Lecture notes:
Statistics, optimization and algorithms in high dimension

Edouard Pauwels

(Version of February 6, 2020)

Forewords:

These are lecture notes for a course on statistics and optimization in high dimensions taught at Université Toulouse 3 Paul Sabatier in 2019. A website for the course is available at

`https://www.math.univ-toulouse.fr/~epauwels/M2RI/`

with supporting slides and practical sessions. The content is adapted from various sources, which I tried to cite as precisely as possible throughout the chapters. I am aware that the text may still contain imprecisions, typos or mistakes and welcome feedback of any kind.

Contents

Chapter 1

Introduction

1.1 Motivation

The motivation for this course is to illustrate the theoretical and practical implications of high dimensionality when working with data. In this context, high dimensionality means a large number of observations or a large number of descriptor variables, or both. The actual meaning of *large* is vague and could be taken as "too big to be ignored", meaning that usual methods may not work directly and necessitate special developments. Two concrete examples

- In a statistical context, the number of observations is not sufficient to propose a conclusive quantitative analysis.
- From a computational perspective, the execution time is bottleneck resource.

These constraints motivated theoretical and practical developments at the interplay between statistical analysis and optimization. These two disciplines meet naturally in high dimensional regimes and the first goal of this course is to provide an illustration of the different issues at stake.

The purpose is not to be exhaustive, we will stick to one of the most simple and well known example in this field, the sparse least squares problem, which will be a running example to illustrate:

- Statistical efficiency issues in high dimension and their resolution
- Computational complexity barriers in high dimensional estimation.
- All purpose solvers for conic programming
- First order methods and composite optimization
- Randomized methods to treat high dimensionality issues from a computational view point.

1.2 Overview of regression and learning

The content of this section is adapted from Philippe Rigollet lectures notes [51].

1.2.1 General setting:

We call a topological space $\mathcal{X}$ a feature space and $\mathcal{Y} \subset \mathbb{R}$ an output space and let (X, Y) be a random variable on $\mathcal{X} \times \mathcal{Y}$ such that Y has finite variance. Our goal will be to predit Y given X. This question has a simple probabilistic solution. We consider a decision theoretic framework which adds a notion of risk on top of probability theory. Let P denote the joint probability distribution

on $\mathcal{X} \times \mathcal{Y}$ for the random variable (X, Y) and consider the following, for any measureable function $f \colon \mathcal{X} \mapsto \mathcal{Y}$, we set

$$R(f) = \mathbb{E}\left[(f(X) - Y)^2\right] = \int_{\mathcal{X}\times\mathcal{Y}} (f(x) - y))^2 P(dx, dy). \tag{1.1}$$

Risk minimization

Our goal is to find the prediction function f having the lowest risk, that is $\min_f R(f)$ where the minimum is taken over all measurable functions. The solution to the minimization of the expected prediction error is given by the regression function

$$f^* \colon x \mapsto \mathbb{E}\left[Y | X = x\right]. \tag{1.2}$$

see [29], for a mathematical treatement of conditional expectation. Indeed, for any measurable function $g \colon \mathcal{X} \mapsto \mathbb{R}$, one has

$$\mathbb{E}\left[(Y - g(X))^2\right] = \mathbb{E}\left[(Y - f^*(X))^2\right] + \mathbb{E}\left[(f^*(X) - g(X))^2\right] + 2\mathbb{E}\left[(Y - f^*(X))(f^*(X) - g(X))\right],$$

and one sees that

$$\begin{aligned}\mathbb{E}\left[(Y - f^*(X))(f^*(X) - g(X))\right] &= \mathbb{E}\left[\mathbb{E}\left[(Y - f^*(X))(f^*(X) - g(X))|X\right]\right] \\ &= \mathbb{E}\left[(f^*(X) - g(X))\mathbb{E}\left[(Y - f^*(X))|X\right]\right] = 0.\end{aligned}$$

From a probabilistic point of view, the problem is solved. The minimal value of the expected prediction error is called *Bayes risk* and is a lower bound on what could potentially be achieved. However, as a statistician, one does not have access to this conditional expectation and it needs to be estimated from a finite sample.

Estimation:

We are given a sample $\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ which consists of independant copies of (X, Y). The goal of regression is to construct an estimator $\hat{f}_n \colon \mathcal{X} \mapsto \mathcal{Y}$ which has a small L_2 risk $R(\hat{f}_n)$. If we denote by P_X the marginal distribution of X, for any $h \colon \mathcal{X} \mapsto \mathbb{R}$, we define

$$\|h\|_2^2 := \|h\|^2_{L^2(dP_X)} = \int_{\mathcal{X}} h^2 dP_X = \mathbb{E}\left[h^2(X)\right] \tag{1.3}$$

then one has

$$R(\hat{f}_n) = R(f^*) + \|\hat{f}_n - f^*\|_2^2.$$

In other words, it is equivalent to study $R(\hat{f}_n)$ and $\|\hat{f}_n - f^*\|_2^2$, we will focus on the second term. Note that both are random quantities since our sample $\mathcal{D}_n$ is random and we need deterministic estimates to quantify convergence speed. We shall use bounds in expectation:

$$\mathbb{E}\left[\|\hat{f}_n - f^*\|_2^2\right] \leq \phi(n), \quad \forall n \in \mathbb{N}$$

or bounds with high probability

$$\mathbb{P}\left[\|\hat{f}_n - f^*\|_2^2 > \psi(n, \delta)\right] \leq \delta, \quad \forall n \in \mathbb{N}, \delta \in (0, 1)$$

where in both cases, the randomness is over the random draw of $\mathcal{D}_n$. Note that this ensures for any $\delta > 0$, with probability at least $1 - \delta$, $\|\hat{f}_n - f^*\|_2^2 \leq \psi(n, \delta)$. Often bounds in probability are deduced from bounds in expectation by concentration of measure. We will mostly focus on sub-gaussian concentration, but the topic is much more vast [18].

Empirical risk minimization:

How is $\hat{f}_n$ constructed? The majority of methods rely on Empirical Risk Minimization (ERM). The law of large numbers entails that for a given $g\colon \mathcal{X} \mapsto \mathcal{Y}$ and for sufficiently large $n \in \mathbb{N}$, the expectation $\mathbb{E}[g(X)]$ may be approximated by an empirical average, $\frac{1}{n}\sum_{i=1}^{n} g(X_i)$ with independent copies of X. We denote by R_n the empirical risk, obtained by replacing the expectation in R by such a discrete sum:

$$R_n(g) = \frac{1}{n}\sum_{i=1}^{n} \left(g(X_i) - Y_i\right)^2$$

One could proceed by minimizing $R_n(g)$ in place of R. Building estimators based on solutions of optimization problems is refered as M-estimation in the statistics litterature. However, the law of large number is not uniform and it is easy to see that there exists functions g (for example polynomials) which satisfy $R_n(g) = 0$ while the Bayes risk remains positive. This phenomenon is called *overfitting* and is a curse which has to be avoided. Common approaches include fixing a function class $\mathcal{G}$ which will restrict the search space for candidate prediction functions, or imposing a penalty Ω on the prediction function in order to favor "simple" objects. This second approach is refered to penalized or regularized empirical risk or inductive bias, we construct decision functions by solving the following problem

$$\min_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n} \left(g(X_i) - Y_i\right)^2 + \Omega(g)$$

This notions of course generalizes to other notions of risk than the square loss. This formulation of the problem of estimating a prediction function underlines the core importance of optimization to compute those estimates.

Linear regression:

In the statistical linear regression setting, a generating process on (X, Y) is assumed such that $f^*\colon x \mapsto \mathbb{E}[Y|X = x]$ has the form $x \mapsto \theta^T x$ for some $\theta \in \mathbb{R}^p$. This hypothesis is very strong and can be seen as invalid or impossible to verify in most pracitcal situations. However, one may view this as a simplified toy model which allow to get hands on high dimensional phenomena in statistics. In learning theory, it is custom not to assume much on the generative process and in this case one can decompose the L_2 error as follows

$$\|\hat{f}_n - f^*\|_2^2 = \|\hat{f}_n - \bar{f}\|_2^2 + \|\bar{f} - f^*\|_2^2$$

where $\bar{f}$ is the projection of f^* on the subspace of $L^2_{dP_X}$ consisting of linear functions. The second term is deterministic so that it is sufficient to consider $\|\hat{f}_n - \bar{f}\|_2^2$ and obtain bounds of the form

$$\mathbb{E}\left[\|\hat{f}_n - f^*\|_2^2\right] \leq \|\bar{f} - f^*\|_2^2 + \phi(n), \quad \forall n \in \mathbb{N}.$$

Such bounds are called oracle inequalities because they refer to the unknown oracle $\bar{f}$ which is the best one can hope when considering only linear models.

Chapter 2

Sub Gaussian random variables

This chapter is mostly based on [51, Chapter 1] and the expository note in [52]. See the tutorial course [33] for a broader view on the topic of concentration of measure.

2.1 Introduction and characterization

2.1.1 Gaussian concentration

The centered Gaussian random variable X on $\mathbb{R}$ with variance $\sigma^2 > 0$ has density given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right).$$

It plays a central role in statistics due to the central limit theorem. It also has a central position in statistical and signal processing estimation problems. Important properties of this distribution is closure under addition of iid replicates and concentration (Mill's inequality), if X is $\mathcal{N}\left(0, \sigma^2\right)$, we have, for any $t > 0$, $\mathbb{P}\left(|X| \geq t\right) \leq \frac{\sigma\sqrt{2}}{t\sqrt{\pi}} \exp\left(\frac{-t^2}{2\sigma^2}\right)$.

Proof.

$$\begin{aligned}
\mathbb{P}\left(|X| \geq t\right) &\leq 2\mathbb{P}\left(X \geq t\right) && \text{(symmetry and union bound)} \\
&= \frac{\sqrt{2}}{\sqrt{\pi\sigma^2}} \int_t^{+\infty} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx \\
&\leq \frac{\sigma^2\sqrt{2}}{\sqrt{\pi\sigma^2}} \int_t^{+\infty} \frac{x}{\sigma^2 t} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx \\
&= \frac{\sigma\sqrt{2}}{t\sqrt{\pi}} \int_t^{+\infty} -\frac{\partial}{\partial x} \exp\left(\frac{-x^2}{2\sigma^2}\right) dx \\
&= \frac{\sigma\sqrt{2}}{t\sqrt{\pi}} \exp\left(\frac{-t^2}{2\sigma^2}\right).
\end{aligned}$$

□

Sub Gaussian random variables are constrained to concentrate in a similar way which is sufficient for many purposes.

2.1.2 Equivalent definitions

The following provides qualitatively equivalent definitions for sub Gaussianity with variance proxy $\sigma^2 > 0$ (up to multiplicative constants).

Theorem 2.1.1. *Let X be a centered random variable on $\mathbb{R}$, each statement bellow implies the next (we take $\sigma^2 > 0$ in the first definition as a variance proxy). The first one can be taken as the definition of sub gaussian random variable.*

- *Laplace transform: for any $s \in \mathbb{R}$, $\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{\sigma^2 s^2}{2}\right)$.*
- *Concentration: for any $t > 0$, $\max\{\mathbb{P}(X \geq t), \mathbb{P}(X \leq -t)\} \leq \exp\left(\frac{-t^2}{2\sigma^2}\right)$.*
- *Moment condition: for any $q \in \mathbb{N}^*$, $\mathbb{E}\left[X^{2q}\right] \leq q!(4\sigma^2)^q$.*
- *Orlicz condition: $\mathbb{E}\left[\exp\left(\frac{X^2}{8\sigma^2}\right)\right] \leq 2$.*
- *Laplace transform: for any $t \in \mathbb{R}$, $\mathbb{E}[\exp(tX)] \leq \exp\left(\frac{24\sigma^2 t^2}{2}\right)$.*

Proof. The first implication is through Chernov's bound which is a consequence of Markov's inequality, for any $s > 0$, $t > 0$:

$$\begin{aligned}\mathbb{P}(X > t) &= \mathbb{P}(\exp(sX) > \exp(st)) \\ &\leq \frac{\mathbb{E}[\exp(sX)]}{\exp(st)} \\ &\leq \exp\left(\frac{\sigma^2 s^2}{2} - st\right),\end{aligned}$$

where the last inequalitie uses the Laplace transform condition. The result follows from the fact that $\min_{s>0} \frac{\sigma^2 s^2}{2} - st = \frac{-t^2}{2\sigma^2}$ attained for $s = t/\sigma^2$. For the second implication, we have, for any $q \in \mathbb{N}$,

$$\begin{aligned}\mathbb{E}\left[X^{2q}\right] &= \int_0^{+\infty} \mathbb{P}\left(Z^{2q} > u\right) du \\ &= \int_0^{+\infty} \mathbb{P}\left(|Z| > u^{1/2q}\right) du \\ &\leq 2\int_0^{+\infty} \exp\left(\frac{-u^{1/q}}{2\sigma^2}\right) du \\ &= \left(2\sigma^2\right)^q 2q \int_0^{+\infty} \exp(-v)\, v^{q-1} dv && v = \frac{u^{1/q}}{2\sigma^2} \\ &= \left(2\sigma^2\right)^q 2q q! \\ &\leq \left(4\sigma^2\right)^q q! && 2q \leq 2^q\end{aligned}$$

The next implication follows from the monotone convergence theorem. We obtain

$$\mathbb{E}\left[\exp\left(\frac{X^2}{8\sigma^2}\right)\right] = \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{X^{2k}}{\left(4\sigma^2\right)^k k!} \frac{1}{2^k}\right] \leq \sum_{k=0}^{\infty} \frac{1}{2^k} = 2$$

Getting back to the first item is done as follows, for any $s \in \mathbb{R}$, using the fact that X is centered,

for any $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(tX\right)\right] &= \mathbb{E}\left[\sum_{k=0}^{+\infty} \frac{(tX)^k}{k!}\right] \\
&= 1 + \mathbb{E}\left[\sum_{k=2}^{+\infty} \frac{(tX)^k}{k!}\right] && \mathbb{E}[X] = 0 \\
&\leq 1 + \frac{t^2}{2}\mathbb{E}\left[X^2 \exp\left(|tX|\right)\right] && \frac{(tX)^k}{k!} \leq \frac{t^2X^2}{2}\frac{|tX|^{k-2}}{(k-2)!},\ k \geq 2 \\
&\leq 1 + \frac{t^2}{2}\exp\left(4\sigma^2t^2\right)\mathbb{E}\left[X^2 \exp\left(\frac{X^2}{16\sigma^2}\right)\right] && \inf_a\left\{\frac{t^2}{2a} + \frac{aX^2}{2}\right\} = t|X|,\ a = \frac{1}{8\sigma^2} \\
&\leq 1 + 4\sigma^2t^2\exp\left(4\sigma^2t^2\right)\mathbb{E}\left[\exp\left(\frac{X^2}{8\sigma^2}\right)\right] && z \leq \exp\left(\frac{z}{2}\right) \\
&\leq \left(1 + 8\sigma^2t^2\right)\exp\left(4\sigma^2t^2\right) \\
&\leq \exp\left(\frac{24\sigma^2t^2}{2}\right) && (1+2z) \leq e^{2z}
\end{aligned}$$

□

A first exercise is to show that if X is sub gaussian with variance proxy σ^2, then aX is sub gaussian with variance proxy $a^2\sigma^2$.

