Quantification d'incertitude pour l'Approximation Stochastique

Gersende Fort

CNRS

Institut de Mathématiques de Toulouse

GRETSI, Lille, Août 2019

En collaboration avec

Stéphane Crepey, LaMME, Univ. d'Evry,
Emmanuel Gobet, CMAP, Ecole Polytechnique,
Uladzislau Stazhynski, CMAP, Ecole Polytechnique.

Travaux disponibles sur HAL :

Uncertainty quantification for Stochastic Approximation Limits using Chaos Expansion

Financements : Labex CIMI 11-LABX-0040-CIMI, via le programme ANR-11-IDEX-0002-02.

Le problème

L'Approximation Stochastique : situation usuelle (1/2)

Robbins et Monro, 1951; Benveniste, Métivier, Priouret, 1990; Kushner et Yin, 1998.

- Algorithme itératif pour la résolution de

$$\theta \in \Theta \subset \mathbb{R}^q, \qquad h(\theta) = 0_{\mathbb{R}^q}$$

lorsque h non explicite mais

$$h(\theta) = \int_{\mathsf{X}} H(\theta, x)\ \mathrm{d}\mu(x), \qquad \mu \text{ probabilité}$$

- Etant donnés une suite de pas $\{\gamma_t\}_t$ et une valeur initiale $\theta_0 \in \mathbb{R}^q$,

$$\theta_{t+1} = \theta_t + \gamma_{t+1} H(\theta_t, X_{t+1}) \qquad X_{t+1} \sim \mu$$

- Extensions classiques bien que possiblement techniques :

- Loi dépendant du paramètre $h(\theta) = \int_{\mathsf{X}} H(\theta, x)\ \mathrm{d}\mu_\theta(x)$
- Tirages approchés : $\mathrm{Loi}(X_{t+1}) \approx \mathrm{d}\mu_{\theta_t}$ (MCMC, Particules, $\cdots$)

L'Approximation Stochastique : situation usuelle (2/2)

- Exemple 1 : Robbins Monro

$$h(\theta) = \int_{\mathsf{X}} H(x) \mathrm{d}\mu(x) - \theta$$

- Exemple 2 : Gradient stochastique pour l'optimisation de $V : \Theta \to \mathbb{R}$

$$h(\theta) = \nabla V(\theta) = \int H(\theta, x) \ \mathrm{d}\mu_\theta(x)$$

▶ Application en statistique : ML dans modèles à données latentes

$$V(\theta) = \log L(\theta; Y_{1:n}) = \log \int p(x, Y_{1:n}; \theta) \ \mathrm{d}\nu(x)$$

$$h(\theta) = \int_{\mathsf{X}} \partial_\theta \log p(x, Y_{1:n}; \theta) \ \mathrm{d}\mu_\theta^{(Y_{1:n})}(x)$$

$\mathrm{d}\mu_\theta^{(Y_{1:n})}$: loi a posteriori des données latentes sachant les obs, dans le modèle statistique indexé par θ

- Exemple 3 : Quantile d'ordre α

$$h(\theta) = \int_{\mathsf{X}} 1_{x \leq \theta} \ \mathrm{d}\mu(x) - \alpha$$

Approximation Stochastique : le problème considéré ici (1/2)

- Modélisation d'une incertitude dans la formulation du problème :

$$h(\theta; w) = \int_{\mathsf{X}} H(\theta, x, w)\ \mu(w, \mathrm{d}x)$$

avec un a priori $\mathrm{d}\pi$ sur w.

- Exemple 1 : Quantile d'ordre α

$$h(\theta, w) = \int_{\mathsf{X}} 1_{x \leq \theta}\ \mu(w, \mathrm{d}x) - \alpha$$

w: la loi μ dépend d'un "paramètre" w dont la valeur n'est pas connue; on ne souhaite pas le fixer mais étudier le rôle de w et quantifier l'incertitude sur sa valeur.

- Exemple 2 : Analyse statistique "bayésienne"

$$h(\theta, w) = \int_{\mathsf{X}} H(\theta, x, w)\, \mathrm{d}\mu_\theta(x)$$

w: le critère d'attache aux données pénalisé / l'a priori sur θ dépend d'une quantité qu'on ne souhaite pas fixer, mais dont on souhaite comprendre le rôle.

