t \in \{t_{r-1}+1, \dots, t_r\}$,

$$\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\big[\ell_{i_r(\boldsymbol{\gamma}),t}\big] = 1/2, \quad \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\big[\ell_{j_r(\boldsymbol{\gamma}),t}\big] = 1/2 - \varepsilon, \quad \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\big[\ell_{k,t}\big] = 1, \, \forall k \notin \{2r-1, 2r\}\,. \tag{5.60}$$

Note also that

$$\forall t \notin \{t_{r-1}+1, \dots, t_r\}, \quad \ell_{i_r(\boldsymbol{\gamma}),t} = \ell_{j_r(\boldsymbol{\gamma}),t} = 1 \qquad \text{a.s.} \tag{5.61}$$

Next we use an induction argument and Pinsker's inequality to show that, for at least one $\boldsymbol{\gamma}$ in the hypercube $\{0,1\}^{\lfloor K/2 \rfloor}$, the expected swap regret under $\mathbb{Q}_{\boldsymbol{\gamma}}$ is at least of the order of $\sqrt{TK}$. The lower bound on individual sequences will then follow by (5.59). First note from the key equality (5.10) of Section 5.2.2 that, for all $\boldsymbol{\gamma} \in \{0,1\}^{\lfloor K/2 \rfloor}$,

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\left[\sum_{t=1}^T \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F \in \mathcal{F}_K} \sum_{t=1}^T \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t\right] &= \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\left[\sum_{i=1}^K \max_{1 \leqslant j \leqslant K} \sum_{t=1}^T p_{i,t}(\ell_{i,t} - \ell_{j,t})\right] \\ &\geqslant \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\left[\sum_{r=1}^{\lfloor K/2 \rfloor} \sum_{i \in \{2r-1, 2r\}} \max_{1 \leqslant j \leqslant K} \sum_{t=1}^T p_{i,t}(\ell_{i,t} - \ell_{j,t})\right] \qquad (5.62) \\ &\geqslant \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\left[\sum_{r=1}^{\lfloor K/2 \rfloor} \sum_{t=t_{r-1}+1}^{t_r} p_{i_r(\boldsymbol{\gamma}),t}(\ell_{i_r(\boldsymbol{\gamma}),t} - \ell_{j_r(\boldsymbol{\gamma}),t})\right]\,. \qquad (5.63) \end{aligned}$$

Inequality (5.62) follows from the fact that the pairs $\{2r-1, 2r\}$, $r = 1, \dots, \lfloor K/2 \rfloor$, are mutually disjoint subsets of $\{1, \dots, K\}$ and from the nonnegativity of $\max_{1 \leqslant j \leqslant K} \sum_{t=1}^T p_{i,t}(\ell_{i,t} - \ell_{j,t})$ for $i = K$ (useful when K is odd). As for (5.63), we only kept for each $r = 1, \dots, \lfloor K/2 \rfloor$ the term corresponding to $i = i_r(\boldsymbol{\gamma})$ (the other one being nonnegative) and used the fact that

[15] With a slight abuse of notation, we denote the identity function on $[0,1]^{TK}$ by $\boldsymbol{\ell}_{1:T} = (\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T)$. The coordinate mappings $\boldsymbol{\ell}_t : [0,1]^{TK} \to [0,1]^K$ can be seen as random vectors defined on the measurable space $[0,1]^{TK}$ (endowed with its Borel σ-algebra).

$\ell_{i_r(\boldsymbol{\gamma}),t} - \ell_{j_r(\boldsymbol{\gamma}),t} = 1 - 1 = 0$ a.s. for all $t \notin \{t_{r-1}+1, \dots, t_r\}$ (by (5.61)).

But, by (5.60) and by conditioning on $(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t-1})$, we have $\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\big[p_{i_r(\boldsymbol{\gamma}),t}(\ell_{i_r(\boldsymbol{\gamma}),t} - \ell_{j_r(\boldsymbol{\gamma}),t})\big] = \varepsilon\, \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\big[p_{i_r(\boldsymbol{\gamma}),t}\big]$. Substituting the last equality in (5.63), we get

$$\begin{aligned}
&\mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\left[\sum_{t=1}^{T} \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F \in \mathcal{F}_K} \sum_{t=1}^{T} \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t\right] \geqslant \varepsilon \sum_{r=1}^{\lfloor K/2 \rfloor} \sum_{t=t_{r-1}+1}^{t_r} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\big[p_{i_r(\boldsymbol{\gamma}),t}\big] \\
&= \varepsilon \sum_{r=1}^{\lfloor K/2 \rfloor} \sum_{t=t_{r-1}+1}^{t_r} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\left[1 - p_{j_r(\boldsymbol{\gamma}),t} - \sum_{k \notin \{2r-1,2r\}} p_{k,t}\right] \\
&\geqslant \varepsilon \sum_{r=1}^{\lfloor K/2 \rfloor} \frac{T}{\lfloor K/2 \rfloor} \left(1 - \underbrace{\frac{\lfloor K/2 \rfloor}{T} \sum_{t=t_{r-1}+1}^{t_r} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\big[p_{j_r(\boldsymbol{\gamma}),t}\big]}_{A_{r,\boldsymbol{\gamma}}} - \underbrace{\frac{\lfloor K/2 \rfloor}{T} \sum_{k \notin \{2r-1,2r\}} \sum_{t=t_{r-1}+1}^{t_r} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}[p_{k,t}]}_{B_{r,\boldsymbol{\gamma}}}\right),
\end{aligned} \tag{5.64}$$

where we used the fact that $t_r - t_{r-1} = T/\lfloor K/2 \rfloor$ for all $r = 1, \dots, \lfloor K/2 \rfloor$.

Next we show that, in non-trivial situations, $A_{r,\boldsymbol{\gamma}} \leqslant 3/4$ and $B_{r,\boldsymbol{\gamma}} \leqslant 1/8$ for an appropriate choice of $\boldsymbol{\gamma} \in \{0,1\}^{\lfloor K/2 \rfloor}$. We start with $B_{r,\boldsymbol{\gamma}}$. By (5.60) and yet another use of the tower rule, we have, for all $r \in \big\{1, \dots, \lfloor K/2 \rfloor\big\}$ and all $k \notin \{2r-1, 2r\}$,

$$\frac{1}{2} \sum_{t=t_{r-1}+1}^{t_r} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}[p_{k,t}] = \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\left[\sum_{t=t_{r-1}+1}^{t_r} p_{k,t}\big(\ell_{k,t} - \ell_{i_r(\boldsymbol{\gamma})}\big)\right] .$$

Following an argument of [Sto05, Theorem 3.3], note that the last expectation can be assumed to be smaller than $c\sqrt{2TK}$ for all $r \in \big\{1, \dots, \lfloor K/2 \rfloor\big\}$ and all $k \notin \{2r-1, 2r\}$. Otherwise, the lower bound of (5.58) would follow straightforwardly by using (5.59) on the sub-interval $\{t_{r-1}+1, \dots, t_r\}$ and by monotonicity[16] of the minimax swap regret in T. Therefore, we can assume that, for all $r \in \big\{1, \dots, \lfloor K/2 \rfloor\big\}$,

$$\frac{\lfloor K/2 \rfloor}{T} \sum_{k \notin \{2r-1,2r\}} \sum_{t=t_{r-1}+1}^{t_r} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}[p_{k,t}] \leqslant \frac{\lfloor K/2 \rfloor}{T} (K-2) 2\, c\sqrt{2TK} \leqslant \sqrt{\frac{2c^2K^5}{T}} \leqslant \frac{1}{8}\,, \tag{5.65}$$

where the last inequality follows from the assumption $T \geqslant 128c^2K^5$.

As for the term $A_{r,\boldsymbol{\gamma}}$, we show in what follows that, by iteratively using Pinsker's inequality, there

[16] To see why the minimax swap regret is nondecreasing in T, it suffices to show that the worst-case swap regret of any strategy S is nondecreasing in T. The latter fact is elementary by associating with each loss sequence $(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T)$ the loss sequence $(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_T, \mathbf{0})$.

exists $\boldsymbol{\gamma} \in \{0,1\}^{\lfloor K/2 \rfloor}$ such that, for all $r \in \big\{1, \dots, \lfloor K/2 \rfloor\big\}$,

$$\frac{\lfloor K/2 \rfloor}{T} \sum_{t=t_{r-1}+1}^{t_r} \mathbb{E}_{\mathbb{Q}_{\boldsymbol{\gamma}}}\big[p_{j_r(\boldsymbol{\gamma}),t}\big] \leqslant \frac{3}{4} \,. \tag{5.66}$$

For this purpose, we introduce an external randomization (as in [Sto05, Theorem 3.3]). Let $(\Omega_{\mathrm{ext}}, \mathcal{B}_{\mathrm{ext}}, \mathbb{Q}_{\mathrm{ext}})$ be a probability space, and let $I_1, \dots, I_T \in \{1, \dots, K\}$ be random variables defined on the augmented space $[0,1]^{TK} \times \Omega_{\mathrm{ext}}$ such that[17] I_t is measurable with respect to the σ-field $\sigma(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t-1}) \otimes \mathcal{B}_{\mathrm{ext}}$, and, for all $\boldsymbol{\gamma} \in \{0,1\}^{\lfloor K/2 \rfloor}$,

$$\forall t \in \{1, \dots, T\}, \quad \forall i \in \{1, \dots, K\}, \quad \mathbb{Q}_{\boldsymbol{\gamma}} \otimes \mathbb{Q}_{\mathrm{ext}}\Big[I_t = i \,\Big|\, (\boldsymbol{\ell}_1, I_1), \dots, (\boldsymbol{\ell}_{t-1}, I_{t-1})\Big] = p_{i,t} \,.$$

By the property above, (5.66) is equivalent to

$$\begin{aligned} &\frac{\lfloor K/2 \rfloor}{T} \sum_{t=t_{r-1}+1}^{t_r} \mathbb{Q}_{\boldsymbol{\gamma}} \otimes \mathbb{Q}_{\mathrm{ext}}\big[I_t = j_r(\boldsymbol{\gamma})\big] \leqslant \frac{3}{4} \,, \\ \text{i.e.,} \qquad &\frac{\lfloor K/2 \rfloor}{T} \sum_{t=t_{r-1}+1}^{t_r} \mathbb{Q}_{(\gamma_1, \dots, \gamma_r, \bullet)} \otimes \mathbb{Q}_{\mathrm{ext}}\big[I_t = 2r - \gamma_r\big] \leqslant \frac{3}{4} \,, \end{aligned} \tag{5.67}$$

where we used the definition of $j_r(\boldsymbol{\gamma}) \triangleq 2r - \gamma_r$, and where $\mathbb{Q}_{(\gamma_1, \dots, \gamma_r, \bullet)}$ denotes the joint distribution of $(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t_r})$ (note that for all $t \leqslant t_r$, I_t is measurable with respect to $\sigma(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t-1}) \otimes \mathcal{B}_{\mathrm{ext}}$ and a fortioti to $\sigma(\boldsymbol{\ell}_1, \dots, \boldsymbol{\ell}_{t_r}) \otimes \mathcal{B}_{\mathrm{ext}}$).

Next we define $\gamma^* = \big(\gamma_1^*, \dots, \gamma_{\lfloor K/2 \rfloor}^*\big) \in \{0,1\}^{\lfloor K/2 \rfloor}$ such that the condition (5.67) holds for all $r \in \big\{1, \dots, \lfloor K/2 \rfloor\big\}$. Fix

$$\gamma_1^* \in \operatorname*{argmin}_{\gamma_1 \in \{0,1\}} \left\{ \sum_{t=t_0+1}^{t_1} \mathbb{Q}_{(\gamma_1, \bullet)}\Big[I_t = 2 - \gamma_1\Big] \right\} \,,$$

and, by induction,

$$\gamma_r^* \in \operatorname*{argmin}_{\gamma_r \in \{0,1\}} \left\{ \sum_{t=t_{r-1}+1}^{t_r} \mathbb{Q}_{(\gamma_1^*, \dots, \gamma_{r-1}^*, \gamma_r, \bullet)} \otimes \mathbb{Q}_{\mathrm{ext}}\Big[I_t = 2r - \gamma_r\Big] \right\} \,.$$

The definition of γ_r^* above is motivated by the use of Pinsker's inequality — see (5.73) below. Let $r = 1, \dots, \lfloor K/2 \rfloor$. Then, using Pinsker's inequality at each $t \in \{t_{r-1}+1, \dots, t_r\}$ (see Lemma A.8 in Appendix A.7), and averaging the resulting bounds, we get, for all $\gamma_r \in \{0,1\}$,

$$\begin{aligned} &\frac{\lfloor K/2 \rfloor}{T} \sum_{t=t_{r-1}+1}^{t_r} \mathbb{Q}_{(\gamma_1^*, \dots, \gamma_{r-1}^*, \gamma_r, \bullet)} \otimes \mathbb{Q}_{\mathrm{ext}}\big[I_t = 2r - \gamma_r\big] \\ &\leqslant \frac{\lfloor K/2 \rfloor}{T} \sum_{t=t_{r-1}+1}^{t_r} \mathbb{P}_{(\gamma_1^*, \dots, \gamma_{r-1}^*, \bullet)} \otimes \mathbb{Q}_{\mathrm{ext}}\big[I_t = 2r - \gamma_r\big] + \frac{\lfloor K/2 \rfloor}{T} \sum_{t=t_{r-1}+1}^{t_r} \sqrt{\frac{\bar{\mathcal{K}}_{r,t}(\gamma_r)}{2}} \,, \end{aligned} \tag{5.68}$$

[17] The random variables I_t can be constructed as follows: at each time $t = 1, \dots, T$, pick $I_t \in \{1, \dots, K\}$ at random such that $I_t = i$ with probability $p_{i,t}$ (conditionally on the past data $(\boldsymbol{\ell}_1, I_1), \dots, (\boldsymbol{\ell}_{t-1}, I_{t-1})$).

where $\mathbb{P}_{(\gamma_1^*,\dots,\gamma_{r-1}^*,\bullet)}$ is defined similarly to $\mathbb{Q}_{(\gamma_1^*,\dots,\gamma_{r-1}^*,\gamma_r,\bullet)}$ except that $\ell_{2r-1,t}$ and $\ell_{2r,t}$ are both $\mathrm{Ber}(1/2)$ under $\mathbb{P}_{(\gamma_1^*,\dots,\gamma_{r-1}^*,\bullet)}$ on the regime $\{t_{r-1}+1,\dots,t_r\}$, and where we set

$$\begin{aligned}
\bar{\mathcal{K}}_{r,t}(\gamma_r) &\triangleq \mathcal{K}\Big(\Big(\mathbb{P}_{(\gamma_1^*,\dots,\gamma_{r-1}^*,\bullet)}\otimes\mathbb{Q}_{\mathrm{ext}}\Big)^{I_t},\Big(\mathbb{Q}_{(\gamma_1^*,\dots,\gamma_{r-1}^*,\gamma_r,\bullet)}\otimes\mathbb{Q}_{\mathrm{ext}}\Big)^{I_t}\Big) && (5.69)\\
&\leqslant \mathcal{K}\Big(\mathbb{P}_{(\gamma_1^*,\dots,\gamma_{r-1}^*,\bullet)}\otimes\mathbb{Q}_{\mathrm{ext}},\ \mathbb{Q}_{(\gamma_1^*,\dots,\gamma_{r-1}^*,\gamma_r,\bullet)}\otimes\mathbb{Q}_{\mathrm{ext}}\Big) && (5.70)\\
&= \frac{T}{\lfloor K/2\rfloor}\mathcal{K}\Big(\mathrm{Ber}(1/2),\mathrm{Ber}(1/2-\varepsilon)\Big) && (5.71)\\
&\leqslant \frac{T}{\lfloor K/2\rfloor}8\ln(4/3)\,\varepsilon^2 \leqslant \frac{32T\ln(4/3)\,\varepsilon^2}{K} && (5.72)
\end{aligned}$$

provided that $\varepsilon \leqslant 1/4$. In (5.69) we denote by $\big(\mathbb{Q}'\otimes\mathbb{Q}_{\mathrm{ext}}\big)^{I_t}$ the law of I_t under $\mathbb{Q}'\otimes\mathbb{Q}_{\mathrm{ext}}$. Inequality (5.70) follows by joint convexity of $\mathcal{K}(\cdot,\cdot)$. To get (5.71), we used the chain rule for the Kullback-Leibler divergence, the independence of the losses $\ell_{i,t}$, $1\leqslant i\leqslant K$, $1\leqslant t\leqslant T$, and the fact that $\mathbb{P}_{(\gamma_1^*,\dots,\gamma_{r-1}^*,\bullet)}$ and $\mathbb{Q}_{(\gamma_1^*,\dots,\gamma_{r-1}^*,\gamma_r,\bullet)}$ only differ on the regime $\{t_{r-1}+1,\dots,t_r\}$ (of length $T/\lfloor K/2\rfloor$) at $i=2r-\gamma_r$. Finally, (5.72) follows from the fact that $\mathcal{K}\big(\mathrm{Ber}(1/2),\mathrm{Ber}(1/2-\varepsilon)\big) = -\ln(1-4\varepsilon^2)/2$ and that $-\ln(1-x)\leqslant 4\ln(4/3)\,x$ for all $x\in(0,1/4)$ (see, e.g., [CBL06, pp. 167-168]), and from the elementary inequality $\lfloor K/2\rfloor \geqslant K/4$ (since $K\geqslant 2$).

Substituting the upper bound of (5.72) (that does not depend on γ_r) in (5.68) and using the definition of γ_r^*, we get

$$\begin{aligned}
&\frac{\lfloor K/2\rfloor}{T}\sum_{t=t_{r-1}+1}^{t_r}\mathbb{Q}_{(\gamma_1^*,\dots,\gamma_r^*,\bullet)}\otimes\mathbb{Q}_{\mathrm{ext}}\big[I_t=2r-\gamma_r^*\big]\\
&\leqslant \underbrace{\min_{\gamma_r\in\{0,1\}}\frac{\lfloor K/2\rfloor}{T}\sum_{t=t_{r-1}+1}^{t_r}\mathbb{P}_{(\gamma_1^*,\dots,\gamma_{r-1}^*,\bullet)}\otimes\mathbb{Q}_{\mathrm{ext}}\big[I_t=2r-\gamma_r\big]}_{\leqslant 1/2}+\sqrt{\frac{16T\ln(4/3)\varepsilon^2}{K}} && (5.73)\\
&\leqslant 3/4\,, && (5.74)
\end{aligned}$$

where the upper bound by $1/2$ in (5.73) follows from the fact that[18]

$$\begin{aligned}
&\min_{\gamma_r\in\{0,1\}}\sum_{t=t_{r-1}+1}^{t_r}\mathbb{P}_{(\gamma_1^*,\dots,\gamma_{r-1}^*,\bullet)}\otimes\mathbb{Q}_{\mathrm{ext}}\big[I_t=2r-\gamma_r\big]\\
&\quad\leqslant \frac{1}{2}\sum_{\gamma_r\in\{0,1\}}\sum_{t=t_{r-1}+1}^{t_r}\mathbb{P}_{(\gamma_1^*,\dots,\gamma_{r-1}^*,\bullet)}\otimes\mathbb{Q}_{\mathrm{ext}}\big[I_t=2r-\gamma_r\big]\leqslant\frac{t_r-t_{r-1}}{2}=\frac{T}{2\lfloor K/2\rfloor}\,,
\end{aligned}$$

and where (5.74) holds provided that $\varepsilon\leqslant 1/4$ and $16T\ln(4/3)\varepsilon^2/K\leqslant 1/16$. For such an ε, and for the choice of $\boldsymbol{\gamma}^*\triangleq\big(\gamma_1^*,\dots,\gamma_{\lfloor K/2\rfloor}^*\big)$ above, we have just proved (5.67) or, equivalently, (5.66).

[18] Note that the minimum of two quantities is smaller than their means, and that $\{I_t=2r\}\cap\{I_t=2r-1\}=\varnothing$.

Combining it with (5.64) and (5.65), we finally get

$$\mathbb{E}_{\mathbb{Q}_{\gamma^*}}\left[\sum_{t=1}^{T} \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F\in\mathcal{F}_K} \sum_{t=1}^{T} \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t\right] \geqslant \varepsilon T/8 \,.$$

Choosing $\varepsilon = 8c\sqrt{2K/T}$ with $c \triangleq 1/\big(16\sqrt{128\ln(4/3)}\big)$, we can check that $16T\ln(4/3)\varepsilon^2/K \leqslant 1/16$ and that $\varepsilon \leqslant 1/4$ (since by assumption $T \geqslant 128c^2K^5 \geqslant (32c)^2 2K$ because $K^4 \geqslant 2^4$). This choice of ε yields the lower bound $c\sqrt{2TK}$ under $\mathbb{Q}_{\gamma^*}$, which in turn yields (5.58) by (5.59). This concludes the proof of the theorem when T is a multiple of $\lfloor K/2\rfloor$.

General case: We no longer assume that T is a multiple of $\lfloor K/2\rfloor$.
We use a reduction to the previous case. Denote by $T' \in \mathbb{N}$ the largest multiple of $\lfloor K/2\rfloor$ smaller than or equal to T. Then, by monotonocity of the worst-case swap regret in T (see Footnote 16 on Page 202), we have

$$\begin{aligned}\sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_T\in[0,1]^K}\left\{\sum_{t=1}^{T} \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F\in\mathcal{F}_K} \sum_{t=1}^{T} \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t\right\} &\geqslant \sup_{\boldsymbol{\ell}_1,\dots,\boldsymbol{\ell}_{T'}\in[0,1]^K}\left\{\sum_{t=1}^{T'} \boldsymbol{p}_t \cdot \boldsymbol{\ell}_t - \min_{F\in\mathcal{F}_K} \sum_{t=1}^{T'} \boldsymbol{p}_t^F \cdot \boldsymbol{\ell}_t\right\} \\ &\geqslant c\sqrt{2T'K}\,,\end{aligned}$$

where the last inequality follows from the previous analysis (see (5.58)). We conclude the proof by noting that, by definition of T',

$$T' > T - \lfloor K/2\rfloor \geqslant T/2$$

(since $\lfloor K/2\rfloor \leqslant T/2$ by the assumption $T \geqslant K$). □

5.B Elementary lemmas

The first lemma of this section follows from elementary manipulations and can be found, e.g., in [EDKMM09, Lemma 3].

Lemma 5.2. *Let $K \geqslant 1$ and $x_1,\dots,x_K \in \mathbb{R}_+$. Then, using the conventions $1/0 = +\infty$, $1/(+\infty) = 0$, and $x + (+\infty) = +\infty$ for all $x \in \mathbb{R}_+$, we have*

$$\min_{\boldsymbol{q}\in\mathcal{X}_K} \max_{1\leqslant i\leqslant K} q_i x_i = \frac{1}{\sum_{i=1}^K 1/x_i}\,.$$

Moreover, for every minimizer $\boldsymbol{q} \in \mathcal{X}_K$ of the above expression, we have, for all $i = 1,\dots,K$,

$$q_i x_i = \frac{1}{\sum_{j=1}^K 1/x_j} \qquad \textit{so that} \qquad q_i = \frac{1/x_i}{\sum_{j=1}^K 1/x_j} \quad \textit{if } x_i > 0\,.$$

The second lemma can be obtained from elementary calculations or directly seen as a consequence of the concavity of $(u_1,\dots,u_K) \in (\mathbb{R}_+^*)^K \mapsto \big(\sum_{j=1}^K 1/u_j\big)^{-1}$ proved, e.g., in [EDKMM09, Lemma 22]. It indicates that the harmonic mean is superadditive.

Lemma 5.3. *Let $K, T \geqslant 1$ and $\boldsymbol{x}_1, \dots, \boldsymbol{x}_T \in \mathbb{R}_+^K$. Then, using the same conventions as above, we have*

$$\sum_{t=1}^{T} \frac{1}{\sum_{j=1}^{K} 1/x_{j,t}} \leqslant \frac{1}{\sum_{j=1}^{K} 1/\left(\sum_{t=1}^{T} x_{j,t}\right)} \,.$$

Chapter 6

Aggregation of nonlinear models

We consider the generalized linear Gaussian framework introduced in [BM01a], which includes as special cases the Gaussian regression model with fixed design and the white noise framework. Given a collection of subsets (or nonlinear models) in a separable Hilbert space, the goal is to estimate the unknown vector almost as well as the best of the least squares estimators associated with the models in the collection. In this setting we analyse a Bayesian variant of the celebrated general model selection procedure of [BM01a, Mas07]. As in [LB06], our procedure is based on exponential weighting, but the models at hand can be arbitrary. In such generality, we use the concentration approach of [Mas07] and derive (non-sharp) oracle-type inequalities with high probability. This work exhibits a natural connection between model aggregation and model selection: our oracle-type inequalities hold for a continuum of estimators ranging from classical model aggregation (where the inverse temperature parameter is small enough) to model selection (where the inverse temperature parameter is infinite). We finally prove a lower bound indicating that aggregation is more robust that model selection in case of linear models. This lower bound suggests that aggregation might benefit from a similar advantage with nonlinear models.

DISCLAIMER: This chapter is a work in progress. In particular, important questions remain open (see Section 6.5). The preliminary results stated thereafter were presented at the workshop StatMathAppli 2011 [Ger11b].

Contents

6.1 Introduction

In this section we briefly introduce our framework and our statistical procedure. We also discuss our main contributions and some related works. For the sake of clarity, we omit some technical details, which are postponed to Section 6.2.

We consider the generalized linear Gaussian framework introduced in [BM01a], i.e., one observes the whole stochastic process $\big(Y_\varepsilon(t)\big)_{t\in\mathbb{H}}$ given by

$$Y_\varepsilon(t) = <s,t> + \varepsilon W(t)\;, \quad t\in\mathbb{H}\;, \tag{6.1}$$

where $\big(\mathbb{H}, <\cdot,\cdot>\big)$ is some separable Hilbert space, where W is an isonormal process on $\mathbb{H}$ (i.e., an isometry from $\mathbb{H}$ onto a centered Gaussian space[1]), where the noise level $\varepsilon>0$ is assumed to be known, and where $s\in\mathbb{H}$ is the unknown vector to be estimated. See Examples 6.1 and 6.2.

To estimate the unknown vector s, the statistician is given a family of least-squares estimators associated with different models, and his goal is to mimic the best of them. More precisely, following the same lines as [Mas07], we fix some at most countable collection $(S_m)_{m\in\mathcal{M}}$ of non-empty subsets of $\mathbb{H}$, which will be refered to as the *models* thereafter. For each $m\in\mathcal{M}$, the closest point to s in S_m (when it exists) is $s_m\in\operatorname{argmin}_{t\in S_m}\|t-s\|^2 = \operatorname{argmin}_{t\in S_m}\big\{\|t\|^2 - 2<s,t>\big\}$. Therefore, a natural estimator of s within the model S_m is a least-squares estimator[2] $\widehat{s}_m\in S_m$ defined by

$$\widehat{s}_m \in \operatorname*{argmin}_{t\in S_m}\gamma_\varepsilon(t)\;, \quad \text{where } \gamma_\varepsilon(t)\triangleq\|t\|^2 - 2Y_\varepsilon(t)\;. \tag{6.2}$$

In this setting, a natural goal is to construct an estimator $\widetilde{s}\in\mathbb{H}$ of s which is almost as good as the best least-squares estimator in the family $\big(\widehat{s}_m\big)_{m\in\mathcal{M}}$. This is the case, when, e.g., $\widetilde{s}$ satisfies a risk bound of the form

$$\mathbb{E}_s\Big[\|\widetilde{s}-s\|^2\Big] \leqslant C\inf_{m\in\mathcal{M}}\mathbb{E}_s\Big[\|\widehat{s}_m-s\|^2\Big]\;, \tag{6.3}$$

where $\mathbb{E}_s$ denotes the expectation with respect to the law of $\big(Y_\varepsilon(t)\big)_{t\in\mathbb{H}}$ (which depends on the unknown vector s), and where $C\geqslant 1$ is a constant.