2.1.3 Examples

Sub Gaussian random variables exist, for example the Gaussian random variable is subgaussian. Hoeffding's Lemma (1963) asserts that bounded random variables are also sub Gaussian.

Lemma 2.1.1. *Let X be a real centered random variable such that $X \in [a, b]$ almost surely. Then $\mathbb{E}[\exp(sX)] \leq \exp\left(s^2\frac{(b-a)^2}{8}\right)$ for any $s \in \mathbb{R}$, or X is sub Gaussian with variance proxy $\frac{(b-a)^2}{4}$.*

Proof. Consider the cumulent generating function $\psi\colon s \mapsto \log\left(\mathbb{E}\left[\exp\left(sX\right)\right]\right)$, we have

$$\psi'(s) = \frac{\mathbb{E}\left[X\exp\left(sX\right)\right]}{\mathbb{E}[\exp\left(sX\right)]} \qquad \psi''(s) = \frac{\mathbb{E}\left[X^2\exp\left(sX\right)\right]}{\mathbb{E}[\exp\left(sX\right)]} - \left(\frac{\mathbb{E}\left[X\exp\left(sX\right)\right]}{\mathbb{E}[\exp\left(sX\right)]}\right)^2$$

and ψ'' is the variance under the law of X reweighted by $\frac{\exp(sX)}{\mathbb{E}[\exp(sX)]}$. For any random variable Z in $[a, b]$, we have $\operatorname{var}[Z] = \operatorname{var}\left[Z - \frac{a+b}{2}\right] \leq \frac{(b-a)^2}{4}$. We can integrate two times using $\psi(0) = \log(1) = 0$ and $\psi'(0) = \mathbb{E}[X] = 0$. □

2.1.4 Sub Gaussian vectors

The definition extends similarly as for the Gaussian case.

Definition 2.1.1. *A random vector $X \in \mathbb{R}^d$ is said to be sub Gaussian with variance proxy σ^2 if it is centered and for any $u \in \mathbb{R}^d$ such that $\|u\| = 1$, the real random variable u^TX is subgaussian with variance proxy σ^2. We write $X \sim \mathrm{subG}(\sigma^2)$.*

There exists such random vectors, for example

Theorem 2.1.2. *Let $X_1, \ldots, X_p$ be independant $\mathrm{subG}(\sigma^2)$ real random variables then the random vector $X \in \mathbb{R}^p$ which i-th coordinates is X_i, is $\mathrm{subG}(\sigma^2)$.*

Proof. For any $u \in \mathbb{R}^p$ such that $\|u\| = 1$, we have for any $s \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left(su^T X\right)\right] = \prod_{i=1}^{p} \mathbb{E}\left[\exp\left(su_i X_i\right)\right] \leq \prod_{i=1}^{p} \exp\left(\frac{\sigma^2 s^2 u_i^2}{2}\right) = \exp\left(\frac{\sigma^2 s^2 \|u\|^2}{2}\right) = \exp\left(\frac{\sigma^2 s^2}{2}\right)$$

□

This allows to obtain various concentration results for sub Gaussian random variables.

2.2 Maximal inequalities

We first provide tail bounds for maximum of a finite number of subgaussian random variables and then over polytopes and Euclidean ball.

Theorem 2.2.1. *Let $X_1, \ldots, X_N$ be N real random variables with $X_i \sim \mathrm{subG}(\sigma^2)$, $i = 1, \ldots, N$, not necessarily independant. Then*

$$\mathbb{E}\left[\max_{i=1,\ldots,N} X_i\right] \leq \sigma\sqrt{2\log(N)} \qquad \textit{and} \qquad \mathbb{E}\left[\max_{i=1,\ldots,N} |X_i|\right] \leq \sigma\sqrt{2\log(2N)}$$

and for any $t > 0$

$$\mathbb{P}\left[\max_{i=1,\ldots,N} X_i > t\right] \leq N\exp\left(\frac{-t^2}{2\sigma^2}\right) \qquad \textit{and} \qquad \mathbb{P}\left[\max_{i=1,\ldots,N} |X_i| > t\right] \leq 2N\exp\left(\frac{-t^2}{2\sigma^2}\right)$$

Proof. For any $s > 0$

$$\begin{aligned}
\mathbb{E}\left[\max_{i=1,\ldots,N} X_i\right] &= \frac{1}{s}\mathbb{E}\left[\log\left(\exp\left(s\max_{i=1,\ldots,N} X_i\right)\right)\right] \\
&\leq \frac{1}{s}\log\left(\mathbb{E}\left[\exp\left(s\max_{i=1,\ldots,N} X_i\right)\right]\right) \qquad \text{(Jensen)} \\
&= \frac{1}{s}\log\left(\mathbb{E}\left[\max_{i=1,\ldots,N} \exp\left(sX_i\right)\right]\right) \\
&\leq \frac{1}{s}\log\left(\mathbb{E}\left[\sum_{i=1}^{N} \exp\left(sX_i\right)\right]\right) \\
&\leq \frac{1}{s}\log\left(\sum_{i=1}^{N} \exp\left(\frac{s^2\sigma^2}{2}\right)\right) \\
&= \frac{\log(N)}{s} + \frac{s\sigma^2}{2}.
\end{aligned}$$

The result follows by taking $s = \sqrt{2\log(N)/\sigma^2}$. The result on the deviation probability is a simple union bound and the results on the absolute value follows by applying the two previous results to the $2N$ random variables $X_1, \ldots, X_N, -X_1, \ldots, -X_N$. □

Remark 2.2.1. *For any $\delta > 0$, by taking $t = \sigma\sqrt{2\log(2N/\delta)}$, it holds with probability at least $1 - \delta$,*

$$\max_{i=1\ldots N} |X_i| \leq \sigma\sqrt{2\log(2N/\delta)}.$$

We will conclude this chapter by providing a bound for the maximum over an L_2 ball: if $X \in \mathbb{R}^p$ is $\mathrm{subG}(\sigma^2)$, can we control $\max_{\|c\|\leq 1} c^T X$? We begin with a Lemma.

Lemma 2.2.1. *For any $\epsilon \in (0,1)$, it is possible to cover the Euclidean unit ball in $\mathbb{R}^p$ by at most $(3/\epsilon)^p$ Euclidean balls of radius ϵ.*

Proof. Build a covering iteratively, start with the unit ball of radius ϵ centered at 0, $\mathcal{S} = \{0\}$ and while there exists x, $\|x\| \leq 1$ and $\text{dist}(x, \mathcal{S}) > \epsilon$, add such an x to $\mathcal{S}$. After $\mathcal{N}$ iterations, call $x_1, \dots, x_{\mathcal{N}}$ the elements of S.

We clearly have that the balls of radius $\epsilon/2$ centered at the points in $\mathcal{S}$ are disjoint and contained in the euclidean ball of radius $1 + \epsilon/2$. Computing volumes, we obtain

$$\mathcal{N}\left(\frac{\epsilon}{2}\right)^p \leq \left(1 + \frac{\epsilon}{2}\right)^p.$$

Hence the process must stop after at most $\left(\frac{2}{\epsilon} + 1\right)^p \leq \left(\frac{3}{\epsilon}\right)^p$ iterations at which point we obtain a cover. □

This allows to prove the following result

Theorem 2.2.2. *Let $X \sim \text{subG}(\sigma^2)$ be a d dimensional random vector. Then*

$$\mathbb{E}\left[\max_{\|c\| \leq 1} c^T X\right] = \mathbb{E}\left[\max_{\|c\| \leq 1} |c^T X|\right] \leq 4\sigma\sqrt{d}$$

and for any $t > 0$

$$\mathbb{P}\left[\max_{\|c\| \leq 1} |c^T X| > t\right] = \mathbb{P}\left[\max_{\|c\| \leq 1} c^T X > t\right] \leq 6^d \exp\left(\frac{-t^2}{8\sigma^2}\right).$$

Proof. Consider a covering of the unit Euclidean ball with at most 6^d balls of radius $1/2$, denote by $x_1, \dots, x_{6^d}$ the centers of these balls. For any c such that $\|c\| \leq 1$, there exists i such that $\|c - x_i\| \leq \frac{1}{2}$. Hence we have

$$\max_{\|c\| \leq 1} c^T X \leq \max_{i=1,\dots,6^d} x_i^T X + \max_{\|c\| \leq 1/2} c^T X = \max_{i=1,\dots,6^d} x_i^T X + \frac{1}{2} \max_{\|c\| \leq 1} c^T X$$

and hence $\max_{\|c\| \leq 1} c^T X \leq \max_{i=1,\dots,6^d} 2x_i^T X$ and the result follows from Theorem 2.2.1 because $2x_i^T X \sim \text{subG}(4\sigma^2)$ and $\log(6) \leq 2$. □

Remark 2.2.2. *For any $\delta > 0$, taking $t = \sqrt{8\log(6)}\sigma\sqrt{p} + 2\sigma\sqrt{2\log(1/\delta)}$, we obtain that with probability $1 - \delta$, it holds that*

$$\max_{\|c\| \leq 1} c^T X = \max_{\|c\| \leq 1} |c^T X| \leq 4\sigma\sqrt{p} + 2\sigma\sqrt{2\log(1/\delta)} = 4\sigma\sqrt{p}\left(1 + \sqrt{\frac{\log(1/\delta)}{2p}}\right).$$

Theorem 2.2.3. *Let P be a polytope, the convex hull of N points, $v^{(1)}, \dots, v^{(N)}$ in $\mathbb{R}^d$. Let $X \in \mathbb{R}^d$ be a random variable such that for all $i = 1, \dots, n$, $[v^{(i)}]^T X \sim \text{subG}(\sigma^2)$, then the conclusion of Theorem 2.2.1 holds*

$$\mathbb{E}\left[\max_{\theta \in P} \theta^T X\right] \leq \sigma\sqrt{2\log(N)} \qquad \textit{and} \qquad \mathbb{E}\left[\max_{\theta \in P} |\theta^T X|\right] \leq \sigma\sqrt{2\log(2N)}$$

and for any $t > 0$

$$\mathbb{P}\left[\max_{\theta \in P} \theta^T X > t\right] \leq N \exp\left(\frac{-t^2}{2\sigma^2}\right) \qquad \textit{and} \qquad \mathbb{P}\left[\max_{\theta \in P} |\theta^T X| > t\right] \leq 2N \exp\left(\frac{-t^2}{2\sigma^2}\right)$$

Exercises

Exercise 2.2.1. *Under the setting of Theorem 2.1.2, show that for any $t > 0$, we have*

$$\mathbb{P}\left[\frac{1}{p}\sum_{i=1}^{p} X_i \geq t\right] \leq \exp\left(\frac{-t^2 p}{2\sigma^2}\right).$$

Exercise 2.2.2. *Let Z be a real random variable with probability measure P_z on $\mathbb{R}$ such that $Z \geq 0$ almost surely. Show that*

$$\mathbb{E}[Z] = \int_0^{+\infty} \mathbb{P}(Z > u) du.$$

(Hint: use Funini's theorem. Beware: we did not assume that $\mathbb{E}[Z]$ is finite).

Exercise 2.2.3. *For $\mathbb{X} \in \mathbb{R}^{n\times d}$ and $Y \in \mathbb{R}^n$, the least squares estimator is written as*

$$\hat{\theta}^{LS} \in \arg\min_{\theta\in\mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2. \tag{2.1}$$

We have $\mathbb{X}^T\mathbb{X}\hat{\theta}^{LS} = \mathbb{X}^T Y$ and one solution is given by $\hat{\theta}^{LS} = (\mathbb{X}^T\mathbb{X})^\dagger \mathbb{X}^T Y$, where $\dagger$ denotes the Moore-Penrose pseusdo inverse. (Hint: First assume that $\mathbb{X}^T\mathbb{X}$ is invertible, the pseudo inverse is then the usual matrix inverse. If you are familiar with convex analysis, the result can be deduced from convexity of the ojective, solving the first order conditions)

Recall that if D is diagonal, then its pseudo inverse is obtained by inverting the non zero diagonal elements (leaving the others unchainged). Pseudo inverse of real symmetric matrices are defined in the same way after diagonalization.

Exercise 2.2.4. *Let X be $\mathcal{N}(0, \sigma^2)$, prove that for any $t > 0$, $\mathbb{P}(|X| \geq t) \leq \frac{\sigma\sqrt{2}}{t\sqrt{\pi}} \exp\left(\frac{-t^2}{2\sigma^2}\right)$. This is called Mill's inequality.*

Exercise 2.2.5. *Let $v^{(1)}, \ldots, v^{(N)} \in \mathbb{R}^d$ and set*

$$P = \mathrm{conv}(v^{(1)}, \ldots, v^{(N)}) = \left\{\sum_{i=1}^{N} \lambda_i v^{(i)}, \quad \lambda_i \geq 0,\, i = 1, \ldots, N,\, \sum_{i=1}^{N} \lambda_i = 1\right\}$$

Show that for any $c \in \mathbb{R}^d$, the problem $\sup_{\theta\in P} c^T\theta$ is attained at $v^{(i)}$ for some $i \in \{1, \ldots, N\}$. Prove Theorem 2.2.3.

Exercise 2.2.6. *Let $X \sim \mathrm{subG}(\sigma^2)$ be a d-dimensional random vector, show that, for any $\delta > 0$, with probability $1 - \delta$,*

$$\sup_{\|\theta\|_1 \leq 1} |\theta^T X| \leq \sigma\sqrt{2\log(2d/\delta)}.$$

Exercise 2.2.7. *Let $A \in \mathbb{R}^{n\times m}$ be a random matrix which entries are iid subgaussian with variance proxy σ^2. The operator norm of A is given by $\|A\|_{op} = \sup_{x\in\mathbb{R}^m} \|Ax\|_2/\|x\|_2$. Show that $\mathbb{E}[\|A\|_{op}] \leq c\sigma\sqrt{m+n}$ for a constant c to be determined.*

Exercise 2.2.8. *Prove Jensen's inequality, if $D \subset \mathbb{R}$ is an interval and $\phi\colon D \mapsto \mathbb{R}$ is concave continuous, if X is a real random variable such that $X \in D$ with probability 1, then $\mathbb{E}[\phi(X)] \leq \phi(\mathbb{E}[X])$.*

Chapter 3

Linear regression

This chapter is mostly based on [51, Chapter 2]. Further reading include [51, Chapter 3,4],

3.1 Introduction

We consider a generative model of the following form $Y_i = f^*(X_i) + \epsilon_i$, $i = 1 \dots, n$, where $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T \sim \text{subG}(\sigma^2)$ and $\mathbb{E}[\epsilon] = 0$. The regression function $f^* \colon x \mapsto \mathbb{E}[Y|X = x]$ is assumed to be of the form $f^* \colon x \mapsto x^T\theta^*$ for an unknown $\theta^* \in \mathbb{R}^d$. This generative model is assumed to hold true throughout the chapter.