Approximation Stochastique : le problème considéré ici (2/2)

Proposer une méthode numérique pour en même temps

- Obj-1 : trouver pour π-presque tout w,

$$\theta(w) \in \Theta \subset \mathbb{R}^q, \qquad \text{racine de } \tau \mapsto h(\tau; w)$$

- Obj-2 : caractériser la distribution de $\theta(W)$ lorsque $W \sim \pi$

Dans la suite, π est une loi sur un espace mesurable $(\mathsf{W}, \mathcal{W})$.

$\hookrightarrow$ on ne veut pas
- déterminer $\theta(w)$ en des valeurs de $W_j \sim \pi$
- "interpoler" pour obtenir $w \mapsto \theta(w)$
- approcher la loi de $\theta(W)$ par la mesure empirique

.

Solution numérique proposée

.

(i) Trouver pour π-presque tout w, $\theta(w)$ racine de $\tau \mapsto h(\tau, w) = 0$ lorsque $h(\tau, w) = \int_{\mathsf{X}} H(\theta, x, w)\ \mu(w, \mathsf{d}x)$;

(ii) caractérider $\theta(W)$ quand $W \sim \pi$

L'approche (1/2)

- Chercher une fonction $w \mapsto \theta(w)$ dans

$$\mathcal{S} := \{\theta_\star : \mathsf{W} \to \mathbb{R}^q, \text{mesurable, tq pour } \pi\text{-presque tout } w,\ h(\theta_\star(w), w) = 0_{\mathbb{R}^q}\}\,.$$

- Réponse en imposant

\- (A) que les q composantes de la fonction θ soient dans $L^2_\pi(\mathbb{R})$

- Csq-1 de (A) : chercher $\{u_i\}_{i\in\mathbb{N}} \in \ell_2(\mathbb{R}^q)$ tels que

$$\theta(w) = \sum_{i\geq 0} u_i\ B_i(w) \qquad \{w \mapsto B_i(w)\}_i \text{ base orthonormée de } L^2_\pi(\mathbb{R})$$

- Csq-2 de (A) : Cas $B_0 = 1$ on connaît tous les moments de $\theta(W)$ lorsque $W \sim \pi$

$$\mathbb{E}\left[\theta(W)\right] = u_0, \qquad \mathrm{Cov}\left[\theta(W)\right] = \sum_{i\geq 1} u_i\, u_i'$$

L'approche (2/2)

- En imposant aussi

- (B) $w \mapsto h(\theta(w), w) \in L^2_\pi(\mathbb{R}^q)$
- (C) il existe $\theta_\star \in \mathcal{S} \cap L^2_\pi(\mathbb{R}^q)$

$$\int_{\mathsf{W}} \langle \theta(w) - \theta_\star(w); h\left(\theta(w), w\right)\rangle_{\mathbb{R}^q} \; \mathrm{d}\pi(w) > 0, \qquad \forall\, \theta \in L^2_\pi(\mathbb{R}^q) \setminus \mathcal{S}.$$

- Csq fondamentale de (C) :

(i) la mise à jour de la règle d'apprentissage des coefficients u_i

(ii) la fonction de Lyapunov est évidente

- Par exemple, (C) se déduit de

$$\langle z - z_\star; h(z,w)\rangle_{\mathbb{R}^q} > 0 \quad \forall z, h(z,w) \neq 0,$$! "classique" en absence d'incertitude !