Example 6.1 (Gaussian regression framework with fixed design)**.** *In this setting, we observe*

$$Y_i = s_i + \sigma\xi_i\;, \quad 1\leqslant i\leqslant n\;,$$

where the ξ_i are independent standard normal random variables, where the noise level $\sigma>0$ is assumed to be known, and where $s\in\mathbb{R}^n$ is the vector to be estimated. As explained in [BM01a], this setting corresponds to $\mathbb{H}=\mathbb{R}^n$ endowed with $<u,v> = n^{-1}\sum_{i=1}^n u_iv_j$, together with $Y(t)=<Y,t>$, $W(t)=\sqrt{n}<\xi,t>$, $s=(s_1,\dots,s_n)$, and $\varepsilon=\sigma/\sqrt{n}$, where $Y\triangleq(Y_1,\dots,Y_n)$ and $\xi\triangleq(\xi_1,\dots,\xi_n)$.

[1] Equivalently, $W=\big(W(t)\big)_{t\in\mathbb{H}}$ is a family of real random variables such that, for all $p\geqslant 1$ and all $t_1,\dots,t_p\in\mathbb{H}$, the random vector $\big(W(t_1),\dots,W(t_p)\big)$ is Gaussian with zero mean and covariance matrix $\big(<t_i,t_j>\big)_{1\leqslant i,j\leqslant p}$.

[2] In this introduction we assume that a least-squares estimator $\widehat{s}_m$ exists for all $m\in\mathcal{M}$. The standard extension to approximate least-squares estimators is presented in Section 6.2.2.

Note that in this setting, the least-squares criterion can be rewritten as $\gamma_\varepsilon(t) = \|t\|^2 - 2<Y,t> = \|Y-t\|^2 - \|Y\|^2$. *Therefore, the least-squares estimators take the more standard form:*

$$\widehat{s}_m \in \operatorname*{argmin}_{t\in S_m} \|Y-t\|^2 \ .$$

Example 6.2 (White noise framework)**.** *In this setting, we observe the whole path of the stochastic process* $\big(\zeta_\varepsilon(u)\big)_{0\leqslant u\leqslant 1}$ *defined by*

$$\zeta_\varepsilon(u) = \int_0^u s(x)\mathrm{d}x + \varepsilon B(u) \ ,$$

where $\big(B(u)\big)_{u\geqslant 0}$ *is a standard Brownian motion, and where* $s \in \mathbb{L}^2([0,1],\mathrm{d}x)$ *is an unknown square-integrable function on* $[0,1]$. *As noted in [BM01a], this setting corresponds to* $\mathbb{H} = \mathbb{L}^2([0,1],\mathrm{d}x)$, $W(t) = \int_0^1 t(x)\mathrm{d}B(x)$, *and* $Y_\varepsilon(t) = \int_0^1 t(x)\mathrm{d}\zeta_\varepsilon(x)$ *provided that* $\mathbb{H}$ *is equipped with the usual inner product* $<s,t> = \int_0^1 s(x)t(x)\mathrm{d}x$.

6.1.1 Model selection

One way to obtain risk bounds of the form (6.3) is to employ the celebrated *model selection via penalization* procedure of [BM01a, Mas07]. In the setting described above, this procedure chooses the estimator $\widetilde{s} = \widehat{s}_{\widehat{m}}$, where $\widehat{m}$ minimizes some penalized least-squares criterion over $\mathcal{M}$. When the models S_m are finite-dimensional linear subspaces of $\mathbb{H}$, and for an appropriately well-chosen penalty function, the estimator $\widetilde{s} = \widehat{s}_{\widehat{m}}$ is shown to satisfy a risk bound of the form (6.3) with a constant $C \geqslant 1$ depending on the collection $(S_m)_{m\in\mathcal{M}}$ (C can be large if there are many models of the same dimension). Following the terminology of [DJ94a], the latter risk bound is called an *oracle inequality*, since it indicates that the estimator $\widetilde{s} = \widehat{s}_{\widehat{m}}$ mimics the oracle $\widehat{s}_{m^*(s)}$, where $m^*(s) \in \operatorname{argmin}_{m\in\mathcal{M}} \mathbb{E}_s\big[\|\widehat{s}_m - s\|^2\big]$ corresponds to the unknown best model in $(S_m)_{m\in\mathcal{M}}$.

First consider the case where all the models S_m are finite-dimensional linear subspaces of $\mathbb{H}$ (in the sequel, such models are refered to as *linear models*). As detailed in [BM01a, Mas07], this includes as important examples the problems of variable selection, curve estimation, and change points detection. In this setting [BM01a] associate with each model S_m a "weight" $L_m \geqslant 0$ such that $\Sigma \triangleq \sum_{m\in\mathcal{M}:D_m>0} e^{-L_m D_m} < \infty$, where D_m denotes the dimension of the linear space S_m. They then select the model $\widehat{m}$ as a minimizer of a penalized least-squares criterion:

$$\widehat{m} \in \operatorname*{argmin}_{m\in\mathcal{M}} \Big\{\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}(m)\Big\} \ , \quad \text{where} \quad \mathrm{pen}(m) \geqslant K\varepsilon^2 D_m \left(1+\sqrt{2L_m}\right)^2$$

for some constant $K > 1$ to be chosen by the statistician. Note that the lower bound on the penalty $\mathrm{pen}(m)$ is proportional to the dimension D_m. As shown in [BM01a], for some constant $C_K > 1$ depending only on K, the selected estimator $\widehat{s}_{\widehat{m}}$ satisfies the oracle-type inequality

$$\mathbb{E}_s\Big[\|\widehat{s}_{\widehat{m}} - s\|^2\Big] \leqslant C_K \left(\inf_{m\in\mathcal{M}} \Big\{\|s_m - s\|^2 + \mathrm{pen}(m)\Big\} + (\Sigma+1)\,\varepsilon^2\right) \tag{6.4}$$

$$\leqslant C'_{K,\Sigma}\left(1 + \sup_{m\in\mathcal{M}} L_m\right) \inf_{m\in\mathcal{M}} \mathbb{E}_s\Big[\|\widehat{s}_m - s\|^2\Big] \ , \tag{6.5}$$

where the last inequality holds true if $\mathrm{pen}(m) = K\varepsilon^2 D_m \left(1 + \sqrt{2L_m}\right)^2$, and where $C'_{K,\Sigma} > 1$ is a constant depending only on K and Σ. In particular (6.5) yields an oracle inequality of the form (6.3) if we can choose the L_m such that $\sup_m L_m < \infty$ and $\Sigma \triangleq \sum_{m \in \mathcal{M}: D_m > 0} e^{-L_m D_m} < \infty$. Examples include the problems of ordered variable selection (where the L_m can be chosen such that $\sup_m L_m$ is independent of the number of variables) and of complete variable selection (where an optimal choice of the L_m is such that $\sup_m L_m$ scales as the logarithm of the number of variables); see Section 6.4.1 for further details. Another important consequence of (6.4) is that, as shown in [BM01b], the model selection procedure $\widehat{s}_{\widehat{m}}$ can be used to perform adaptive estimation in an (approximately) minimax sense for various problems (e.g., variable selection, curve estimation).

The case where the models S_m are not necessarily linear was addressed by [Mas07] via a notion of generalized dimension D_m (defined through a suitable weighted empirical process – see Section 6.2.2). The family of models $(S_m)_{m \in \mathcal{M}}$ is associated with "weights" $x_m \geqslant 0$ such that $\Sigma \triangleq \sum_{m \in \mathcal{M}} e^{-x_m} < \infty$. The selected model $\widehat{m}$ is then defined by

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}(m) \right\}, \quad \text{where} \quad \mathrm{pen}(m) \geqslant K\varepsilon^2 \left(\sqrt{D_m} + \sqrt{2x_m} \right)^2$$

for some constant $K > 1$ to be chosen by the statistician. As shown in [Mas07], the selected estimator $\widehat{s}_{\widehat{m}}$ satisfies the risk bound

$$\mathbb{E}_s \left[\|\widehat{s}_{\widehat{m}} - s\|^2 \right] \leqslant C_K \left(\inf_{m \in \mathcal{M}} \left\{ \|s_m - s\|^2 + \mathrm{pen}(m) \right\} + \varepsilon^2 (\Sigma + 1) \right) .$$

The above risk bound cannot be rewritten in the form (6.3) in general. It is thus called an *oracle-type inequality*. Note however that it still leads to adaptivity properties (see, e.g., [Mas07, Section 4.4.2] about adaptation to Besov ellipsoids). In this chapter, we derive risk bounds of the same form but for a more general procedure based on exponential weighting.

Some references on model selection

The model selection via penalization approach of [BM01a] was inspired from the pioneering paper [BC91]. It was first introduced by [BM97] in the context of density estimation and later developed in [BBM99, BM01a] for the density estimation and regression problems. Several extensions and refinements were later addressed, e.g., by [Bar00] in non-Gaussian settings, by [Bar02, Bir04] for the random-design case, by [BM07a, AM09] for the study of minimal penalties and the corresponding slope's heuristics, by [BGH09] in the case of an unknown variance, and by [BGH11] for the wider problem of estimator selection. We refer the reader to [Mas07] for a comprehensive introduction to the topic. Detailed historical references can also be found in [MCL98].

6.1.2 Aggregation: main contributions and related works

In this chapter we are interested in the connections between model selection and aggregation, in the same vein as in [FG00] and, more recently, [LB06]. The key idea, already pointed out in [BM01a, Section 3.4], is to interpret the selected model $\widehat{m}$ in Theorem 6.1 as the mode of some

posterior probability distribution in a Bayesian context[3]. More precisely, putting all convergence issues aside (see Section 6.3), we have

$$\widehat{m} \in \operatorname*{argmax}_{m \in \mathcal{M}} \widehat{\rho}_m^{(\eta)} \,,$$

where, for some inverse temperature parameter $\eta > 0$, the posterior probability distribution $\widehat{\rho}^{(\eta)} = \big(\widehat{\rho}_m^{(\eta)}\big)_{m \in \mathcal{M}}$ is defined by

$$\widehat{\rho}_m^{(\eta)} = \frac{\exp\Big[- \eta\Big(\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}(m)\Big)\Big]}{\sum_{m' \in \mathcal{M}} \exp\Big[- \eta\Big(\gamma_\varepsilon(\widehat{s}_{m'}) + \mathrm{pen}(m')\Big)\Big]} \,, \quad m \in \mathcal{M} \,. \tag{6.6}$$

In this chapter, we consider the following Bayesian variant of the model selection procedure of [BM01a, Mas07]. Instead of estimating s with $\widehat{s}_{\widehat{m}}$ where $\widehat{m}$ is the mode of $\widehat{\rho}^{(\eta)}$, we estimate s with the convex combination of the estimators $\widehat{s}_m$ given by $\widehat{\rho}^{(\eta)}$, i.e., with the estimator

$$\widetilde{s}^{(\eta)} = \sum_{m \in \mathcal{M}} \widehat{\rho}_m^{(\eta)} \widehat{s}_m \,. \tag{6.7}$$

(In the sequel, we also allow pen to depend on η.)

Aggregation (or mixing) via exponential weighting has now quite a long history in both the machine learning and the statistical literatures. In machine learning, the exponentially weighted average forecaster has received a considerable attention from the seminal works [Vov90, LW94] to more recent parameter-tuning-oriented papers such as [CBMS07]; see Chapter 2 and [CBL06] for an introduction to the subject.

As for the statistical literature, *progressive mixture rules*[4] for the regression model with random design have been thoroughly studied by [Cat99, Cat04] and later, e.g., by [Yan00, Yan01, Yan03, Yan04] and [Aud07]. In this batch i.i.d. setting, aggregation via exponential weighting can also be carried out in a non-sequential way, i.e., by computing the exponential weights only once, with the whole sample — as in (6.6)–(6.7). Versions of such procedures were first analysed under the name of *Gibbs estimators* by [Cat04], where they are proved to satisfy sharp oracle-type inequalities (i.e., with a leading constant equal to 1). Subsequent contributions include, among other papers, [Aud04b] and [Alq08, AL11]. Most results of the aforementioned works are obtained for families of base estimators that are deterministic — or random, but independent of the sample used for the aggregation task (the last situation corresponds to the so-called *sample splitting* trick).

In the regression framework with fixed design (cf. Example 6.1), sharp PAC-Bayesian oracle-type inequalities for deterministic base estimators were derived by [DT08] under weak assumptions on the noise distribution. However, if the aggregated estimators are random and constructed

[3]When $\mathbb{H} = \mathbb{R}^n$, [BM01a] show in Section 3.4 that $\widehat{m}$ is the mode of the posterior probability distribution in a Bayesian framework with an improper "uniform" prior on each model S_m and a prior on the collection $\mathcal{M}$ which is proportional to $\exp\big(-\mathrm{pen}(m)/(2\varepsilon^2)\big)$, $m \in \mathcal{M}$. See also [FG00, Theorem 1], [AG10], and [AGS11, Part IV].

[4]In the terminology of online learning, a progressive mixture rule is the result of the standard online-to-batch conversion (see Section 2.5.1) when applied to an exponentially weighted average forecaster computed on the sample.

on a fraction of the whole sample, it is not possible in general to combine them via the remaining subsample so as to ensure a low expected empirical squared error on the whole sample (since the two subsamples are no longer identically distributed). In a word, the sample splitting trick is not appropriate in this setting.

The model selection via penalization procedure of [BM01a] is a way to overcome this limitation: the selected model $\widehat{m}$ is chosen as a function of the same sample on which the least-squares estimators $\widehat{s}_m$ are constructed. Another way is to use the Bayesian variant introduced above. A key contribution in this respect was carried out by [LB06] when the (finite) family of estimators consists of least-squares estimators on linear models S_m (still in the framework of Example 6.1). Using our notations, and denoting by D_m the dimension of S_m, their procedure is of the form of (6.6)–(6.7) with $\mathrm{pen}(m) = 2D_m\sigma^2/n + x_m/\eta$, where the x_m are such that $\Sigma \triangleq \sum_{m\in\mathcal{M}} e^{-x_m} < +\infty$. Then, [LB06] show that for all $\eta \leqslant n/(4\sigma^2)$, the estimator $\widetilde{s}^{(\eta)}$ defined in (6.6)–(6.7) satisfies the sharp oracle-type inequality

$$\mathbb{E}_s\left[\left\| \sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)}\widehat{s}_m - s \right\|^2\right] \leqslant \inf_{m\in\mathcal{M}} \left\{ \mathbb{E}_s\left[\|\widehat{s}_m - s\|^2\right] + \frac{x_m}{\eta} \right\} + \frac{\ln\Sigma}{\eta}\,. \tag{6.8}$$

The above risk bound has proved useful, e.g., in high-dimensional linear regression. As shown by [LB06, RT11, AL11], for a proper choice of the prior weights e^{-x_m}/Σ, this risk bound leads to sharp sparsity oracle inequalities without any assumption on the dictionary at hand (cf. Chapter 2, Section 2.6).

First contribution: aggregation of nonlinear models

The risk bound (6.8) of [LB06] was derived under the assumption that the models $S_m \subset \mathbb{R}^n$ are linear and that the $\widehat{s}_m$ are the associated least-squares estimators (i.e., the orthogonal projections of $Y \in \mathbb{R}^n$ on the S_m). This work was further extended in two directions. On the one hand, the case of an unknown variance was addressed by [Gir08]. On the other hand, [DS11] replaced the family of projection estimators $\left(\widehat{s}_m\right)_{m\in\mathcal{M}}$ with an arbitrary family of affine estimators; this wider class of estimators includes, e.g., diagonal filters, kernel ridge regression, and multiple kernel learning.

In this chapter, we extend the work of [LB06] in a third direction: we still consider projection estimators, but the models $S_m \subset \mathbb{R}^n$ can be almost arbitrary (or *nonlinear*). (Actually, we consider the more general case $S_m \subset \mathbb{H}$, where $\mathbb{H}$ is possibly infinite-dimensional). In such generality, the use of the key Stein's unbiased risk formula of [Ste81] as carried out in [LB06, DT08, DS11] seems difficult. Instead we follow the concentration approach of [Mas07] to derive oracle-type inequalities with high probability (with, however, leading constants larger than 1).

Second contribution: continuum of estimators from model aggregation to model selection

This work exhibits a natural connection between model aggregation and model selection: our oracle-type inequalities hold for a continuum of estimators $\left\{\widetilde{s}^{(\eta)} : \eta > 0\right\}$ ranging from classical model aggregation (where η is at most of the order[5] of $1/\varepsilon^2$) to model selection (where $\eta = +\infty$).

[5] In the Gaussian regression framework with fixed design, [LB06, DS11] assume that $\eta \leqslant n/(4\sigma^2)$, where n and σ^2 denote the sample size and the variance of the noise respectively. This condition can be rewritten in our setting (6.1) as $\eta \leqslant 1/(4\varepsilon^2)$ (cf. the correspondence $\varepsilon = \sigma/\sqrt{n}$ in Example 6.1).

In particular, for an appropriate choice of the penalty, our aggregating estimator $\widetilde{s}^{(\eta)}$ converges almost surely to the selected estimator $\widehat{s}_{\widehat{m}}$ as $\eta \to \infty$ (if $\widehat{m}$ is unique); see Corollary 6.1. This fact is not surprising since our analysis is mostly based on the arguments of [Mas07, Theorem 4.18] — the main change, however, consists in using a key duality formula for the Kullback-Leibler divergence instead of the definition of $\widehat{m}$. The present chapter thus shows that the analysis of [Mas07] for nonlinear models can be extended to arbitrary values of $\eta > 0$, instead of just taking $\eta = +\infty$.

As of now, we do not know whether aggregation outperforms model selection for classical nonlinear models (e.g., Besov ellipsoids, ℓ^1-balls, neural networks). However, our risk bounds suggest that it might be the case because of the presence of a Jensen-type nonnegative term — see Corollary 6.2. Another reason is that, even in the case of linear models, there are situations where aggregation is more robust than model selection in terms of excess risks — see Section 6.4.2 for a lower bound on model selection procedures. This lower bound suggests that aggregation might benefit from a similar advantage with nonlinear models.

Outline of the chapter

This chapter is organized as follows. In Section 6.2 we formally describe our statistical framework together with the model-selection and aggregation procedures at hand. In Section 6.3 we prove our main oracle-type inequality for aggregation of nonlinear models. Then, in Section 6.4, we derive several corollaries in classical examples and explain in which situations convexification may be useful (as compared to model selection). Finally some important questions raised by this work in progress are stated in Section 6.5.

6.2 Framework and statistical procedures at hand

In this section we recall the framework mentioned in the introduction and give a precise description of the collection of models at hand. Then we recall with our notations the main theorem of [Mas07] for model selection with nonlinear models. Finally we define our aggregation procedure.

6.2.1 The generalized linear Gaussian framework

We consider the generalized linear Gaussian framework introduced in [BM01a], i.e., one observes the whole stochastic process $\big(Y_\varepsilon(t)\big)_{t\in\mathbb{H}}$ given by

$$Y_\varepsilon(t) = <s,t> + \varepsilon W(t)\,, \quad t \in \mathbb{H}\,, \tag{6.9}$$

where $\big(\mathbb{H}, <\cdot,\cdot>\big)$ is some separable Hilbert space, where W is an isonormal process on $\mathbb{H}$ (i.e., an isometry from $\mathbb{H}$ onto a centered Gaussian space[6]), where the noise level $\varepsilon > 0$ is assumed to be known, and where $s \in \mathbb{H}$ is the unknown vector to be estimated.

In the sequel, $\mathbb{P}_s$ denotes the law of $\big(Y_\varepsilon(t)\big)_{t\in\mathbb{H}}$ (which depends on the unknown vector s), and $\mathbb{E}_s$ denotes the corresponding expectation.

[6]Equivalently, $W = \big(W(t)\big)_{t\in\mathbb{H}}$ is a family of real random variables such that, for all $p \geqslant 1$ and all $t_1, \ldots, t_p \in \mathbb{H}$, the random vector $\big(W(t_1), \ldots, W(t_p)\big)$ is Gaussian with zero mean and covariance matrix $\big(<t_i, t_j>\big)_{1\leqslant i,j\leqslant p}$.

6.2.2 Collection of models

We follow below the same lines as in [Mas07]. Fix some at most countable collection $(S_m)_{m\in\mathcal{M}}$ of subsets of $\mathbb{H}$, which will be refered to as the *models* thereafter. We assume that for every $m \in \mathcal{M}$, there exists some almost-surely continuous version of the isonormal process W on the closure $\overline{S_m}$ of S_m, still denoted by W.

[Mas07] associates with each model S_m a generalized dimension D_m defined as follows (see Section 6.4.1 for some examples). We assume that for all $m \in \mathcal{M}$, their exists a nondecreasing continuous function $\varphi_m : [0, +\infty) \to \mathbb{R}_+$ such that $x \mapsto x^{-1}\varphi_m(x)$ is nonincreasing on $\mathbb{R}_+^*$ and[7]

$$2\,\mathbb{E}\left[\sup_{t\in S_m}\left(\frac{W(t)-W(u)}{\|t-u\|^2+x^2}\right)\right] \leqslant x^{-2}\varphi_m(x)\;, \tag{6.10}$$

for all $x > 0$ and all $u \in S_m$. We let $\tau_m = 1$ if S_m is closed and convex and $\tau_m = 2$ otherwise. Under the assumptions above, we associate with each model S_m a generalized dimension $D_m \geqslant 0$ defined by:

- if $\varphi_m \equiv 0$, then $D_m \triangleq 0$;
- if $\varphi_m(x_0) > 0$ for some $x_0 \geqslant 0$, then D_m is the unique positive solution of the equation

$$\varphi_m\big(\tau_m\varepsilon\sqrt{D_m}\big) = \varepsilon D_m\;. \tag{6.11}$$

Legitimate definition: Next we explain why D_m is well-defined when there is $x_0 \geqslant 0$ such that $\varphi_m(x_0) > 0$. Note that (6.11) has a unique solution $D_m > 0$ if and only if the equation $x^{-2}\varphi_m(x) = 1/(\tau_m^2\varepsilon)$ has a unique solution $x > 0$ (by the change of variables $x = \tau_m\varepsilon\sqrt{D_m}$).

But, since $x \mapsto x^{-1}\varphi_m(x)$ is nonincreasing on $(0, x_0]$ and since φ_m is nondecreasing on $[x_0, +\infty)$, we can see that $\varphi_m(x) > 0$ for all $x > 0$.

Therefore, $x \mapsto x^{-2}\varphi_m(x)$ is a product of two continuous, positive, and nonincreasing functions $x \mapsto x^{-1}\varphi_m(x)$ and $x \mapsto x^{-1}$, one of them being decreasing. The function $x \mapsto x^{-2}\varphi_m(x)$ is thus continuous and decreasing on $(0, +\infty)$. Since in addition $\lim_{x\to 0} x^{-2}\varphi_m(x) = +\infty$ and $\lim_{x\to+\infty} x^{-2}\varphi_m(x) = 0$ (because $x \mapsto x^{-1}\varphi_m(x)$ is positive and nonincreasing), we get from the intermediate value theorem that the equation $x^{-2}\varphi_m(x) = 1/(\tau_m^2\varepsilon)$ has a unique solution $x > 0$. This implies that (6.11) has a unique solution $D_m > 0$.

We associate with each $m \in \mathcal{M}$ a real number $x_m \geqslant 0$ such that $\Sigma \triangleq \sum_{m\in\mathcal{M}} e^{-x_m} < \infty$. The sequence $(e^{-x_m}/\Sigma)_{m\in\mathcal{M}}$ can be seen as a prior probability distribution on the sequence of models $(S_m)_{m\in\mathcal{M}}$. We will denote thereafter by $\Delta(\mathcal{M})$ the set of all probability distributions on

[7] To avoid any measurability issues, the supremum $\sup_{t\in S_m}$ in (6.10) should be understood as a supremum $\sup_{t\in A_m}$ over any at most countable dense subset $A_m \subset S_m$. In the same way, the infimum $\inf_{t\in S_m}$ in (6.13) can be replaced with an infimum $\inf_{t\in A_m}$; the resulting weaker assumption ensures that their always exists $(\widehat{s}_m)_{m\in\mathcal{M}}$ such that $\widehat{s}_m$ and $W(\widehat{s}_m)$ are measurable. Though we do not focus on measurablity issues in this chapter, all stated results remain true under the aforementioned slight modification (by density of A_m and since W admits an almost-sure continuous version on $\overline{S_m}$).

$\mathcal{M}$ (endowed with its discrete σ-algebra).

We also set, for all $t \in \mathbb{H}$,

$$\gamma_\varepsilon(t) = \|t\|^2 - 2Y_\varepsilon(t) \; . \tag{6.12}$$

Given $\mu > 0$, we associate with this empirical contrast some collection of μ-least-squares estimators $(\widehat{s}_m)_{m\in\mathcal{M}}$ (or *μ-LSEs* for short). This means that for all $m \in \mathcal{M}$,

$$\widehat{s}_m \in S_m \qquad \text{and} \qquad \gamma_\varepsilon(\widehat{s}_m) \leqslant \inf_{t\in S_m} \gamma_\varepsilon(t) + \mu \quad \text{almost surely.} \tag{6.13}$$

Finally, for all $m \in \mathcal{M}$, we denote by $d(s, S_m) \triangleq \inf_{t\in S_m} \|s - t\|$ the distance between s and the model S_m.

6.2.3 Model selection via penalization

In the framework described above, the general model selection theorem by [Mas07, Theorem 4.18] reads as follows. Note that it holds for possibly nonlinear models S_m (i.e., the $S_m \subset \mathbb{H}$ are not necessarily linear subspaces of $\mathbb{H}$).

Theorem 6.1 (A general model selection theorem by [Mas07] for nonlinear models)**.**
Consider the framework given in (6.9). Let $K > 1$ be some constant and $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ be such that, for all $m \in \mathcal{M}$,

$$\mathrm{pen}(m) \geqslant K\varepsilon^2 \left(\sqrt{D_m} + \sqrt{2x_m}\right)^2 \; , \tag{6.14}$$

where D_m is defined in (6.10) – (6.11) and where $(x_m) \in \mathbb{R}_+^{\mathcal{M}}$ is such that $\Sigma \triangleq \sum_{m\in\mathcal{M}} e^{-x_m} < \infty$. Then, almost surely, there is a minimizer

$$\widehat{m} \in \operatorname*{argmin}_{m\in\mathcal{M}} \big\{\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}(m)\big\} \; . \tag{6.15}$$

Defining a penalized μ-LSE as $\widetilde{s} = \widehat{s}_{\widehat{m}}$, we have, for some constant $C_K > 1$ depending only on K, for all $s \in \mathbb{H}$,

$$\mathbb{E}_s\Big[\|\widetilde{s} - s\|^2\Big] \leqslant C_K \left(\inf_{m\in\mathcal{M}} \Big\{d^2(s, S_m) + \mathrm{pen}(m)\Big\} + \varepsilon^2(\Sigma + 1) + \mu\right) \; . \tag{6.16}$$

6.2.4 Our aggregation procedure

Our aggregation procedure is defined as follows. We refer the reader to the introduction for related references.