Design points:

The sample points $X_1, \dots, X_n$ are called *design* points. Depending on the nature of these points one may consider different ways to measure the quality of an estimate.

Random design: The design points are random, given $\mathcal{D}_n$ and a new observation X_{n+1}, one would like to build a predictor $\hat{f}_n$ for Y_{n+1}. In this case $R(\hat{f}_n)$ is a relevant measure.

Fixed design: If the design points are not random, one talks about fixed design and we denote the design points by $x_1, \dots, x_n$. In this situation, there is not much interest in talking about risk or expected prediction errror, since there is no expectation to consider. In this situation, we will consider for any g the mean squared error:

$$\text{MSE}(g) = \frac{1}{n}\sum_{i=1}^{n} (g(x_i) - f^*(x_i))^2$$

We denote by $\mathbb{X} \in \mathbb{R}^{n \times d}$ the design matrix for which each row is one of the design points. Our model can then be expressed as follows:

$$Y = \mathbb{X}\theta^* + \epsilon \tag{LM}$$

where $Y = (Y_1, \dots, Y_n)^T$ and $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$. In the sequel, we will focus on fixed designs. In this case, the mean squared error is given for any $\theta \in \mathbb{R}^d$, by

$$\text{MSE}(\theta) = \frac{1}{n}\|\mathbb{X}(\theta - \theta^*)\|_2^2.$$

3.2 Least squares and constrained least squares with fixed design

Least squares estimator

The least squares estimator is given by

$$\hat{\theta}^{LS} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - Y\|_2^2 \tag{3.1}$$

where we use the Euclidean norm. We start with an algebraic expression for $\hat{\theta}^{LS}$.

Lemma 3.2.1. *We have*

$$\mathbb{X}^T\mathbb{X}\hat{\theta}^{LS} = \mathbb{X}^T Y$$

and one solution is given by $\hat{\theta}^{LS} = (\mathbb{X}^T\mathbb{X})^\dagger \mathbb{X}^T Y$*, where* $\dagger$ *denotes the Moore-Menrose pseusdo inverse.*

Proof. The matrix $\mathbb{X}^T\mathbb{X}$ is positive semidefinite so that the objective in (3.1) is a convex quadratic function of θ. A necessary and sufficient condition for global optimality is that the gradient vanishes. This is the first claim and the second one follows from properties of the pseudoinverse. □

3.2.1 Constrained least squares estimator

Let K denote a closed subset of $\mathbb{R}^d$, the K constrained least squares estimator is given by

$$\hat{\theta}_K^{LS} \in \arg\min_{\theta \in K} \|\mathbb{X}\theta - Y\|_2^2 \tag{3.2}$$

where we use the Euclidean norm. The following lemma will be useful to prove finite sample bounds for $\hat{\theta}_K^{LS}$. The difficulty in bounding mean squared errors comes from the randomness of $\hat{\theta}^{LS}$, here we bound the MSE by a product of the noise and a quantity which can be controled uniformly. The question of how to compute constrained least squares estimates will be the topic of further chapters.

3.3 Finite sample bounds for least squares

We start with a general Lemma for constrained least squares estimators.

Lemma 3.3.1. *Let* $K \subset \mathbb{R}^d$ *be closed and* $g\colon \mathbb{R}^d \mapsto \mathbb{R}$ *denote any function. Assume that model* (LM) *holds and that* $\theta^* \in K$*, and set, assuming that the infimum is attained*

$$\hat{\theta}_{Kg}^{LS} \in \arg\min_{\theta \in K} \|\mathbb{X}\theta - Y\|_2^2 + g(\theta).$$

Then, almost surely

$$\|\mathbb{X}(\hat{\theta}_{Kg}^{LS} - \theta^*)\|_2^2 \leq 2\epsilon^T\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*) + g(\theta^*) - g(\hat{\theta}_{Kg}^{LS}).$$

Proof. Since $\theta^* \in K$ and we have by definition of $\hat{\theta}_{Kg}^{LS}$,

$$\|\mathbb{X}\hat{\theta}_{Kg}^{LS} - Y\|_2^2 + g(\hat{\theta}_{Kg}^{LS}) \leq \|\mathbb{X}\theta^* - Y\|_2^2 + g(\theta^*) = \|\epsilon\|_2^2 + g(\theta^*).$$

Furthermore, it holds that

$$\begin{aligned}\|\mathbb{X}\hat{\theta}_{Kg}^{LS} - Y\|_2^2 &= \|\mathbb{X}\hat{\theta}_{Kg}^{LS} - \mathbb{X}\theta^* - \epsilon\|_2^2 \\ &= \|\mathbb{X}\hat{\theta}_{Kg}^{LS} - \mathbb{X}\theta^*\|_2^2 - 2\epsilon^T\mathbb{X}(\hat{\theta}_{Kg}^{LS} - \theta^*) + \|\epsilon\|_2^2\end{aligned}$$

So that

$$\begin{aligned}\|\mathbb{X}(\hat{\theta}_{Kg}^{LS} - \theta^*)\|_2^2 &= \|\mathbb{X}\hat{\theta}_{Kg}^{LS} - Y\|_2^2 - \|\epsilon\|_2^2 + 2\epsilon^T\mathbb{X}(\hat{\theta}_{Kg}^{LS} - \theta^*) \\ &\leq 2\epsilon^T\mathbb{X}(\hat{\theta}_{Kg}^{LS} - \theta^*) + g(\theta^*) - g(\hat{\theta}_{Kg}^{LS}).\end{aligned}$$

□

Unconstrained least squares

The following result provides mean squared error estimates for the least squares estimator.

Theorem 3.3.1. *Assume that* (LM) *holds with* $\epsilon \sim \text{subG}(\sigma^2)$*, then*

$$\mathbb{E}\left[\text{MSE}(\hat{\theta}^{LS})\right] \leq 16\sigma^2\frac{r}{n}$$

where $r = \text{rank}(\mathbb{X}^T\mathbb{X})$*, furthermore, for any* $\delta > 0$*, with probability at least* $1 - \delta$*,*

$$\text{MSE}(\hat{\theta}^{LS}) \leq \frac{64\sigma^2\left(2r + \log(1/\delta)\right)}{n}$$

Proof. Denote by $\Phi \in \mathbb{R}^{n\times r}$ a matrix which column constitute an orthonormal basis of the column span of $\mathbb{X}$. One may write $\mathbb{X}\left(\hat{\theta}^{LS} - \theta^*\right) = \Phi\nu$ where $\nu \in \mathbb{R}^r$. We have

$$\frac{\epsilon^T\mathbb{X}(\hat{\theta}^{LS} - \theta^*)}{\|\mathbb{X}(\hat{\theta}^{LS} - \theta^*)\|_2} = \frac{\epsilon^T\Phi\nu}{\Phi\nu} = \left(\epsilon^T\Phi\right)\frac{\nu}{\|\nu\|_2} \leq \|\Phi^T\epsilon\|_2.$$

Applying Lemma 3.3.1, with $K = \mathbb{R}^d$, we have

$$\|\mathbb{X}(\hat{\theta}^{LS} - \theta^*)\|_2^2 \leq 4\left(\frac{\epsilon^T\mathbb{X}(\hat{\theta}^{LS} - \theta^*)}{\|\mathbb{X}(\hat{\theta}^{LS} - \theta^*)\|_2}\right)^2 \leq 4\|\Phi^T\epsilon\|_2^2 = 4\sum_{i=1}^{r}(\Phi_i^T\epsilon)^2,$$

where Φ_i denotes the i-th column of Φ, $i = 1, \ldots, n$. Note that $\Phi_i^T\epsilon \sim \text{subG}(\sigma^2)$ by orthonormality of the Columns of Φ and Theorem 2.1.2 for $i = 1\ldots, r$ and hence using Theorem 2.1.1, we have

$$\mathbb{E}\left[\text{MSE}(\hat{\theta}^{LS})\right] \leq \frac{4}{n}\sum_{i=1}^{r}(\Phi_i^T\epsilon)^2 \leq \frac{16r\sigma^2}{n}.$$

This concludes the bound in expectation. For the bound in probability, we remakr that $\|\Phi^T\epsilon\|_2 = \max_{\|u\|\leq 1} u^T\Phi^T\epsilon$ where $\Phi^T\epsilon \sim \text{subG}(\sigma^2)$. Theorem 2.2.2 and Remark 2.2.2 entails for any $\delta > 0$, with probability at least $1 - \delta$,

$$\begin{aligned}\text{MSE}(\hat{\theta}^{LS}) &\leq \frac{4}{n}\left(4\sigma\sqrt{r} + 2\sigma\sqrt{2\log(1/\delta)}\right)^2 \\ &\leq \frac{64\sigma^2\left(2r + \log(1/\delta)\right)}{n}\end{aligned}$$

□

Optimality and high dimensional setting

A natural question arising about Theorem 3.3.1 is "could we do better?". If d is the number of variables an $\mathbb{X}$ has full possible rank, then $r = \min(n, d) = d$ assuming $n \geq d$. We obtain a rate of the order of $\sigma^2 d/n$. In this case, we have

$$\text{MSE}(\hat{\theta}^{LS}) = (\hat{\theta}^{LS} - \theta^*)^T\frac{\mathbb{X}^T\mathbb{X}}{n}(\hat{\theta}^{LS} - \theta^*) \geq \lambda_{\min}\left(\frac{\mathbb{X}^T\mathbb{X}}{n}\right)\|\hat{\theta}^{LS} - \theta^*\|_2^2.$$

It turns out that this rate is optimal in a precise minimax sense.

Theorem 3.3.2. *Suppose that $Y = \xi + \theta$ where $\theta \in \mathbb{R}^d$ and $\xi_i \sim \mathcal{N}(0, \sigma^2/n)$, $i = 1, \dots, d$. Then, for any $\alpha \in (0, 1/4)$:*

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{P}_\theta \left(\|\hat{\theta} - \theta\|_2^2 \geq \frac{\alpha}{256} \frac{\sigma^2 d}{n} \right) \geq \frac{1}{2} - 2\alpha$$

where the infimum is taken over all measurable functions of Y.

The proof of this statement, can be done by reduction to statistical hypothesis testing and use known impossibility results to discriminate between two close hyptheses (See Chapter 4 of Philippe Rigollet's notes). Note that in the specific Gaussian sequence model proposed in the Theorem, the order of decay predicted by Theorem 3.3.1 is precisely $\sigma^2 d/n$. The theorem essentially says that for any estimator, there is a statistical setting for which this rate is attained. This type of result is called *minimax*. The conclusion is that the least squares estimator is optimal among all estimators without any prior knowledge.

In the high dimensional setting, we have $d \geq n$ and in this case, the bound of Theorem 3.3.1 remains bounded away from zero. Since this bound is optimal, it seems that there is no hope to solve high dimensional statistical problems. This is in fact not true, if one has for example prior knowledge that θ^* is in a certain ball of radius δ, then imposing that our estimaor $\hat{\theta}$ is in the same ball allows to estimate θ^* such that $\|\hat{\theta} - \theta^*\|^2 \leq \delta^2$. If δ is small, this may improve over the estimate of Theorem 3.3.1.

How is this compatible with Theorem 3.3.2? In the inf sup expression, the sup is taken over $\mathbb{R}^d$ and considering smaller subsets of $\mathbb{R}^d$ would reduce the right hand side.

ℓ_1 constrained least squares

We let B_1 denote the unit ball of the ℓ_1 norm in $\mathbb{R}^d$,

$$B_1 = \left\{ x \in \mathbb{R}^d, \sum_{i=1}^d |x_i| \leq 1 \right\}.$$

This is a polytope with $2d$ vertices given by the elements of the canonical basis and their oposite. The following result shows that under prior knowledge on θ^*, one can hope for better rates.

Theorem 3.3.3. *Let $K = B_1$ and $d \geq 2$. Assume that model* (LM) *holds with $\epsilon \sim \mathrm{subG}(\sigma^2)$ and $\theta^* \in K$. Assume also that the columns of $\mathbb{X}$ are normalized such that $\|\mathbb{X}_j\| \leq \sqrt{n}$, $j =, 1 \dots, d$. Then, it holds that*

$$\mathbb{E}\left[\mathrm{MSE}(\hat{\theta}_K^{LS}) \right] \leq \frac{4\sigma}{\sqrt{n}} \sqrt{2 \log(2d)}$$

and for any $\delta > 0$, with probability at least $1 - \delta$, it holds that

$$\mathrm{MSE}(\hat{\theta}_K^{LS}) \leq \sigma \sqrt{\frac{32 \log(2d/\delta)}{n}}.$$

Proof. Invoking Lemma 3.3.1, we have

$$\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2^2 \leq 2\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*).$$

Note that since $\|\hat{\theta}_K^{LS}\|_1 \leq 1$ and $\|\theta^*\|_1 \leq 1$, we have $\|\hat{\theta}_K^{LS} - \theta^*\|_1 \leq 2$ so that

$$\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2^2 \leq 2 \sup_{\|v\|_1 \leq 2} \epsilon^T \mathbb{X} v = 4 \sup_{\|v\|_1 \leq 1} \epsilon^T \mathbb{X} v = 4 \sup_{u \in \mathbb{X}K} \epsilon^T u.$$

Now $\mathbb{X}K$ by linearity if v is not an extreme point of K then $\mathbb{X}v$ is not an extreme point of $\mathbb{X}K$. Hence $\mathbb{X}K$ is a polytope with at most $2d$ vertices which are taken among the columns of $\mathbb{X}$. The normalization of the columns of $\mathbb{X}$ ensures that on each of these vertices, $\mathbb{X}_j^T\epsilon \sim \text{subG}(\sigma^2 n)$. Applying Theorem 2.2.3, we have

$$\mathbb{E}\left[\text{MSE}(\hat{\theta}_K^{LS})\right] \leq \frac{4}{n}\sqrt{n}\sigma\sqrt{2\log(2d)} = \frac{4\sigma}{\sqrt{n}}\sqrt{2\log(2d)}.$$

Furthermore, for any $t > 0$, we have

$$\mathbb{P}\left[\text{MSE}(\hat{\theta}_K^{LS}) \geq t\right] \leq \mathbb{P}\left[\sup_{u\in\mathbb{X}K} \epsilon^T u \geq \frac{nt}{4}\right] \leq 2de^{-\frac{nt^2}{32\sigma^2}}.$$

Given any $\delta \geq 0$, one has

$$2de^{-\frac{nt^2}{32\sigma^2}} \leq \delta \quad \Leftrightarrow t^2 \geq \frac{32\sigma^2}{n}\log\left(\frac{2d}{\delta}\right),$$

and the conclusion follows. □

ℓ_0 constrained least squares

We refer to the ℓ_0 norm as the cardinality of the set of non zero coordinates of a vector $\theta \in \mathbb{R}^d$. Note that this is an abuse of notations since this is not a norm. For any $\theta \in \mathbb{R}^d$,

$$\|\theta\|_0 = \sum_{i=1}^{d} \mathbb{I}(\theta_j \neq 0).$$

A vector with small ℓ_0 norm is called sparse. The support of a vector is the set of indices of its nonzero coordinates:

$$\text{supp}(\theta) = \{j \in \{1,\dots,d\}, \theta_j \neq 0\},$$

so that $\|\theta\|_0 = \text{card}(\text{supp}(\theta))$. By extension, for any $k = 1,\dots,d$, we denote by $B_0(k)$ the set of k-sparse vectors.