L'algorithme Rappel: $\{u_i\}_i \in \ell_2(\mathbb{R}^q)$ tq

$$\int_{\mathsf{X}} H\left(\sum_{i\geq 0} u_i B_i(w), x, w\right) \; \mu(w, \mathrm{d}x) = 0_{\mathbb{R}^q}, \qquad \text{pour } \pi\text{-presque tout } w$$

- Se donner

- une suite de pas stt positifs $\{\gamma_t\}_t$
- une suite d'entiers $\{L_t\}_t \uparrow +\infty$
- une valeur initiale pour les premières composantes: $u_i^0, \quad i = 0, \cdots, L_0$

- Répéter jusqu'à convergence :

- Echantillonner $W_{t+1} \sim \mathrm{d}\pi, \qquad X_{t+1}|W_{t+1} \sim \mu(W_{t+1}, \mathrm{d}x)$
- Mettre à jour les composantes déjà en cours d'apprentissage :

$$i = 0, \cdots, L_t : \quad u_i^{t+1} = u_i^t + \gamma_{t+1} H\left(\sum_{j=0}^{L_t} u_j^t B_j(W_{t+1}),\ X_{t+1},\ W_{t+1}\right) B_i(W_{t+1})$$

- Apprendre de nouvelles composantes :

$$i = L_t + 1, \cdots, L_{t+1} : \quad \text{même formule, avec } u_i^t = 0_{\mathbb{R}^q}$$

Orginalité de l'approche

Rappel: $\theta : \mathsf{W} \to \mathbb{R}^q$ tq

$$\int_{\mathsf{X}} H(\theta(w), x, w)\, \mu(w, \mathrm{d}x) = 0_{\mathbb{R}^q}, \qquad \text{pour } \pi\text{-presque tout } w$$

Dans la littérature, pour apprendre une fonction par Approx Sto :

- Double Monte Carlo : (a) tirer des valeurs $W_k \sim \pi$; (b) apprendre $\theta(W_k) \in \mathbb{R}^q$ par Approx Sto; (c) interpolation

- Troncation : choisir L et chercher $u_0, \cdots, u_L$ tq $\theta = \sum_{i=0}^{L} u_i B_i$.

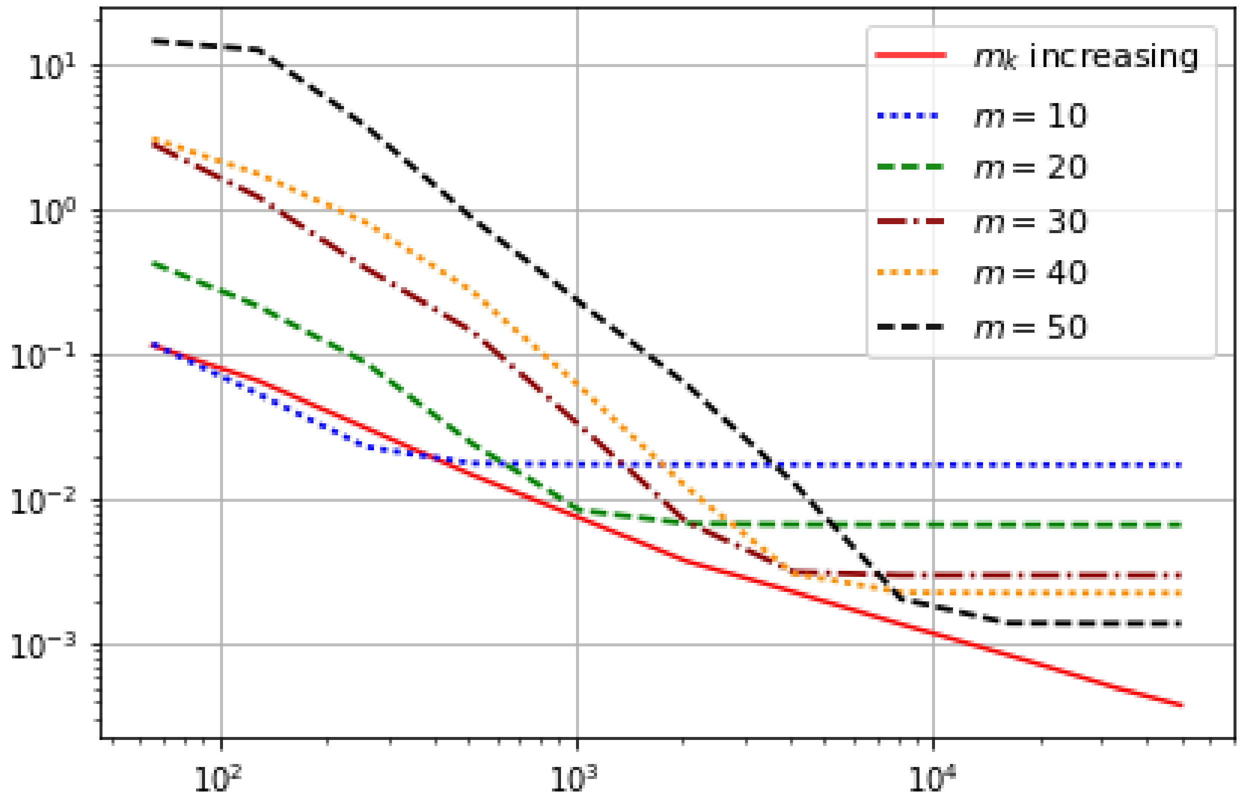
- Faussement en dimension infinie h est un critère empirique, on recherche dans un RKHS