Given an inverse temperature parameter $\eta > 0$ and a penalty function $\mathrm{pen}^{(\eta)} : \mathcal{M} \to \mathbb{R}_+$, we define the associated Gibbs posterior $\widehat{\rho}^{(\eta)} = \big(\widehat{\rho}_m^{(\eta)}\big)_{m\in\mathcal{M}} \in \Delta(\mathcal{M})$ by

$$\widehat{\rho}_m^{(\eta)} = \frac{\exp\Big[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}^{(\eta)}(m)\Big)\Big]}{\sum_{m'\in\mathcal{M}} \exp\Big[-\eta\Big(\gamma_\varepsilon(\widehat{s}_{m'}) + \mathrm{pen}^{(\eta)}(m')\Big)\Big]} \; , \quad m \in \mathcal{M} \; . \tag{6.17}$$

We will later see sufficient conditions for the almost sure convergence of the series in the denominator above, in which case $\widehat{\rho}^{(\eta)}$ is well-defined. If this is the case, and if almost surely the next series converges absolutely in $\mathbb{H}$, then we define our estimator $\widetilde{s}^{(\eta)}$ by

$$\widetilde{s}^{(\eta)} = \sum_{m \in \mathcal{M}} \widehat{\rho}_m^{(\eta)} \widehat{s}_m \,. \tag{6.18}$$

Note that $\widehat{\rho}^{(\eta)}$ and $\widetilde{s}^{(\eta)}$ are always well-defined when $\mathcal{M}$ is finite.

6.3 Model aggregation with nonlinear models

In this section we prove an oracle-type inequality for the aggregation procedure $\widetilde{s}^{(\eta)}$. To avoid convergence technicalities, all the proofs in this section are given for the case of a finite collection $\mathcal{M}$; the extensions to any at most countable collection are postponed to Section 6.A.

We first fix some notations. We set $\ln_+(x) \triangleq \max\bigl\{\ln(x), 0\bigr\}$ for all $x \in \mathbb{R}$. Moreover, for all $\rho, \pi \in \Delta(\mathcal{M})$, the Kullback-Leibler divergence $\mathcal{K}(\rho, \pi)$ between ρ and π is defined by

$$\mathcal{K}(\rho, \pi) \triangleq \begin{cases} \displaystyle\sum_{m \in \mathcal{M}} \rho_m \ln\left(\frac{\rho_m}{\pi_m}\right) & \text{if } \rho \text{ is absolutely continuous with respect to } \pi; \\ +\infty & \text{otherwise.} \end{cases}$$

Finally, for all $\rho \in \Delta(\mathcal{M})$ such that $\sum_{m \in \mathcal{M}} \rho_m \|\widehat{s}_m\| < \infty$ almost surely (which is the case when, e.g., $\mathcal{M}$ is finite), we set

$$\mathcal{J}(\rho) \triangleq \sum_{m \in \mathcal{M}} \rho_m \|\widehat{s}_m - s\|^2 - \left\| \sum_{m \in \mathcal{M}} \rho_m \widehat{s}_m - s \right\|^2 . \tag{6.19}$$

Note that $\mathcal{J}(\rho) \geqslant 0$ by Jensen's inequality.

The main result of this section is the following theorem. See also Remark 6.1 below for a risk bound with high probability.

Theorem 6.2. *Consider the framework given in (6.9). Assume that $\mathcal{M}$ is at most countable. Let $\eta > 0$ and $K > 1$ be some constants and take $\mathrm{pen}^{(\eta)} : \mathcal{M} \to \mathbb{R}_+$ such that, for all $m \in \mathcal{M}$,*

$$\mathrm{pen}^{(\eta)}(m) \geqslant K\varepsilon^2 \left(\sqrt{D_m} + \sqrt{2x_m}\right)^2 + \frac{x_m}{\eta} \,, \tag{6.20}$$

where D_m is defined in (6.10)–(6.11) and where $(x_m) \in \mathbb{R}_+^{\mathcal{M}}$ is such that $\Sigma \triangleq \sum_{m \in \mathcal{M}} e^{-x_m} < \infty$. Then, almost surely,

- *$\sum_{m \in \mathcal{M}} \exp\Bigl[-\eta\Bigl(\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}^{(\eta)}(m)\Bigr)\Bigr] < \infty$, so that $\widehat{\rho}^{(\eta)}$ is well defined in (6.17);*
- *$\sum_{m \in \mathcal{M}} \widehat{\rho}_m^{(\eta)} \|\widehat{s}_m\| < \infty$, so that $\widetilde{s}^{(\eta)} = \sum_{m \in \mathcal{M}} \widehat{\rho}_m^{(\eta)} \widehat{s}_m$ is well defined in (6.18).*

Moreover, defining the prior $\pi \in \Delta(\mathcal{M})$ by $\pi_m \triangleq e^{-x_m}/\Sigma$ for all $m \in \mathcal{M}$, we have, for some

constant $C_K > 1$ depending only on K (cf. (6.34)), for all $s \in \mathbb{H}$,

$$\begin{aligned}
&\mathbb{E}_s\left[\left\|\widetilde{s}^{(\eta)} - s\right\|^2\right] \\
&\leqslant C_K\left(\inf_{\rho\in\Delta(\mathcal{M})}\left\{\sum_{m\in\mathcal{M}}\rho_m\left[d^2(s,S_m) + \mathrm{pen}^{(\eta)}(m) - \frac{x_m}{\eta}\right] + \frac{\mathcal{K}(\rho,\pi)}{\eta}\right\}\right. \\
&\qquad\left. + \varepsilon^2(\ln_+(\Sigma)+1) + \mu\right) - \mathbb{E}_s\left[\mathcal{J}\left(\widehat{\rho}^{(\eta)}\right)\right] \qquad (6.21)\\
&\leqslant C_K\left(\inf_{m\in\mathcal{M}}\left\{d^2(s,S_m) + \mathrm{pen}^{(\eta)}(m)\right\} + \frac{\ln\Sigma}{\eta} + \varepsilon^2(\ln_+(\Sigma)+1) + \mu\right) - \mathbb{E}_s\left[\mathcal{J}\left(\widehat{\rho}^{(\eta)}\right)\right]. \qquad (6.22)
\end{aligned}$$

Remark 6.1 (High-probability bound)**.**
The above oracle-type inequalities are only stated in expectation. The proof of Theorem 6.2 also leads to the following high-probability bound: for all $z > 0$, with probability at least $1 - \Sigma^2 e^{-z}$,

$$\left\|\widetilde{s}^{(\eta)} - s\right\|^2 \leqslant C'_K\left(\inf_{m\in\mathcal{M}}\left\{d^2(s,S_m) + \mathrm{pen}^{(\eta)}(m)\right\} + \frac{\ln\Sigma}{\eta} + \varepsilon^2(z+1) + \mu\right) - \mathcal{J}\left(\widehat{\rho}^{(\eta)}\right)$$

for some absolute constant $C'_K > 0$ depending only on K (cf. (6.35)). See Remark 6.2 later for a proof of this bound. Note that it implies a bound in expectation similar to (6.22) — via Lemma A.7 in Appendix A.6 — but that is in general not comparable to (6.21).

Before proving Theorem 6.2, we first make some comments on the procedure $\widetilde{s}^{(\eta)}$ and state two corollaries of Theorem 6.2.

Next we compare $\widetilde{s}^{(\eta)}$ with the standard model selection procedure $\widehat{s}_{\widehat{m}}$ recalled in Theorem 6.1; we assume for simplicity that $\mathcal{M}$ is finite. For this purpose, let $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ be a penalty function, and consider the case when $\mathrm{pen}^{(\eta)}(m) = \mathrm{pen}(m) + x_m/\eta$ for all $m \in \mathcal{M}$. Assume that, for all $m \in \mathcal{M}$, we have $\mathrm{pen}(m) \geqslant K\varepsilon^2\left(\sqrt{D_m} + \sqrt{2x_m}\right)^2$, or, equivalently, $\mathrm{pen}^{(\eta)}(m) \geqslant K\varepsilon^2\left(\sqrt{D_m} + \sqrt{2x_m}\right)^2 + x_m/\eta$. Then, by the fact that $\widehat{\rho}_m^{(\eta)}$ is proportional to

$$\exp\left[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}^{(\eta)}(m)\Big)\right] = \exp\left[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}(m)\Big) - x_m\right], \qquad (6.23)$$

it is easy to see that, as $\eta \to +\infty$, the probability distribution $\widehat{\rho}^{(\eta)}$ converges almost surely to $\widehat{\rho}^{(\infty)} \in \Delta(\mathcal{M}) \subset \mathbb{R}^{|\mathcal{M}|}$ defined by

$$\forall m \in \mathcal{M}\,, \qquad \widehat{\rho}_m^{(\infty)} \triangleq \begin{cases} e^{-x_m}/\widehat{Z} & \text{if } m \in \widehat{\mathcal{M}}\,, \\ 0 & \text{if } m \notin \widehat{\mathcal{M}}\,, \end{cases} \qquad (6.24)$$

where $\widehat{\mathcal{M}} \triangleq \mathrm{argmin}_{m\in\mathcal{M}}\left\{\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}(m)\right\} \subset \mathcal{M}$ and where $\widehat{Z} \triangleq \sum_{m\in\widehat{\mathcal{M}}} e^{-x_m}$. In particular, if there is a unique minimizer $\widehat{m} \in \mathrm{argmin}_{m\in\mathcal{M}}\left\{\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}(m)\right\}$, then $\widehat{\rho}^{(\eta)}$ tends to the Dirac distribution at $\widehat{m}$, so that $\widetilde{s}^{(\eta)} \to \widehat{s}_{\widehat{m}}$ almost surely. This is stated formally in Corollary 6.1 below.

On the contrary, when $\eta \to 0$, we can see from (6.23) that, almost surely, the probability distribution $\widehat{\rho}^{(\eta)}$ tends to the prior $\pi \triangleq (e^{-x_m}/\Sigma)_{m\in\mathcal{M}}$, so that $\widetilde{s}^{(\eta)} \to \sum_{m\in\mathcal{M}} \pi_m \widehat{s}_m$. Therefore,

the continuous family $\{\widetilde{s}^{(\eta)} : \eta > 0\}$ contains Bayesian variants of the estimator $\widehat{s}_{\widehat{m}}$ ranging from pure model aggregation (when $\eta \to 0$) to pure model selection (when $\eta \to \infty$).

How to choose the value of the tuning parameter η? Though this important question is beyond the scope of the present chapter, two values of η seem reasonable. As indicated above (and formally stated in Corollary 6.1 below), letting $\eta \to +\infty$ is never a bad choice, since we recover the theoretical guarantees of the standard model selection procedure $\widehat{s}_{\widehat{m}}$ stated in Theorem 6.1. Another interesting choice is suggested in Corollary 6.2. For smaller values of $\eta \approx 1/\varepsilon^2$, the estimator $\widetilde{s}^{(\eta)}$ still satisfies a risk bound comparable to that of $\widehat{s}_{\widehat{m}}$, but it can also benefit from the convexification phenomenon due to aggregation — see the brief comment after Corollary 6.2 and the discussion in Section 6.4.2.

Note that for the two choices of η mentioned above, the estimator $\widetilde{s}^{(\eta)}$ is built as a function of the noise level ε, which may be unknown in practice. In the case $\eta = +\infty$, adaptation to ε was first tackled by [BM07a, AM09] for histogram-based regression via the so-called slope's heuristics. As for smaller values of $\eta \approx 1/\varepsilon^2$, adaptation to ε was addressed by [Gir08] for linear models with a Mallows' C_p-type penalty. However, it is not clear whether the choices $\eta = +\infty$ or $\eta \approx 1/\varepsilon^2$ mentioned above are the best ones, so that two important questions remain open. First, can we identify an optimal choice of η (in a reasonable sense) at least for classical prediction problems? Second, if such an optimal (and theoretical) choice is identified, is it possible to tune our aggregating procedure in an automatic and near-optimal way?

Corollary 6.1 (Choice of $\eta \to +\infty$)**.** *Consider the framework given in (6.9). Assume that $\mathcal{M}$ is at most countable. Let $K > 1$ be a constant and take $\mathrm{pen} : \mathcal{M} \to \mathbb{R}_+$ such that $\mathrm{pen}(m) \geqslant K\varepsilon^2 \left(\sqrt{D_m} + \sqrt{2x_m}\right)^2$ for all $m \in \mathcal{M}$, where D_m is defined in (6.10) – (6.11) and where $(x_m) \in \mathbb{R}_+^{\mathcal{M}}$ is such that $\Sigma \triangleq \sum_{m\in\mathcal{M}} e^{-x_m} < \infty$.*

Set $\mathrm{pen}^{(\eta)}(m) \triangleq \mathrm{pen}(m) + x_m/\eta$. Then, as $\eta \to +\infty$, the estimator $\widetilde{s}^{(\eta)}$ defined in (6.17) – (6.18) converges almost surely to the estimator

$$\widetilde{s}^{(\infty)} \triangleq \sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\infty)} \widehat{s}_m \,,$$

where $\widehat{\rho}^{(\infty)}$ is defined in (6.24) and has almost surely a finite support $\widehat{\mathcal{M}}$. Moreover, for some constant $C_K > 1$ depending only on K, we have, for all $s \in \mathbb{H}$,

$$\mathbb{E}_s\left[\left\|\widetilde{s}^{(\infty)} - s\right\|^2\right] \leqslant C_K\left(\inf_{m\in\mathcal{M}}\left\{d^2(s,S_m) + \mathrm{pen}(m)\right\} + \varepsilon^2(\ln_+(\Sigma)+1) + \mu\right) . \quad (6.25)$$

The proof of the last corollary is immediate when $\mathcal{M}$ is finite: the almost sure convergence $\|\widetilde{s}^{(\eta)} - \widetilde{s}^{(\infty)}\| \to 0$ was already explained in the previous paragraphs, and the risk bound above follows from Fatou's lemma and from (6.22) in Theorem 6.2. The proof in the general case of an at most countable collection $\mathcal{M}$ is postponed to Appendix 6.A.2.

Note that the oracle-type inequality above is identical to that of [Mas07, Theorem 4.18] recalled in Theorem 6.1 (except for the term $\ln_+(\Sigma)$ that is smaller than Σ, but this improvement can also be made for Theorem 6.1). This is not surprising since our proof of Theorem 6.2 follows

the same lines as that of Theorem 6.1. The present chapter thus shows that the analysis of [Mas07] for nonlinear models can be extended to arbitrary values of $\eta > 0$, instead of just taking $\eta = +\infty$.

Corollary 6.2 (Choice of $\eta = c/\varepsilon^2$)**.** *Consider the framework given in (6.9). Assume that $\mathcal{M}$ is at most countable. Let $c > 0$ and $K > 1$ be some constants and take $\mathrm{pen}_c : \mathcal{M} \to \mathbb{R}_+$ such that, for all $m \in \mathcal{M}$,*

$$\mathrm{pen}_c(m) \geqslant K\varepsilon^2 \left(\sqrt{D_m} + \sqrt{2x_m}\right)^2 + c^{-1}\varepsilon^2 x_m \;, \tag{6.26}$$

where D_m is defined in (6.10) – (6.11) and where $(x_m) \in \mathbb{R}_+^{\mathcal{M}}$ is such that $\Sigma \triangleq \sum_{m\in\mathcal{M}} e^{-x_m} < \infty$.

Then, for some constant $C_K > 1$ depending only on K, the estimator $\widetilde{s}^{(c/\varepsilon^2)}$ defined in (6.17) – (6.18) satisfies, for all $s \in \mathbb{H}$,

$$\begin{aligned}&\mathbb{E}_s\left[\left\|\widetilde{s}^{(c/\varepsilon^2)} - s\right\|^2\right]\\ &\leqslant C_K\left(\inf_{m\in\mathcal{M}}\left\{d^2(s,S_m) + \mathrm{pen}_c(m)\right\} + \frac{\varepsilon^2\ln\Sigma}{c} + \varepsilon^2(\ln_+(\Sigma)+1) + \mu\right) - \mathbb{E}_s\left[\mathcal{J}\left(\widehat{\rho}^{(c/\varepsilon^2)}\right)\right].\end{aligned}$$

The last corollary is an immediate consequence of Theorem 6.2, so the proof is omitted. Note that the risk bound above is at most of the same order[8] as the bound of Corollary 6.1 if in each corollary the penalty functions are chosen as the smallest penalties allowed by the assumptions. However, choosing $\eta = c/\varepsilon^2$ instead of $\eta \to \infty$ enables to reduce the risk bound by the additive term $\mathbb{E}_s\left[\mathcal{J}\left(\widehat{\rho}^{(c/\varepsilon^2)}\right)\right]$. We have not investigated yet the extent to which the above risk bound improves — via the term $\mathbb{E}_s\left[\mathcal{J}\left(\widehat{\rho}^{(c/\varepsilon^2)}\right)\right]$ — on the bound of Theorem 6.1 for model selection. We however briefly explain in Section 6.4.2 in which situations the convexification phenomenon due to aggregation may be useful.

Proof (of Theorem 6.2): To avoid any convergence technicalities, we assume in the sequel that $\mathcal{M}$ is finite (in particular, the first two claims of the theorem are straightforward). The proof in the countably infinite case is postponed to Appendix 6.A.1.

The proof follows the same lines as the general model selection theorem of [Mas07, Theorem 4.18] recalled in Theorem 6.1. The key point is to replace the line stating that $\widehat{m}$ minimizes some penalized empirical risk over $\mathcal{M}$ by the fact that $\widehat{\rho}^{(\eta)}$ minimizes some penalized average empirical risk over $\Delta(\mathcal{M})$. Indeed, first note that $\widehat{\rho}^{(\eta)}$ defined in (6.17) can be rewritten as

$$\widehat{\rho}_m^{(\eta)} = \frac{\exp\left[-\eta\left(\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}_2^{(\eta)}(m)\right)\right]\pi_m}{\sum_{m'\in\mathcal{M}}\exp\left[-\eta\left(\gamma_\varepsilon(\widehat{s}_{m'}) + \mathrm{pen}_2^{(\eta)}(m')\right)\right]\pi_{m'}}\;, \quad m \in \mathcal{M}\;,$$

where $\pi_m \triangleq e^{-x_m}/\Sigma$ and $\mathrm{pen}_2^{(\eta)}(m) \triangleq \mathrm{pen}^{(\eta)}(m) - x_m/\eta$ for all $m \in \mathcal{M}$. We can thus use a key duality formula on the Kullback-Leibler divergence proved, e.g., in [Cat04, pp. 159-160] and that

[8] Indeed, for all $m \in \mathcal{M}$, the bound of Corollary 6.2 is larger than the bound of Corollary 6.1 by at most the additive term $c^{-1}\varepsilon^2 x_m + c^{-1}\varepsilon^2\ln\Sigma$. Therefore, the overall bound of Corollary 6.2 is smaller than $(1 + c^{-1})$ times that of Corollary 6.1 (note that $c^{-1}\varepsilon^2 x_m \leqslant c^{-1}K\varepsilon^2\left(\sqrt{D_m} + \sqrt{2x_m}\right)^2$ since $K > 1 \geqslant 1/2$).

we recall in Proposition A.1 of Appendix A.1. Applying Proposition A.1 with $E = \mathcal{M}$ and the (random) function $h : \mathcal{M} \to \mathbb{R}$ defined by $h(m) = \eta\big(\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}_2^{(\eta)}(m)\big)$ for all $m \in \mathcal{M}$, we can see that h is almost surely bounded on $\mathcal{M}$ (since $\mathcal{M}$ is finite) and that $\widehat{\rho}^{(\eta)} = \pi_{-h}^{\exp}$; therefore, for all $\rho \in \Delta(\mathcal{M})$ and all families $(s_m)_{m\in\mathcal{M}}$ of elements in $\mathbb{H}$, almost surely,

$$\begin{aligned}\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)}\big(\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}_2^{(\eta)}(m)\big) + \frac{\mathcal{K}\left(\widehat{\rho}^{(\eta)}, \pi\right)}{\eta} &\leqslant \sum_{m\in\mathcal{M}} \rho_m\big(\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}_2^{(\eta)}(m)\big) + \frac{\mathcal{K}(\rho,\pi)}{\eta} \\ &\leqslant \sum_{m\in\mathcal{M}} \rho_m\big(\gamma_\varepsilon(s_m) + \mathrm{pen}_2^{(\eta)}(m)\big) + \frac{\mathcal{K}(\rho,\pi)}{\eta} \\ &\quad + \mu\,, \end{aligned} \tag{6.27}$$

where the last inequality follows by definition of $\widehat{s}_m$ in (6.13) and by the fact that $\sum_{m\in\mathcal{M}} \rho_m = 1$. In the sequel we fix $\delta > 0$ and choose $s_m \in S_m$ such that $\|s - s_m\|^2 \leqslant d^2(s, S_m) + \delta^2$ for all $m \in \mathcal{M}$. At the end of the proof, we will let $\delta \to 0$.

We also fix $\rho \in \Delta(\mathcal{M})$. We can assume that $\mathcal{K}(\rho,\pi) < \infty$ (since otherwise, ρ does not participate to the infimum in (6.21)). Therefore, by (6.27), we also have $\mathcal{K}\left(\widehat{\rho}^{(\eta)}, \pi\right) < \infty$, so that all terms in (6.27) are finite. Moreover, note that for all $t \in \mathbb{H}$, by definition of $\gamma_\varepsilon(t)$ and $Y_\varepsilon(t)$,

$$\begin{aligned}\gamma_\varepsilon(t) &\triangleq \|t\|^2 - 2Y_\varepsilon(t) = \|t\|^2 - 2\big(<s,t> + \varepsilon W(t)\big) \\ &= \|t - s\|^2 - \|s\|^2 - 2\varepsilon W(t)\,. \end{aligned} \tag{6.28}$$

Substituting the last equality in (6.27), and noting that two terms $-\|s\|^2$ cancel out, we get

$$\begin{aligned}\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} \|\widehat{s}_m - s\|^2 \leqslant \sum_{m\in\mathcal{M}} \rho_m\big(\|s_m - s\|^2 + \mathrm{pen}_2^{(\eta)}(m)\big) + \frac{\mathcal{K}(\rho,\pi)}{\eta} - \frac{\mathcal{K}\left(\widehat{\rho}^{(\eta)},\pi\right)}{\eta} + \mu \\ + 2\varepsilon\left(\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} W(\widehat{s}_m) - \sum_{m\in\mathcal{M}} \rho_m W(s_m)\right) - \sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} \mathrm{pen}_2^{(\eta)}(m)\,. \end{aligned} \tag{6.29}$$

But since $W : \mathbb{H} \to \mathbb{L}^2\big(\mathbb{P}_s\big)$ is linear (by definition of an isonormal process), we have, almost surely, $\sum_{m\in\mathcal{M}} \rho_m W(s_m) = W\big(\sum_{m\in\mathcal{M}} \rho_m s_m\big) = W\big(s_\rho\big)$, where we set $s_\rho \triangleq \sum_{m\in\mathcal{M}} \rho_m s_m$. Therefore, using the fact that $\sum_m \widehat{\rho}_m^{(\eta)} = 1$, we get, almost surely,

$$\begin{aligned}\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} W(\widehat{s}_m) - \sum_{m\in\mathcal{M}} \rho_m W(s_m) &= \sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)}\big(W(\widehat{s}_m) - W(s_\rho)\big) \\ &\leqslant \sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} w_m(\widehat{s}_m) \sup_{t\in S_m}\left(\frac{W(t) - W(s_\rho)}{w_m(t)}\right)\,, \end{aligned}$$

where, for all $m \in \mathcal{M}$ and $t \in \mathbb{H}$, we set $w_m(t) \triangleq (1/2)\left(\big[\|s - s_\rho\| + \|s - t\|\big]^2 + y_m^2\right)$ for some real numbers y_m to be chosen later. Substituting the above inequality in (6.29), and neglecting the term $\mathcal{K}\left(\widehat{\rho}^{(\eta)},\pi\right)/\eta \geqslant 0$, we get, almost surely,

$$\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} \|\widehat{s}_m - s\|^2 \leqslant \sum_{m\in\mathcal{M}} \rho_m\big(\|s_m - s\|^2 + \mathrm{pen}_2^{(\eta)}(m)\big) + \frac{\mathcal{K}(\rho,\pi)}{\eta} + \mu$$

$$+\sum_{m\in\mathcal{M}}\widehat{\rho}_m^{(\eta)}\left(2\varepsilon w_m(\widehat{s}_m)\underbrace{\sup_{t\in S_m}\left(\frac{W(t)-W(s_\rho)}{w_m(t)}\right)}_{\triangleq V_m}-\mathrm{pen}_2^{(\eta)}(m)\right). \tag{6.30}$$

The rest of the proof follows the same lines as the proof of [Mas07, Theorem 4.18] (recalled in Theorem 6.1). First, we use the fact that with large probability, for all $m\in\mathcal{M}$, the penalty $\mathrm{pen}_2^{(\eta)}(m)$ is large enough to annihilate the fluctuations $2\varepsilon w_m(\widehat{s}_m)V_m$. This is proved in [Mas07, Theorem 4.18] between Equations (4.78) and (4.82); Lemma 6.2 in Appendix 6.B.2 is a straightforward variant of this argument. We apply it with $a=s_\rho$.

To that end, let $z>0$ and $K'>1$, and set $y_m \triangleq K'\varepsilon\left(\sqrt{D_m}+\sqrt{2x_m}+(2\pi)^{-1/2}+\sqrt{2z}\right)$ for all $m\in\mathcal{M}$. Then, Lemma 6.2 indicates that on some event $\Omega_{z,K'}$ of probability $\mathbb{P}_s\big(\Omega_{z,K'}\big)\geqslant 1-\Sigma e^{-z}$, for all $m\in\mathcal{M}$,

$$\begin{aligned}2\varepsilon w_m(\widehat{s}_m)V_m \leqslant K'^2\varepsilon^2\big(\sqrt{D_m}+\sqrt{2x_m}\big)^2+\frac{2K'^2\varepsilon^2}{K'-1}\left(\frac{1}{2\pi}+2z\right)\\ +\frac{1}{\sqrt{K'}}\left(\|s-\widehat{s}_m\|^2+\frac{\|s-s_\rho\|^2}{\sqrt{K'}-1}\right).\end{aligned}$$

Now, we choose $K'\triangleq\sqrt{K}$. Therefore, from the last inequality and from the fact that $\mathrm{pen}_2^{(\eta)}(m)\triangleq \mathrm{pen}^{(\eta)}(m)-x_m/\eta\geqslant K\varepsilon^2\left(\sqrt{D_m}+\sqrt{2x_m}\right)^2$ by Assumption (6.20), we can see that on the event $\Omega_{z,K'}$, for all $m\in\mathcal{M}$,

$$2\varepsilon w_m(\widehat{s}_m)V_m-\mathrm{pen}_2^{(\eta)}(m)\leqslant\frac{2K\varepsilon^2}{\sqrt{K}-1}\left(\frac{1}{2\pi}+2z\right)+\frac{1}{K^{1/4}}\left(\|s-\widehat{s}_m\|^2+\frac{\|s-s_\rho\|^2}{K^{1/4}-1}\right). \tag{6.31}$$

Substituting the last inequality in (6.30), and using again the fact that $\sum_m\widehat{\rho}_m^{(\eta)}=1$, we get, on the event $\Omega_{z,K'}$,

$$\begin{aligned}\sum_{m\in\mathcal{M}}\widehat{\rho}_m^{(\eta)}\|\widehat{s}_m-s\|^2\leqslant\sum_{m\in\mathcal{M}}\rho_m\big(\|s_m-s\|^2+\mathrm{pen}_2^{(\eta)}(m)\big)+\frac{\mathcal{K}(\rho,\pi)}{\eta}+\mu\\ +\frac{1}{K^{1/4}}\sum_{m\in\mathcal{M}}\widehat{\rho}_m^{(\eta)}\|s-\widehat{s}_m\|^2+\frac{2K\varepsilon^2}{\sqrt{K}-1}\left(\frac{1}{2\pi}+2z\right)+\frac{\|s-s_\rho\|^2}{K^{1/4}\big(K^{1/4}-1\big)}.\end{aligned}$$

Noting that $\|s-s_\rho\|^2\leqslant\sum_{m\in\mathcal{M}}\rho_m\|s-s_m\|^2$ by definition of s_ρ and by convexity of $\|\cdot\|^2$, and reordering the terms of the inequality above, we get, with probability at least $1-\Sigma e^{-z}$,

$$\begin{aligned}&\left(1-\frac{1}{K^{1/4}}\right)\sum_{m\in\mathcal{M}}\widehat{\rho}_m^{(\eta)}\|\widehat{s}_m-s\|^2\\ &\leqslant(1+A_K)\left(\sum_{m\in\mathcal{M}}\rho_m\big(\|s_m-s\|^2+\mathrm{pen}_2^{(\eta)}(m)\big)+\frac{\mathcal{K}(\rho,\pi)}{\eta}+\mu\right)+\frac{2K\varepsilon^2}{\sqrt{K}-1}\left(\frac{1}{2\pi}+2z\right),\end{aligned} \tag{6.32}$$

where we set $A_K\triangleq 1/\big(K^{1/4}(K^{1/4}-1)\big)$. Dividing both sides by $1-K^{-1/4}$, recalling that

$\|s - s_m\|^2 \leqslant d^2(s, S_m) + \delta^2$ for all $m \in \mathcal{M}$, and integrating the resulting high-probability bound via Lemma A.7 in Appendix A.6 (see Example A.1), we get that

$$\begin{aligned}\mathbb{E}_s\left[\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} \|\widehat{s}_m - s\|^2\right] &\\ \leqslant \frac{1+A_K}{1-K^{-1/4}}\left(\sum_{m\in\mathcal{M}} \rho_m\big(d^2(s,S_m) + \delta^2 + \mathrm{pen}_2^{(\eta)}(m)\big) + \frac{\mathcal{K}(\rho,\pi)}{\eta} + \mu\right)&\\ + \frac{2K\varepsilon^2}{\big(1-K^{-1/4}\big)\big(\sqrt{K}-1\big)}\left(\frac{1}{2\pi} + 2\big(\ln_+(\Sigma)+1\big)\right).&\end{aligned} \tag{6.33}$$

Since the last inequality holds for all $\delta > 0$ and all $\rho \in \Delta(\mathcal{M})$ such that $\mathcal{K}(\rho,\pi) < \infty$, we conclude the proof of (6.21) by letting $\delta \to 0$, by recalling that $\mathrm{pen}_2^{(\eta)}(m) = \mathrm{pen}^{(\eta)}(m) - x_m/\eta$, by setting

$$C_K \triangleq \max\left\{\frac{1+A_K}{1-K^{-1/4}}\,,\ \frac{2K}{\big(1-K^{-1/4}\big)\big(\sqrt{K}-1\big)}\left(\frac{1}{2\pi}+2\right)\right\} \geqslant 1\,, \tag{6.34}$$

and by using the definitions of $\mathcal{J}\big(\widehat{\rho}^{(\eta)}\big)$ and of $\widetilde{s}^{(\eta)}$.