Theorem 3.3.4. *For any $k \in \mathbb{N}^*$, $k \leq d/2$, let $K = B_0(k)$ and assume that model* (LM) *holds with $\epsilon \sim \text{subG}(\sigma^2)$ and $\theta^* \in K$. Then, for any $\delta > 0$, with probability $1-\delta$, it holds*

$$\text{MSE}(\hat{\theta}_K^{LS}) \leq \frac{32\sigma^2}{n}\left(\log\left(\binom{d}{2k}\right) + 2k\log(6) + \log(1/\delta)\right).$$

Furthermore, we have

$$\mathbb{E}\left[\text{MSE}(\hat{\theta}_K^{LS})\right] \leq \frac{32\sigma^2}{n}\left(1 + \log\left(\binom{d}{2k}\right) + 2k\log(6)\right)$$

Proof. Using Lemma 3.3.1, we have

$$\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2^2 \leq 4\frac{\left(\epsilon^T\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\right)^2}{\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2^2}.$$

We have $\|\hat{\theta}_K^{LS} - \theta^*\|_0 \leq 2k$ and we set $\hat{S} = \text{supp}\left(\hat{\theta}_K^{LS} - \theta^*\right)$, we have $|\hat{S}| \leq 2k$. We repeat similar steps as for the unconstrained least squares. For any $S \subset \{1,\dots,d\}$, denote by $\mathbb{X}_S \in \mathbb{R}^{n\times|S|}$

the matrix composed of the columns of $\mathbb{X}$ indexed by S, by r_S the rank of $\mathbb{X}_S$ and by Φ_S an orthonormal basis of the span of the collumns of $\mathbb{X}$. There exists $\nu \in \mathbb{R}^{r_{\hat{S}}}$, such that

$$\frac{\epsilon^T \mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)}{\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2} = \frac{\epsilon \Phi_{\hat{S}} \nu}{\|\nu\|} \leq \max_{|S|=2k} \max_{u \in \mathbb{R}^{r_S}, \|u\|_2 \leq 1} u^T \Phi_S^T \epsilon.$$

Using Theorem 2.1.2, for any S, $\Phi_S^T \epsilon \sim \text{subG}(\sigma^2)$. Using a union bound, and Theorem 2.2.2, for any $t > 0$, we have

$$\begin{aligned}
\mathbb{P}\left[\|\mathbb{X}(\hat{\theta}_K^{LS} - \theta^*)\|_2^2 \geq 4t\right] &\leq \mathbb{P}\left[\max_{|S|=2k} \max_{u \in \mathbb{R}^{r_S}, \|u\|_2 \leq 1} \left(u^T \Phi_S^T \epsilon\right)^2 > t\right] \\
&\leq \mathbb{P}\left[\max_{|S|=2k} \max_{u \in \mathbb{R}^{r_S}, \|u\|_2 \leq 1} \left|u^T \Phi_S^T \epsilon\right| > \sqrt{t}\right] \\
&\leq \sum_{|S|=2k} \mathbb{P}\left[\max_{u \in \mathbb{R}^{r_S}, \|u\|_2 \leq 1} |u^T \Phi_S^T \epsilon| > \sqrt{t}\right] \\
&\leq \sum_{|S|=2k} 6^{|S|} e^{\frac{-t}{8\sigma^2}} \\
&\leq \binom{d}{2k} 6^{2k} e^{\frac{-t}{8\sigma^2}}.
\end{aligned}$$

We deduce that

$$\mathbb{P}\left[\text{MSE}(\hat{\theta}^{LS}) \geq \frac{4t}{n}\right] \leq \binom{d}{2k} 6^{2k} e^{\frac{-t}{8\sigma^2}}$$

and we choose t such that the right hand side is bounded by δ, that is

$$t \geq 8\sigma^2 \left(\log\left(\binom{d}{2k}\right) + 2k\log(6) + \log(1/\delta)\right)$$

and the bound in probability follows. The expectation is deduced from the bound in probability. We have, for any $H \geq 0$, using

$$\begin{aligned}
\mathbb{E}\left[\text{MSE}(\hat{\theta}_K^{LS})\right] &= \int_0^{+\infty} \mathbb{P}\left[\text{MSE}(\hat{\theta}_K^{LS}) > u\right] du \\
&\leq H + \int_0^{+\infty} \mathbb{P}\left[\text{MSE}(\hat{\theta}_K^{LS}) \geq (u + H)\right] du \\
&\leq H + \binom{d}{2k} 6^{2k} \int_0^{+\infty} e^{\frac{-n(u+H)}{32\sigma^2}} du \\
&= H + \binom{d}{2k} 6^{2k} e^{\frac{-nH}{32\sigma^2}} \frac{32\sigma^2}{n}.
\end{aligned}$$

Inverting the relation

$$\binom{d}{2k} 6^{2k} e^{\frac{-nH}{32\sigma^2}} = 1,$$

we obtain

$$H = \frac{32\sigma^2}{n} \left(\log\left(\binom{d}{2k}\right) + 2k\log(6)\right)$$

and the result follows. □

Lemma 3.3.2. *For any $1 \leq k \leq n$, it holds*

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k$$

Proof. This is a simple recursion. □

As a consequence, the order of the bounds which we obtain is $\frac{\sigma^2 k}{n} \log\left(\frac{ed}{2k}\right)$. This also turns out to be minimax optimal for sparse estimation.

3.4 Penalized estimators

Adaptivity

Theorem 3.3.3 and 3.3.4 are very attractive since they provide fast decrease of the mean squared error in high dimensional settings. However, they require the knowledge of properties of the unknown θ^*. It is possible to produce adaptive estimators which do not require such knowledge.

Consider the sub-gaussian sequence model: $y = \theta^* + \xi \in \mathbb{R}^d$, where $\xi \sim \mathrm{subG}(\sigma^2/n)$. This allows to capture the intuition about penalization. Using Theorem 2.2.1 and Remark 2.2.1, we have for any $\delta > 0$, with probability at least $1 - \delta$

$$\max_{1 \leq i \leq d} |\xi_i| \leq \sigma \sqrt{\frac{2\log(2d/\delta)}{n}} = \tau.$$

If $|y_j| \gg \tau$ for some j, then it must correspond to $\theta_j^* \neq 0$. On the other hand, if $|y_j| \leq \tau$, then $|\theta_j^*| \leq |y_j| + |\xi_j| \leq 2\tau$ with high probability. This motivates the use of the following estimator, called the hard-thresholding estimator:

$$\hat{\theta}_j^{HT} = y_j \mathbb{I}(|y_j| \geq 2\tau), \quad j = 1, \ldots, d.$$

Indeed, conditioning on the event:

$$\mathcal{A} = \left\{ \max_i |\xi_i| \leq \tau \right\},$$

we have for all j, $|y_j| \geq 2\tau \Rightarrow |\theta_j^*| \geq \tau$ and $|y_j| \leq 2\tau \Rightarrow |\theta_j^*| \leq 3\tau$ and

$$\begin{aligned}
\|\hat{\theta}^{RT} - \theta^*\|^2 &= \sum_{i=1}^d \left(|y_i - \theta_i^*| \mathbb{I}(|y_i| \geq 2\tau) + |\theta_i^*| \mathbb{I}(|y_i| < 2\tau)\right)^2 \\
&\leq \sum_{i=1}^d \left(\tau \mathbb{I}(|\theta_i^*| \geq \tau) + (\theta_i^*) \mathbb{I}(|\theta_i^*| < 3\tau)\right)^2 \\
&\leq \sum_{i=1}^d \left(4 \min\left\{|\theta_j^*|^2, \tau\right\}\right)^2 \leq 16 \|\theta^*\|_0 \tau^2 = \frac{32 \|\theta\|_0 \sigma^2 \log(2d/\delta)}{n}.
\end{aligned}$$

Furthermore, if $\min_{j \in \mathrm{supp}(\theta^*)} |\theta_j^*| \geq 3\tau$, then $\mathrm{supp}(\hat{\theta}^{HT}) = \mathrm{supp}(\theta^*)$.

It turns out that $\hat{\theta}^{HT}$ is obtained by penalization using ℓ_0 pseudo norm ball:

$$\hat{\theta}^{HT} = \arg\min_{\theta \in \mathbb{R}^d} \|y - \theta\|^2 + 4\tau^2 \|\theta\|_0.$$

This is easily seen as if $|y_i| < 2\tau$ for some j, then $4\tau^2 \mathbb{I}(\theta_j \neq 0) > y_j^2$. This motivates the use of penalized estimators which are more adaptative to unknown properties of θ^*.

Under model (LM), we set, for any $\lambda \geq 0$,

$$\hat{\theta}^{\ell_0} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n}\|\mathbb{X}\theta - Y\|^2 + \lambda\|\theta\|_0$$
$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n}\|\mathbb{X}\theta - Y\|^2 + \lambda\|\theta\|_1$$

The second estimator is commonly called the Lasso estimator.

ℓ_0 penalized least squares

Theorem 3.4.1. *Assume that model* (LM) *holds with* $\epsilon \sim \mathrm{subG}(\sigma^2)$ *then choosing* $\lambda = 8\log(6)\sigma^2/n + 16\sigma^2\log(ed)/n$, *we have for any* $\delta > 0$ *with probability at least* $1-\delta$,

$$\mathrm{MSE}(\hat{\theta}^{\ell_0}) \leq \frac{32\sigma^2\left(2\|\theta^*\|_0\left(\log(6)+\log(ed)\right)+\log(1/\delta)+\log(2)\right)}{n}$$

Proof. We have by definition

$$\frac{1}{2n}\|\mathbb{X}\hat{\theta}^{\ell_0} - Y\|^2 + \lambda\|\hat{\theta}^{\ell_0}\|_0 \leq \frac{1}{2n}\|\mathbb{X}\theta^* - Y\|^2 + \lambda\|\theta^*\|_0.$$

Similarly as in Lemma 3.3.2, we have

$$\|\mathbb{X}\hat{\theta}^{\ell_0} - \mathbb{X}\theta^*\|^2 \leq 2\epsilon^T\mathbb{X}\left(\hat{\theta}^{\ell_0} - \theta^*\right) + 2n\lambda(\|\theta^*\|_0 - \|\hat{\theta}^{\ell_0}\|_0).$$

For any $a, b \in \mathbb{R}^d$, we have

$$2a^Tb = 2a^T\frac{b}{\|b\|_2}\|b\|_2 \leq 2\left(a^T\frac{b}{\|b\|_2}\right)^2 + \frac{1}{2}\|b\|_2^2,$$

and hence

$$\|\mathbb{X}\hat{\theta}^{\ell_0} - \mathbb{X}\theta^*\|^2 \leq 4\left(\frac{\epsilon^T\mathbb{X}\left(\hat{\theta}^{\ell_0} - \theta^*\right)}{\left\|\mathbb{X}\left(\hat{\theta}^{\ell_0} - \theta^*\right)\right\|_2^2}\right)^2 + 4n\lambda(\|\theta^*\|_0 - \|\hat{\theta}^{\ell_0}\|_0). \tag{3.3}$$

Setting $\mathcal{U}(\hat{\theta}^{\ell_0} - \theta^*) = \mathbb{X}\left(\hat{\theta}^{\ell_0} - \theta^*\right) / \left\|\mathbb{X}\left(\hat{\theta}^{\ell_0} - \theta^*\right)\right\|_2$, we have

$$\begin{aligned}
\left(\epsilon^T\mathcal{U}(\hat{\theta}^{\ell_0} - \theta^*)\right)^2 - n\lambda\|\hat{\theta}^{\ell_0}\|_0 &\leq \sup_{\theta\in\mathbb{R}^d}\left(\epsilon^T\mathcal{U}(\theta-\theta^*)\right)^2 - n\lambda\|\theta\|_0 \\
&\leq \max_{0\leq k\leq d}\max_{|S|=k}\sup_{\mathrm{supp}(\theta)=S}\left(\epsilon^T\mathcal{U}(\theta-\theta^*)\right)^2 - n\lambda k \\
&\leq \max_{0\leq k\leq d}\max_{|S|=k}\sup_{u\in\mathbb{R}^{r_{S*}},\|u\|_2\leq 1}\left(\epsilon^T\Phi_{S*}u\right)^2 - n\lambda k
\end{aligned}$$

where $\Phi_{S*} \in \mathbb{R}^{n\times r_{S*}}$ denotes an orthonormal basis of the span of the columns of $\mathbb{X}$ indexed by $S \cup \mathrm{supp}(\theta^*)$, and $r_{S*} \leq |S| + \|\theta^*\|_0$. For any $t > 0$, k and S with $|S| = k$, we have using Theorem 2.2.2.

$$\begin{aligned}
\mathbb{P}\left[4\sup_{u\in\mathbb{R}^{r_{S*}},\|u\|_2\leq 1}\left(\epsilon^T\Phi_{S*}u\right)^2 - 4n\lambda k > t\right] &= \mathbb{P}\left[\sup_{u\in\mathbb{R}^{r_{S*}},\|u\|_2\leq 1}\left\|\epsilon^T\Phi_{S*}u\right\| > \sqrt{\frac{t}{4}+n\lambda k}\right] \\
&\leq 6^{r_{S*}}\exp\left(-\frac{\frac{t}{4}+n\lambda k}{8\sigma^2}\right) \\
&\leq \exp\left(-\frac{t}{32\sigma^2} - \frac{n\lambda k}{8\sigma^2} + (k+\|\theta^*\|_0)\log(6)\right). \qquad (3.4)
\end{aligned}$$

Using the definition of λ, we have

$$\begin{aligned}-\frac{n\lambda k}{8\sigma^2} + (k + \|\theta^*\|_0)\log(6) &= -k\log(6) - 2k\log(ed) + (k + \|\theta^*\|_0)\log(6)\\ &= -2k\log(ed) + \|\theta^*\|_0\log(6).\end{aligned}$$

Using a union bound with (3.3) and (3.4), we obtain, for any $t > 0$,

$$\begin{aligned}&\mathbb{P}\left[\|\mathbb{X}\hat{\theta}^{\ell_0} - \theta^*\|_2^2 \geq 4n\lambda\|\theta^*\|_0 + t\right]\\ &\leq \sum_{k=0}^{d}\sum_{|S|=k}\exp\left(-\frac{t}{32\sigma^2} - 2k\log(ed) + \|\theta^*\|_0\log(6)\right)\\ &\leq \exp\left(-\frac{t}{32\sigma^2} + \|\theta^*\|_0\log(6)\right) + \sum_{k=1}^{d}\binom{d}{k}\exp\left(-\frac{t}{32\sigma^2} - 2k\log(ed) + \|\theta^*\|_0\log(6)\right)\\ &\leq \exp\left(-\frac{t}{32\sigma^2} + \|\theta^*\|_0\log(6)\right) + \sum_{k=1}^{d}\exp\left(-\frac{t}{32\sigma^2} - k\log(ed) + \|\theta^*\|_0\log(6)\right) \qquad \text{Lemma 3.3.2}\\ &\leq \exp\left(-\frac{t}{32\sigma^2} + \|\theta^*\|_0\log(6)\right) + \sum_{k=1}^{d}(ed)^{-k}\exp\left(-\frac{t}{32\sigma^2} + \|\theta^*\|_0\log(6)\right)\\ &\leq 2\exp\left(-\frac{t}{32\sigma^2} + \|\theta^*\|_0\log(6)\right)\end{aligned}$$

Choosing $t = 32\sigma^2\left(\log(1/\delta) + \|\theta^*\|_0\log(6) + \log(2)\right)$, the right hand side is equal to δ and we obtain that with probability $1-\delta$,

$$\begin{aligned}\|\mathbb{X}\hat{\theta}^{\ell_0} - \mathbb{X}\theta^*\|_2^2 &\leq 4n\lambda\|\theta^*\|_0 + t\\ &= 32\sigma^2\left(\|\theta^*\|_0\left(\log(6) + 2\log(ed)\right) + \left(\log(1/\delta) + \|\theta^*\|_0\log(6) + \log(2)\right)\right)\\ &= 32\sigma^2\left(2\|\theta^*\|_0\left(\log(6) + \log(ed)\right) + \log(1/\delta) + \log(2)\right)\end{aligned}$$

□

This is a very strong result as it provides an estimator which completely adapts to unknwon support, including its size.