- Algorithme à implémenter en dimension infinie : en pratique ??

- Algorithme (irréaliste) en dimension finie mais croissante "the sieve approach" : disposer de produits scalaires dans L^2_π - non, en pratique.

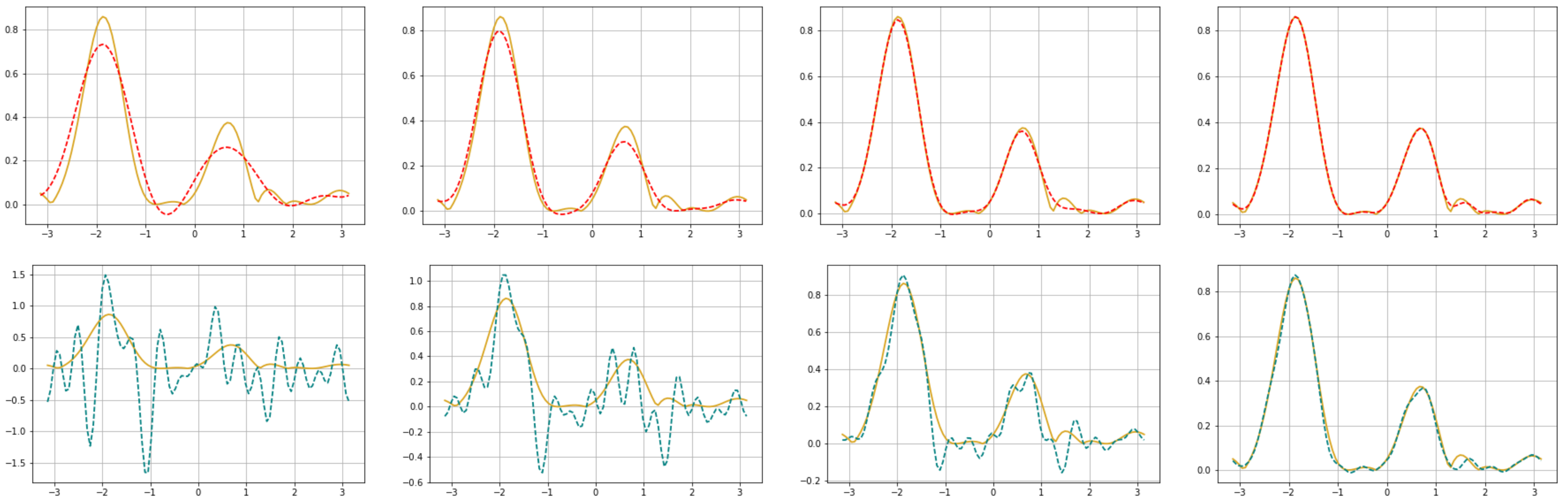
Aurait-on perdu à restreindre à un sous-espace (de dim m) de $L^2_\pi(\mathbb{R}^q)$?