The upper bound (6.22) then follows straightforwardly by restricting the infimum over all $\rho \in \Delta(\mathcal{M})$ to the Dirac distributions $\rho = \delta_m$ at $m \in \mathcal{M}$, and by noting that $\mathcal{K}(\delta_m,\pi) = \ln(1/\pi_m) = x_m + \ln\Sigma$ (since $\pi_m \triangleq e^{-x_m}/\Sigma$). □

Remark 6.2 (Bound with high probability)**.** *We can derive oracle-type inequalities with high probability instead of risk bounds in expectation (as in [AM09, MM11] for instance). Indeed, note that* (6.32) *in the proof above holds with probability at least* Σe^{-z} *for any fixed* $\rho \in \Delta(\mathcal{M})$*, and in particular for any fixed Dirac distribution* $\delta_{m'}$ *at* $m' \in \mathcal{M}$*. Therefore, by a union-bound over* $\mathcal{M}$ *and by the equality* $\mathcal{K}(\delta_{m'},\pi) = x_{m'} + \ln\Sigma$*, we get that, with probability at least* $1 - \Sigma^2 e^{-z}$*, for all* $m' \in \mathcal{M}$*,*

$$\begin{aligned}\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} \|\widehat{s}_m - s\|^2 \leqslant \frac{1+A_K}{1-K^{-1/4}}\left(\|s_{m'} - s\|^2 + \mathrm{pen}_2^{(\eta)}(m') + \frac{x_{m'}+\ln\Sigma}{\eta} + \mu\right)&\\ + \frac{2K\varepsilon^2}{\big(1-K^{-1/4}\big)\big(\sqrt{K}-1\big)}\left(\frac{1}{2\pi} + 2(z + x_{m'})\right).&\end{aligned}$$

Therefore, using again $\|s - s_{m'}\|^2 \leqslant d^2(s,S_{m'}) + \delta^2$ *and* $\mathrm{pen}_2^{(\eta)}(m') = \mathrm{pen}^{(\eta)}(m') - x_{m'}/\eta$*, and letting*[9] $\delta \to 0$*, we get that, with probability at least* $1 - \Sigma^2 e^{-z}$*, for all* $m' \in \mathcal{M}$*,*

$$\begin{aligned}\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} \|\widehat{s}_m - s\|^2 \leqslant \frac{1+A_K}{1-K^{-1/4}}\left(d^2(s,S_{m'}) + \mathrm{pen}^{(\eta)}(m') + \frac{\ln\Sigma}{\eta} + \mu\right) + \frac{A'_K 2K\varepsilon^2 x_{m'}}{1-K^{-1/4}}&\\ + \frac{2K\varepsilon^2}{\big(1-K^{-1/4}\big)\big(\sqrt{K}-1\big)}\left(\frac{1}{2\pi} + 2z\right),&\end{aligned}$$

[9] Note that the probability $\Sigma^2 e^{-z}$ is independent of δ and that $\mathbb{P}[Z > a + \delta] \uparrow \mathbb{P}[Z > a]$ as $\delta \downarrow 0$ for any real random variable Z and any constant $a \in \mathbb{R}$.

where we set $A'_K \triangleq 2/\big(\sqrt{K}-1\big)$. *Finally, noting that* $2K\varepsilon^2 x_{m'} \leqslant \mathrm{pen}^{(\eta)}(m')$ *(by Assumption (6.20)), using the definitions of* $\mathcal{J}\big(\widehat{\rho}^{(\eta)}\big)$ *and of* $\widetilde{s}^{(\eta)}$, *and setting*

$$C'_K \triangleq \max\left\{\frac{1+A_K+A'_K}{1-K^{-1/4}}\,,\ \frac{4K}{\big(1-K^{-1/4}\big)\big(\sqrt{K}-1\big)}\right\}\,, \tag{6.35}$$

we get the bound stated in Remark 6.1: with probability at least $1-\Sigma^2 e^{-z}$,

$$\left\|\widetilde{s}^{(\eta)}-s\right\|^2 \leqslant C'_K \inf_{m'\in\mathcal{M}}\left\{d^2(s,S_{m'})+\mathrm{pen}^{(\eta)}(m')+\frac{\ln\Sigma}{\eta}+\varepsilon^2(z+1)+\mu\right\}-\mathcal{J}\Big(\widehat{\rho}^{(\eta)}\Big)\ .$$

6.4 Examples

In this section we apply Theorem 6.2 to classical problems such as aggregation of linear models, of finite models, and of ℓ^1-balls. The resulting oracle-type inequalities are comparable to those obtained with the model selection procedure of [Mas07, Theorem 4.18] (cf. Theorem 6.1). In a second part, we briefly explain why aggregation might outperform model selection in some situations where convexification is useful.

6.4.1 Application to some classical problems

Next we derive several corollaries of Theorem 6.2 in classical settings. They follow in a straightforward manner from the latter theorem and from the computations of the various generalized dimensions D_m that are carried out in [Mas07, Chapter 4]. We only present a few of them (linear models, finite models, and ℓ^1-balls) but all examples treated in [Mas07, Chapter 4] could also be addressed here (e.g., aggregation of Besov ellipsoids).

Aggregation of linear models

As explained in [BM01a, Mas07], the particular case of linear models already includes important practical problems such as variable selection, curve estimation, and change points detection.

The next corollary is an immediate consequence of Theorem 6.2 (with $\mu=0$ and $\eta \geqslant c/\varepsilon^2$) and of the fact that, for any finite dimensional linear subspace S_m of $\mathbb{H}$, its generalized dimension D_m coincides with its (classical) dimension – see [Mas07, p. 130].

Corollary 6.3 (Linear models)**.** *Fix some constant* $c>0$. *Consider the framework given in (6.9), and assume that* $(S_m)_{m\in\mathcal{M}}$ *is an at most countable collection of linear subspaces of* $\mathbb{H}$ *with finite dimensions* D_m *respectively. For all* $m\in\mathcal{M}$, *let* $\widehat{s}_m \in \mathrm{argmin}_{t\in S_m}\big\{\|t\|^2-2Y_\varepsilon(t)\big\}$ *be the least-squares estimator on* S_m. *Finally, let* $\eta \geqslant c/\varepsilon^2$ *and* $K>1$ *be some constants and take* $\mathrm{pen}^{(\eta)}:\mathcal{M}\to\mathbb{R}_+$ *such that, for all* $m\in\mathcal{M}$,

$$\mathrm{pen}^{(\eta)}(m) \geqslant K\varepsilon^2\left(\sqrt{D_m}+\sqrt{2x_m}\right)^2+\frac{x_m}{\eta}\,, \tag{6.36}$$

where $(x_m)\in\mathbb{R}_+^{\mathcal{M}}$ *is such that* $\Sigma \triangleq \sum_{m\in\mathcal{M}} e^{-x_m}<\infty$.

Then, for some constant $C_K > 1$ depending only on K, the estimator $\widetilde{s}^{(\eta)}$ defined in (6.17) – (6.18) satisfies, for all $s \in \mathbb{H}$,

$$\mathbb{E}_s\left[\left\|\widetilde{s}^{(\eta)} - s\right\|^2\right] \leqslant C_K\left(\inf_{m\in\mathcal{M}}\left\{d^2(s,S_m) + \mathrm{pen}^{(\eta)}(m)\right\} + \frac{\ln\Sigma}{\eta} + \varepsilon^2(\ln_+(\Sigma)+1)\right)$$
$$\leqslant C_K\left(\inf_{m\in\mathcal{M}}\left\{d^2(s,S_m) + \mathrm{pen}^{(\eta)}(m)\right\} + \varepsilon^2\big((1+c^{-1})\ln_+(\Sigma)+1\big)\right) .$$

A possible choice for the constant C_K above is given in (6.34). This choice, which follows from our analysis for nonlinear models, can of course be tightened in the particular case of linear models. A tighter analysis for refined penalties was indeed carried out for selection of linear models in [BM07a]. As for the aggregation of linear models, [LB06, Corollary 6] proved an oracle-type inequality with leading constant 1 (instead of $C_K > 1$) for an inverse temperature parameter corresponding to $\eta \leqslant 1/(4\varepsilon^2)$. They use a Mallow's C_p-type penalty (i.e., $\mathrm{pen}^{(\eta)}(m) = 2\varepsilon^2 D_m + x_m/\eta$) and analyse it through Stein's unbiased risk formula [Ste81].

An interesting open question is thus the following: in the particular case of linear models, and for a proper choice of the penalty, can we extend the refined risk bounds of [BM07a, LB06] through a single analysis to all $\eta > 0$ (instead of $\eta = +\infty$ or $\eta \leqslant 1/(4\varepsilon^2)$)? In particular, what are the performance of our aggregation procedure with a Mallow's C_p-type penalty when $\eta > 1/(4\varepsilon^2)$? Stein's unbiased risk formula could also be useful in this case.

Next we rewrite the risk bound above for the particular problems of ordered and complete variable selection in the Gaussian regression framework with fixed design. Recall from Example 6.1 in Section 6.1 that, in this setting, we observe

$$Y_i = s_i + \sigma\xi_i\,, \quad 1 \leqslant i \leqslant n\,,$$

where the ξ_i are independent standard normal random variables, where the noise level $\sigma > 0$ is assumed to be known, and where $s \in \mathbb{R}^n$ is the vector to be estimated. Recall also from Example 6.1 that for all $m \in \mathcal{M}$, the least-squares estimator $\widehat{s}_m \in \mathrm{argmin}_{t\in S_m}\left\{\|t\|^2 - 2Y_\varepsilon(t)\right\}$ can be rewritten in a more standard way:

$$\widehat{s}_m \in \operatorname*{argmin}_{t\in S_m}\|Y-t\|^2 .$$

Let $(\varphi_j)_{1\leqslant j\leqslant p}$ be a family of linearly independent vectors in $\mathbb{H} = \mathbb{R}^n$. Let $\mathcal{M}$ be a finite collection of subsets of $\{1,\dots,p\}$. For all $m \in \mathcal{M}$, define S_m as the linear span of $\{\varphi_j : j \in m\}$, i.e.,

$$S_m \triangleq \left\{\sum_{j=1}^p u_j\varphi_j : \boldsymbol{u}\in\mathbb{R}^p;\ \forall j\notin m\,,\ u_j = 0\right\}\,,$$

and denote by $D_m = |m|$ the dimension of S_m. Following the same lines as [BM01a, Mas07], we take $x_m \triangleq x(D_m)$ defined by

$$x(D) \triangleq \alpha D + \ln\left|\mathcal{M}_D\right|\,, \quad 0 \leqslant D \leqslant p\,, \tag{6.37}$$

where $\mathcal{M}_D \triangleq \{m \in \mathcal{M} : D_m = D\}$ and where $\alpha > 0$ is an absolute constant. Then we get $\Sigma = \sum_{D=0}^{p} |\mathcal{M}_D| e^{-x(D)} = 1/(1-e^{-\alpha}) < \infty$. With this choice of $(x_m)_{m\in\mathcal{M}}$, the choice of $\mathrm{pen}^{(\eta)}(m) = K\varepsilon^2\big(\sqrt{D_m} + \sqrt{2x(D_m)}\big)^2 + x(D_m)/\eta$, and the choice of $\eta \geqslant c/\varepsilon^2$ (as in Corollary 6.2), we get the risk bound

$$\mathbb{E}_s\left[\left\|\widetilde{s}^{(\eta)} - s\right\|^2\right] \leqslant C'_K \min_{0\leqslant D\leqslant p}\left\{d^2\big(s, S_{(D)}\big) + \varepsilon^2 D + \varepsilon^2\left(\ln\big|\mathcal{M}_D\big| + 1\right)\right\}, \tag{6.38}$$

for some constant $C'_K > 1$ depending only on (K, α, c). In the last inequality, $S_{(D)} \triangleq \bigcup_{m\in\mathcal{M}_D} S_m$ so that

$$d^2\big(s, S_{(D)}\big) = \inf_{m\in\mathcal{M}:D_m=D} d^2(s, S_m),$$

where the right-hand side equals $+\infty$ by convention if no model S_m has dimension D.

In the problem of *ordered variable selection*, $\mathcal{M} \triangleq \big\{\{1,\dots,D\} : D = 1,\dots,p\big\}$, so that, for all $m = \{1,\dots,D\} \in \mathcal{M}$, we have $S_m = \mathrm{span}\big(\varphi_1,\dots,\varphi_D\big)$ and $D_m = D$. Moreover, $\big|\mathcal{M}_D\big| = 1$ for all $D = 1,\dots,p$. Therefore, the risk bound (6.38) reads

$$\begin{aligned}\mathbb{E}_s\left[\left\|\widetilde{s}^{(\eta)} - s\right\|^2\right] &\leqslant C'_K\left(\min_{1\leqslant D\leqslant p}\left\{d^2\big(s, \mathrm{span}(\varphi_1,\dots,\varphi_D)\big) + \varepsilon^2 D\right\} + \varepsilon^2\right)\\ &= C'_K\left(\min_{m\in\mathcal{M}} \mathbb{E}_s\left[\|\widehat{s}_m - s\|^2\right] + \varepsilon^2\right).\end{aligned} \tag{6.39}$$

The last equality follows from the well-known bias-variance decomposition $\mathbb{E}_s\big[\|\widehat{s}_m - s\|^2\big] = d^2\big(s, \mathrm{span}(\varphi_1,\dots,\varphi_D)\big) + \varepsilon^2 D$ for all $m = \{1,\dots,D\}$, $D = 1,\dots,p$ (see, e.g., [Mas07, Section 4.2]). It indicates that the estimator $\widetilde{s}^{(\eta)}$ mimics the oracle $\widehat{s}_{m^*(s)}$, where $m^*(s) \in \mathrm{argmin}_{m\in\mathcal{M}} \mathbb{E}_s\big[\|\widehat{s}_m - s\|^2\big]$. It is thus an oracle inequality.

We now turn to the problem of *complete variable selection*, where $\mathcal{M} \triangleq \mathcal{P}\big(\{1,\dots,N\}\big)$. For all $D = 0,\dots,p$, we have[10] $\ln\big|\mathcal{M}_D\big| = \ln\binom{p}{D} \leqslant D\ln(\mathrm{e}p/D)$ by, e.g., [Mas07, Proposition 2.5]. Therefore, the risk bound (6.38) reads

$$\mathbb{E}_s\left[\left\|\widetilde{s}^{(\eta)} - s\right\|^2\right] \leqslant C'_K \min_{0\leqslant D\leqslant p}\left\{d^2\big(s, S_{(D)}\big) + \varepsilon^2 D\left(2 + \ln\frac{p}{D}\right) + \varepsilon^2\right\}, \tag{6.40}$$

where $S_{(D)} \triangleq \bigcup_{D_m=D} S_m$. Note that the above risk bound can be rewritten in a way similar to (6.39) but with a leading constant of the order of $\ln p$. However, if $(\varphi_j)_{1\leqslant j\leqslant p}$ is an orthonormal system in $\mathbb{R}^n$, then, by the lower bound for complete variable selection of [Mas07, Corollary 4.12], the risk bound above cannot be improved on any $S_{(D)}$ more than by constant factors. In particular, the estimator $\widetilde{s}^{(\eta)}$ is minimax optimal (up to constant factors) on each $S_{(D)}$, $D = 1,\dots,p$.

Note that (6.40) yields the following risk bound. This bound, which is due to [BM01a] for $\eta = +\infty$, is one of the first sparsity oracle inequalities — see Section 2.6 in Chapter 2 for an

[10] We use the natural convention $0\ln(A/0) = 0$ for all $A > 0$.

introduction. Recall that $\|\boldsymbol{u}\|_0 \triangleq \big|\{j : u_j \neq 0\}\big|$ for all $\boldsymbol{u} \in \mathbb{R}^p$.

$$\mathbb{E}_s\left[\left\|\widetilde{s}^{(\eta)} - s\right\|^2\right] \leqslant C'_K \min_{\boldsymbol{u}\in\mathbb{R}^p}\left\{\left\|\sum_{j=1}^p u_j\varphi_j - s\right\|^2 + \frac{\sigma^2}{n}\|\boldsymbol{u}\|_0\left(2 + \ln\frac{p}{\|\boldsymbol{u}\|_0}\right) + \frac{\sigma^2}{n}\right\}. \tag{6.41}$$

To see why (6.40) leads to (6.41), it suffices to note that for all $\boldsymbol{u} \in \mathbb{R}^d$, we have $\boldsymbol{u} \in S_{m(\boldsymbol{u})}$ where $m(\boldsymbol{u}) \triangleq \{j : u_j \neq 0\}$. Then, upper bounding the minimum in (6.40) by its argument in $D_{m(\boldsymbol{u})} \leqslant |m(\boldsymbol{u})| \triangleq \|\boldsymbol{u}\|_0$, noting that $x \mapsto x\big(2 + \ln(p/x)\big)$ is nondecreasing[11] on $[0, ep]$ (and a fortiori on $[0, p]$), and using $\boldsymbol{u} \in S_{m(\boldsymbol{u})} \subset S_{(D_{m(\boldsymbol{u})})}$ and $\varepsilon = \sigma/\sqrt{n}$ concludes the proof of (6.41).

Remark 6.3. *The linear independence assumption on the dictionary $(\varphi_j)_{1\leqslant j\leqslant p}$ is not necessary to derive the sparsity oracle inequality* (6.41)*. This assumption is only useful to reinterpret* (6.38) *as an oracle inequality of the form* (6.3)*, e.g., in* (6.39)*. If $(\varphi_j)_{1\leqslant j\leqslant p}$ is abitrary, then the penalty $\mathrm{pen}^{(\eta)}(m) = K\varepsilon^2\big(\sqrt{|m|} + \sqrt{2x'(|m|)}\big)^2 + x'(|m|)/\eta$ with $x'(D) \triangleq \alpha D + D\ln(ep/D)$ still satisfies* (6.36)*–*(6.37) *(since $|m| \geqslant D_m$ and since $D \mapsto x(D)$ is nondecreasing and such that $x'(D) \geqslant x(D)$). Applying Corollary 6.3 then also yields* (6.41)*.*

Aggregation of finite models

Next we rewrite Theorem 6.2 in the case of an at most countable collection of finite models. The interest of such models is commented on after the proof of the corollary.

Corollary 6.4 (Finite models)**.** *Consider the framework given in* (6.9)*. Assume that $(S_m)_{m\in\mathcal{M}}$ is an at most countable collection of non-empty finite subsets of $\mathbb{H}$, and denote their cardinalities by $|S_m|$. Let $\eta > 0$ and $K > 1$ be some constants and take $\mathrm{pen}^{(\eta)} : \mathcal{M} \to \mathbb{R}_+$ such that, for all $m \in \mathcal{M}$,*

$$\mathrm{pen}^{(\eta)}(m) \geqslant K\varepsilon^2\left(\sqrt{8\ln|S_m|} + \sqrt{2x_m}\right)^2 + \frac{x_m}{\eta},$$

where $(x_m) \in \mathbb{R}_+^{\mathcal{M}}$ is such that $\Sigma \triangleq \sum_{m\in\mathcal{M}} e^{-x_m} < \infty$.

Then, for some constant $C_K > 1$ depending only on K, the estimator $\widetilde{s}^{(\eta)}$ defined in (6.17) *–* (6.18) *satisfies, for all $s \in \mathbb{H}$,*

$$\mathbb{E}_s\left[\left\|\widetilde{s}^{(\eta)} - s\right\|^2\right] \leqslant C_K\left(\inf_{m\in\mathcal{M}}\left\{d^2(s, S_m) + \mathrm{pen}^{(\eta)}(m)\right\} + \frac{\ln\Sigma}{\eta} + \varepsilon^2(\ln_+(\Sigma) + 1)\right).$$

Proof: In view of Theorem 6.2 the only thing to prove is that for all $m \in \mathcal{M}$, a valid value[12] for the generalized dimension D_m of S_m is given by $D_m = 8\ln|S_m|$. By (6.11) it is sufficient to prove that the assumption (6.10) is satisfied with $\varphi_m(x) = x\sqrt{2\ln|S_m|}$.

For this purpose, let $x > 0$ and $u \in S_m$. Then, by the elementary inequality $a^2 + b^2 \geqslant 2ab$, and

[11] See Footnote 10.

[12] If S_m has cardinality one, then $\tau_m = 1$ so that the value $8\ln|S_m|$ exhibited below is only an upper bound on the solution of (6.11). This is not an issue since Theorem 6.2 only assumes a lower bound on the penalty.

by linearity[13] of $t \mapsto W(t)$, we get

$$2\,\mathbb{E}\left[\sup_{t\in S_m}\left(\frac{W(t)-W(u)}{\|t-u\|^2+x^2}\right)\right] \leqslant x^{-1}\mathbb{E}\left[\sup_{t\in S_m}\left(\frac{W(t)-W(u)}{\|t-u\|}\right)\right] = x^{-1}\mathbb{E}\left[\max_{t\in S_m} W\left(\frac{t-u}{\|t-u\|}\right)\right]$$
$$\leqslant x^{-1}\sqrt{2\ln|S_m|} = x^{-2}x\sqrt{2\ln|S_m|}\,,$$

where the last upper bound follows from a maximal inequality for subgaussian random variables stated in [Mas07, Lemma 2.3]. This fact is recalled in Lemma A.3 in Appendix A.5; we used it here with $T = |S_m| \geqslant 1$ and with $v = 1$ (since the random variables $W\big((t-u)/\|t-u\|\big)$ all have standard Gaussian distribution).

Therefore, the choice of $\varphi_m(x) = x\sqrt{2\ln|S_m|}$ satisfies (6.10), which in turns yields $D_m = 8\ln|S_m|$ by (6.11). This concludes the proof. □

Note that the above proof slightly improves on a computation carried out in [Mas07, Section 4.4.3] through a peeling argument. The author first remarks that Dudley's bound for metric entropy combined with this peeling argument immediately yields (6.10) with $\varphi_m(x) = \kappa x\sqrt{\ln|S_m|}$ for some absolute constant $\kappa > 0$. He then mentions that this is true for $\kappa = 8\sqrt{2}$ by Lemma A.3. The proof above shows that, unsurprisingly, we get a better constant $\kappa = \sqrt{2}$ via a global argument (i.e., without using a peeling argument, which is unnecessarily involved here). This is in the same spirit as in [MM11, Theorem A.1].

Finite models can be useful in at least two situations. First, we can see from the above corollary that deterministic estimators s_m of s are associated with the 0-dimensional nonlinear models $S_m = \{s_m\}$. Thus, to aggregate deterministic or *frozen* estimators in the Gaussian regression framework with fixed design, it is sufficient to use the penalty $\mathrm{pen}^{(\eta)}(m) = 2\big(K\varepsilon^2 + 1/\eta\big)x_m$. In this case, up to a small additive term of the order of $\ln_+(\Sigma)$, the price $\mathrm{pen}^{(\eta)}(m)$ to pay for aggregation is only proportional to the logarithm of the inverse of the prior probability mass e^{-x_m}/Σ assigned to $m \in \mathcal{M}$. This is what is expected when aggregating deterministic estimators (see, e.g., [DT08]).

The second situation for which finite models are useful (at least from a theoretical viewpoint) is when the models at hand are arbitrary compact subsets of $\mathbb{H}$. In this case, each model can be approximated by a finite set, so that selecting the best model in the collection approximately amounts to selecting the best associated finite set. This remark is one of the ideas that underly the metric point of view advocated by [Bir06] for adaptive estimation. We refer the reader to [Mas07, Section 4.4.3] for further details.

Aggregation of ℓ^1-balls

Next we derive another corollary of Theorem 6.2 when the models are associated to ℓ^1-balls. We consider the Gaussian regression framework with fixed design described in Example 6.1. Let $p \geqslant 1$ and $\boldsymbol{\varphi} = (\varphi_j)_{1\leqslant j\leqslant p}$ be a family of vectors in $\mathbb{R}^n$. We denote $\|\boldsymbol{u}\|_1 \triangleq \sum_{j=1}^p |u_j|$ and

[13] More precisely, we use the fact that, for all $t \in S_m$, $\big(W(t)-W(u)\big)/\|t-u\| = W\big((t-u)/\|t-u\|\big)$ almost surely (since by definition of an isonormal process $t \mapsto W(t)$ is a linear function from $\mathbb{H}$ into a space of square-integrable random variables). Since S_m is finite, the latter equality holds almost surely simultaneously for all $t \in S_m$.