ℓ_1 penalized least squares

Theorem 3.4.2. *Assume that model* (LM) *holds with* $\epsilon \sim \mathrm{subG}(\sigma^2)$. *Moreover assume that the columns of* $\mathbb{X}$ *have norm at most* $\sqrt{n}$. *Then, for any* $\delta > 0$, *choosing* $\lambda = \sigma/\sqrt{n}\left(\sqrt{2\log(2d)} + \sqrt{2\log(1/\delta)}\right)$, *we have for any* $\delta > 0$ *with probability at least* $1-\delta$,

$$\mathrm{MSE}(\hat{\theta}^{\ell_1}) \leq \frac{4\|\theta^*\|_1\sigma}{\sqrt{n}}\left(\sqrt{2\log(2d)} + \sqrt{2\log(1/\delta)}\right).$$

Proof. We have by definition

$$\frac{1}{2n}\|\mathbb{X}\hat{\theta}^{\ell_1} - Y\|^2 + \lambda\|\hat{\theta}^{\ell_1}\|_1 \leq \frac{1}{2n}\|\mathbb{X}\theta^* - Y\|^2 + \lambda\|\theta^*\|_1.$$

Similarly as in Lemma 3.3.2, we have

$$\|\mathbb{X}\hat{\theta}^{\ell_1} - \mathbb{X}\theta^*\|^2 \leq 2\epsilon^T\mathbb{X}\left(\hat{\theta}^{\ell_1} - \theta^*\right) + 2n\lambda(\|\theta^*\|_1 - \|\hat{\theta}^{\ell_1}\|_1).$$

Hölder's inequality states that for any $a, b \in \mathbb{R}^d$, we have $a^T b \leq \|a\|_\infty \|b\|_1$, and hence

$$\frac{\|\mathbb{X}\hat{\theta}^{\ell_1} - \mathbb{X}\theta^*\|_2^2}{2} \leq \|\epsilon^T \mathbb{X}\|_\infty \left(\|\hat{\theta}^{\ell_1}\|_1 + \|\theta^*\|_1\right) + n\lambda(\|\theta^*\|_1 - \|\hat{\theta}^{\ell_1}\|_1) \tag{3.5}$$

$$= \|\hat{\theta}^{\ell_1}\|_1(\|\epsilon^T \mathbb{X}\|_\infty - \lambda n) + \|\theta^*\|_1(\|\epsilon^T \mathbb{X}\|_\infty + \lambda n). \tag{3.6}$$

Now for any $t > 0$, and any column $\mathbb{X}_j$ of $\mathbb{X}$, we have that $\mathbb{X}_j^T \epsilon \sim \text{subG}(\sigma^2 n)$ and from Theorem 2.2.3

$$\mathbb{P}\left[\|\mathbb{X}^T \epsilon\|_\infty > t\right] \leq 2de^{-\frac{t^2}{2n\sigma^2}}.$$

Taking $t = \sigma(\sqrt{2n\log(2d)} + \sqrt{2n\log(1/\delta)}) = n\lambda$, we have that $2de^{-\frac{t^2}{2n\sigma^2}} \leq \delta$, we obtain using (3.6), that with probability $1 - \delta$,

$$\|\mathbb{X}\hat{\theta}^{\ell_1} - \mathbb{X}\theta^*\|_2^2 \leq 4n\lambda\|\theta^*\|_1.$$

□

3.5 Incoherence and fast rates for Lasso

Incoherence, random matrices and cone condition

Definition 3.5.1. *A matrix $\mathbb{X} \in \mathbb{R}^{n\times d}$ is said to have incoherence $k \in \mathbb{N}^*$, if*

$$\left\|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right\|_\infty \leq \frac{1}{32k},$$

where $\|\cdot\|_\infty$ denotes the largest absolute value of a matrix.

For $k \to \infty$ this entails that $\mathbb{X}$ is orthonormal and prevents situations where $d > n$. However, finite values of k, amount to relax this constraint and allow for much larger d.

Proposition 3.5.1. *Let $\mathbb{A} \in \mathbb{R}^{n\times d}$ be a random matrix which entries are independent Rademacher variables (± 1 with probability $1/2$). Then, for any $\delta > 0$, if $n \geq 2^{11}k^2\log(1/\delta) + 2^{13}k^2\log(d)$, with probability $1 - \delta$ over the random draw of its entries, $\mathbb{A}$ has incoherence k.*

Proof. The diagonal entries of $\mathbb{A}^T\mathbb{A}$ are equal to n and the off-diagonal elements are sum of n independant Rademacher random variables. From Hoeffding's lemma (2.1.1), Rademarcher random variables are sub gaussian with variance proxy 1 and using Theorem 2.1.2, their sum is subG(n). Using a union bound, we have, for any $t \geq 0$, using Theorem 2.1.1 and summing over the d^2 entries of $\mathbb{A}^T\mathbb{A}$,

$$\mathbb{P}\left[\left\|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right\|_\infty > t\right] \leq 2d^2 e^{\frac{-nt^2}{2}}.$$

Choosing $t = 1/(32k)$, we have

$$\mathbb{P}\left[\left\|\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right\|_\infty > \frac{1}{32k}\right] \leq e^{\log(2)+2\log(d)-\frac{n}{2^{11}k^2}} \leq \delta,$$

for the choice of n which has been made.

□

The k^2 term can actually be improved to k. For any $\theta \in \mathbb{R}^d$, $S \subset \{1, \ldots, d\}$, we denote by θ_S, the vector which support is S and which entries agree with those of θ on S. We have $\|\theta\|_1 = \|\theta_S\|_1 + \|\theta_{S^c}\|_1$.

Lemma 3.5.1. *For any $k \leq d$ and $\mathbb{X}$ having incoherence k, any S with $|S| \leq k$ and any $\theta \in \mathbb{R}^d$ satisfying the cone condition:*

$$\|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1,$$

we have $\|\theta\|_2^2 \leq 2\frac{\|\mathbb{X}\theta\|_2^2}{n}$.

Proof. We have $\theta = \theta_S + \theta_{S^c}$, and hence

$$\|\mathbb{X}\theta\|_2^2 = \|\mathbb{X}\theta_S\|_2^2 + \|\mathbb{X}\theta_{S^c}\|_2^2 + 2\theta_S\mathbb{X}^T\mathbb{X}\theta_{S^c}.$$

From the incoherence condition, we have

$$\|\mathbb{X}\theta_S\|_2^2 = n\|\theta_S\|_2^2 + n\theta_S^T\left(\frac{\mathbb{X}^T\mathbb{X}}{n} - I_d\right)\theta_S \geq n\|\theta_S\|_2^2 - n\frac{\|\theta_S\|_1^2}{32k}.$$

This also holds for θ_{S^c} and using the cone condition, we obtain

$$\|\mathbb{X}\theta_{S^c}\|_2^2 \geq n\|\theta_{S^c}\|_2^2 - n\frac{\|\theta_{S^c}\|_1^2}{32k} \geq n\|\theta_{S^c}\|_2^2 - 9n\frac{\|\theta_S\|_1^2}{32k}.$$

Using the incoherence property again as well as Hölder's inequality and the fact that S and S^c are disjoint, we obtain

$$2\left|\theta_S^T\mathbb{X}^T\mathbb{X}\theta_{S^c}\right| \leq \frac{2n}{32k}\|\theta_S\|_1\|\theta_{S^c}\|_1 \leq \frac{6n}{32k}\|\theta_S\|_1^2.$$

Finnally, from Cauchy-Schwartz inequality, one has $\|\theta_S\|_1^2 \leq |S|\|\theta_S\|_2^2 \leq k\|\theta_S\|_2^2$ and

$$\frac{\|\mathbb{X}\theta\|_2^2}{n} \geq \|\theta_S\|_2^2 + \|\theta_{S^c}\|_2^2 - \frac{16|S|\|\theta_S\|_2^2}{32k} \geq \frac{\|\theta_S\|_2^2}{2}.$$

□

Fast rate for the Lasso estimator

Theorem 3.5.1. *For $n \neq 2$, assume that model LM holds with $\epsilon \sim \mathrm{subG}(\sigma^2)$. Assume that $\|\theta_0\|_0 \leq k$ and that $\mathbb{X}$ has incoherence k. Then, for any $\delta > 0$, the Lasso estimator $\hat{\theta}^{\ell_1}$ with $\lambda = 8\sigma/n(\sqrt{\log(2d)} + \sqrt{\log(1/\delta)})$ satisfies with probability $1 - \delta$*

$$\mathrm{MSE}(\hat{\theta}^{\ell_1}) \leq (2^{12})\frac{k\sigma^2\log(2d/\delta)}{n}$$
$$\|\hat{\theta}^{\ell_1} - \theta^*\|_2^2 \leq (2^{13})\frac{k\sigma^2\log(2d/\delta)}{n}$$

Proof. We have by definition

$$\frac{1}{2n}\|\mathbb{X}\hat{\theta}^{\ell_1} - Y\|^2 \leq \frac{1}{2n}\|\mathbb{X}\theta^* - Y\|^2 + \lambda(\|\theta^*\|_1 - \|\hat{\theta}^{\ell_1}\|_1).$$

and similarly as in Lemma 3.3.1,

$$\|\mathbb{X}\hat{\theta}^{\ell_1} - \mathbb{X}\theta^*\|^2 + n\lambda\|\hat{\theta}^{\ell_1} - \theta^*\|_1 \leq 2\epsilon^T\mathbb{X}(\hat{\theta}^{\ell_1} - \theta^*) + n\lambda\|\hat{\theta}^{\ell_1} - \theta^*\|_1 + 2n\lambda(\|\theta^*\|_1 - \|\hat{\theta}^{\ell_1}\|_1).$$

Similarly as in the proof of Theorem 3.4.2, $\mathbb{X}$ has columns satisfying $\|\mathbb{X}_j\|_2^2 \leq n + \frac{1}{32k} \leq 2n$ from the incoherence condition. Hence, for any $t > 0$,

$$\mathbb{P}\left[\|\mathbb{X}^T\epsilon\|_\infty > t\right] \leq 2de^{-\frac{t^2}{4n\sigma^2}}.$$

Taking $t = 2\sigma(\sqrt{n\log(2d)} + \sqrt{n\log(1/\delta)}) = n\frac{\lambda}{4}$, the right hand side is smaller than δ, and we obtain that with probability $1-\delta$,

$$\begin{aligned}\epsilon^T\mathbb{X}(\hat{\theta}^{\ell_1} - \theta^*) &\leq \|\mathbb{X}^T\epsilon\|_\infty \|\hat{\theta}^{\ell_1} - \theta^*\|_1 \\ &\leq \frac{n\lambda}{4}\|\hat{\theta}^{\ell_1} - \theta^*\|_1.\end{aligned}$$

Setting S the support of θ^* and noting that $\|\hat{\theta}^{\ell_1} - \theta^*\|_1 - \|\hat{\theta}^{\ell_1}\|_1 = \|\hat{\theta}^{\ell_1}_S - \theta^*\|_1 - \|\hat{\theta}^{\ell_1}_S\|_1$, we obtain, with probability $1-\delta$

$$\begin{aligned}\|\mathbb{X}\hat{\theta}^{\ell_1} - \mathbb{X}\theta^*\|^2 + n\lambda\|\hat{\theta}^{\ell_1} - \theta^*\|_1 &\leq 2n\lambda\|\hat{\theta}^{\ell_1} - \theta^*\|_1 + 2n\lambda(\|\theta^*\|_1 - \|\hat{\theta}^{\ell_1}\|_1) &(3.7)\\ &\leq 2n\lambda\|\hat{\theta}^{\ell_1}_S - \theta^*\|_1 + 2n\lambda(\|\theta^*\|_1 - \|\hat{\theta}^{\ell_1}_S\|_1) &(3.8)\\ &\leq 4n\lambda\|\hat{\theta}^{\ell_1}_S - \theta^*\|_1. &(3.9)\end{aligned}$$

In particular, we have

$$\|\hat{\theta}^{\ell_1}_{S^c} - \theta^*_{S^c}\|_1 \leq 3\|\hat{\theta}^{\ell_1}_S - \theta^*\|_1$$

which is the cone condition of Lemma 3.5.1. Using this and Cauchy-Schwartz inequality, we obtain

$$\|\hat{\theta}^{\ell_1}_S - \theta^*\|_1 \leq \sqrt{|S|}\|\hat{\theta}^{\ell_1}_S - \theta^*\|_2 \leq \sqrt{|S|}\|\hat{\theta}^{\ell_1} - \theta^*\|_2 \leq \sqrt{\frac{2k}{n}}\left\|\mathbb{X}\left(\hat{\theta}^{\ell_1} - \theta^*\right)\right\|_2.$$

Combining with (3.9), we have

$$\left\|\mathbb{X}\left(\hat{\theta}^{\ell_1} - \theta^*\right)\right\|_2^2 \leq 32nk\lambda^2 \leq (2^{12})k\sigma^2\log(2d/\delta).$$

The secon inequality follows because from Lemma 3.5.1, we have $\|\hat{\theta}^{\ell_1} - \theta^*\|_2^2 \leq 2\text{MSE}(\hat{\theta}^{\ell_1})$. □

For the proof, we only used Lemma 3.5.1 and more precisely

$$\inf_{|S|\leq k}\inf_{\theta\in C_S}\frac{\|\mathbb{X}\theta\|_2^2}{n\|\theta\|_2^2} \geq \frac{1}{2},$$

where C_S is the cone defined by $\|\theta_{S^c}\|_1 \leq 3\|\theta_S\|_1$. This condition is called restricted eigenvalue condition. It can be seen as a lower bound on the eigenvalues of $\mathbb{X}$ when restricted to sparse eigen vectors. In particular it implies that the smallest singular value of $\mathbb{X}_S$ is at least $n/2$ for all $|S| \leq k$. To summarize, Proposition 3.5.1 and Theorem 3.5.1 en sure that there exists design matrices $\mathbb{X}$ such that the Lasso estimators has a fast convergence rate in high dimensions.