Exemple jouet : $\theta_\star$ est unique et connu; $\pi \equiv \mathcal{U}([-\pi, \pi])$; base trigonométrique sur $L^2_\pi(\mathbb{R})$



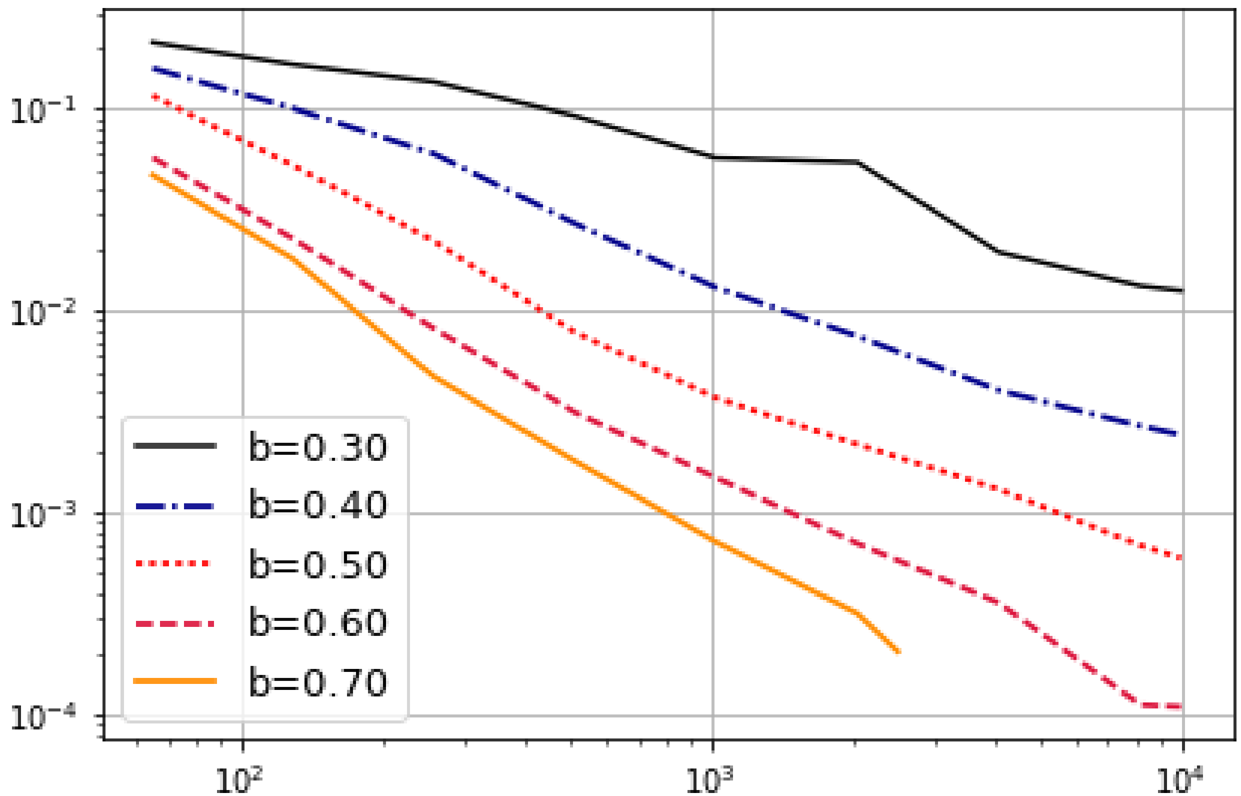
On trace $t \mapsto \mathbb{E}\left[\|u^t - u_\star\|^2_{\ell_2}\right]$, où l'espérance est estimée par Monte Carlo sur des runs indépendants de l'algorithme. On envisage différentes valeurs de l'indice de troncation m, ainsi que la stratégie de non-troncation.