$\boldsymbol{u} \cdot \boldsymbol{\varphi} \triangleq \sum_{j=1}^{p} u_j \varphi_j \in \mathbb{R}^n$ for all $\boldsymbol{u} \in \mathbb{R}^p$. For all $U > 0$ we set

$$\widehat{\boldsymbol{u}}(U) \in \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_1 \leqslant U} \left\| Y - \boldsymbol{u} \cdot \boldsymbol{\varphi} \right\|^2 . \tag{6.42}$$

Let $(U_m)_{m \in \mathcal{M}}$ be an at most countable family of positive real numbers. Following (6.17), given $\eta > 0$ and a penalty function $\mathrm{pen}^{(\eta)} : \mathcal{M} \to \mathbb{R}$, we set, for all $m \in \mathcal{M}$,

$$\widehat{\rho}_m^{(\eta)} = \frac{\exp\left[-\eta \Big(\|Y - \widehat{\boldsymbol{u}}(U_m) \cdot \boldsymbol{\varphi}\|^2 + \mathrm{pen}^{(\eta)}(m) \Big) \right]}{\sum_{m' \in \mathcal{M}} \exp\left[-\eta \Big(\|Y - \widehat{\boldsymbol{u}}(U_{m'}) \cdot \boldsymbol{\varphi}\|^2 + \mathrm{pen}^{(\eta)}(m') \Big) \right]} . \tag{6.43}$$

The next corollary upper bounds the risk of the aggregated estimator $\sum_{m \in \mathcal{M}} \widehat{\rho}_m^{(\eta)} \big(\widehat{\boldsymbol{u}}(U_m) \cdot \boldsymbol{\varphi} \big)$. As in the previous corollaries, it essentially relies on the computation of generalized dimensions carried out by [MM11, Theorem 3.1]. A key contribution of the last paper was to interpret the Lasso estimator as a selected projection estimator $\widehat{s}_{\widehat{m}}$ and hence to derive ℓ^1-oracle-type inequalities on the Lasso without any assumption on the dictionary $\boldsymbol{\varphi}$. Next we show that, unsurprisingly, the Bayesian variant $\sum_{m \in \mathcal{M}} \widehat{\rho}_m^{(\eta)} \big(\widehat{\boldsymbol{u}}(U_m) \cdot \boldsymbol{\varphi} \big)$ satisfies a similar risk bound. We however do not claim that this estimator should be prefered to the Lasso since, unlike this efficient algorithm, it involves the computation of possibly many exponential weights[14].

Corollary 6.5 (Aggregation of ℓ^1-balls)**.** *Consider the Gaussian regression framework with fixed design described in Example 6.1. Let $(\varphi_j)_{1 \leqslant j \leqslant p}$ be a family of vectors in $\mathbb{R}^n$, and $(U_m)_{m \in \mathcal{M}}$ be an at most countable family of positive real numbers. Let $\eta > 0$ and $K > 1$ be some constants and take $\mathrm{pen}^{(\eta)} : \mathcal{M} \to \mathbb{R}_+$ such that, for all $m \in \mathcal{M}$,*

$$\mathrm{pen}^{(\eta)}(m) \geqslant 4KU_m \gamma\sigma \sqrt{\frac{2\ln(2p)}{n}} + \left(\frac{4K\sigma^2}{n} + \frac{1}{\eta} \right) x_m , \tag{6.44}$$

where $(x_m)_{m \in \mathcal{M}} \in \mathbb{R}_+^{\mathcal{M}}$ is such that $\Sigma \triangleq \sum_{m \in \mathcal{M}} e^{-x_m} < \infty$ and where $\gamma \triangleq \max_{1 \leqslant j \leqslant p} \|\varphi_j\| = \max_{1 \leqslant j \leqslant p} \big(n^{-1} \sum_{i=1}^n \varphi_{i,j}^2 \big)^{1/2}$.

Then, the estimator $\widetilde{s}^{(\eta)} = \sum_{m \in \mathcal{M}} \widehat{\rho}_m^{(\eta)} \big(\widehat{\boldsymbol{u}}(U_m) \cdot \boldsymbol{\varphi} \big)$ given by (6.42)–(6.43) is well-defined and satisfies, for some constant $C_K > 1$ depending only on K, for all $s \in \mathbb{R}^n$,

$$\begin{aligned} &\mathbb{E}_s \left[\left\| \widetilde{s}^{(\eta)} - s \right\|^2 \right] \\ &\qquad \leqslant C_K \left(\inf_{m \in \mathcal{M}} \left\{ \min_{\boldsymbol{u} : \|\boldsymbol{u}\|_1 \leqslant U_m} \|\boldsymbol{u} \cdot \boldsymbol{\varphi} - s\|^2 + \mathrm{pen}^{(\eta)}(m) \right\} + \frac{\ln \Sigma}{\eta} + \frac{\sigma^2}{n} (\ln_+(\Sigma) + 1) \right) . \end{aligned}$$

Before proving the corollary, note that if the penalty is chosen as the right-hand side of (6.44) and if η is at least of the order of n/σ^2, then the risk bound of Corollary 6.5 scales for each $m \in \mathcal{M}$ approximately as $U_m \gamma\sigma \sqrt{\ln(2p)/n}$. It is therefore very similar to the regret bounds derived on ℓ^1-balls in the online linear regression setting (see Chapters 2 and 4). This similarity is not surprising

[14] Note however that, letting $\widehat{\boldsymbol{u}}^{\mathrm{LSE}}$ be any least-squares estimator in $\mathbb{R}^p$, we can choose $\widehat{\boldsymbol{u}}(U_m) = \widehat{\boldsymbol{u}}^{\mathrm{LSE}}$ for all $U_m > \big\| \widehat{\boldsymbol{u}}^{\mathrm{LSE}} \big\|_1$. Thus, for U_m of the form $U_m = 2^m \sigma/(\gamma\sqrt{n})$ and for $x_m = m$, the infinite sum in (6.43) and the estimator $\widetilde{s}^{(\eta)}$ can be computed exactly and with a computational complexity which is linear in $\log_2 \big(\big\| \widehat{\boldsymbol{u}}^{\mathrm{LSE}} \big\|_1 \gamma\sqrt{n}/\sigma \big)$.

in view of the connections between the online linear regression setting and the stochastic batch setting (e.g., as shown in Section 4.2.1, a Maurey-type argument can be used in both settings).

The connections between these settings are actually deeper. Indeed, take $\mathcal{M} = \mathbb{N}$ and set $U_m \triangleq 2^m\sigma/(\sqrt{n}\gamma)$ and $x_m \triangleq m$ for all $m \in \mathbb{N}$ in a way similar to [MM11]; choose the penalty as the right-hand side of (6.44). Then, we can see that for all η at least of the order of n/σ^2, Corollary 6.5 yields a risk bound of the form

$$\mathbb{E}_s\left[\left\|\widetilde{s}^{(\eta)} - s\right\|^2\right] \leqslant C'_K\left(\inf_{\boldsymbol{u}\in\mathbb{R}^p}\left\{\|\boldsymbol{u}\cdot\boldsymbol{\varphi} - s\|^2 + \|\boldsymbol{u}\|_1\,\gamma\sigma\sqrt{\frac{\ln(2p)}{n}}\right\} + \frac{\sigma^2}{n}(\ln_+(\Sigma)+1)\right),$$

where $C'_K > 0$ is an absolute constant depending only on K. In Chapter 4 we proved a regret bound of a similar form as far as adaptation to U was concerned (see Section 4.4).

Proof (of Corollary 6.5): First note that the estimators $\widehat{\boldsymbol{u}}(U_m)\cdot\boldsymbol{\varphi} \in \mathbb{R}^n$ are nothing but the projection estimators associated to the models

$$S_{U_m} \triangleq \left\{\boldsymbol{u}\cdot\boldsymbol{\varphi} : \boldsymbol{u}\in\mathbb{R}^p, \|\boldsymbol{u}\|_1 \leqslant U_m\right\},$$

i.e., $\left(\widehat{\boldsymbol{u}}(U_m)\cdot\boldsymbol{\varphi}\right) \in \operatorname{argmin}_{t\in S_{U_m}}\|Y-t\|^2$ for all $m\in\mathcal{M}$. Note also that, by (6.44) and by the elementary inequality $2(a^2+b^2) \geqslant (a+b)^2$ for all $a,b\in\mathbb{R}$, we have, for all $m\in\mathcal{M}$,

$$\operatorname{pen}^{(\eta)}(m) \geqslant K\frac{\sigma^2}{n}\left(\sqrt{\frac{2U_m\gamma}{\sigma}\sqrt{2n\ln(2p)}} + \sqrt{2x_m}\right)^2 + \frac{x_m}{\eta}\,.$$

Theorefore, in view of Theorem 6.2, the only thing to prove is that for all $m\in\mathcal{M}$, a valid value for the generalized dimension D_m of S_{U_m} is given by $D_m = (2U_m\gamma/\sigma)\sqrt{2n\ln(2p)}$. By (6.11) and by the fact that $\varepsilon = \sigma/\sqrt{n}$, it is sufficient to prove that the assumption (6.10) is satisfied with $\varphi_m(x) = 2U_m\gamma\sqrt{2\ln(2p)}$. This fact is essentially proved in [MM11, Theorem 3.1] via the linearity of $t\mapsto W(t)$ on the polytope S_{U_m} and via the maximal inequality for subgaussian random variables of [Mas07, Lemma 2.3] (cf. Lemma A.3 in Appendix A.5). □

6.4.2 A situation where convexification is useful

In all the examples addressed in the previous section, we always neglected the nonnegative term $\mathbb{E}_s\left[\mathcal{J}\left(\widehat{\rho}^{(\eta)}\right)\right]$ appearing in the risk bound of Theorem 6.2. As a consequence, all bounds of the previous section are comparable to the bounds that would derive from Theorem 6.1 for the model selection procedure, but they do not show any improvement over them. Since the nonnegative term $\mathbb{E}_s\left[\mathcal{J}\left(\widehat{\rho}^{(\eta)}\right)\right]$ is a gap in a Jensen-type inequality, it suggests that in some favorable situations, combining the base estimators $\widehat{s}_m$ instead of selecting one of them may result in better performance. Next we describe a typical situation in which convexification is indeed useful and we prove a simple toy lower bound for model selection indicating that the latter is less robust than aggregation in large-bias situations.

The benefits of convexification were already pointed out in various settings in the past (see, e.g., the introduction of [Yan01]). For example, the *bagging* method introduced by [Bre96] was shown to improve the performance of unstable base estimators. For the regression problem, the

latter technique consists in computing several predictions over independent bootstrap samples, and then in averaging the resulting predictions. As indicated by the author, this averaging is useful when the base predictions are unstable (because of a large gap in a Jensen-type inequality).

The improvement of model aggregation over model selection was formally proved by, e.g., [Cat99, Section 8] for a distribution estimation problem and by [Cat04, Section 4.7] and [Aud07, JRT08, LM09] for the regression problem. A key idea in these works is that aggregation procedures can improve over selection procedures if the base (deterministic) estimators at hand are far away from the prediction target and if several of them are quasi-optimal and well-separated. The aforementioned works address the case of deterministic (frozen) base estimators.

Next we show that a similar improvement is also possible in our setting, where the base estimators $\widehat{s}_m$ are random (recall that no sample splitting is allowed in this fixed-design context).

Proposition 6.1 (A lower bound on model selection)**.** *Consider the regression framework with fixed design described in Example 6.1. Then, there exists a collection of linear models $(S_m)_{m\in\mathcal{M}}$ in $\mathbb{R}^n$ with $|\mathcal{M}| = 2$ such that for all $n \geqslant 16/(\sqrt{2}-1)^2$,*

$$\forall s \in \mathbb{R}^n, \quad \mathbb{E}_s\Big[\big\|\widetilde{s}^{(\eta)} - s\big\|^2\Big] \leqslant \inf_{m\in\mathcal{M}} \mathbb{E}_s\Big[\big\|\widehat{s}_m - s\big\|^2\Big] + \frac{4\ln(2)\,\sigma^2}{n}\,, \tag{6.45}$$

$$\forall \widehat{m}, \quad \exists s \in \mathbb{R}^n, \quad \mathbb{E}_s\Big[\big\|\widehat{s}_{\widehat{m}} - s\big\|^2\Big] \geqslant \inf_{m\in\mathcal{M}} \mathbb{E}_s\Big[\big\|\widehat{s}_m - s\big\|^2\Big] + \frac{\sigma^2}{4\sqrt{n}}\,, \tag{6.46}$$

where $\widetilde{s}^{(\eta)}$ is defined in (6.17)-(6.18) with $\eta = n/(4\sigma^2)$ and $\mathrm{pen}^{(\eta)}(m) = 2\dim(S_m)\sigma^2/n$ (i.e., $\widetilde{s}^{(\eta)}$ is the aggregating estimator of [LB06] with the largest allowed inverse temperature parameter), and where (6.46) holds for all measurable functions $\widehat{m} : \mathbb{R}^n \to \mathcal{M}$ (i.e., all data-driven selectors).

Therefore, there are situations where the aggregation procedure of [LB06] has a risk smaller than that of the oracle up to an additive term at most of the order of $1/n$, while any model selection procedure cannot beat the oracle at a rate faster than $1/\sqrt{n}$ uniformly over all $s \in \mathbb{R}^n$.

In the toy example exhibited in the subsequent proof, the bias of the estimators $\widehat{s}_1$ and $\widehat{s}_2$ is large (of the order of σ^2). Therefore, the lower bound (6.46) does not contradict the fact that model selection procedures are minimax optimal (up to constant factors) in many classical problems for which the prediction target s lies within a model (see, e.g., [Mas07, Chapter 4]). However, this lower bound indicates that, at least for linear models, if the target vector is far from all the models at hand (hence a large bias) and if a few models are nearly-optimal and well-separated, then there is an aggregation procedure whose excess risk is much smaller than that of any model selection procedure (compare the rates $1/n$ and $1/\sqrt{n}$). In this sense, model aggregation can be thought of as more robust than model selection.

In the general case of nonlinear models, all our oracle-type inequalities were obtained with leading constants larger than 1 (contrary to [LB06]). Therefore, it is not clear for the moment whether the aforementioned robustness property also holds true for our aggregation procedure. However, in view of the simplicity of the example exhibited in the following proof, we tend to think that aggregation might benefit from a similar advantage with nonlinear models. This important question remains open.

Proof (of Proposition 6.1): The upper bound (6.45) follows directly from [LB06, Corollary 6]. As for the lower bound (6.46), it can be proved with the following toy example, which is inspired from a lower bound of [Cat04, pages 134–135] in a slightly different setting (in our case, the base estimators $\widehat{s}_m$ are random). Define the two models $S_1 \subset \mathbb{R}^n$ and $S_2 \subset \mathbb{R}^n$ by

$$S_1 \triangleq \mathbb{R} \times \{0\} \times \ldots \times \{0\} \qquad \text{and} \qquad S_2 \triangleq \{0\} \times \mathbb{R} \times \{0\} \times \ldots \times \{0\} \,.$$

Consider the following two potential target vectors (one of which is going to be badly estimated by the model selection procedure $\widehat{m}$ at hand):

$$s_\alpha \triangleq \sigma\sqrt{n}\left(1 + \frac{c\alpha}{\sqrt{n}}, 1 - \frac{c\alpha}{\sqrt{n}}, 0, \ldots, 0\right) \in \mathbb{R}^n \,, \quad \alpha \in \{-1, 1\} \,,$$

where $c \in (0,1)$ is an absolute constant to be determined by the analysis. If the true vector is s_α, the statistician observes the n-dimensional vector $Y = s_\alpha + \sigma\xi$ with $\xi = (\xi_1, \ldots, \xi_n)$, where the ξ_i are i.i.d. standard normal random variables (cf. Example 6.1). In the sequel we denote the law of $s_\alpha + \sigma\xi$ by $\mathbb{P}_{s_\alpha}$ (i.e., $\mathbb{P}_{s_\alpha} = \mathcal{N}(s_\alpha, \sigma^2 I_n)$) and the corresponding expectation by $\mathbb{E}_{s_\alpha}$.

Let $\widehat{m} : \mathbb{R}^n \to \{1, 2\}$ be any measurable function. The rest of the proof is dedicated to show that, for all $c \in (0,1)$,

$$\max_{\alpha \in \{-1,1\}} \left\{ \mathbb{E}_{s_\alpha}\left[\left\| \widehat{s}_{\widehat{m}} - s_\alpha \right\|^2 \right] - \inf_{m \in \{1,2\}} \mathbb{E}_{s_\alpha}\left[\left\| \widehat{s}_m - s_\alpha \right\|^2 \right] \right\} \geqslant \frac{4c\sigma^2}{\sqrt{n}} \left(\frac{1}{2} - \frac{c}{\sqrt{2}} - \frac{1}{4c\sqrt{n}} \right) . \tag{6.47}$$

First note that if $Y = s_\alpha + \sigma\xi$, then, by definition of $\widehat{s}_m \in \operatorname{argmin}_{t \in S_m} \|Y - t\|^2$,

$$\widehat{s}_1 = \left(\sigma\sqrt{n}\left(1 + \frac{c\alpha}{\sqrt{n}}\right) + \sigma\xi_1, 0, 0, \ldots, 0 \right) \,,$$
$$\widehat{s}_2 = \left(0, \sigma\sqrt{n}\left(1 - \frac{c\alpha}{\sqrt{n}}\right) + \sigma\xi_2, 0, \ldots, 0 \right) \,.$$

Recall from Example 6.1 that we set $\|u\|^2 \triangleq n^{-1}\sum_{i=1}^n u_i^2$ for all $u \in \mathbb{R}^n$. By the two equalities above, if $Y = s_\alpha + \sigma\xi$, then for all $m \in \{1, 2\}$,

$$\|\widehat{s}_m - s_\alpha\|^2 = \begin{cases} \dfrac{\sigma^2\xi_1^2}{n} + \sigma^2\left(1 - \dfrac{c\alpha}{\sqrt{n}}\right)^2 & \text{if } m = 1 \,, \\ \sigma^2\left(1 + \dfrac{c\alpha}{\sqrt{n}}\right)^2 + \dfrac{\sigma^2\xi_2^2}{n} & \text{if } m = 2 \,. \end{cases}$$

The last equality yields, on the one hand,

$$\inf_{m \in \{1,2\}} \mathbb{E}_{s_\alpha}\left[\left\| \widehat{s}_m - s_\alpha \right\|^2 \right] = \sigma^2\left(1 - \frac{c}{\sqrt{n}}\right)^2 + \frac{\sigma^2}{n} \,, \tag{6.48}$$

and, on the other hand, almost surely,

$$\|\widehat{s}_{\widehat{m}} - s_\alpha\|^2 \geqslant \mathbb{I}_{\{\widehat{m}=1\}}\,\sigma^2\left(1 - \frac{c\alpha}{\sqrt{n}}\right)^2 + \mathbb{I}_{\{\widehat{m}=2\}}\,\sigma^2\left(1 + \frac{c\alpha}{\sqrt{n}}\right)^2$$

$$= \sigma^2 \left(1 + \frac{c}{\sqrt{n}}\right)^2 - \sigma^2 \left[\left(1 + \frac{c}{\sqrt{n}}\right)^2 - \left(1 - \frac{c}{\sqrt{n}}\right)^2\right] \mathbb{I}_{\{\widehat{m}=m_\alpha\}} \,,$$

where we set $m_1 \triangleq 1$ and $m_{-1} \triangleq 2$ and where used the fact that $\mathbb{I}_{\{\widehat{m}=1\}} + \mathbb{I}_{\{\widehat{m}=2\}} = 1$ almost surely. Taking the expectations of both sides of the last inequality, substracting (6.48), and using the fact that $\big(1 + c/\sqrt{n}\big)^2 - \big(1 - c/\sqrt{n}\big)^2 = 4c/\sqrt{n}$, we get

$$\begin{aligned} &\max_{\alpha\in\{-1,1\}} \left\{ \mathbb{E}_{s_\alpha}\Big[\|\widehat{s}_{\widehat{m}} - s_\alpha\|^2\Big] - \inf_{m\in\{1,2\}} \mathbb{E}_{s_\alpha}\Big[\big\|\widehat{s}_m - s_\alpha\big\|^2\Big] \right\} \\ &\qquad \geqslant \frac{4c\sigma^2}{\sqrt{n}} \left(1 - \min_{\alpha\in\{-1,1\}} \mathbb{P}_{s_\alpha}\big[\widehat{m} = m_\alpha\big]\right) - \frac{\sigma^2}{n} \,. \end{aligned} \tag{6.49}$$

To prove (6.47), it suffices to upper bound the minimum in the last inequality. But, using Pinsker's inequality (cf. Lemma A.8 in Appendix A.7), we can see that, if $\mathbb{P}_{s_0}$ denotes the law of the n-dimensional random vector $s_0 + \sigma\xi$ with $s_0 \triangleq \sigma\sqrt{n}(1, 1, 0, \ldots, 0)$, then

$$\begin{aligned} \min_{\alpha\in\{-1,1\}} \mathbb{P}_{s_\alpha}\big[\widehat{m} = m_\alpha\big] &\leqslant \min_{\alpha\in\{-1,1\}} \left\{ \mathbb{P}_{s_0}\big[\widehat{m} = m_\alpha\big] + \sqrt{\frac{\mathcal{K}\big(\mathbb{P}_{s_\alpha}, \mathbb{P}_{s_0}\big)}{2}} \right\} \\ &\leqslant \min_{\alpha\in\{-1,1\}} \mathbb{P}_{s_0}\big[\widehat{m} = m_\alpha\big] + \sqrt{\frac{\max_{\alpha\in\{-1,1\}} \mathcal{K}\big(\mathbb{P}_{s_\alpha}, \mathbb{P}_{s_0}\big)}{2}} \\ &\leqslant \frac{1}{2} + \frac{c}{\sqrt{2}} \,, \end{aligned}$$

where the last inequality follows from the fact that $\{\widehat{m} = m_1\} \cap \{\widehat{m} = m_2\} = \varnothing$ and from the elementary equalities $\mathcal{K}\big(\mathbb{P}_{s_\alpha}, \mathbb{P}_{s_0}\big) = \sum_{i=1}^n \big(s_{i,\alpha} - s_{i,0}\big)^2/(2\sigma^2) = 2\sigma^2c^2/(2\sigma^2) = c^2$.

Substituting the last upper bound in the right-hand side of (6.49) directly yields the lower bound (6.47). We conclude the proof of (6.46) by choosing $c = 1/(2\sqrt{2})$ and by using the assumption that $n \geqslant 16/(\sqrt{2} - 1)^2$. □

6.5 Future works

As mentioned earlier, this chapter is a work in progress. In particular, important open questions remain open. Among the issues raised throughout this chapter, we ask the following:

- Our oracle-type inequalities are only obtained with leading constants larger than 1. Is this a consequence of the concentration approach — which however yields risk bounds with high probability — or of the generality of the models? In particular, when the models are linear, it could be interesting to recover via a single analysis the tighter bounds of [LB06] and of [BM07a] obtained for model aggregation and model selection respectively.

- The important problem of the tuning of η is left open. Is it possible to identify — at least for classical problems — an optimal choice of η? If so, can we tune η in an automatic and nearly-optimal way?

- Finally, investigating classical examples of nonlinear models (e.g., Besov ellipsoids, ℓ^1-balls, neural networks) could help to compare the model selection procedure of [Mas07] with our aggregation procedure.

6.A Proofs

6.A.1 Proof of Theorem 6.2 when $\mathcal{M}$ is countably infinite

Theorem 6.2 is stated for an at most countable collection $\mathcal{M}$. In Section 6.3 we only proved it under the assumption that $\mathcal{M}$ is finite. Next we provide a proof in the other and more technical case, i.e., when $\mathcal{M}$ is countably infinite.

Proof (of Theorem 6.2, $\mathcal{M}$ countably infinite): We assume in the sequel that $\mathcal{M}$ in countably infinite. The proof consists of three steps. In Steps 1 and 2, we check that the two sums over $\mathcal{M}$ appearing in the definition of $\widetilde{s}^{(\eta)}$ are convergent. In Step 3, we then employ a reduction to the case of a finite collection to prove the oracle-type inequality (6.21).

Step 1: We prove that

$$Z \triangleq \sum_{m\in\mathcal{M}} \exp\Big[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}^{(\eta)}(m)\Big)\Big] < \infty \quad \text{almost surely.}$$

Recall from (6.28) that $\gamma_\varepsilon(\widehat{s}_m)$ can be rewritten as

$$\gamma_\varepsilon(\widehat{s}_m) = \|\widehat{s}_m - s\|^2 - \|s\|^2 - 2\varepsilon W\big(\widehat{s}_m\big)\;, \quad m\in\mathcal{M}\;. \tag{6.50}$$

We also set $\mathrm{pen}_2^{(\eta)}(m) \triangleq \mathrm{pen}^{(\eta)}(m) - x_m/\eta$ for all $m\in\mathcal{M}$. Therefore, we have, for all $m\in\mathcal{M}$,

$$\begin{aligned}&\exp\Big[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}^{(\eta)}(m)\Big)\Big]\\ &\quad= e^{\eta\|s\|^2}e^{-x_m}\exp\Big[-\eta\,\|\widehat{s}_m-s\|^2\Big]\exp\Big[\eta\Big(2\varepsilon W\big(\widehat{s}_m\big)-\mathrm{pen}_2^{(\eta)}(m)\Big)\Big]\;.\end{aligned} \tag{6.51}$$

Next we use Lemma 6.2 to upper bound the quantity $2\varepsilon W\big(\widehat{s}_m\big)-\mathrm{pen}_2^{(\eta)}(m)$ with high probability. We follow the same arguments that led to (6.31) in the proof of Theorem 6.2 for a finite collection (see Section 6.3). Namely, let $z>0$, and set, for all $m\in\mathcal{M}$ and $t\in\mathbb{H}$,

$$\begin{aligned} y_m &\triangleq \sqrt{K}\varepsilon\left(\sqrt{D_m}+\sqrt{2x_m}+(2\pi)^{-1/2}+\sqrt{2z}\right)\;,\\ w_m(t) &\triangleq \frac{1}{2}\left(\big[\|s\|+\|s-t\|\big]^2+y_m^2\right)\;.\end{aligned}$$

Applying Lemma 6.2 in Appendix 6.B.2 with $K'=\sqrt{K}$, $a=0$, and $V_m=\sup_{t\in S_m}\big\{W(t)/w_m(t)\big\}$, we get, on some event $\Omega_{z,K'}$ of probability $\mathbb{P}_s\big(\Omega_{z,K'}\big)\geqslant 1-\Sigma e^{-z}$, that for all $m\in\mathcal{M}$,

$$\begin{aligned}2\varepsilon W(\widehat{s}_m) \leqslant 2\varepsilon w_m(\widehat{s}_m)V_m &\leqslant K\varepsilon^2\big(\sqrt{D_m}+\sqrt{2x_m}\big)^2+\frac{2K\varepsilon^2}{\sqrt{K}-1}\left(\frac{1}{2\pi}+2z\right)\\ &\quad+K^{-1/4}\left(\|s-\widehat{s}_m\|^2+\frac{\|s\|^2}{K^{1/4}-1}\right)\;.\end{aligned}$$

Therefore, from the last inequality and from the fact that $\mathrm{pen}_2^{(\eta)}(m) \triangleq \mathrm{pen}^{(\eta)}(m) - x_m/\eta \geqslant K\varepsilon^2\left(\sqrt{D_m}+\sqrt{2x_m}\right)^2$ by Assumption (6.20), we can see that, on the event $\Omega_{z,K'}$, for all $m \in \mathcal{M}$,

$$2\varepsilon W(\widehat{s}_m) - \mathrm{pen}_2^{(\eta)}(m) \leqslant \frac{2K\varepsilon^2}{\sqrt{K}-1}\left(\frac{1}{2\pi}+2z\right) + K^{-1/4}\left(\|s-\widehat{s}_m\|^2 + \frac{\|s\|^2}{K^{1/4}-1}\right) . \quad (6.52)$$

Substituting the last inequality in (6.51), we get, on the event $\Omega_{z,K'}$, that for all $m \in \mathcal{M}$,

$$\begin{aligned}&\exp\Big[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}^{(\eta)}(m)\Big)\Big] \\ &\quad\leqslant A_{\eta,\varepsilon,s,K}\exp\left[\frac{4\eta K\varepsilon^2 z}{\sqrt{K}-1}\right]e^{-x_m}\exp\Big[-\eta\big(1-K^{-1/4}\big)\|\widehat{s}_m-s\|^2\Big] , \end{aligned} \quad (6.53)$$

where we set $A_{\eta,\varepsilon,s,K} \triangleq e^{\eta\|s\|^2}\exp\Big[\big(\pi^{-1}\eta K\varepsilon^2\big)/\big(\sqrt{K}-1\big)+\big(\eta K^{-1/4}\|s\|^2\big)/\big(K^{1/4}-1\big)\Big]$. Upper bounding the last exponential in (6.53) by 1 (since $K>1$), summing the resulting inequality over $m \in \mathcal{M}$, and using the assumption $\sum_{m\in\mathcal{M}} e^{-x_m} < \infty$ (see Section 6.2.2), we can see that

$$Z \triangleq \sum_{m\in\mathcal{M}} \exp\Big[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}^{(\eta)}(m)\Big)\Big]$$

is finite on all events $\Omega_{z,K'}$, $z>0$. Applying, e.g., Borel-Cantelli's lemma to the family of complementary events $\big(\Omega_{2\ln k}^c\big)_{k\in\mathbb{N}^*}$, we deduce from $\sum_{k=1}^\infty \mathbb{P}_s\big(\Omega_{2\ln k}^c\big) \leqslant \sum_{k=1}^\infty \Sigma/k^2 < \infty$ that, almost surely, the defining conditions of the events $\Omega_{2\ln k}$ are satisfied for all k large enough. Therefore, Z is almost surely finite, which proves that $\widehat{\rho}^{(\eta)}$ is well-defined.