3.6 Compressed sensing

High dimentional statistics have an important intersection with compressed sensing [28, 23] in signal processing. Traditional approaches separate signal aquisition and signal compression which is performed on a signal which is fully characterized in the memory of a device (or at least very accurately described). The field of compressed sensing emerge as different approach for this problem based on two observations.

- Natural signals such as speach, sounds, images, are not generic or completely random and they have a strong intrinsic strucutre.
- If this structure was known it should be possible to take advantage in a signal aquisition / compression scheme.

Compressed sensing emerged as a development of the preceeding observation based on two ideas.

- the undelying structure of natural signals is captured by sparsity patterns in a certain basis.
- if a signal is sparse in a given basis, one could probably mix the aquisition and compression phase by aquiring only a very limited number of measurements.

We describe a signal recovery result from random measurements relying on linear programming. Further readings on the topic include [22, 23, 25].

Signal recovery

Althouth the notations will be the same as in the high dimensional statistics context, the viewpoint is a bit different. The signal to be recovered is $\theta^* \in \mathbb{R}^d*$ which is unknown and assumed to be sparse, that is $\|\theta^*\|_0 = k < d$. The operator has the possibility to choose a sensing matrix $\mathbb{X} \in \mathbb{R}_{n\times d}$ which will result in the following measurements:

$$\mathbb{X}\theta^* = y \tag{3.10}$$

The goal of compressed sensing is to establish methods and conditions ensuring large classes of values of θ^* can be infered accurately only from the knowledge of y and $\mathbb{X}$. Other questions of interest include robustness to noise and exact recovery of $\mathrm{supp}(\theta^*)$. For simplicity we will only touch the noiseless setting in (3.10). We will deduce compressed sensing type results from MSE estimates of the previous sections.

Exact recovery using ℓ_0 minimization

We introduce the estimator

$$\hat{\theta}^{\ell_0}_{CS} \in \min_{\theta\in\mathbb{R}_d} \quad \|\theta\|_0 \quad \text{s.t.} \quad \mathbb{X}\theta = y. \tag{3.11}$$

under mild assumption on the sensing matrix $\mathbb{X}$, this estimator deterministically recovers the unknown signal θ^*.

Proposition 3.6.1. *Given $k \in \mathbb{N}$, $k \le d/2$, assume that $\|\theta^*\|_0 \le k$, and assume that for any S, $|S| \le 2k$, that $\mathbb{X}_S$ has full column rank. Then, the solution of (3.11) is unique and is equal to θ^*.*

Proof. Assume that $\hat{\theta}^{\ell_0}_{CS} \neq \theta^*$. We have $\|\hat{\theta}^{\ell_0}_{CS}\|_0 \le \|\theta^*\|_0 = k$. Set $S = \mathrm{supp}(\theta^* - \hat{\theta}^{\ell_0}_{CS})$. We have $|S| \le 2k$ and $\mathbb{X}(\theta^* - \hat{\theta}^{\ell_0}_{CS}) = 0$ and hence $\theta^* = \hat{\theta}^{\ell_0}_{CS}$. □

Exact recovery from random measurements with ℓ_1 minimization

Intuitively if one is interested in signal recovery over large classes of signals using ℓ_1 norm, the sensing matrix in (3.10) should not have structure fooling the ℓ_1 norm. This happens if $\mathbb{X}$ is generic in some sense. One way to achieve this is to use random measurements. This amounts to choose a random $\mathbb{X}$ in (3.10) such as the one described in Proposition 3.5.1 for example. Furthermore, since there is no noise, in the measurements, the least squares approach does not really make sense. We introduce an estimator.

$$\hat{\theta}^{\ell_1}_{CS} \in \min_{\theta\in\mathbb{R}_d} \quad \|\theta\|_1 \quad \text{s.t.} \quad \mathbb{X}\theta = y. \tag{3.12}$$

Corollary 3.6.1. *Given $k \in \mathbb{N}$, $k \le d$, and $\delta > 0$, assume that $\mathbb{X}$ is a Rademacher matrix with $n \ge 2^{11}k^2\log(1/\delta) + 2^{13}k^2\log(d)$. Assume furthermore that $\|\theta^*\|_0 \le k$ in (3.10). Then with probability $1-\delta$ over the random draw of $\mathbb{X}$, the solution of (3.12) is unique and is equal to θ^*.*

Proof. Assume that $\hat{\theta}^{\ell_1}_{CS} \neq \theta^*$ and set $d = \hat{\theta}^{\ell_1}_{CS} - \theta^*$. We have $\|\hat{\theta}^{\ell_1}_{CS}\|_1 \le \|\theta^*\|_1$ and $\mathbb{X}d = 0$. Set $S = \mathrm{supp}(\theta^*)$, we have

$$\|\theta^*\|_1 \ge \|\hat{\theta}^{\ell_1}_{CS}\|_1 = \|d_{S^c}\|_1 + \|d_S + \theta^*\|_1 \ge \|d_{S^c}\|_1 + \|\theta^*\|_1 - \|d_S\|_1.$$

As a result, we have $\|d_S\|_1 \ge \|d_{S^c}\|_1$ and $\mathbb{X}d = 0$. Lemma 3.5.1 implies that $d = 0$ with probability $1-\delta$ corresponding to the event $\mathbb{X}$ having incoherence at level k. □

This result shows that it is possible to recover θ^* with high probability only from the order of $O(k^2 \log(d))$ measurements provided that $\|\theta^*\|_0 \leq k$. The k^2 term can be improved further. In the context of noisy measurements, conditioning both on the realization of $\mathbb{X}$ and the realization of the noise, one can obtain results similar to Theorem 3.5.1 for signal processing.

Exercises

Exercise 3.6.1. *Given $x \in \mathbb{R}^d$ and $\lambda > 0$, show that the solution to the problem*

$$\min_{y \in \mathbb{R}^p} \frac{1}{2}\|y - x\|_2^2 + \lambda \|y\|_1$$

is given by coordinatwise application of $p_\lambda \colon \mathbb{R} \mapsto \mathbb{R}$ to x, where, for any $s \in \mathbb{R}$

$$p_\lambda(s) = \begin{cases} s - \lambda, & \text{if } s > \lambda \\ 0, & \text{if } |s| \leq \lambda \\ s + \lambda, & \text{if } s < \lambda \end{cases}.$$

This is the soft-thresholding operation and the result is called the proximity operator of the function $\lambda\|\cdot\|_1$. Give a graphical representation of p_λ and compare it to the hard-thresholding operator given by $t \mapsto t\mathbb{I}(|t| \geq \lambda)$.

Exercise 3.6.2. *Let $X = (1, Z, \dots, Z^d)^T \in \mathbb{R}^{d+1}$ be a random vector where Z is a real random variable. Show that $\mathbb{E}\left[XX^T\right] \in \mathbb{R}^{d+1 \times d+1}$ is positive definite when Z admits a density with respect to Lebesgue measure on $\mathbb{R}$. Provide a counter example for which $\mathbb{E}\left[XX^T\right]$ is singular.*

Exercise 3.6.3. *Under the linear model* (LM)*,*

- *Assuming that $\mathbb{X}^T\mathbb{X}$ is invertible and $\mathbb{E}[\epsilon] = 0$, show that $\mathbb{E}[\theta_{LS}] = \theta^*$.*
- *Assuming in addition that $\epsilon \sim \text{subG}(\sigma^2)$, show that $\theta_{LS} - \theta^* \sim \text{subG}\left(\frac{\sigma^2}{\lambda_{\min}}\right)$ where $\lambda_{\min}$ denotes the smallest eigenvalue of $\mathbb{X}^T\mathbb{X}$. Propose a generalization of the result when the invertibility assumption is dropped.*
- *If $\mathbb{X}^T\mathbb{X}$ is not invertible, show that $\theta_{LS} = \arg\min_\theta \|\theta\|_2$, such that $\mathbb{X}^T\mathbb{X}\theta = \mathbb{X}^T Y$.*

Exercise 3.6.4. *We consider the model* (LM)*, and define the ridge regression estimator, for any $\lambda > 0$*

$$\hat{\theta}^{\ell_2} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2n}\|\mathbb{X}\theta - Y\|^2 + \lambda\|\theta\|_2^2.$$

- *Show that $\hat{\theta}^{\ell_2}$ is indeed uniquely defined and propose a closed form expression for it.*
- *Compute the bias: $\mathbb{E}\left[\hat{\theta}^{\ell_2} - \theta^*\right]$ and show that it is bounded by $\|\theta^*\|_2^2$.*
- *Show that $\hat{\theta}^{\ell_2} - \mathbb{E}\left[\hat{\theta}^{\ell_2}\right] \sim \text{subG}\left(\frac{\sigma^2}{\lambda}\right)$.*
- *Show the bias variance decomposition identity:*

$$\mathbb{E}\left[\|\hat{\theta}^{\ell_2} - \theta^*\|_2^2\right] = \mathbb{E}\left[\left\|\hat{\theta}^{\ell_2} - \mathbb{E}\left[\hat{\theta}^{\ell_2}\right]\right\|_2^2\right] + \left\|\mathbb{E}\left[\hat{\theta}^{\ell_2} - \theta^*\right]\right\|_2^2.$$

- *Using the previous exercise, suggest as situation for which*

$$\mathbb{E}\left[\|\hat{\theta}^{\ell_2} - \theta^*\|_2^2\right] < \mathbb{E}\left[\|\hat{\theta}^{LS} - \theta^*\|_2^2\right]$$

Chapter 4

Computation, Complexity, Conic Programming

Source: most of the content of this chapter is described in Ben-Tal and Nemirowski's book on "Modern Convex Opitmization" [10]. An extensive treatment of the linear programming problem is found in Schrijver's book on linear programming [57]. Further reading include Bertsekas's book [13] and Boyd and Vandenberghe [20] (freely available) which content is a bit wider than our interest here. An interesting discussion between complexity theory and mathematics is given by [59]. Physical implications of complexity theory are given in [1]. Oracle complexity is extensively discussed in [43] and a more recent treatment is given in [45].

4.1 Introduction

When considering high dimensional statistics, computation has to be taken into account because the size of the problems to be adressed does not allow to ignore the computational cost of handling them. In particular, one may prefer a less statistically accurate estimator if it can be computed efficiently. Although very intuitive, the mathematical definition of "computation" is highly non trivial and has very strong connections to logics, physics and phylosophy. We start with a brief overview of theoretical computer science concepts which enlights computational properties of the statistical estimators we considered.

The second part of this chapter presents an overview of convex optimization as developped in the 90's. This resulted in classification of families of tractable convex optimization problems for which general purpose numerical solvers were developped. In the context of high dimensional statistics, these elements are mostly interesting for historical purposes as prefered methods for modern data analysis do not fall in the category of methods described in this chapter.

4.2 Computation over $\mathbb{Q}$.

We first provide an overview of computation formalism. Most of this is borowed from Schrijver's book [57].

4.2.1 Computation over a finite alphabet and complexity over $\mathbb{Q}$

Alphabet, words, size: We consider a finite set Σ (usually $\Sigma = \{0, 1\}$), which is called an alphabet and its elements are called *letters*. An ordered finite sequence of elements in Σ is called a *word*. The set of words is denoted by Σ^*. The *size* of a string is the number of its components. The zero length string is the empty string $\emptyset$.

Strings can be used to represent rational numbers, vectors, matrices, and combinatorial structures such as graphs and trees. There are standard ways to encode these objects over a finite alphabet such as $\{0,1\}$, depending on the chosen way this induces a concept of size for these objects. For example if $\alpha = p/q$ (where p and q are relatively prime integers), $c = (c_1, \ldots, c_n)$ a rational vector and $A = (a_{ij})_{i=1\ldots m, j=1\ldots n}$ a rational matrix, we have

$$\begin{aligned} \text{size}(\alpha) &= 1 + \lceil \log_2(p) \rceil + \lceil \log_2(q) \rceil \\ \text{size}(c) &= n + \sum_{i=1}^{n} \text{size}(c_i) \\ \text{size}(A) &= nm + \sum_{i=1}^{m} \sum_{j=1}^{n} \text{size}(a_{ij}) \end{aligned}$$

Size of linear inequalities, or equalities are defined in a similar way.

Problems: A *(search) problem* is a subset $\Pi \subset \Sigma^* \times \Sigma^*$, the corresponding meta-mathematical problem read as follows:

$$\text{Given } z \in \Sigma^*, \text{ find } y \in \Sigma^* \text{ such that } (z,y) \in \Pi \text{ or decide that there exists no such } y.$$

An example of a search problem is given a matrix $A \in \mathbb{Q}^{m \times n}$ and a vector $b \in \mathbb{Q}^m$, find $x \in \mathbb{Q}^n$ such that $Ax \leq b$ (where the inequality is understood elementwise). A decision problem is a problem which output is either 0 or 1. For example, given A and b, is there an x such that $Ax \leq b$? A decision problem is often identified with $\mathcal{L} \subset \Sigma^*$, the set of inputs such that the output is 1.

Algorithm and running time: An algorithm is a list of instruction to solve a problem. A *Turing machine* is a thought experiment object which formalizes the notion of algorithm. The *Church-Turing thesis* is a founding hypothesis of computer science stating that functions of natural numbers computable by humans using pen and pencil, following an algorithm are precisely the ones which can be computed by a Turing machine. One can view a Turing machine as a device which performs pen and paper computation automatically and take it as a rigorous formalization of what it means "*to compute*". There exists equivalent formalizations such as recursive functions, lambda calculus, circuits which lead to equivalent notions of computation all of them are called *Turing complete*.

For a given input Σ^*, an algorithm for problem Π determines an output y such that (z,y) is in Π, or stops without delivering an output if there exists no such y. An algorithm can have the shape of a computer program, which is a finite string of symbols from a finite alphabet. Hence, an algorithm can be defined as a finite string A of 0's and 1's. One says that A solves problem Π, if for any instance z of Σ^*, when giving the string (A,z) to a *universal Turing machine* (a Turing machine which could simulate any other Turing machine, in particular, a Turing machine implementing A), the machine stops after a finite number of steps, and delivers y with $(z,y) \in \Pi$, or no string in the case where such a string y does not exist.