Période de chauffe : avantage à n'apprendre que peu de coefficients



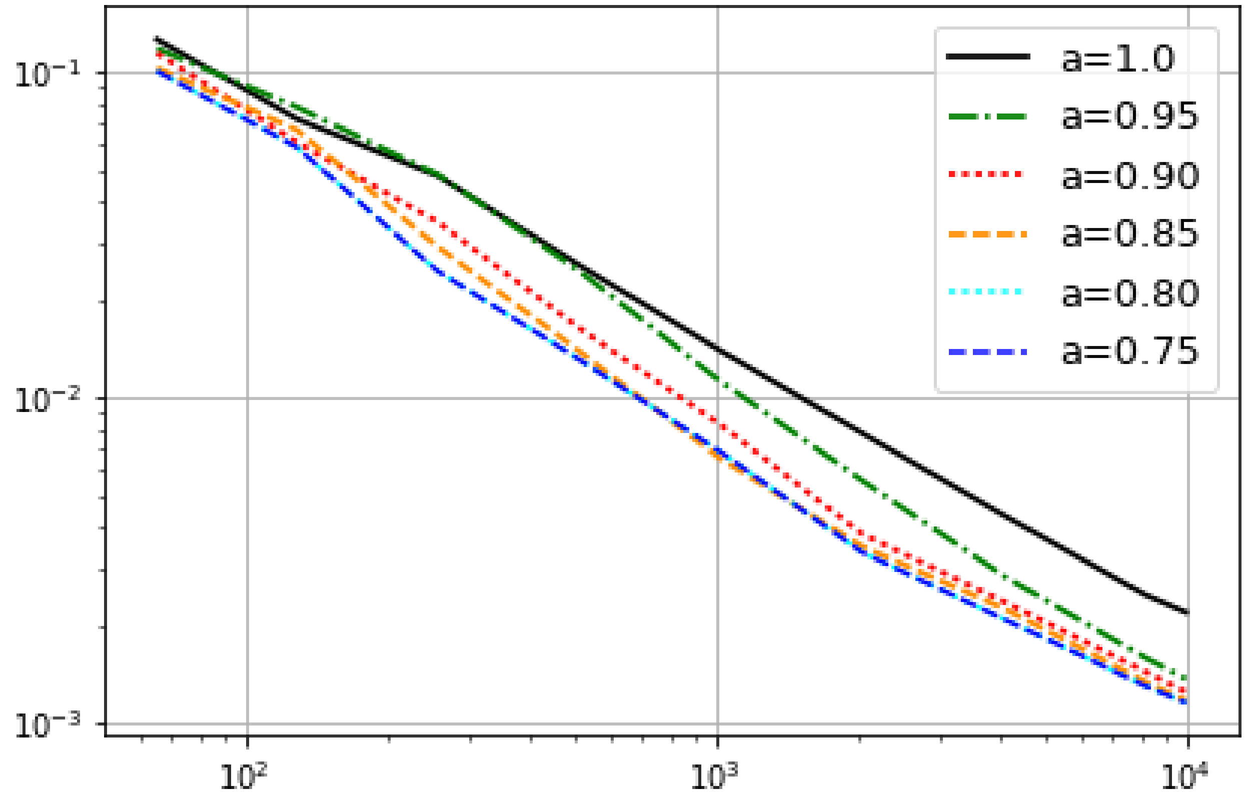
En trait plein : la fonction $w \mapsto \theta_\star(w)$
En pointillés, pour la stratégie troncation $m = 30$ fixe (bas) et la stratégie de non troncation (haut), on trace $w \mapsto \theta^t(w)$ pour t petit (en phase transiente, $t \leq 1024$).

Taille des blocs L_t



On compare le rôle de b lorsque l'on prend $L_t = O(t^b)$ sur l'erreur $t \mapsto \mathbb{E}\left[\|u^t - u_\star\|_{\ell_2}^2\right]$ estimée par MC sur des runs indépendants. ATTENTION au coût computationnel (mise à jour de vecteurs de grande taille au cours des itérations) qui n'est pas analysé ici.