Step 2: We prove that $\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)}\|\widehat{s}_m\| < \infty$ almost surely.
We follow the same lines as in Step 1. Noting that $\widehat{\rho}_m^{(\eta)} = Z^{-1}\exp\Big[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}^{(\eta)}(m)\Big)\Big]$ where $Z>0$ is the random normalization constant studied above, we get, by the triangle inequality and by the equality $\sum_{m\in\mathcal{M}}\widehat{\rho}_m^{(\eta)}=1$, that on the event $\Omega_{z,K'}$ introduced in Step 1,

$$\begin{aligned}&\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)}\|\widehat{s}_m\| \\ &\quad\leqslant \|s\| + \sum_{m\in\mathcal{M}}\widehat{\rho}_m^{(\eta)}\|\widehat{s}_m-s\| \\ &\quad= \|s\| + \frac{1}{Z}\sum_{m\in\mathcal{M}}\exp\Big[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}^{(\eta)}(m)\Big)\Big]\|\widehat{s}_m-s\| \\ &\quad\leqslant \|s\| + \frac{A_{\eta,\varepsilon,s,K}}{Z}\exp\left[\frac{4\eta K\varepsilon^2 z}{\sqrt{K}-1}\right]\sum_{m\in\mathcal{M}} e^{-x_m}\exp\Big[-\eta\big(1-K^{-1/4}\big)\|\widehat{s}_m-s\|^2\Big]\|\widehat{s}_m-s\| , \end{aligned}$$

where the last inequality follows from (6.53). Using the fact that $\sup_{t\in\mathbb{R}_+}\big\{e^{-At^2}t\big\} < \infty$ for all $A>0$, and in particular for $A \triangleq \eta\big(1-K^{-1/4}\big)$ (note that $A>0$ since $K>1$), we get that

$$\sum_{m\in\mathcal{M}}\widehat{\rho}_m^{(\eta)}\|\widehat{s}_m\| < \infty \quad \text{on all events } \Omega_{z,K'} \ , \ z>0 \ .$$

Therefore, using the same argument as at the end of Step 1 (e.g., Borel-Cantelli's lemma), we can see that $\sum_{m\in\mathcal{M}}\widehat{\rho}_m^{(\eta)}\|\widehat{s}_m\|$ is almost surely finite. Since $\mathbb{H}$ is complete (by definition of a Hilbert

space), this proves that $\widetilde{s}^{(\eta)} = \sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} \widehat{s}_m$ is well-defined.

Step 3: Reduction to a finite collection.

Since $\mathcal{M}$ is countably infinite, we can assume (up to a one-to-one mapping) that $\mathcal{M} = \mathbb{N}$. In order to derive (6.21), we employ a reduction to the finite collections $(S_m)_{0\leqslant m\leqslant M}$, where $M \in \mathbb{N}$. Applying Theorem 6.2 in the finite case will then conclude the proof.

To that end, we set, for all $M \in \mathbb{N}$ and all $m \in \{0,\dots,M\}$,

$$\widehat{\rho}_m^{(\eta,M)} \triangleq \frac{\exp\Big[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}(m)\Big)\Big]}{\sum_{m'=0}^{M} \exp\Big[-\eta\Big(\gamma_\varepsilon(\widehat{s}_{m'}) + \mathrm{pen}(m')\Big)\Big]} \tag{6.54}$$

$$\pi_m^{(M)} \triangleq \frac{e^{-x_m}}{\sum_{m'=0}^{M} e^{-x_{m'}}} \,. \tag{6.55}$$

Thus $\widehat{\rho}^{(\eta,M)} \triangleq \big(\widehat{\rho}_m^{(\eta,M)}\big)_{0\leqslant m\leqslant M} \in \Delta\big(\{0,\dots,M\}\big)$ is the Gibbs distribution associated with the finite collection $(S_m)_{0\leqslant m\leqslant M}$ (compare to (6.17)), and $\pi^{(M)} \triangleq \big(\pi_m^{(M)}\big)_{0\leqslant m\leqslant M} \in \Delta\big(\{0,\dots,M\}\big)$ is the associated prior. The corresponding estimator is defined by

$$\widetilde{s}^{(\eta,M)} = \sum_{m=0}^{M} \widehat{\rho}_m^{(\eta,M)} \widehat{s}_m \,. \tag{6.56}$$

Now we apply the conclusions of Theorem 6.2 to the finite collection $(S_m)_{0\leqslant m\leqslant M}$. Setting $B_m \triangleq d^2(s,S_m) + \mathrm{pen}^{(\eta)}(m) - x_m/\eta$ for all $m \in \mathbb{N}$, we get from (6.21) and by definition of $\mathcal{J}\big(\widehat{\rho}^{(\eta,M)}\big)$ that for all $M \in \mathbb{N}$,

$$\begin{aligned}
&\mathbb{E}_s\left[\sum_{m=0}^{M} \widehat{\rho}_m^{(\eta,M)} \,\|\widehat{s}_m - s\|^2\right] \\
&\leqslant C_K\left(\inf_{\rho\in\Delta(\{0,\dots,M\})}\left\{\sum_{m=0}^{M} \rho_m B_m + \frac{\mathcal{K}\big(\rho,\pi^{(M)}\big)}{\eta}\right\} + \varepsilon^2\left(\ln_+\big(\Sigma^{(M)}\big) + 1\right) + \mu\right) \\
&\leqslant C_K\left(\inf_{\rho\in\Delta(\{0,\dots,M\})}\left\{\sum_{m=0}^{+\infty} \widetilde{\rho}_m B_m + \frac{\mathcal{K}\big(\widetilde{\rho},\pi\big)}{\eta}\right\} + \varepsilon^2(\ln_+(\Sigma) + 1) + \mu\right),
\end{aligned} \tag{6.57}$$

where we set $\Sigma^{(M)} \triangleq \sum_{m'=0}^{M} e^{-x_{m'}} \leqslant \Sigma$ and where for all $M \in \mathbb{N}$ and $\rho \in \Delta(\{0,\dots,M\})$, we defined $\widetilde{\rho} \in \Delta(\mathbb{N})$ by $\widetilde{\rho}_m = \rho_m$ if $0 \leqslant m \leqslant M$ and by $\widetilde{\rho}_m = 0$ if $m > M$. Inequality (6.57) follows by noting that $\pi_m^{(M)} \geqslant \pi_m$ so that

$$\mathcal{K}\big(\rho,\pi^{(M)}\big) = \sum_{m=0}^{M} \rho_m \ln\big(\rho_m/\pi_m^{(M)}\big) \leqslant \sum_{m=0}^{M} \rho_m \ln\big(\rho_m/\pi_m\big) = \mathcal{K}\big(\widetilde{\rho},\pi\big) \,.$$

Now, noting that almost surely $\sum_{m=0}^{M} \widehat{\rho}_m^{(\eta,M)} \,\|\widehat{s}_m - s\|^2 \to \sum_{m=0}^{+\infty} \widehat{\rho}_m^{(\eta)} \,\|\widehat{s}_m - s\|^2$ as $M \to +\infty$

(see Explanation 1 below), we get from (6.57) and from Fatou's lemma[15] that

$$\begin{aligned}
&\mathbb{E}_s\left[\sum_{m=0}^{+\infty} \widehat{\rho}_m^{(\eta)} \left\|\widehat{s}_m - s\right\|^2\right] \\
&\quad \leqslant C_K \liminf_{M\to+\infty} \inf_{\rho\in\Delta(\{0,\dots,M\})} \left\{\sum_{m=0}^{+\infty} \widetilde{\rho}_m B_m + \frac{\mathcal{K}(\widetilde{\rho},\pi)}{\eta}\right\} + C_K\left(\varepsilon^2\left(\ln_+(\Sigma)+1\right)+\mu\right) \\
&\quad = C_K \inf_{\substack{\rho\in\Delta(\mathbb{N}) \\ \mathrm{supp}(\rho)<\infty}} \left\{\sum_{m=0}^{+\infty} \rho_m B_m + \frac{\mathcal{K}(\rho,\pi)}{\eta}\right\} + C_K\left(\varepsilon^2\left(\ln_+(\Sigma)+1\right)+\mu\right) \qquad (6.58) \\
&\quad = C_K \inf_{\rho\in\Delta(\mathbb{N})} \left\{\sum_{m=0}^{+\infty} \rho_m B_m + \frac{\mathcal{K}(\rho,\pi)}{\eta}\right\} + C_K\left(\varepsilon^2\left(\ln_+(\Sigma)+1\right)+\mu\right)\ , \qquad (6.59)
\end{aligned}$$

where the infimum in (6.58) is taken over all $\rho \in \Delta(\mathbb{N})$ whose support $\mathrm{supp}(\rho) \triangleq \{m \in \mathbb{N} : \rho_m > 0\}$ is finite ((6.58) is straightforward), and where (6.59) follows from Explanation 2 below. Combining (6.59) with the definition of $\mathcal{J}\left(\widehat{\rho}^{(\eta)}\right)$ concludes the proof.

Explanation 1: We show below that, almost surely,

$$\sum_{m=0}^{M} \widehat{\rho}_m^{(\eta,M)} \left\|\widehat{s}_m - s\right\|^2 \to \sum_{m=0}^{+\infty} \widehat{\rho}_m^{(\eta)} \left\|\widehat{s}_m - s\right\|^2 \qquad \text{as } M \to +\infty.$$

First note that, by the same arguments[16] as in Step 2, $\sum_{m=0}^{+\infty} \widehat{\rho}_m^{(\eta)} \left\|\widehat{s}_m - s\right\|^2$ is almost surely finite. Moreover, by the triangle inequality, for all $M \in \mathbb{N}$,

$$\begin{aligned}
&\left|\sum_{m=0}^{M} \widehat{\rho}_m^{(\eta,M)} \left\|\widehat{s}_m - s\right\|^2 - \sum_{m=0}^{+\infty} \widehat{\rho}_m^{(\eta)} \left\|\widehat{s}_m - s\right\|^2\right| \\
&\qquad \leqslant \sum_{m=0}^{M} \left|\widehat{\rho}_m^{(\eta,M)} - \widehat{\rho}_m^{(\eta)}\right| \left\|\widehat{s}_m - s\right\|^2 + \sum_{m=M+1}^{+\infty} \widehat{\rho}_m^{(\eta)} \left\|\widehat{s}_m - s\right\|^2\ . \qquad (6.60)
\end{aligned}$$

The second sum $\sum_{m=M+1}^{+\infty} \widehat{\rho}_m^{(\eta)} \left\|\widehat{s}_m - s\right\|^2$ converges almost surely to 0 as $M \to +\infty$ since $\sum_{m=0}^{+\infty} \widehat{\rho}_m^{(\eta)} \left\|\widehat{s}_m - s\right\|^2$ is finite almost surely. As for the first sum, note by definitions of $\widehat{\rho}_m^{(\eta,M)}$ and $\widehat{\rho}_m^{(\eta)}$ that it can be rewritten as

$$\begin{aligned}
&\sum_{m=0}^{M} \left|\widehat{\rho}_m^{(\eta,M)} - \widehat{\rho}_m^{(\eta)}\right| \left\|\widehat{s}_m - s\right\|^2 \\
&\quad = \left(\frac{\sum_{m'=0}^{+\infty} \exp\Big(-\eta\big(\gamma_\varepsilon(\widehat{s}_{m'}) + \mathrm{pen}(m')\big)\Big)}{\sum_{m'=0}^{M} \exp\Big(-\eta\big(\gamma_\varepsilon(\widehat{s}_{m'}) + \mathrm{pen}(m')\big)\Big)} - 1\right) \sum_{m=0}^{M} \widehat{\rho}_m^{(\eta)} \left\|\widehat{s}_m - s\right\|^2
\end{aligned}$$

[15] Note that we only work in expectation here. However, it is of course also possible to follow similar arguments to derive a high-probability bound from the high-probability bound obtained with finite collections (Remark 6.2).

[16] We use $\sup_{t\in\mathbb{R}_+}\left\{e^{-At^2}t^2\right\} < \infty$ for all $A > 0$ instead of $\sup_{t\in\mathbb{R}_+}\left\{e^{-At^2}t\right\} < \infty$.

$$\leqslant \left(\frac{\sum_{m'=0}^{+\infty} \exp\Big(-\eta\big(\gamma_\varepsilon(\widehat{s}_{m'}) + \mathrm{pen}(m')\big)\Big)}{\sum_{m'=0}^{M} \exp\Big(-\eta\big(\gamma_\varepsilon(\widehat{s}_{m'}) + \mathrm{pen}(m')\big)\Big)} - 1 \right) \sum_{m=0}^{+\infty} \widehat{\rho}_m^{(\eta)} \, \|\widehat{s}_m - s\|^2 \ .$$

Since $\sum_{m=0}^{+\infty} \widehat{\rho}_m^{(\eta)} \, \|\widehat{s}_m - s\|^2$ is almost surely finite, the last upper bound converges almost surely to 0 as $M \to +\infty$.

Therefore, combining (6.60) with the above remarks, we get that, almost surely, $\sum_{m=0}^{M} \widehat{\rho}_m^{(\eta,M)} \, \|\widehat{s}_m - s\|^2 \to \sum_{m=0}^{+\infty} \widehat{\rho}_m^{(\eta)} \, \|\widehat{s}_m - s\|^2$ as $M \to +\infty$.

Explanation 2: We show that

$$\inf_{\substack{\rho \in \Delta(\mathbb{N}) \\ |\mathrm{supp}(\rho)| < \infty}} \left\{ \sum_{m=0}^{+\infty} \rho_m B_m + \frac{\mathcal{K}(\rho,\pi)}{\eta} \right\} = \inf_{\rho \in \Delta(\mathbb{N})} \left\{ \sum_{m=0}^{+\infty} \rho_m B_m + \frac{\mathcal{K}(\rho,\pi)}{\eta} \right\} . \tag{6.61}$$

Note that it suffices to show that, for all $\rho \in \Delta(\mathbb{N})$,

$$\inf_{\substack{\rho' \in \Delta(\mathbb{N}) \\ |\mathrm{supp}(\rho')| < \infty}} \left\{ \sum_{m=0}^{+\infty} \rho'_m B_m + \frac{\mathcal{K}(\rho',\pi)}{\eta} \right\} \leqslant \sum_{m=0}^{+\infty} \rho_m B_m + \frac{\mathcal{K}(\rho,\pi)}{\eta} \ . \tag{6.62}$$

Let $\rho \in \Delta(\mathbb{N})$ and $\delta > 0$. We can assume that $\mathcal{K}(\rho,\pi) < \infty$ (otherwise, the inequality obviously holds true). Then, since $\sum_{m=0}^{+\infty} \rho_m = 1$ and $\sum_{m=0}^{+\infty} \rho_m \ln(\rho_m/\pi_m) = \mathcal{K}(\rho,\pi) < \infty$, we can fix $M \in \mathbb{N}$ such that

$$S \triangleq \sum_{m=0}^{M} \rho_m \geqslant \frac{1}{1+\delta} \quad \text{and} \quad \sum_{m=M+1}^{+\infty} \rho_m \ln(\rho_m/\pi_m) \geqslant -\delta \ . \tag{6.63}$$

Define $\rho' \in \Delta(\mathbb{N})$ by $\rho'_m \triangleq \rho_m / S$ for all $m \in \{0,\dots,M\}$ and by $\rho'_m \triangleq 0$ for all $m \geqslant M+1$, so that $|\mathrm{supp}(\rho')| < \infty$. Moreover, we have

$$\begin{aligned} \sum_{m=0}^{+\infty} \rho'_m B_m + \frac{\mathcal{K}(\rho',\pi)}{\eta} &= \frac{1}{S} \sum_{m=0}^{M} \rho_m B_m + \frac{1}{\eta S} \sum_{m=0}^{M} \rho_m \ln\left(\frac{\rho_m}{\pi_m}\right) - \frac{\ln S}{\eta} \\ &\leqslant (1+\delta) \left(\sum_{m=0}^{M} \rho_m B_m + \frac{1}{\eta}\big(\mathcal{K}(\rho,\pi) + \delta\big) \right) + \frac{\ln(1+\delta)}{\eta} \ , \end{aligned}$$

where the last inequality follows from (6.63). Letting $\delta \to 0$, we get (6.62), which in turn yields (6.61). □

6.A.2 Proof of Corollary 6.1

Proof (of Corollary 6.1): We prove that $\widetilde{s}^{(\eta)} \underset{\eta \to \infty}{\longrightarrow} \widetilde{s}^{(\infty)}$ almost surely. The bound (6.25) will then follow directly from Fatou's lemma by letting $\eta \to \infty$ in (6.22) of Theorem 6.2.

First note that the random set $\widehat{\mathcal{M}} \triangleq \mathrm{argmin}_{m \in \mathcal{M}} \big\{ \gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}(m) \big\} \subset \mathcal{M}$ is almost surely non-empty and finite. This fact is proved in [Mas07, p. 130] under the assumption that $\mathrm{pen}(m) \geqslant$

$K\varepsilon^2\left(\sqrt{D_m}+\sqrt{2x_m}\right)^2$ for all $m\in\mathcal{M}$. Therefore, the probability distribution $\widehat{\rho}^{(\infty)}$ defined in (6.24) has almost surely a finite support $\widehat{\mathcal{M}}$, so that the estimator $\widetilde{s}^{(\infty)}\triangleq\sum_{m\in\mathcal{M}}\widehat{\rho}_m^{(\infty)}\widehat{s}_m$ is well-defined. Moreover, by the triangle inequality,

$$\begin{aligned}\left\|\widetilde{s}^{(\eta)}-\widetilde{s}^{(\infty)}\right\| &= \left\|\sum_{m\in\mathcal{M}}\left(\widehat{\rho}_m^{(\eta)}-\widehat{\rho}_m^{(\infty)}\right)\widehat{s}_m\right\| \leqslant \sum_{m\in\mathcal{M}}\left|\widehat{\rho}_m^{(\eta)}-\widehat{\rho}_m^{(\infty)}\right|\|\widehat{s}_m\| \\ &\leqslant \sum_{m\in\widehat{\mathcal{M}}}\left|\widehat{\rho}_m^{(\eta)}-\frac{e^{-x_m}}{\widehat{Z}}\right|\|\widehat{s}_m\| + \sum_{m\in\mathcal{M}\setminus\widehat{\mathcal{M}}}\widehat{\rho}_m^{(\eta)}\|\widehat{s}_m\|\,, \end{aligned} \tag{6.64}$$

where the last inequality follows by definition of $\widehat{\rho}_m^{(\infty)}$ in (6.24), and where $\widehat{Z}\triangleq\sum_{m\in\widehat{\mathcal{M}}}e^{-x_m}$.

We start by proving that the first sum above goes to 0 as $\eta\to+\infty$. First note from (6.17) and from the equality $\mathrm{pen}^{(\eta)}(m)=\mathrm{pen}(m)+x_m/\eta$ that $\widehat{\rho}_m^{(\eta)}$ can be rewritten as

$$\widehat{\rho}_m^{(\eta)}=\frac{\exp\left[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}(m)-B\Big)-x_m\right]}{\sum_{m'\in\mathcal{M}}\exp\left[-\eta\Big(\gamma_\varepsilon(\widehat{s}_{m'})+\mathrm{pen}(m')-B\Big)-x_{m'}\right]}\,, \tag{6.65}$$

where we set $B\triangleq\min_{m\in\mathcal{M}}\big\{\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}(m)\big\}$. By definition of B and $\widehat{\mathcal{M}}$, the quantity $\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}(m)-B$ is equal to zero if and only if $m\in\widehat{\mathcal{M}}$, and it is positive otherwise. Therefore, we get, almost surely, for all $m\in\mathcal{M}$,

$$\exp\left[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}(m)-B\Big)-x_m\right]\underset{\eta\to\infty}{\longrightarrow}\begin{cases}e^{-x_m} & \text{if } m\in\widehat{\mathcal{M}}\,,\\ 0 & \text{if } m\notin\widehat{\mathcal{M}}\,.\end{cases} \tag{6.66}$$

Since the exponential above is nonincreasing in η and is bounded by e^{-x_m} for all $m\in\mathcal{M}$, we get by Lebesgue's dominated convergence theorem that, almost surely,

$$\sum_{m\in\mathcal{M}}\exp\left[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}(m)-B\Big)-x_m\right]\downarrow\sum_{m\in\widehat{\mathcal{M}}}e^{-x_m}\triangleq\widehat{Z}\quad\text{as }\eta\to\infty\,. \tag{6.67}$$

Combining (6.65), (6.66), and (6.67), we get that, almost surely, $\widehat{\rho}_m^{(\eta)}\underset{\eta\to\infty}{\longrightarrow}e^{-x_m}/\widehat{Z}$ for all $m\in\mathcal{M}$. Since $\widehat{\mathcal{M}}$ is almost surely finite, we can conclude that $\sum_{m\in\widehat{\mathcal{M}}}\big|\widehat{\rho}_m^{(\eta)}-e^{-x_m}/\widehat{Z}\big|\,\|\widehat{s}_m\|\underset{\eta\to\infty}{\longrightarrow}0$ almost surely.

We now show that $\sum_{m\in\mathcal{M}\setminus\widehat{\mathcal{M}}}\widehat{\rho}_m^{(\eta)}\|\widehat{s}_m\|\underset{\eta\to\infty}{\longrightarrow}0$ almost surely. First note from (6.65) and from (6.67) that, almost surely,

$$0\leqslant\sum_{m\in\mathcal{M}\setminus\widehat{\mathcal{M}}}\widehat{\rho}_m^{(\eta)}\|\widehat{s}_m\|\leqslant\frac{1}{\widehat{Z}}\sum_{m\in\mathcal{M}\setminus\widehat{\mathcal{M}}}\exp\left[-\eta\Big(\gamma_\varepsilon(\widehat{s}_m)+\mathrm{pen}(m)-B\Big)-x_m\right]\|\widehat{s}_m\|\,.$$

The last sum above is convergent for all $\eta>0$ (see, e.g., Step 2 in the proof of Theorem 6.2). Since in addition the summand decreases to 0 as $\eta\to\infty$ (by (6.66)), we get by Lebesgue's dominated

convergence theorem that $\sum_{m\in\mathcal{M}\setminus\widehat{\mathcal{M}}} \widehat{\rho}_m^{(\eta)} \|\widehat{s}_m\| \underset{\eta\to\infty}{\longrightarrow} 0$ almost surely.

Putting everything together, we get from (6.64) that, almost surely, $\widetilde{s}^{(\eta)} \to \widetilde{s}^{(\infty)}$ as $\eta \to \infty$. □

6.B Useful lemmas

In this section we recall two results that prove to be useful throughout this chapter.

6.B.1 A classical concentration inequality for Gaussian processes

We recall below a well-known concentration inequality for the suprema of Gaussian processes that follows straightforwardly from [Led01, Theorem 7.1] by the almost-sure continuity assumption and by separability arguments (see also [Mas07, Proposition 3.19]).

Lemma 6.1. *Let $(X_t)_{t\in S}$ be a centered Gaussian process indexed by a separable topological space S such that $t \mapsto X_t$ is almost surely continuous on S. Then, for all $x \geqslant 0$,*

$$\mathbb{P}\left[\sup_{t\in S} X_t \leqslant \mathbb{E}\left[\sup_{t\in S} X_t\right] + x\right] \geqslant 1 - \exp\bigl(-x^2/(2\nu)\bigr) ,$$

where $\nu \triangleq \sup_{t\in S} \mathbb{E}\bigl[X_t^2\bigr]$, where $\mathbb{E}\bigl[\sup_{t\in S} X_t\bigr] \in (-\infty, +\infty]$ is always well-defined[17] *(since X_{t_0} is integrable for any $t_0 \in S$), and where we used the conventions $+\infty \leqslant +\infty$ and $(+\infty) + x = +\infty$ for all $x \in \mathbb{R}$.*

6.B.2 An upper bound on some fluctuations

Next we recall a result due to [Mas07]. We use it in Theorem 6.2 to show that with large probability, for all $m \in \mathcal{M}$, the penalty $\mathrm{pen}_2^{(\eta)}(m)$ is large enough to annihilate the fluctuations $2\varepsilon w_m(\widehat{s}_m)V_m$. When $\mathcal{M}$ is infinite, it is also useful to show that the normalizing constant $\sum_{m\in\mathcal{M}} \exp\bigl[-\eta\bigl(\gamma_\varepsilon(\widehat{s}_m) + \mathrm{pen}_2^{(\eta)}(m)\bigr)\bigr]$ and the sum $\sum_{m\in\mathcal{M}} \widehat{\rho}_m^{(\eta)} \|\widehat{s}_m\|$ are almost surely finite.

Lemma 6.2. *As in Section 6.2.2, we assume that for every $m \in \mathcal{M}$, there exists some almost-surely continuous version of the isonormal process W on the closure $\overline{S_m}$ of S_m and that (6.10) holds true for some nondecreasing continuous function $\varphi_m : [0, +\infty) \to \mathbb{R}_+$ such that $x \mapsto x^{-1}\varphi_m(x)$ is nonincreasing on $\mathbb{R}_+^*$. We let $\tau_m = 1$ if S_m is closed and convex and $\tau_m = 2$ otherwise and define the generalized dimension D_m of S_m as in (6.11).*

Let $z > 0$ and $K' > 1$, and set $y_m \triangleq K'\varepsilon\bigl(\sqrt{D_m} + \sqrt{2x_m} + (2\pi)^{-1/2} + \sqrt{2z}\bigr)$ for all $m \in \mathcal{M}$. Let $a \in \mathbb{H}$, define $w_m(t) \triangleq (1/2)\left(\bigl[\|s-a\| + \|s-t\|\bigr]^2 + y_m^2\right)$ for all $m \in \mathcal{M}$ and $t \in \mathbb{H}$, and set

$$V_m \triangleq \sup_{t\in S_m}\left(\frac{W(t) - W(a)}{w_m(t)}\right) , \quad m \in \mathcal{M} .$$

[17] As previously mentioned, if $\sup_{t\in S} X_t$ is not measurable, we consider $\sup_{t\in A} X_t$ instead, where A is any at most countable dense subset of S; the last two suprema are almost surely equal by the almost-sure continuity of $t \mapsto X_t$.