The running time of an algorithm is number of elementary operations during the execution of the algorithm. It depends on the precise implementation considered. One way to formalize this is the number of moves of the head of a universal Turing machine before stoping given the input (A,z). Formally, the runing time function of an algorithm $f\colon \mathbb{N} \mapsto \mathbb{N}$ can be given by

$$f(\sigma) = \max_{\text{size}(z) \leq \sigma} (\text{running time of } A \text{ for input } z).$$

Polynomial algorithm and computation over $\mathbb{Q}$ An algorithm is called *polynomial time*, if its time function is upper bounded by a polynomial. A problem is calle *polynomially solvable* if there exists a polynomial time algorithm to solve it.

The elementary operations such as adding, substracting, multiplying, dividing, comparing numbers can be executed in polynomial time. Note that for computation over $\mathbb{Q}$, we use the (polynomial

time) Euclidean algorithm to obtain a unique representation of these numbers. Therefore, in order to show that a numerical algorithm is polynomial time, it suffices to show that it applies a number of elementary operations which is polynomial in the size of the input and that the size of the intermediate numbers to which these elementary operations are polynomially bounded by the size of the input.

Note that any numerical software, such as the ones used for statistical estimation, actually perform computation over $\mathbb{Q}$ as they implement finite precision arithmetic. This amounts to consider choose a finite precision $\epsilon \in \mathbb{Q}$, $\epsilon > 0$ and perform all numerical operations by rounding over a discrete grid $\{n\epsilon\}_{n\in\mathbb{Z}} \subset \mathbb{Q}$.

The classes $\mathcal{P}$ and $\mathcal{NP}$ and $\mathrm{co}-\mathcal{NP}$ The class of decision problems solvable in polynomial time is called $\mathcal{P}$. The class $\mathcal{NP}$ is central in complexity analysis and corresponds to decisions problems for which there is an easy to check verification, that is, which have a polynomial size proof. More formaly a decision problem $\mathcal{L} \subset \Sigma^*$ belongs to $\mathcal{NP}$ if there exists a polynomialy solvable decision problem $\mathcal{L}' \subset \Sigma^* \times \Sigma^*$ and a polynomial ϕ such that

$$z \in \mathcal{L} \quad \Leftrightarrow \quad \exists y \in \Sigma^*, (z, y) \in \mathcal{L}' \text{ and } \mathrm{size}(y) \leq \phi(\mathrm{size}(z)).$$

The crucial point here is that it is not required that y is found in polynomial time, but if it was given, the proof could be checked in polynomial time. The string y is called a certificate. Brute force search over all possible strings of a given length provides an algorithm showing that for any problem in $\mathcal{NP}$ there exists a polynomial ψ such that the solution for input z can be found in time at most $2^{\psi(\mathrm{size}(z))}$.

Example 4.2.1. *Given a set of cities and distances between cities (in $\mathbb{Q}$), the traveling salesman problem is in $\mathcal{NP}$:*

Given $d \in \mathbb{Q}$, decide if there is a path visiting all the cities of total length at most d.

Indeed, if such a path exists, it has the same length as the total number of cities so that checking that it passes through all cities and that its length is less than d can be done in polynomial time. Therefore, if the decision problem admits a solution, it has a polynomial time certificate.

Example 4.2.2. *Given $A \in \mathbb{Q}^{n\times d}$ and $b \in \mathbb{Q}^n$, consider the problem of deciding if $Ax \leq b$ has a solution over $\mathbb{Q}^n$. It can be shown (See Schiver's book chapter 10) that if such a solution exists, then there should be a solution which size is polynomially bounded by the size of A and b. Hence this decision problem is in $\mathcal{NP}$.*

The class of decision problems $\mathcal{L} \subset \Sigma^*$ which complement in Σ^* is in $\mathcal{NP}$ is denoted by $\mathrm{co}-\mathcal{NP}$. The class $\mathcal{NP} \cap \mathrm{co}-\mathcal{NP}$ consists of those decision problems which answer (positive or negative) have a polynomial length proof. We have $\mathcal{P} \subset \mathcal{NP}$ and $\mathcal{P} \subset \mathrm{co}-\mathcal{NP}$ and it is not known wether these inclusions are strict (there is a million dollars price on these questions).

The term $\mathcal{NP}$ comes from "Non deterministic Polynomial time". This means that a lucky algorithm which has the possibility to "guess" in polynomial time a good certificate over a set with polynomial size can solve the corresponding decision problem.

4.3 Karp reduction and $\mathcal{NP}$ completeness

A decision problem $\mathcal{L} \in \Sigma^*$ is *Karp* reducible to a decision problem $\mathcal{L}' \subset \Sigma^*$ if there exists a polynomial time algorithm such that, for any input string $z \in \Sigma^*$, A delivers a string x such that

$$z \in \mathcal{L} \quad \Leftrightarrow \quad x \in \mathcal{L}'$$

This can be denoted as $\mathcal{L} \leq \mathcal{L}'$ as an algorithm for solving $\mathcal{L}'$ would provide an algorithm for solving $\mathcal{L}$ with an added computational cost which is at most polynomial.

Example 4.3.1. *For any boolean formula there is a formula over linearly more variable in conjunctive normal form, which preserves satisfiability. The size of the new formula is at most linear in the size of the original formula, using Tseytin transformation for example. We obtain a formula of the form*

$$(a \vee b \vee c \vee d) \wedge (\bar{a} \vee e \vee f \vee \bar{g} \vee d) \dots$$

Then any disjunction can be reduced to a conjunction of disjunctions of size at most 3 *by adding variables. For example*

$$\begin{aligned} & q \vee r \vee s \vee t \vee u \\ \Leftrightarrow \quad & (q \vee r \vee a) \wedge (\bar{a} \vee s \vee b) \wedge (\bar{b} \vee t \vee u). \end{aligned}$$

Thus if $\mathcal{L}$ denotes the boolean formula satisfiability problem (SAT) and $\mathcal{L}'$ denotes the satisfiability problem of boolean formula in 3 conjunctive normal form (3-SAT), we have shown that $\mathcal{L} \leq \mathcal{L}'$.

Similarly, if $\mathcal{L}'$ belongs to $\mathcal{NP}$ and $\mathcal{L} \leq \mathcal{L}'$, then $\mathcal{L}$ also belongs to $\mathcal{NP}$. A problem $\mathcal{L}$ is called *$\mathcal{NP}$-hard*, if each problem in $\mathcal{NP}$ is reducible to $\mathcal{L}$ and if furthermore, $\mathcal{L}$ is in $\mathcal{NP}$, then $\mathcal{L}$ is called *$\mathcal{NP}$-complete*. As we have seen, we have an exponential time algorithm to solve problems in $\mathcal{NP}$, this is a brute force search algorithm. It is widely believed that for a given $\mathcal{NP}$-complete problem, this is the most efficient algorithm to solve all the possible instances. Indeed, a polynomial time algorithm for any $\mathcal{NP}$ complete problem would provide a proof that $\mathcal{P} = \mathcal{NP}$ which is widely believed to be false. This is undelying the $\mathcal{P} \neq \mathcal{NP}$ conjecture. It is important to note that the notion of $\mathcal{NP}$-hardness is a *worst case* notion.

- $\mathcal{NP}$-complete problems are considered to be hard as there si no known polynomial time algorithm to solve them and it is believed that no such algorithm exists.
- This concept relies on Karp reduction which only underlines that some instances are hard, not necessarily all of them.
- There is no notion of constant or exponent in these concepts so that an algorithm in $\mathcal{P}$ may still be intractable in practice. The notion is mostly used to prove computational difficulty of certain problems.

From optimization to decision An optimization problem is the minimization of an objective function c over a finite set or over rational numbers. An efficient algorithm to solve an optimization problem provide an algorithm to decide if there exists a sequence of input with cost less or equal to α, for any α. For example given $A \in \mathbb{Q}^{n \times d}$, $b \in \mathbb{Q}^n$, $c \in \mathbb{Q}^d$, computing

$$\rho = \inf_{Ax \leq B} c^T x$$

provides an algorithm to decide whether $Ax \leq b$ and $c^T x \leq \alpha$ has a solution for any $\alpha \in \mathbb{Q}$. As a result, optimization objectives involving $\mathcal{NP}$-complete problems are considered as hard.

Examples:

Example 4.3.2 (Cook's Theorem)**.** *The boolean satisfiability problem (SAT) consists of decision problem over boolean variables involving boolean formulas in conjunctive normal form: the variables are augmented with their negations, and the formula consists of a conjunction of disjunctions (all clauses made using "or" and are aggregated with an "and"). Example*

$$(x_1 \text{ and } x_2 \text{ and } x_6) \text{ or } (\bar{x_2} \text{ and } x_3 \text{ and } \bar{x_7}) \text{ or } \dots$$

This is the first problem proved to be $\mathcal{NP}$-complete by Cook in 1971.

The idea of the proof is as follows. First the problem is clearly in $\mathcal{NP}$ as it suffices to exhibit a an instance of boolean values which satisfy the formula. The problem is $\mathcal{NP}$-hard because because

a polynomial time verifier implemented on a Turing machine can be shown to be equivalent to a boolean formula (this is the technical bulk of the proof). Finaly there is a polynomial time reduction from any boolean formula to a formula of the above form where each disjunction involves at most 3 *variables.*

This problem remains $\mathcal{NP}$-complete if we restrict the disjunctions to involve at most 3 *variables (as in the example) by the* 3*-SAT reduction argument. This shows that* 3*-SAT is $\mathcal{NP}$-complete.*

Example 4.3.3. *An important list of $\mathcal{NP}$-complete problems can be found in the classic book, Computers and Intractability: A Guide to the Theory of NP-Completeness.*

Theorem 4.3.1. *Consider the decision problem with input $A \in \mathbb{Q}^{m\times n}$, $b \in \mathbb{Q}^m$, does there exist $x \in \mathbb{Q}^n$ such that $Ax = b$ and $\|x\|_0 \leq m/3$. This problem is $\mathcal{NP}$-hard.*

Proof. We reproduce the proof given in [42]. First, the problem is clearly in $\mathcal{NP}$. Completness is shown by reduction is to "cover by 3 sets" which is an $\mathcal{NP}$-complete problem:

Given a set S and a set C which elements consists of subsets of S of size 3. Decide if there is $\hat{C} \subset C$ such that each elements of S occurs exactly once in $\hat{C}$.

Assume that $S = \{s_1, \ldots, s_m\}$ and $C = \{c_1, \ldots, c_n\}$ and assume that m is a multiple of 3. Set $b = (1, \ldots, 1)^T \in \mathbb{Q}^m$ and $A \in \mathbb{R}^{m\times n}$ which column i is zero except at the j, k, l where $(s_j, s_k, s_l) = c_i$, $i = 1, \ldots, n$. We show that there exists $x \in \mathbb{Q}^n$ such that $Ax = b$ with $\|x\|_0 \leq m/3$ if an only if the answer to the "cover by 3 set problem" is positive.

On the one hand given $\hat{C}$, choosing $x_i = 1$ if $c_i \in \hat{C}$ and zero otherwise. We have $Ax = b$ and x has $m/3$ non zero entries. On the other hand if one finds $x \in \mathbb{Q}^n$ with $Ax = b$ and $\|x\|_0 \leq m/3$. The entries of Ax must be in 1. Since x has at most $m/3$ non zero entries and each column has at most 3 nonzero entries, it means that x has exactly $m/3$ nonzero entries. The nonzero entries of x solves the "cover by 3 set" problem. □

The theorem generalizes to real inputs and real variables in the computation model of infinite precision RAM model or computation over the reals and approximate solutions $\|Ax - b\|_2^2 \leq \epsilon$, see [42]. The implication of these results is that solving problems involving the $\|\cdot\|_0$ pseudonorm is hard. For example computing $\hat{\theta}^{\ell_0}$ can be done by solving 2^d unconstrained least squares problems, and, unless $\mathcal{P} = \mathcal{NP}$, no algorithm can do significantly better on all possible instances for all values of d. This underlines the value of the question $\mathcal{P} = \mathcal{NP}$? This kind of statement is very common in computer science: if there is a reduction from a given problem to another problem which is proved (or largely believed) to be hard, then the original problem must be hard.

4.4 Computation over the reals

The statistical estimators which are defined in previous chapters, are given over the real field and we only mentioned computation over the rationals so far. The difference may look innocuous at first sight but it actually has tremendous implications. Furthermore most of the optimization theory which we are going to describe is given for algorithms over the reals, and therefore, it is worth mentioning models of computation over the reals. The content of this section is mostly theoretical since real arithmetic is not realisable in the physical word [1] (it would break well accepted physical impossibility principles).

Computable number: Computer Algebra Systems use symbolic programing to perform operations on algebraic objects. However the set of real numbers which can be described by such systems is only denumerable and therefore, miss most of the reals. Another definition of computable number concerns the possibility to approximate it up to an arbitrary precision.

Definition 4.4.1. *A number $a \in \mathbb{R}$ is called computable if there is a terminating algorithm A such that for any $\epsilon \in \mathbb{Q}$, $\epsilon > 0$, $|A(\epsilon) - a| \leq \epsilon$.*

Intuitively, there are only countably many terminating algorithms and the set of computable numbers is therefore only countable. Hence, most real numbers are not computable in this sense.

Real machines: In 1989 Blum, Shub and Smale described a theoretical machine for real computation [15]. This is refered to as *BSD* machine or *real RAM* machine and leads to the theory of algebraic complexity. Roughly speaking such a machine manipulates real numbers instead of element of a finite alphabet and is able to perform addition, multiplication, division and comparison over real numbers.

This is a canonical model for computation over the reals. Although not realisable in the physiscal word [1], this constitutes an interesting thought experiment. For example, we will see that the linear programming (LP) problem (4.7) is polynomially solvable over the rationals, but it is not known if it is polynomialy solvable over the reals [59]. The main difference between computation over $\mathbb{Q}$ and over $\mathbb{R}$ is that in the first case, the size of the input (number of bits required to decribe it using standard encoding) provides a bound on the accuracy level required to obtain provably correct rounding schemes. In the real case, the size of the input is only the number of entries. For example the condition number of a matrix A depends on its size over $\mathbb{Q}$ while it does not depend on its size over $\mathbb{R}$.

Oracle complexity: The computational model underlying continuous optimization mixes real computation and unknown primitives which are provided by an oracle. For example, if one wishes to minimize a differentiable function f over $\mathbb{R}^d$, one can construct an algorithm which is allowed to querry sequentially the value of f and its gradient ∇f at different points in $\mathbb{R}^d$. The running time of an algorithm is given by the number of call to the oracle and the number of real arithmetic operations performed by the algorithm. Complexity is then given by worst case bounds on the number of operations required solve a specific problem and it usually depends on properties of f such as conditioning. Depending on the oracle of choice, one may define different notions of running time and complexity for optimization algorithms. Note that although this model is quite intuitive, it is actually very far from what is performed in practice when using physical computers. Nemirovski Yudin [43] introduced this notion of complexity as a systematic way to study continous optimization algorithms, further comments and a more recent exposition can be found in the book of Nesterov [45]. An interesting discussion about the connections between such a model and classical complexity theory is found in [10].