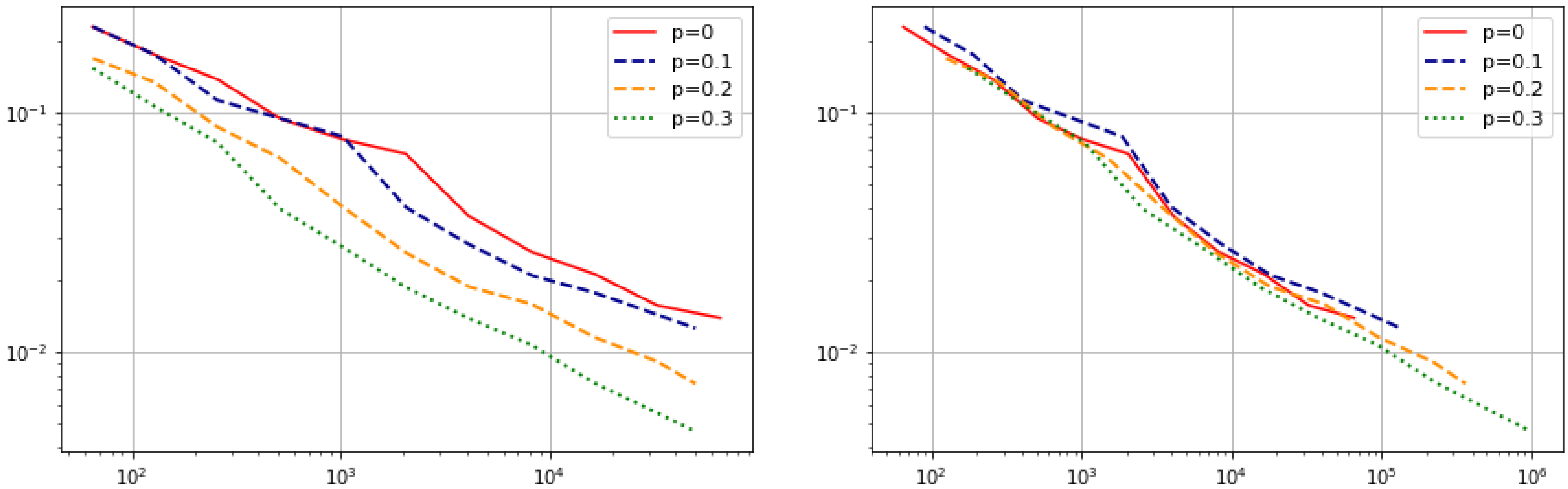
Comment choisir le pas γ_t ?



On compare le rôle de a lorsque l'on prend $\gamma_t = O(t^{-a})$ sur l'erreur $t \mapsto \mathbb{E}\left[\|u^t - u_\star\|_{\ell_2}^2\right]$

estimée par MC sur des runs indépendants.

Gagne-t-on à faire plus d'un tirage Monte Carlo par itération ?



On regarde l'effet sur l'erreur $\mathbb{E}\left[\|u^t - u_\star\|^2_{\ell_2}\right]$ estimée par MC sur des runs indépendants lorsqu'à chaque itération de l'algorithme, on ne fait qu'un seul tirage MC ($p = 0$) ou un nombre qui augmente avec le nombre d'itérations en $O(t^p)$. A gauche, l'erreur en fonction du nombre d'itérations t et à droite, en fonction du nombre total de MC.

Bien-fondé de la méthode

Principe des preuves de convergence en Approx Sto

- Etape 1 : une fonction de Lyapunov i.e. $V: \ L^2_\pi(\mathbb{R}^q) \to \mathbb{R}^+$ tq

$$V(\theta^{t+1}) \leq V(\theta^t) - \gamma_{t+1} \ \phi^2(\theta_t) + \underbrace{\mathcal{E}_{t+1}}_{\text{sorte de mesure des erreurs cumulées}}$$

- Etape 2 : Sous la condition que $\sum_t \mathcal{E}_t < \infty$ p.s., on déduit

- la stabilité de la suite $\{\theta_t\}_t$
- puis la convergence de la suite $\{\theta^t\}_t$ vers $\{\theta : \phi^2(\theta) = 0\}$.

- Ici, on établit pour $u_\star \in \mathcal{S} \cap \ell_2$,

$$\|u^{t+1} - u_\star\|^2_{\ell_2} \leq \|u^t - u_\star\|^2 - 2\gamma_{t+1} \int \left\langle \theta^t(w) - \theta_\star(w); h(\theta^t(w), w) \right\rangle_{\mathbb{R}^q} \mathrm{d}\pi(w) + \mathcal{E}_{t+1}$$

et on propose un jeu de conditions

- assurant $\sum_t \mathcal{E}_t < \infty$ p.s.
- garantissant que les zeros du terme de rappel sont dans $\mathcal{S}$.