Then, on some event $\Omega_{z,K'}$ of probability $\mathbb{P}_s\big(\Omega_{z,K'}\big) \geqslant 1-\Sigma e^{-z}$, we have, for all $m \in \mathcal{M}$,

$$2\varepsilon w_m(\widehat{s}_m) V_m \leqslant K'^2\varepsilon^2\big(\sqrt{D_m}+\sqrt{2x_m}\big)^2 + \frac{2K'^2\varepsilon^2}{K'-1}\left(\frac{1}{2\pi}+2z\right) + \frac{1}{\sqrt{K'}}\left(\|s-\widehat{s}_m\|^2 + \frac{\|s-a\|^2}{\sqrt{K'}-1}\right) .$$

Proof: This upper bound is proved in [Mas07, Theorem 4.18] between Equations (4.78) and (4.82) therein[18]. We only recall its proof for the convenience of the reader.

We first make the following assumption. The general case will be addressed at the end of the proof.

Assumption 6.1. *S_m is closed for all $m \in \mathcal{M}$.*

Step 1: High-probability bound on εV_m.
Recall that we fixed $z > 0$. Next we apply a well-known concentration inequality for the suprema of Gaussian processes that can essentially be found, e.g., in [Led01, Theorem 7.1] or [Mas07, Proposition 3.19], and that is recalled in Lemma 6.1 above. By definition of V_m and since the centered Gaussian process $\big([W(t)-W(a)]/w_m(t)\big)_{t\in S_m}$ is almost surely continuous on S_m (by the almost-sure continuity of $t \mapsto W(t)$ on S_m), Lemma 6.1 ensures that, for all $m \in \mathcal{M}$,

$$\mathbb{P}\Big[V_m \leqslant \mathbb{E}[V_m] + \sqrt{2v_m(x_m+z)}\Big] \geqslant 1-e^{-x_m}e^{-z} , \tag{6.68}$$

where

$$v_m \triangleq \sup_{t\in S_m} \mathbb{E}\left[\left(\frac{W(t)-W(a)}{w_m(t)}\right)^2\right] = \sup_{t\in S_m}\frac{\|t-a\|^2}{w_m^2(t)} .$$

But we have $w_m(t) \geqslant \|t-a\|\, y_m$ by definition of $w_m(t) \triangleq (1/2)\Big(\big[\|s-a\|+\|s-t\|\big]^2 + y_m^2\Big)$, by the triangle inequality, and by the the fact that $2ab \leqslant a^2+b^2$ for all $a, b \in \mathbb{R}$. Therefore, $v_m \leqslant y_m^{-2}$ for all $m \in \mathcal{M}$. Substituting the latter inequality in (6.68) and using a union-bound over $\mathcal{M}$, we get, on some event $\Omega_{z,K'}$ of probability $\mathbb{P}_s\big(\Omega_{z,K'}\big) \geqslant 1-\Sigma e^{-z}$,

$$\forall m \in \mathcal{M} , \quad V_m \leqslant \mathbb{E}[V_m] + y_m^{-1}\sqrt{2(x_m+z)} . \tag{6.69}$$

The rest of this step is dedicated to upper bounding the expectation $\mathbb{E}[V_m]$. First note that by definition of V_m,

$$\mathbb{E}[V_m] \leqslant \mathbb{E}\left[\sup_{t\in S_m}\left(\frac{W(t)-W(s_m)}{w_m(t)}\right)\right] + \mathbb{E}\left[\frac{\big(W(s_m)-W(a)\big)_+}{\inf_{t\in S_m} w_m(t)}\right] . \tag{6.70}$$

We upper bound each term of the right-hand side separately. Let $\delta > 0$. Recall that, by definition, $\tau_m = 1$ if S_m is closed and convex, and $\tau_m = 2$ otherwise. In all cases, by Assumption 6.1, we can fix for all $m \in \mathcal{M}$ a point $s_m \in S_m$ such that the two following conditions are satisfied:

$$\|s-s_m\| \leqslant (1+\delta)\, d(s, S_m) , \tag{6.71}$$

[18]The slight improvement with respect to [Mas07, Theorem 4.18] (with however, the exact same proof) is that we let a be arbitrary.

$$\|s_m - t\| \leqslant \tau_m(1+\delta)\,\|s-t\| \;, \quad \text{for all } t \in S_m \;. \tag{6.72}$$

Details:
Indeed, if S_m is closed and convex, then we can take s_m as the projection of s onto S_m, so that $\|s - s_m\| = d(s, S_m)$. Moreover, since the projection is a contraction, we have $\|s_m - t\| \leqslant \|s - t\|$ for all $t \in S_m$, which yields (6.72) with $\tau_m = 1$.

But if the S_m are arbitrary, then there always exists $s_m \in S_m$ such that (6.71) holds true (by Assumption 6.1 if $d(s, S_m) = 0$, obvious if $d(s, S_m) > 0$). The property (6.72) then follows by noting that, for all $t \in S_m$, $\|s_m - t\| \leqslant \|s_m - s\| + \|s - t\| \leqslant 2(1+\delta)\,\|s - t\|$.

By definition of $w_m(t)$ and by (6.72), we have $2w_m(t) \geqslant \tau_m^{-2}(1+\delta)^{-2}\,\|t - s_m\|^2 + y_m^2$. Using the assumption (6.10) with $u = s_m$, we get

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in S_m}\left(\frac{W(t)-W(s_m)}{w_m(t)}\right)\right] &\leqslant y_m^{-2}\varphi_m\big(\tau_m(1+\delta)y_m\big) \\ &\leqslant y_m^{-1}\left(\varepsilon\sqrt{D_m}\right)^{-1}\varphi_m\left(\tau_m(1+\delta)\varepsilon\sqrt{D_m}\right) \;, \end{aligned}\tag{6.73}$$

where the last inequality follows from the lower bound $y_m \geqslant \varepsilon\sqrt{D_m}$ and from the fact that $x \mapsto x^{-1}\varphi_m\big(\tau_m(1+\delta)x\big)$ in nonincreasing on $\mathbb{R}_+^*$ (since it is the case for $x \mapsto x^{-1}\varphi_m(x)$ by assumption).
As for the second term in (6.70), we can see from $w_m(t) \triangleq (1/2)\left(\big[\|s-a\| + \|s-t\|\big]^2 + y_m^2\right)$, from $\|s-t\| \geqslant (1+\delta)^{-1}\,\|s - s_m\|$ for all $t \in S_m$ (by (6.71)), and from the triangle inequality that

$$\inf_{t\in S_m} w_m(t) \geqslant (1/2)\left((1+\delta)^{-2}\,\|a - s_m\|^2 + y_m^2\right) \geqslant (1+\delta)^{-1}\,\|s_m - a\|\,y_m \;,$$

where the last inequality follows from the fact that $2ab \leqslant a^2 + b^2$ for all $a, b \in \mathbb{R}$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\frac{\big(W(s_m) - W(a)\big)_+}{\inf_{t\in S_m} w_m(t)}\right] &\leqslant (1+\delta)\,y_m^{-1}\,\mathbb{E}\left[\left(\frac{W(s_m)-W(a)}{\|s_m - a\|}\right)_+\right] \\ &= (1+\delta)\,y_m^{-1}(2\pi)^{-1/2} \;,\end{aligned}$$

where, to get the last equality, we used the fact that $\big(W(s_m) - W(a)\big)/\,\|s_m - a\|$ is a standard normal random variable.
Substituting the last upper bound and (6.73) in (6.70), we get

$$\mathbb{E}[V_m] \leqslant y_m^{-1}\left(\varepsilon\sqrt{D_m}\right)^{-1}\varphi_m\left(\tau_m(1+\delta)\varepsilon\sqrt{D_m}\right) + (1+\delta)y_m^{-1}(2\pi)^{-1/2} \;.$$

Letting $\delta \to 0$, we get by continuity of φ_m on $\mathbb{R}_+$ that

$$\begin{aligned}\mathbb{E}[V_m] &\leqslant y_m^{-1}\left(\varepsilon\sqrt{D_m}\right)^{-1}\varphi_m\left(\tau_m\varepsilon\sqrt{D_m}\right) + y_m^{-1}(2\pi)^{-1/2} \\ &\leqslant y_m^{-1}\left(\sqrt{D_m} + (2\pi)^{-1/2}\right) \;,\end{aligned}\tag{6.74}$$

where the last inequality follows from the fact that $\varphi_m\big(\tau_m\varepsilon\sqrt{D_m}\big) = \varepsilon D_m$ by definition of D_m

(see (6.11)). Substituting the last upper bound in (6.69), we get, on the event $\Omega_{z,K'}$,

$$\forall m \in \mathcal{M}\,, \quad \varepsilon V_m \leqslant \varepsilon y_m^{-1} \left(\sqrt{D_m} + (2\pi)^{-1/2} + \sqrt{2x_m + 2z} \right) \leqslant K'^{-1}\,, \tag{6.75}$$

where the last inequality follows by definition of $y_m \triangleq K'\varepsilon\big(\sqrt{D_m} + \sqrt{2x_m} + (2\pi)^{-1/2} + \sqrt{2z}\big)$.

Step 2: Upper bound on $2w_m(\widehat{s}_m)$.
Let $m \in \mathcal{M}$. Next we bound from above the quantity $2w_m(\widehat{s}_m) \triangleq \big(\|s-a\| + \|s-\widehat{s}_m\|\big)^2 + y_m^2$. Using repeatedly the elementary inequality

$$(a+b)^2 \leqslant (1+\theta)a^2 + (1+\theta^{-1})b^2\,, \quad a,b \in \mathbb{R}\,,$$

for various values of $\theta > 0$, we get, on the one hand (with $\theta = \sqrt{K'} - 1$),

$$\big(\|s-\widehat{s}_m\| + \|s-a\|\big)^2 \leqslant \sqrt{K'}\left(\|s-\widehat{s}_m\|^2 + \frac{\|s-a\|^2}{\sqrt{K'}-1} \right)\,,$$

and, on the other hand (first with $\theta = K'-1$, and then with $\theta = 1$),

$$\begin{aligned} y_m^2 &\triangleq K'^2\varepsilon^2 \left(\sqrt{D_m} + \sqrt{2x_m} + (2\pi)^{-1/2} + \sqrt{2z} \right)^2 \\ &\leqslant K'^2\varepsilon^2 \left[K' \left(\sqrt{D_m} + \sqrt{2x_m} \right)^2 + \frac{2K'}{K'-1} \left(\frac{1}{2\pi} + 2z \right) \right]\,. \end{aligned}$$

Combining the two inequalities above, we get

$$\begin{aligned} 2w_m(\widehat{s}_m) \leqslant \sqrt{K'} &\left(\|s-\widehat{s}_m\|^2 + \frac{\|s-a\|^2}{\sqrt{K'}-1} \right) + K'^3\varepsilon^2 \left(\sqrt{D_m} + \sqrt{2x_m} \right)^2 \\ &+ \frac{2K'^3\varepsilon^2}{K'-1} \left(\frac{1}{2\pi} + 2z \right)\,. \end{aligned} \tag{6.76}$$

Step 3: Putting everything together.
Combining (6.75) and (6.76), we get

$$\begin{aligned} 2\varepsilon w_m(\widehat{s}_m) V_m \leqslant \frac{1}{\sqrt{K'}} &\left(\|s-\widehat{s}_m\|^2 + \frac{\|s-a\|^2}{\sqrt{K'}-1} \right) \\ &+ K'^2\varepsilon^2 \big(\sqrt{D_m} + \sqrt{2x_m}\big)^2 + \frac{2K'^2\varepsilon^2}{K'-1} \left(\frac{1}{2\pi} + 2z \right)\,, \end{aligned}$$

which conludes the proof under Assumption 6.1.

General case: We no longer assume that S_m is closed for all $m \in \mathcal{M}$.
We employ a reduction to the case studied above. Namely, we use the above analysis to the modified collection of models $\big(\overline{S_m}\big)_{m\in\mathcal{M}}$.
By contruction, the collection $\big(\overline{S_m}\big)_{m\in\mathcal{M}}$ satisfies Assumption 6.1. Let us check that it also satisfies the assumptions of Lemma 6.2. First, the existence of an almost-sure continuous version of W on $\overline{S_m} = \overline{S_m}$ is straightforward. Moreover, (6.10) holds true for $\overline{S_m}$ since it holds on the restricted

set S_m (by assumption) and since W has an almost-sure continuous version on $\overline{S_m}$.

Second, setting $\tau'_m = 1$ if $\overline{S_m}$ is convex and $\tau'_m = 2$ otherwise, we note that $\tau'_m \leqslant \tau_m$. Therefore, the generalized dimension D'_m of $\overline{S_m}$ defined as in (6.11) satisfies $D'_m \leqslant D_m$. (The last inequality follows from the lower bound $1/({\tau'}_m^2 \varepsilon) \geqslant 1/(\tau_m^2 \varepsilon)$ and from the same arguments as those used after (6.11), e.g., from the fact that $x \mapsto x^{-2}\varphi_m(x)$ is nonincreasing on $\mathbb{R}_+^*$.).

We can thus apply the conclusion of Step 3 to the collection $\left(\overline{S_m}\right)_{m\in\mathcal{M}}$. To conclude the proof, it suffices to use the fact that $D'_m \leqslant D_m$ for all $m \in \mathcal{M}$ and to note that

$$V_m \leqslant V'_m \triangleq \sup_{t\in\overline{S_m}} \left(\frac{W(t) - W(a)}{w_m(t)}\right) .$$

□

Appendix A

Statistical background

Contents

In the sequel we use the following notation. For all probability distributions ρ, π on a given measurable space $(E, \mathcal{B})$, the Kullback-Leibler divergence $\mathcal{K}(\rho, \pi)$ between ρ and π is defined by

$$\mathcal{K}(\rho,\pi) \triangleq \begin{cases} \displaystyle\int_E \ln\left(\frac{\mathrm{d}\rho}{\mathrm{d}\pi}\right)\mathrm{d}\rho & \text{if } \rho \text{ is absolutely continuous with respect to } \pi; \\ +\infty & \text{otherwise.} \end{cases}$$

A.1 A duality formula for the Kullback-Leibler divergence

We recall below a key duality formula satisfied by the Kullback-Leibler divergence and whose proof can be found, e.g., in [Cat04, pp. 159–160] (see also [DZ98, p. 264]).

Proposition A.1. *For any measurable space $(E, \mathcal{B})$, any probability distribution π on $(E, \mathcal{B})$, and any measurable function $h : E \to [a, +\infty)$ bounded from below (by some $a \in \mathbb{R}$), we have*

$$-\ln \int_E e^{-h}\mathrm{d}\pi = \inf_{\rho \in \mathcal{M}_1^+(E)} \left\{ \int_E h \,\mathrm{d}\rho \,+\, \mathcal{K}(\rho,\pi) \right\} ,$$

where $\mathcal{M}_1^+(E)$ denotes the set of all probability distributions on $(E, \mathcal{B})$, and where the expectations $\int_E h \,\mathrm{d}\rho \in [a, +\infty]$ are always well defined since h is bounded from below.

Moreover, the above infimum is achieved at $\rho = \pi_{-h}^{\exp}$, where $\pi_{-h}^{\exp} \in \mathcal{M}_1^+(E)$ is absolutely continuous with respect to π and is given by

$$\mathrm{d}\pi_{-h}^{\exp} \triangleq \frac{e^{-h}}{\int_E e^{-h}\mathrm{d}\pi}\,\mathrm{d}\pi \,.$$

The above duality formula can be equivalently reformulated as follows (just apply it with $-h$). For any measurable function $h : E \to (-\infty, a]$ bounded from above (by some $a \in \mathbb{R}$),

$$\ln \int_E e^h \mathrm{d}\pi = \sup_{\rho \in \mathcal{M}_1^+(E)} \left\{ \int_E h \, \mathrm{d}\rho \, - \, \mathcal{K}(\rho, \pi) \right\} .$$

This more classical statement indicates that the log-moment generating function can be thought of as the Legendre transform of the Kullback-Leibler divergence.

A.2 Exp-concavity of the square loss

Next we recall the notion of exp-concavity and the elementary fact that the square loss is $1/(8B^2)$-exp-concave on $[-B, B]$. See, e.g., [KW99] or [CBL06, Chapter 3] for a reference on exp-concave losses.

Definition A.1 (Exp-concavity)**.**
Let $\mathcal{D}$ be a convex subset of a real vector space. A function $h : \mathcal{D} \to \mathbb{R}$ is said to be exp-concave *for a given $\eta > 0$ (or simply η-*exp-concave*) if the function $H_\eta \triangleq e^{-\eta h}$ is concave on $\mathcal{D}$.*

Noting that $H_{\eta'} = H_\eta^{\eta'/\eta}$ and $h = -\frac{1}{\eta} \ln H_\eta$, we can see that if $h : \mathcal{D} \to \mathbb{R}$ is η-exp-concave, then

- h is η'-exp-concave for all $0 < \eta' \leqslant \eta$ (since $x \mapsto x^{\eta'/\eta}$ is concave and non-decreasing);
- h is convex (since $x \mapsto -\frac{1}{\eta} \ln x$ is convex and non-increasing).

The next elementary result can be found, e.g., in [KW99, Proof of Theorem 2] or in [Vov01, Remark 3].

Proposition A.2. *The square loss is $1/(8B^2)$-exp-concave on $[-B, B]$ in the sense that, for all $y \in [-B, B]$, the function $x \in [-B, B] \mapsto (y - x)^2$ is $1/(8B^2)$-exp-concave. (Moreover, the constant $1/(8B^2)$ is not improvable.)*

A.3 A version of von Neumann's minimax theorem

We recall below a version of von Neumann's minimax theorem due to [Kne52] and [Fan53]. The next statement is a straightforward consequence of [Fan53, Theorem 2] (see also [Sio58, Theorem 4.2]). Our assumptions are slightly stronger (concave/convex instead of concave-like/convex-like, and continuous instead of upper semi-continuous), but they are sufficient for our purposes.

Lemma A.1 (A version of von Neumann's minimax theorem)**.**
Let $\mathcal{X}$ and $\mathcal{Y}$ be convex subsets of vector spaces and let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a function such that $f(\,\cdot\,, y)$ is concave for all $y \in \mathcal{Y}$, and $f(x, \,\cdot\,)$ is convex for all $x \in \mathcal{X}$. Assume also that $\mathcal{X}$ is endowed with a topology that makes it Haussdorff and compact, and that $f(\,\cdot\,, y)$ is continuous for all $y \in \mathcal{Y}$. Then,

$$\sup_{x \in \mathcal{X}} \inf_{y \in \mathcal{Y}} f(x, y) = \inf_{y \in \mathcal{Y}} \sup_{x \in \mathcal{X}} f(x, y) \, .$$

A.4 An elementary lemma to solve for the cumulative loss

The next elementary lemma is due to [CBLS05, Appendix III]. It is useful to compute an upper bound on the cumulative loss $\widehat{L}_T$ of a forecaster when $\widehat{L}_T$ satisfies an inequality of the form (A.1).

Lemma A.2. *Let $a, b \geqslant 0$. Assume that $x \geqslant 0$ satisfies the inequality*

$$x \leqslant a + b\sqrt{x}\,. \tag{A.1}$$

Then,

$$x \leqslant a + b\sqrt{a} + b^2\,.$$

A.5 Some concentration inequalities and a maximal inequality

The next maximal inequality was proved by [Mas07, Lemma 2.3] through an argument of [Pis83] to control the expected supremum of random variables that belong to a given Orlicz space.

Lemma A.3 (A maximal inequality)**.** *Let Z_t, $1 \leqslant t \leqslant T$, be centered real random variables for which there exists $v \in \mathbb{R}_+$ such that $\mathbb{E}\big[e^{\lambda Z_t}\big] \leqslant e^{\lambda^2 v/2}$ for all $t \in \{1, \dots, T\}$ and all $\lambda > 0$ (we say that the Z_t are subgaussian with common variance factor v). Then,*

$$\mathbb{E}\left[\max_{1 \leqslant t \leqslant T} Z_t\right] \leqslant \sqrt{2\, v\, \ln T}\,.$$

The next two lemmas are due to [Hoe63]. The first lemma is stated for a single random variable. The second lemma is a direct extension of the first one by independence of the random variables Z_t, $t = 1, \dots, T$.

Lemma A.4 (Hoeffding's lemma)**.** *Let Z be a real random variable such that $a \leqslant Z \leqslant b$ almost surely, where $a < b \in \mathbb{R}$ are deterministic constants. Then $Z - \mathbb{E}[Z]$ is subgaussian with variance factor $(b-a)^2/4$, i.e.,*

$$\forall \lambda \in \mathbb{R}, \qquad \mathbb{E}\left[\exp\Big(\lambda\big(Z - \mathbb{E}[Z]\big)\Big)\right] \leqslant \exp\left(\frac{\lambda^2}{8}(b-a)^2\right)\,.$$

The above bound can also be rewritten as $\ln\Big(\mathbb{E}\big[e^{\lambda Z}\big]\Big) \leqslant \lambda \mathbb{E}[Z] + \lambda^2 (b-a)^2/8$.

Lemma A.5 (Hoeffding's inequality)**.** *Let Z_t, $1 \leqslant t \leqslant T$, be independent real random variables such that $Z_t \in [a_t, b_t]$ a.s. for all $t \in \{1, \dots, T\}$, where $a_t, b_t \in \mathbb{R}$ are some constants. Then the sum $\sum_{t=1}^T \big(Z_t - \mathbb{E}[Z_t]\big)$ is subgaussian with variance factor $\sum_{t=1}^T (b_t - a_t)^2/4$, i.e.*

$$\forall \lambda \in \mathbb{R}, \qquad \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^T \big(Z_t - \mathbb{E}[Z_t]\big)\right)\right] \leqslant \exp\left(\frac{\lambda^2}{8}\sum_{t=1}^T (b_t - a_t)^2\right)\,.$$

As a consequence, for all $\delta \in (0,1)$,

$$\mathbb{P}\left[\sum_{t=1}^T \big(Z_t - \mathbb{E}[Z_t]\big) > \sqrt{\frac{1}{2}\sum_{t=1}^T (b_t - a_t)^2 \ln\left(\frac{1}{\delta}\right)}\right] \leqslant \delta$$

and

$$\mathbb{P}\left[\left|\sum_{t=1}^{T}\big(Z_t - \mathbb{E}[Z_t]\big)\right| > \sqrt{\frac{1}{2}\sum_{t=1}^{T}(b_t - a_t)^2 \ln\left(\frac{2}{\delta}\right)}\right] \leqslant \delta\,.$$

The next lemma is an extension of Hoeffding's inequality to martingales with zero-mean and bounded increments. It is due to [Hoe63] and [Azu67].

We first need the following definition. A sequence of random variables $(X_t)_{t\in\mathbb{N}^*}$ on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is a *martingale difference sequence* with respect to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$ if and only if, for all $t \geqslant 1$, X_t is $\mathcal{F}_t$-measurable, integrable, and satisfies, almost surely,

$$\mathbb{E}\big[X_t \,\big|\, \mathcal{F}_{t-1}\big] = 0\,.$$

Lemma A.6 (The Hoeffding-Azuma inequality)**.** *Let $(X_t)_{t\in\mathbb{N}^*}$ be a martingale difference sequence with respect to a filtration $(\mathcal{F}_t)_{t\in\mathbb{N}}$. Assume that for all $t \geqslant 1$, there exists a $\mathcal{F}_{t-1}$-measurable random variable A_t and a nonnegative constant c_t such that $X_t \in [A_t, A_t + c_t]$ almost surely. Then, the martingale $(S_t)_{t\geqslant 1}$ defined by $S_t \triangleq \sum_{s=1}^{t} X_s$ satisfies, for all $t \geqslant 1$,*

$$\forall \lambda \in \mathbb{R}\,, \qquad \mathbb{E}\Big[e^{\lambda S_t}\Big] \leqslant \exp\left(\frac{\lambda^2}{8}\sum_{s=1}^{t} c_s^2\right)\,.$$

In other words, S_t is subgaussian with variance factor $\big(\sum_{s=1}^{t} c_s^2\big)/4$. As a consequence,

$$\forall x > 0\,, \qquad \mathbb{P}\left[\max_{1\leqslant t'\leqslant t} S_{t'} > x\right] \leqslant \exp\left(\frac{-2x^2}{\sum_{s=1}^{t} c_s^2}\right)\,.$$

A.6 Integration of high-probability bounds

The next elementary lemma is useful to derive bounds in expectation from bounds in high probability. We then specialize it to two examples that are used throughout the manuscript.

Lemma A.7 (Integration of high-probability bounds)**.**
Let Z be a real random variable such that, for some constants $a, \Sigma > 0$ and $b \in \mathbb{R}$, and for some increasing and continuous function $f : \mathbb{R}_+ \to \mathbb{R}_+$, whose inverse we denote by f^{-1},

$$\forall z > 0\,, \qquad \mathbb{P}\big(Z \leqslant a f(z) + b\big) \geqslant 1 - \Sigma e^{-z}\,. \tag{A.2}$$

Assume that $x \mapsto \exp\big(-f^{-1}(x)\big)$ is integrable on $\big(f(0), \lim_\infty f\big)$, where $\lim_\infty f \triangleq \lim_{u\to+\infty} \uparrow f(u)$. Then, $\mathbb{E}[Z] \in [-\infty, +\infty)$ is well-defined and

$$\mathbb{E}[Z] \leqslant a\left(f\big(\ln_+(\Sigma)\big) + \Sigma\int_{f(\ln_+(\Sigma))}^{\lim_\infty f} \exp\big(-f^{-1}(x)\big)\,\mathrm{d}x\right) + b\,, \tag{A.3}$$

where $\ln_+(\Sigma) \triangleq \max\big\{\ln(\Sigma), 0\big\}$.

Example A.1. *Let $a, \Sigma > 0$ and $b \in \mathbb{R}$. Let Z be a real random variable such that $Z \leqslant az + b$ with probability at least $1 - \Sigma e^{-z}$ for all $z > 0$. This example corresponds to $f(z) = z$ with the notations above. Therefore we get that $\mathbb{E}[Z] \leqslant a\big(\ln_+(\Sigma) + 1\big) + b$.*

Example A.2. *Let $a, c > 0$ and $b \in \mathbb{R}$. Let Z be a real random variable such that, for all $\delta \in (0,1)$, we have $Z \leqslant a \exp\big(c\sqrt{\ln(1/\delta)}\big) + b$ with probability at least $1 - \delta$. This example corresponds to $f(z) = a\exp\big(c\sqrt{z}\big) + b$ and $\Sigma = 1$ with the notations above. Therefore, we get from (A.3) and from elementary manipulations that $\mathbb{E}[Z] \leqslant a\big(\exp(2c^2) + 1\big) + b$.*

Proof (of Lemma A.7): First note from the intermediate value theorem that since f is increasing and continuous, it is a one-to-one mapping from $\mathbb{R}_+$ to $\big[f(0), \limsup_\infty f\big)$. Moreover, setting $x_+ \triangleq \max\{x, 0\}$ for all $x \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}\bigg[\bigg(\frac{Z-b}{a}\bigg)_+\bigg] &= \int_0^{+\infty} \mathbb{P}\bigg[\bigg(\frac{Z-b}{a}\bigg)_+ > x\bigg]\,\mathrm{d}x \\
&= \int_0^{+\infty} \mathbb{P}\big(Z > ax + b\big)\mathrm{d}x \qquad \text{(A.4)} \\
&\leqslant f\big(\ln_+(\Sigma)\big) + \int_{f(\ln_+(\Sigma))}^{\limsup_\infty f} \Sigma \exp\big(-f^{-1}(x)\big)\,\mathrm{d}x\,, \qquad \text{(A.5)}
\end{aligned}$$

where (A.4) follows from the fact that $\big\{((Z-b)/a)_+ > x\big\} = \big\{(Z-b)/a > x\big\}$ for all $x > 0$ and where we proceeded as follows to get (A.5). We split the integral into three terms and upper bounded $\mathbb{P}\big(Z > ax + b\big)$ separately. We first used the crude[1] upper bound $\mathbb{P}\big(Z > ax + b\big) \leqslant 1$ for all $0 \leqslant x \leqslant f\big(\ln_+(\Sigma)\big)$. Second, for $x \in \big(f(\ln_+(\Sigma)), \limsup_\infty f\big)$ we used the fact that $\mathbb{P}\big(Z > ax + b\big) = \mathbb{P}\big(Z > af\big(f^{-1}(x)\big) + b\big) \leqslant \Sigma \exp\big(-f^{-1}(x)\big)$ by assumption (A.2). Finally, by (A.2) again, and by the fact that f is increasing, we get that, for all $x > \limsup_\infty f = \sup_{u \geqslant 0} f(u)$,

$$\mathbb{P}\big(Z > ax + b\big) \leqslant \inf_{z>0} \mathbb{P}\big(Z > af(z) + b\big) \leqslant \inf_{z>0}\big\{\Sigma e^{-z}\big\} = 0\,.$$

Since $x \mapsto \exp\big(-f^{-1}(x)\big)$ is integrable on $\big(f(0), \limsup_\infty f\big)$, then $\mathbb{E}\big[Z_+\big] \leqslant \mathbb{E}\big[(Z-b)_+\big] + b_+ < +\infty$ by (A.5), so that $\mathbb{E}\big[Z\big] \in [-\infty, +\infty)$ is well-defined. Using the fact that $(Z-b)/a \leqslant ((Z-b)/a)_+$ and rearranging the terms of (A.5) concludes the proof. □

A.7 Some information-theoretic tools

The next inequality is due to Pinsker [Pin64] (and to [CK81] for the optimal constant $1/\sqrt{2}$).