Conditions suffisantes pour la convergence

- Espace L^2_π : on cherche $\theta \in L^2_\pi(\mathbb{R}^q)$ et on suppose $\theta \mapsto h(\theta(\cdot), \cdot) \in L^2_\pi(\mathbb{R}^q)$.

- Espace des solutions : $\mathcal{S}$ est un compact non vide de $L^2_\pi(\mathbb{R}^q)$

- Fct de Lyapunov : pour tout $\theta^\star \in \mathcal{S}$,

$$\int \langle \theta(w) - \theta^\star(w), h(\theta(w), w) \rangle_{\mathbb{R}^q} \, \pi(\mathrm{d}w) > 0, \quad \forall \theta \in L^2_\pi(\mathbb{R}^q) \setminus \mathcal{S}.$$

- Traitement ds perturbation : il existe $\theta^\star = \sum_i u_i^\star B_i \in \mathcal{S}$ tq

$$\sum_t \gamma_t = +\infty, \quad \sum_t \gamma_t^{1+\kappa} < \infty, \quad \sum_t \gamma_t^{1-\kappa} \left(\sum_{i \geq L_t} u_{i,\star}^2 \right) < \infty \quad \text{où } \kappa \in]0,1[$$

$$\exists C, \forall z \in \mathbb{R}^q, \quad \sup_{w \in \mathsf{W}} \int_{\mathsf{X}} |H(z,x,w)|^2 \mu(w, \mathrm{d}x) \leq C \left(1 + |z|^2\right)$$

$$\sum_t \gamma_t^2 \left(\sup_w \sum_{i \leq L_t} |B_i(w)|^2 \right) < \infty$$

(les deux dernières conditions sont couplées; l'avant-dernière à fixer en fonction du pbm traité)

Stabilité, Convergence L^p, Convergence p.s. faible et forte (Crepey, F., Gobet, Stazhynski, 2018)

Sous ces hypothèses, il existe une v.a. $\theta^\star$ à valeur dans $\mathcal{S} \cap L^2_\pi(\mathbb{R}^q)$ tq

- Stabilité :

$$\lim_t \|\theta^t - \theta^\star\|_{L^2_\pi(\mathbb{R}^q)} < \infty \ p.s. \qquad \sup_t \mathbb{E}\left[\|\theta^t - \theta^\star\|^2_{L^2_\pi(\mathbb{R}^q)}\right] < \infty.$$

- Convergence L^p, $p \in]0, 2[$

$$\lim_t \mathbb{E}\left[\|\theta^t - \theta^\star\|^p_{L^2_\pi(\mathbb{R}^q)}\right] = 0$$

- Convergence faible avec probabilité 1: si continuité pour la topologie faible de $\theta \in L^2_\pi \mapsto h(\theta, \cdot) \in L^2_\pi$

θ^t converge faiblement vers $\theta^\star$ p.s.

- Convergence forte avec probabilité 1: différents jeux d'hyp. possibles dont par ex. l'ajout d'une projection sur un **compact** $u_i^{t+1} = \Pi_A(u_i^t + \cdots)$:

$$\lim_t \|\theta^t - \theta^\star\|_{L^2_\pi(\mathbb{R}^q)} = 0, \qquad \text{p.s.}$$

Conclusions

Extensions

Loi des tirages

Self-stabilité

Cas du gradient: autres stratégies et sur des arguments d'analyse de convergence similaires

$$\theta_{t+1} = \theta_t + \gamma_{t+1} \widehat{\nabla f(\theta_t)}$$

avec une approximation stochastique du gradient :
- "sample average" i.e. plusieurs tirages plutôt qu'un seul : $m_{t+1}^{-1} \sum_{j=1}^{m_{t+1}} H(\theta_t, X_{j,t+1})$
- réduction de variance (par variables de contrôle; par AS double niveau; $\cdots$)

Bémols

La difficulté de trouver une base orthonormée - cas multivarié, et pour des lois π exotiques.

Néanmoins

Méthode pour répondre à la dimension infinie sans tronquer;

et plus généralement, méthode pour apprendre une suite dénombrable.