Lemma A.8 (Pinsker's inequality)**.**
Let P and Q be two probability distributions on a given measurable space $(E, \mathcal{B})$. Then,

$$\|P - Q\|_{TV} \leqslant \sqrt{\frac{\mathcal{K}(P,Q)}{2}}\,,$$

where $\|P - Q\|_{TV} \triangleq \sup_{B \in \mathcal{B}} \big|P(B) - Q(B)\big|$ is the total variation distance between P and Q.

[1] Note that if $\ln_+(\Sigma) > 0$, then this crude upper bound is smaller than $\Sigma \exp\big(-f^{-1}(x)\big) > 1$ for all $x < f\big(\ln_+(\Sigma)\big)$ since f^{-1} is increasing.

The next lemma is a version of Fano's lemma due to [Bir05] (see also [Mas07, Corollary 2.18] for the statement given below). We denote by $\mathbb{N}^* = \{1, 2, \ldots\}$ the set of positive integers.

Lemma A.9 (Fano's lemma — Birgé's version)**.**
Let $(E, \mathcal{B})$ be a measurable space and $N \in \mathbb{N}^*$. Let $(A_0, \ldots, A_N)$ be a measurable partition of $(E, \mathcal{B})$ and $(\mathbb{P}_0, \ldots, \mathbb{P}_N)$ be a family of probability distributions on $(E, \mathcal{B})$. Then we have

$$\min_{0 \leqslant i \leqslant N} \mathbb{P}_i(A_i) \leqslant \max\left\{\kappa, \frac{\bar{\mathcal{K}}}{\ln(N+1)}\right\},$$

where $\kappa > 0$ is an absolute constant such that $\kappa \leqslant 2e/(2e+1)$ and where $\bar{\mathcal{K}} \triangleq \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \mathcal{K}(\mathbb{P}_i, \mathbb{P}_0)$.

The next lemma is an extension of Lemma A.9 to convex combinations of probability distributions. This extension was proved (with different constants) in [CBLS05, Lemma 18] through a simple adaptation of the arguments of [Bir05]. Another way to prove it is to use Lemma A.9 on the augmented space $\Omega \times \{1, \ldots, S\}$ and to rewrite the resulting bound via the law of total probability and the chain rule for the Kullback-Leibler divergence.

Lemma A.10 (Fano's lemma for convex combinations)**.**
Let $(E, \mathcal{B})$ be a measurable space and $N, S \in \mathbb{N}^*$. Let $\big\{(A_{s,0}, \ldots, A_{s,N}) : s = 1, \ldots, S\big\}$ be a family of measurable partitions of $(E, \mathcal{B})$ and $\big\{\mathbb{P}_{s,j} : s = 1, \ldots, S, j = 1, \ldots, N\big\}$ be a family of probability distributions on $(E, \mathcal{B})$. Let $\alpha_1, \ldots, \alpha_S \in \mathbb{R}_+$ be such that $\sum_{s=1}^{S} \alpha_s = 1$. Then,

$$\min_{0 \leqslant i \leqslant N} \sum_{s=1}^{S} \alpha_s \mathbb{P}_{s,j}\big[A_{s,j}\big] \leqslant \max\left\{\kappa, \frac{\bar{\mathcal{K}}}{\ln(N+1)}\right\},$$

where $\kappa > 0$ is an absolute constant such that $\kappa \leqslant 2e/(2e+1)$ and where

$$\bar{\mathcal{K}} \triangleq \frac{1}{N}\sum_{i=1}^{N}\sum_{s=1}^{S} \alpha_s \mathcal{K}(\mathbb{P}_{s,i}, \mathbb{P}_{s,0}).$$

Bibliography

[AABR09] J. Abernethy, A. Agarwal, P. Bartlett, and A. Rakhlin. A stochastic view of optimal regret through minimax duality. In *Proceedings of the 22th Annual Conference on Learning Theory (COLT'09)*, 2009.

[AB09] J.-Y. Audibert and S. Bubeck. Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT'09)*, 2009.

[AB10] J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.*, 11:2785–2836, 2010.

[ABDJ06] F. Abramovich, Y. Benjamini, D.L. Donoho, and I.M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653, 2006.

[AC11] J.-Y. Audibert and O. Catoni. Robust linear least squares regression. *Ann. Statist.*, 2011. In press. Available at http://arxiv.org/abs/1010.0074.

[ACBF02] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47:235–256, 2002.

[ACBFS02] P. Auer, N. Cesa-Bianchi, Y. Freund, and R.E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

[ACBG02] P. Auer, N. Cesa-Bianchi, and C. Gentile. Adaptive and self-confident on-line learning algorithms. *J. Comp. Sys. Sci.*, 64:48–75, 2002.

[AG10] F. Abramovich and V. Grinshtein. MAP model selection in Gaussian regression. *Electron. J. Statist.*, 4:932–949, 2010.

[AGS11] P. Alquier, E. Gautier, and G. Stoltz, editors. *Inverse Problems and High-Dimensional Estimation, Stats in the Château Summer School, August 31 – September 4, 2009*, volume 203 of *Lecture Notes in Statistics*. Springer, 2011.

[Aka71] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281, 1971.

[AL11] P. Alquier and K. Lounici. PAC-Bayesian bounds for sparse regression estimation with exponential weights. *Electron. J. Stat.*, 5:127–145, 2011.

[Alq08] P. Alquier. PAC-Bayesian bounds for randomized empirical risk minimizers. *Math. Methods Statist.*, 17(4):279–304, 2008.

[AM09] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10(Feb):245–279, 2009.

[ANN04] C. Allenberg-Neeman and B. Neeman. Full information game with gains and losses. In *Proceedings of the 15th International Conference on Algorithmic Learning Theory (ALT'04)*, pages 264–278, 2004.

[Aud04a] J. Y. Audibert. *PAC-Bayesian Statistical Learning Theory*. PhD thesis, Université Paris VI, 2004.

[Aud04b] J.Y. Audibert. Aggregated estimators and empirical complexity for least square regression. *Ann. Inst. Henri Poincaré Probab. Stat.*, 40(6):685–736, 2004.

[Aud06] J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. Technical Report 06-20, CERTIS, 2006.

[Aud07] J.-Y. Audibert. Progressive mixture rules are deviation suboptimal. In *Advances in Neural Information Processing Systems 20 (NIPS'07)*, pages 41–48, 2007.

[Aud09] J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *Ann. Statist.*, 37(4):1591–1646, 2009.

[AW01] K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Mach. Learn.*, 43(3):211–246, 2001.

[Azu67] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Math. Journ.*, 19(3):357–367, 1967.

[Bar00] Y. Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.

[Bar02] Y. Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.

[Bar11] P. Bartlett. Online prediction. Lectures at IHP in May 2011. Slides available at http://www.stat.berkeley.edu/∼bartlett/talks/ihp-may-2011.pdf., 2011.

[BBGO10] G. Biau, K. Bleakley, L. Györfi, and G. Ottucsák. Nonparametric sequential prediction of time series. *J. Nonparametr. Stat.*, 22(3–4):297–317, 2010.

[BBM99] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999.

[BC91] A.R. Barron and T.M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054, 1991.

[BGH09] Y. Baraud, C. Giraud, and S. Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672, 2009.

[BGH11] Y. Baraud, C. Giraud, and S. Huet. Estimator selection in the Gaussian setting. Technical report, 2011. Avalaible at http://arxiv.org/abs/1007.2096.

[BHR08] P. Bartlett, E. Hazan, and A. Rakhlin. Adaptive online gradient descent. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS'07)*, pages 65–72. MIT Press, Cambridge, MA, 2008.

[Bir01] L. Birgé. A new look at an old result: Fano's lemma. Technical Report PMA-632, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI, 2001. Available at http://www.proba.jussieu.fr/mathdoc/textes/PMA-632.dvi.

[Bir04] L. Birgé. Model selection for Gaussian regression with random design. *Bernoulli*, 10(6):1039–1051, 2004.

[Bir05] L. Birgé. A new lower bound for multiple hypothesis testing. *IEEE Trans. Inf. Th.*, 51:1611–1615, 2005.

[Bir06] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. Henri Poincaré*, 42(3):273–325, 2006.

[Bla56] D. Blackwell. An analog of the minimax theorem for vector payoffs. *Pacific J. Math.*, 6(1):1–8, 1956.

[BM97] L. Birgé and P. Massart. From model selection to adaptive estimation. In *Festschrift for Lucien Le Cam*, pages 55–87. Springer, New York, 1997.

[BM01a] L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001.

[BM01b] L. Birgé and P. Massart. A generalized Cp criterion for Gaussian model selection. Technical Report PMA-647, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI, 2001. Available at http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2001.

[BM07a] L. Birgé and P. Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields*, 138:33–73, 2007.

[BM07b] A. Blum and Y. Mansour. From external to internal regret. *J. Mach. Learn. Res.*, 8:1307–1324, 2007.

[BMSS11] S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvari. X-armed bandits. *J. Mach. Learn. Res.*, 12:1587–1627, 2011.

[BN08] F. Bunea and A. Nobel. Sequential procedures for aggregating arbitrary estimators of a conditional mean. *IEEE Trans. Inform. Theory*, 54(4):1725–1735, 2008.

[Bre96] L. Breiman. Bagging predictors. *Mach. Learn.*, 24:123–140, 1996.

[BRT09] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[BT03] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31:167–175, 2003.

[BTW04] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for regression learning. Technical report, 2004. Available at http://arxiv.org/abs/math/0410214.

[BTW06] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation and sparsity via ℓ_1 penalized least squares. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT'06)*, pages 379–391, 2006.

[BTW07a] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[BTW07b] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194, 2007.

[Cat99] O. Catoni. Universal aggregation rules with exact bias bounds. Technical Report PMA-510, Laboratoire de Probabilités et Modèles Aléatoires, CNRS, Paris, 1999.

[Cat04] O. Catoni. *Statistical learning theory and stochastic optimization*. Springer, New York, 2004.

[Cat07] O. Catoni. *Pac-Bayesian supervised classification: the thermodynamics of statistical learning*. Institute of Mathematical Statistics Lecture Notes—Monograph Series, 56. Institute of Mathematical Statistics, Beachwood, OH, 2007.

[CB99] N. Cesa-Bianchi. Analysis of two gradient-based algorithms for on-line regression. *J. Comput. System Sci.*, 59(3):392–411, 1999.

[CBCG04] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Trans. Inform. Theory*, 50(9):2050–2057, 2004.

[CBFH+97] N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, D. Haussler, R. Schapire, and M.K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

[CBG08] N. Cesa-Bianchi and C. Gentile. Improved risk tail bounds for on-line algorithms. *IEEE Trans. Inform. Theory*, 54(1):386–390, 2008.

[CBL99] N. Cesa-Bianchi and G. Lugosi. On prediction of individual sequences. *Ann. Statist.*, 27:1865–1895, 1999.

[CBL03] N. Cesa-Bianchi and G. Lugosi. Potential-based algorithms in on-line prediction and game theory. *Mach. Learn.*, 51(3):239–261, 2003.

[CBL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[CBLS05] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label efficient prediction. *IEEE Trans. Inform. Theory*, 51(6), 2005.

[CBLS06] N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Regret minimization under partial monitoring. *Math. Oper. Res.*, 31(3):562–580, 2006.

[CBLW96] N. Cesa-Bianchi, P.M. Long, and M.K. Warmuth. Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Trans. Neural Networks*, 7(3):604–619, 1996.

[CBMS07] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved second-order bounds for prediction with expert advice. *Mach. Learn.*, 66(2/3):321–352, 2007.

[CK81] I. Csiszar and J. Kôrner. *Information Theory: Coding Theorems for discrete Memory-less Systems*. Academic Press, New York, 1981.

[CT07] E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[DHS10] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT'10)*, pages 257–269, 2010.

[DJ94a] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[DJ94b] D.L. Donoho and I.M. Johnstone. Minimax risk over ℓ_p-balls for ℓ_q-error. *Probab. Theory Relat. Fields*, 99:277–303, 1994.

[DS06] O. Dekel and Y. Singer. Data-driven online to batch conversions. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18 (NIPS'05)*, pages 267–274. MIT Press, Cambridge, MA, 2006.

[DS11] A. Dalalyan and J. Salmon. Optimal aggregation of affine estimators. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT'11)*, 2011.

[DSSST10] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT'10)*, pages 14–26, 2010.

[DT07] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Proceedings of the 20th Annual Conference on Learning Theory (COLT'07)*, pages 97–111, 2007.

[DT08] A. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Mach. Learn.*, 72(1-2):39–61, 2008.

[DT09] A. Dalalyan and A. B. Tsybakov. Sparse regression learning by aggregation and Langevin Monte-Carlo. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT'09)*, pages 83–92, 2009.

[DT11] A. Dalalyan and A. B. Tsybakov. Mirror averaging with sparsity priors. *Bernoulli*, 2011. To appear. Available at http://hal.archives-ouvertes.fr/hal-00461580/.

[DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Springer, New York, 1998.

[EDKMM09] E. Even Dar, R. Kleinberg, S. Mannor, and Y. Mansour. Online learning for global cost functions. In *Proceedings of the 22th Annual Conference on Learning Theory (COLT'09)*, 2009.

[EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2):407–499, 2004.

[Fan53] K. Fan. Minimax theorems. *Proc. Nat. Acad. Sci.*, 39:42–47, 1953.

[FG94] D. P. Foster and E. I. George. The risk inflation criterion for multiple regression. *Ann. Statist.*, 22(4):1947–1975, 1994.

[FG00] D.P. Foster and E.I. George. Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747, 2000.

[FL95] D. Fudenberg and D.K. Levine. Universal consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19:1065–1089, 1995.

[FL99] D. Fudenberg and D.K. Levine. Universal conditional consistency. *Games Econom. Behavior*, 29:104–130, 1999.

[FMG92] M. Feder, N. Merhav, and M. Gutman. Universal prediction of individual sequences. *IEEE Trans. Inform. Theory*, 38:1258–1270, 1992.

[For99] J. Forster. On relative loss bounds in generalized linear regression. In *Proceedings of the 12th International Symposium on Fundamentals of Computation Theory*, volume 1684 of *Lecture Notes in Computer Science*, pages 269–280. Springer-Verlag, Berlin, 1999.

[Fos91] D. Foster. Prediction in the worst-case. *Ann. Statist.*, 19:1084–1090, 1991.

[Fre75] D.A. Freedman. On tail probabilities for martingales. *Ann. Probab.*, 3:100–118, 1975.

[FS97] S. Freund and R.E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.*, 55(1):119–139, 1997.

[FSSW97] Y. Freund, R. E. Schapire, Y. Singer, and M. K. Warmuth. Using and combining predictors that specialize. In *Proceedings of the 29th annual ACM Symposium on Theory of Computing (STOC'97)*, pages 334–343, 1997.

[FV97] D. Foster and R. Vohra. Calibrated learning and correlated equilibrium. *Games Econom. Behavior*, 21:40–45, 1997.

[FV98] D. Foster and R. Vohra. Asymptotic calibration. *Biometrika*, 85:379–390, 1998.

[FV99] D. Foster and R. Vohra. Regret in the on-line decision problem. *Games Econom. Behavior*, 29:7–35, 1999.

[Gen03] C. Gentile. The robustness of the p-norm algorithms. *Mach. Learn.*, 53(3):265–299, 2003.

[Ger10a] Gerchinovitz, S. Minimax rate of internal regret in prediction of individual sequences. Talk at the StatMathAppli 2010 workshop, Fréjus, France, 2010.

[Ger10b] Gerchinovitz, S. Vitesse minimax du regret interne en prédiction de suites individuelles. Talk at 42èmes Journées de Statistique, Marseille, France. Extended abstract (in french) available at http://hal.archives-ouvertes.fr/inria-00494716_v1/, 2010.

[Ger11a] S. Gerchinovitz. Sparsity regret bounds for individual sequences in online linear regression. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT'11)*, 2011.

[Ger11b] Gerchinovitz, S. Aggregation of nonlinear models. Talk at the StatMathAppli 2011 workshop, Fréjus, France, 2011.

[Gir08] C. Giraud. Mixing least-squares estimators when the variance is unknown. *Bernoulli*, 14(4):1089–1107, 2008.

[GJ03] A. Greenwald and A. Jafari. A general class of no-regret learning algorithms and game-theoretic equilibria. In *Proceedings of the 16th Annual Conference on Computational Learning Theory (COLT'03) and 7th Kernel Workshop*, pages 2–12. Springer, 2003.

[GKKW02] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

[GL99] C. Gentile and N. Littlestone. The robustness of the p-norm algorithms. In *Proceedings of the 12th Annual Conference on Learning Theory (COLT'99)*, pages 1–11, 1999.

[GLS01] A.J. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. *Mach. Learn.*, 43(3):173–210, 2001.

[GO07] L. Györfi and G. Ottucsák. Sequential prediction of unbounded stationary time series. *IEEE Trans. Inform. Theory*, 53(5):1866–1872, 2007.

[GY11] S. Gerchinovitz and J.Y. Yu. Adaptive and optimal online linear regression on ℓ^1-balls. In *Proceedings of the 22nd International Conference on Algorithmic Learning Theory (ALT'11)*, 2011. In press.

[Han57] J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the theory of games*, 3:97–139, 1957.

[HK70] A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[HK08] H. Hazan and S. Kale. Extracting certainty from uncertainty: regret bounded by variation in costs. In *Proceedings of the 21st Annual Conference on Learning Theory (COLT'08)*, pages 57–67, 2008.

[HKW98] D. Haussler, J. Kivinen, and M. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Trans. Inform. Theory*, 44:1906–1925, 1998.

[HMC00] S. Hart and A. Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68:1127–1150, 2000.

[HMC01] S. Hart and A. Mas-Colell. A general class of adaptive strategies. *J. Econom. Theory*, 98:26–54, 2001.

[Hoe63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58(301):13–30, 1963.

[HP97] D. Helmbold and S. Panizza. Some label efficient learning results. In *Proceedings of the 10th Annual Conference on Computational Learning Theory (COLT'97)*, pages 218–230. ACM Press, 1997.

[HvdG11] M. Hebiri and S. van de Geer. The Smooth-Lasso and other $\ell^1 + \ell^2$-penalized methods. *Electron. J. Stat.*, 5:1184–1226, 2011.

[JRT08] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Ann. Statist.*, 36(5):2183–2206, 2008.

[KLT11] V. Koltchinskii, K. Lounici, and A.B. Tsybakov. Nuclear norm penalization and optimal rates for noisy low rank matrix completion. *Ann. Statist.*, 2011. To appear.

[Kne52] H. Kneser. Sur un théorème fondamental de la théorie des jeux. *C. R. Acad. Sci. Paris*, 234:2418–2420, 1952.

[Kol09a] V. Koltchinskii. Sparse recovery in convex hulls via entropy penalization. *Ann. Statist.*, 37(3):1332–1359, 2009.

[Kol09b] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Ann. Inst. Henri Poincaré Probab. Stat.*, 45(1):7–57, 2009.

[KT09] S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems 21 (NIPS'08)*, pages 801–808. 2009.

[KW97] Jyrki Kivinen and Manfred K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Inform. and Comput.*, 132(1):1–63, 1997.

[KW99] J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *Proceedings of the 4th European Conference on Computational Learning Theory (EuroCOLT'99)*, pages 153–167, 1999.

[KW01] J. Kivinen and M. Warmuth. Relative loss bounds for multidimensional regression problems. *Mach. Learn.*, 45(3):301–329, 2001.

[LB06] G. Leung and A. R. Barron. Information theory and mixing least-squares regressions. *IEEE Trans. Inform. Theory*, 52(8):3396–3410, 2006.

[Led01] M. Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

[Lit89] N. Littlestone. From on-line to batch learning. In *Proceedings of the 2nd Annual Conference on Learning Theory (COLT'89)*, pages 269–284, 1989.

[LLW95] N. Littlestone, P.M. Long, and M.K. Warmuth. On-line learning of linear functions. *Comput. Complexity*, 5(1):1–23, 1995.

[LLZ09] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *J. Mach. Learn. Res.*, 10:777–801, 2009.

[LM09] G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probab. Theory Related Fields*, 145:591–613, 2009.

[LMS08] G. Lugosi, S. Mannor, and G. Stoltz. Strategies for prediction under imperfect monitoring. *Math. Oper. Res.*, 33(3):513–528, 2008.

[Lou07] K. Lounici. Generalized mirror averaging and D-convex aggregation. *Math. Methods Statist.*, 2007.

[LPvdGT11] K. Lounici, M. Pontil, S. van de Geer, and A.B. Tsybakov. Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.*, 2011. To appear.

[LW94] N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Inform. and Comput.*, 108:212–261, 1994.

[Mal73] C. L. Mallows. Some Comments on Cp. *Technometrics*, 15(4):661–675, 1973.

[Mas07] P. Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007.

[MCL98] A.D.R. McQuarrie and Tsai C.-L. *Regression and Time Series Model Selection*. World Scientific, Singapore, 1998.

[MM11] P. Massart and C. Meynet. The Lasso as an ℓ^1-ball model selection procedure. *Electron. J. Stat.*, 5:669–687, 2011.

[MS10] H.B. McMahan and M. Streeter. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT'10)*, pages 244–256, 2010.

[Nem00] A. Nemirovski. *Topics in Non-Parametric Statistics*. Springer, Berlin/Heidelberg/New York, 2000.

[NY83] A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons Inc., New York, 1983.

[Pin64] M.S. Pinsker. *Information and information stability of random variables and processes*. Translated and edited by Amiel Feinstein. Holden-Day Inc., San Francisco, Calif., 1964.

[Pis83] G. Pisier. Some applications of the metric entropy condition to harmonic analysis. In Ron Blei and Stuart Sidney, editors, *Banach Spaces, Harmonic Analysis, and Probability Theory*, volume 995 of *Lecture Notes in Mathematics*, pages 123–154. Springer Berlin / Heidelberg, 1983.

[Rob52] H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 55:527–535, 1952.

[RST10] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: Random averages, combinatorial parameters, and learnability. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R.S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23 (NIPS'10)*, pages 1984–1992. 2010.

[RST11] A. Rakhlin, K. Sridharan, and A. Tewari. Online learning: beyond regret. In *Proceedings of the 24th Annual Conference on Learning Theory (COLT'11)*, 2011.

[RT11] P. Rigollet and A. B. Tsybakov. Exponential Screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2):731–771, 2011.

[Rus99] A. Rustichini. Minimizing regret: The general case. *Games Econom. Behav.*, 29:224–243, 1999.

[RWY11] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ^q-balls. *IEEE Trans. Inform. Theory*, 57(10):6976–6994, 2011.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6:461–464, 1978.

[Sch03] K. Schlag. How to minimize maximum regret in repeated decision-making. Technical report, Universitat Pompeu Fabra, 2003.

[See08] M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *J. Mach. Learn. Res.*, 9:759–813, 2008.

[Sio58] M. Sion. On general minimax theorems. *Pacific J. Math.*, 8(1):171–176, 1958.

[SL05] G. Stoltz and G. Lugosi. Internal regret in on-line portfolio selection. *Mach. Learn.*, 59:125–159, 2005.

[SL07] G. Stoltz and G. Lugosi. Learning correlated equilibria in games with compact sets of strategies. *Games Econom. Behavior*, 59:187–208, 2007.

[SSSSS09] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT'09)*, pages 177–186, 2009.

[SSSZ10] S. Shalev-Shwartz, N. Srebro, and T. Zhang. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM J. Optim.*, 20(6):2807–2832, 2010.

[SST09] S. Shalev-Shwartz and A. Tewari. Stochastic methods for ℓ^1-regularized loss minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 929–936, 2009.

[Ste81] C.M. Stein. Estimation of the mean of a multivariate distribution. *Ann. Statist.*, 9(6):1135–1151, 1981.

[Sto05] G. Stoltz. *Incomplete information and internal regret in prediction of individual sequences*. PhD thesis, Paris-Sud XI University, 2005.

[Sto10a] G. Stoltz. Agrégation séquentielle de prédicteurs : méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique. *Journal de la Société Française de Statistique*, 151(2):66–106, 2010.

[Sto10b] G. Stoltz. Prédiction de suites individuelles. Lectures at Paris-Sud XI University, 2010.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[Tsy03] A. B. Tsybakov. Optimal rates of aggregation. In *Proceedings of the 16th Annual Conference on Learning Theory (COLT'03)*, pages 303–313, 2003.

[vdG08] S. A. van de Geer. High-dimensional generalized linear models and the Lasso. *Ann. Statist.*, 36(2):614–645, 2008.

[vdGB09] S. A. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.

[Ver10] N. Verzelen. Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons. Technical report, 2010. See http://arxiv.org/abs/1008.0526.

[Vov90] V. Vovk. Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory (COLT'90)*, pages 371–383, 1990.

[Vov98] V. Vovk. A game of prediction with expert advice. *J. Comput. System Sci.*, 56(2):153–173, 1998.

[Vov01] V. Vovk. Competitive on-line statistics. *Internat. Statist. Rev.*, 69:213–248, 2001.

[WH60] B. Widrow and M.E. Hoff. Adaptive switching circuits. In *IRE WESCON Convention Record, Part 4*, pages 96–104, 1960.

[WJ98] M. Warmuth and A.K. Jagota. Continuous and discrete-time nonlinear gradient descent: Relative loss bounds and convergence. In *Proceedings of the 5th International Symposium on Artificial Intelligence and Mathematics*, 1998.

[Xia10] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, 11:2543–2596, 2010.

[Yan00] Y. Yang. Combining different procedures for adaptive regression. *J. Multivariate Anal.*, 75:135–161, 2000.

[Yan01] Y. Yang. Adaptive regression by mixing. *J. Amer. Statist. Assoc.*, 96(454):574–588, 2001.

[Yan03] Y. Yang. Regression with multiple candidate models: selecting or mixing? *Statistica Sinica*, 13:783–809, 2003.

[Yan04] Y. Yang. Aggregating regression procedures to improve performance. *Bernoulli*, 10(1):25–47, 2004.

[Zha05] T. Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *Proceedings of the 18th Annual Conference on Learning Theory (COLT'05)*, pages 173–187, 2005.

[Zin03] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML'03)*, pages 928–936, 2003.