4.5 Recap on convexity

We limit ourselves to the finite dimensional setting which is sufficient for our purpose. Most notions given here generalize to infinite dimensions [39, 8]. The content of this section is mostly related to [10, 20].

Convex sets and functions

A subset of a vector space $\mathcal{X}$ is convex if it is closed under convex combinations.

Definition 4.5.1. *Let $\mathcal{X} \subset \mathbb{R}^d$, we say that $\mathcal{X}$ is convex if for any $x, y \in \mathcal{X}$, $\alpha \in [0,1]$, $\alpha x + (1-\alpha)y \in \mathcal{X}$. A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is convex if its epigraph is convex in $\mathbb{R}^{d+1}$. Recall that $\mathrm{epi}(f) = \big\{(x,z) \in \mathbb{R}^{d+1},\, z \geq f(x)\big\}$. Equivalently, for any $x, y \in \mathbb{R}^d$, and any $\alpha \in [0,1]$, $f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$.*

Convex sets are closed under many set operations including, interior, closure, intersection, (Minkowski) addition, affine mapping and inverse affine mapping. There is a well defined notion of dimenision for convex set $\mathcal{X}$, it is simply the dimension of the smallest affine set containing $\mathcal{X}$.

Lemma 4.5.1. *For any convex set $\mathcal{X} \subset \mathbb{R}^d$ we have*

- *The closure of $\mathcal{X}$ is convex.*
- *The interior of $\mathcal{X}$ is convex.*
- *For any $u \in \mathrm{int}(\mathcal{X})$ and $v \in \mathrm{cl}(\mathcal{X})$, $[u, v) \subset \mathrm{int}(\mathcal{X})$.*

- *If the interior of $\mathcal{X}$ is non empty, then* $\mathrm{cl}(\mathcal{X}) = \mathrm{cl}(\mathrm{int}(\mathcal{X}))$.
- *The interior of $\mathcal{X}$ is empty if and only if it is contained in a lower dimensional affine subspace.*

Characterization of convex functions

We have the following characterizations of convexity

Theorem 4.5.1. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$:*

1. *If f is continuously differentiable, then f is convex if and only if or any $x, y \in \mathbb{R}^d$, $f(y) \geq f(x) + \nabla f(x)^T(y - x)$.*
2. *If f is continuously differentiable, then f is convex if and only if or any $x, y \in \mathbb{R}^d$, $(\nabla f(x) - \nabla f(y))^T(x - y) \geq 0$.*
3. *If f is twice continuously differentiable, then f is convex if and only if or any $x \in \mathbb{R}^d$, $\nabla^2 f(x)$ is positive semidefinite.*

One has the following consequence which is a central motivation for studying convex optimization problems

Corollary 4.5.1. *Let $f\colon \mathbb{R}^d \to \mathbb{R}$ be a convex continuously differentiable function, then the following are equivalent*

- *x is a global minimizer of f.*
- *$\nabla f(x) = 0$.*

Example 4.5.1. *Consider the least squares linear regression estimate $\hat{\theta}^{LS} \in \arg\min_{\theta \in \mathbb{R}^d} \|\mathbb{X}\theta - y\|_2^2$. The hessian matrix of the objective is $\mathbb{X}^T\mathbb{X}$ which is positive semidefinite so that the objective is convex and first order conditions are sufficient for optimality.*

Separating hyperplane and supporting hyperplane

Theorem 4.5.2 (Separating hyperplane)**.** *Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ be two disjoint closed convex sets, then there exists a vector $v \in \mathbb{R}^d$, $v \neq 0$ and a number $c \in \mathbb{R}$ such that $x^T v > c$ for all $x \in \mathcal{X}$ and $y^T v < c$ for all $y \in \mathcal{Y}$.*

Proof. Set $S = \mathcal{X} - \mathcal{Y} = \{s = x - y,\, x \in \mathcal{X},\, y \in \mathcal{Y}\}$, S is convex and closed. Since $\mathcal{X}$ and $\mathcal{Y}$ are disjoint, $0 \notin S$. Let $\bar{s}$ denote any minimal norm element of s. For any $s \in S$, and $t \in [0, 1]$,

$$0 < \|\bar{s}\|_2^2 \leq \|\bar{s} + t(s - \bar{s})\|_2^2 = \|\bar{s}\|_2^2 + 2t\bar{s}^T(s - \bar{s}) + t^2\|(s - \bar{s})\|_2^2.$$

The right hand side is differentiable for $t \in \mathbb{R}$ and the derivative at 0 must be non negative. Hence, for any $s \in S$, $s^T\bar{s} \geq \|\bar{s}\|_2^2 > 0$. We deduce that

$$\inf_{x \in \mathcal{X}} \bar{s}^T x = \|\bar{s}\|_2^2 + \sup_{y \in \mathcal{Y}} \bar{s}^T y$$

which shows that we can choose $v = \bar{s}$ and any $c \in \left(\inf_{x \in \mathcal{X}} \bar{s}^T x, \sup_{y \in \mathcal{Y}} \bar{s}^T y\right)$ where the interval is non empty. □

We deduce the following which is a weak finite dimensional form of the Hahn Banach theorem.

Theorem 4.5.3 (Supporting hyperplane)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex sets such that $0 \notin \mathcal{X}$, then there exists a vector $v \in \mathbb{R}^d$, $v \neq 0$ such that $v^T x \geq 0$, for all $x \in \mathcal{X}$.*

Proof. If $0 \notin \mathrm{cl}(\mathcal{X})$ then the result follows immediately from the separating hyperplane theorem. If $0 \in \mathrm{cl}(\mathcal{X})$, since $0 \notin \mathcal{X}$, $0 \notin \mathrm{int}(\mathcal{X})$, and 0 is on the boundary of $\mathcal{X}$. Hence 0 is in the closure of the complement of $\mathrm{cl}(\mathcal{X})$ and there exists a sequence $\{z_k\}_{k\in\mathbb{N}}$ not in $\mathrm{cl}(\mathcal{X})$. Which converges to 0. Applying the separating hyperplane theorem to each element of the sequence ensures that there exists a sequences $\{v_k\}_{k\in\mathbb{N}}$ non zero in $\mathbb{R}^d$ and $\{c_k\}_{k\in\mathbb{N}}$ in $\mathbb{R}$, such that for all $k \in \mathbb{N}$ and all $x \in \mathcal{X}$,

$$\frac{v_k^T z_k}{\|v_k\|} < \frac{c_k}{\|v_k\|} < \frac{v_k^T x}{\|v_k\|}.$$

Let v be any accumulation point of $\frac{v_k}{\|v_k\|}$, the left hand side of tends to 0 hence $\liminf_{k\to\infty} \frac{c_k}{\|v_k\|} \geq 0$ and $v^T x \geq 0$ for all $x \in \mathcal{X}$. □

If $0 \in \mathrm{cl}(\mathcal{X})$, the vector v defines a supporting hyperplane which provides a notion of tangent to a set convex set.

Theorem 4.5.4 (Supporting hyperplane)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex set such that 0 is on the boundary of $\mathcal{X}$, then there exists a vector $v \in \mathbb{R}^d$, $v \neq 0$ such that $v^T x \geq 0$, for all $x \in \mathcal{X}$.*

Proof. If $0 \notin \mathcal{X}$ then the result follows from Theorem 4.5.3, we assume that $0 \in \mathcal{X}$.

If the interior of $\mathcal{X}$ is empty, then, by Lemma 4.5.1, $\mathcal{X}$ is contained in a lower dimensional affine space which turns out to be a linear subspace since $0 \in \mathrm{cl}(\mathcal{X})$ any vector orthogonal to this subspace will work.

If $\mathrm{int}(\mathcal{X})$ is not empty since it is convex by Lemma 4.5.1, we may apply Theorem 4.5.3 to 0 and $\mathrm{int}(\mathcal{X})$ and obtain $v \in \mathbb{R}^p$ such that $v^T x \geq 0$ for all $x \in \mathrm{int}(\mathcal{X})$. Lemma 4.5.1 ensures that $\mathrm{cl}(\mathrm{int}(\mathcal{X})) = \mathrm{cl}(\mathcal{X})$ so that $v^T x \geq 0$ for all $x \in \mathrm{cl}(\mathcal{X}) \supset \mathcal{X}$ and the result follows. □

More generally, a supporting hyperplane of $\mathcal{X}$ at x is a closed half space which contains $\mathcal{X}$ and x on its boundary. There is a partial converse.

Theorem 4.5.5. *Let $\mathcal{X}$ be a closed set with nonempty interior, such that for every point x on the boundary of $\mathcal{X}$ admits a supporting hyperplane. Then $\mathcal{X}$ is convex.*

Proof. $\mathcal{X}$ is contained in the set S consisting of intersection of all the half spaces given by all the supporting hyperplanes at each point of the boundary of $\mathcal{X}$. S is convex and closed as the intersection of closed convex sets. Fix any $s \in S$, and assume that $s \notin \mathcal{X}$, choose x in the interior of $\mathcal{X}$, the line segment between s an x crosses the boundary of $\mathcal{X}$ at $y \in \mathcal{X}$. The supporting hyperplane at y provides an affine function A which is positive on $\mathcal{X}$. Restriction of this affine function to the line segment $[x, s]$ is still affine with $A(x) > 0$, $A(y) = 0$ and $y \in (x, s)$, and hence $A(s) < 0$ which contradicts the fact that $s \in S$. Therefore, $S = \mathcal{X}$. □

There is a stronger notion of separating hyperplane.

Theorem 4.5.6 (Separating hyperplane)**.** *Let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ be two disjoint convex sets, then there exists a vector $v \in \mathbb{R}^d$, $v \neq 0$ and a number $c \in \mathbb{R}$ such that $x^T v \geq c$ for all $x \in \mathcal{X}$ and $y^T v \leq c$ for all $y \in \mathcal{Y}$.*

Extreme points, polyhedra and polytopes

Definition 4.5.2. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex set and $x \in \mathcal{X}$. x is an extreme point of $\mathcal{X}$ if for any $x_1, x_2 \in \mathcal{X}$, $x = (x_1 + x_2)/2$ implies that $x_1 = x_2 = x$.*

Any nonempty compact convex subset of $\mathbb{R}^d$ contains at least one extreme point (any point of maximal norm). The convex hull of a set S is the set of all convex combinations of elements of S, denoted by

$$\mathrm{conv}(S) = \left\{ x,\ \exists n \in \mathbb{N}^*, (x_i)_{i=1}^n \in S^n,\ (\lambda_i)_{i=1}^n \in \mathbb{R}_+^n,\ \sum_{i=1}^n \lambda_i = 1,\ x = \sum_{i=1}^n \lambda_i x_i \right\}.$$

It is seen from the definition that the extreme points of $\text{conv}(S)$ are contained in S. The interest of extreme points is that linear optimization attains its optima at extreme points.

Lemma 4.5.2. *Let $\mathcal{X}$ be a closed convex set, $x \in \mathcal{X}$ such that there exists $c \neq 0$ and $c^T x = \inf_{y \in \mathcal{X}} c^T y$. Then setting $A = \left\{y \in \mathcal{X},\, y^T c = x^T c\right\}$, any extreme point of A is an extreme point of $\mathcal{X}$.*

Proof. Take $\tilde{p}$ to be one extreme point of A, and suppose that we have $x_1, x_2 \in \mathcal{X}$ such that $(x_1 + x_2)/2 = \tilde{p}$. We have $x_1^T c \geq x^T c$, $x_2^T c \geq x^T c$ and $\frac{1}{2}(x_1 + x_2)^T c = \tilde{p}^T c = x^T c$, the average of two non negative numbers is 0 if and only if both are null and hence $x_1^T c = x_2^T c = \tilde{p}^T c$ and $x_1 \in A$ and $x_2 \in A$. Hence $x_1 = x_2 = \tilde{p}$. □

Lemma 4.5.3. *Let $c \in \mathbb{R}^d$, $c \neq 0$ and $\mathcal{X}$ be a convex and compact set. Then $\min x \in \mathcal{X} c^T x$ is attained then the optimum is attained at an extreme point $\bar{x} \in \mathcal{X}$.*

Proof. If $c = 0$, any extreme point of $\mathcal{X}$ is a solution. If $c \neq 0$, by the compactness of $\mathcal{X}$, the optimum of the problem is attained. Take x^* to be one solution. The set $\{x \in \mathbb{R}^d,\, x^T c = (x^*)^T c\}$ is compact and convex, it contains an extreme point which by Lemma 4.5.2 is an extreme point of $\mathcal{X}$. □

Theorem 4.5.7 (Krein Millman)**.** *Let $\mathcal{X}$ be a compact convex set, then $\mathcal{X} \subset \mathbb{R}^d$ is the convex hull of its extreme points.*

Proof. Let S denote the set of extreme points of $\mathcal{X}$, we have $\text{conv}(S) \subset \mathcal{X}$. Let $x \in \mathcal{X}$, we show that x is in $\text{conv}(S)$. First, we may assume that $\mathcal{X}$ has non empty interior, by reducing the ambiant space to the smallest affine subspace containing $\mathcal{X}$. The proof is now by recursion on d. For $d = 0$, the result is obvious, assume that the result holds for $\mathbb{R}^{d-1}$, $d \geq 1$. Consider any line passing through x, the restriction of $\mathcal{X}$ to this line is compact convex set of dimension 1, that is a segment of the form $[a, b]$ where $a \in \mathcal{X}$, $b \in \mathcal{X}$. Both a and b are on the boundary of $\mathcal{X}$. There is a supporting hyperplane H_a at a and H_b at b. Both sets $\mathcal{X} \cap H_a$ and $\mathcal{X} \cap H_b$ are compact convex sets of dimension $d - 1$. The induction hypothesis ensures that both a and b are convex combinations extreme points of H_a and H_b which are extreme points of $\mathcal{X}$ by Lemma 4.5.2 and the result follows because m is a convex combination of a and b. □

Definition 4.5.3. *A polyhedra is a set $\mathcal{X} \subset \mathbb{R}^d$ which can be described by linear equalities: there exists $A \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$ such that $\mathcal{X} = x \in \mathbb{R}^d$, $Ax \leq b$, where the inequality is understood entry-wise. This representation is called canonical form.*

Adding slack variables $s \in \mathbb{R}^m$ and considereing x_+ and x_- the entry-wise positive and negative part of x, one may equivalently describe $\mathcal{X} = \{(x_+, x_-, s) \in \mathbb{R}^{2n+m},\, s = b - A(x_+ - x_-),\, s \geq 0, x_+ \geq 0, x_- \geq 0\}$. Hence, one may equivalently consider polyhedra expressed as $\mathcal{X} = \{x \in \mathbb{R}^d, Ax = b, x \geq 0\}$ for a matix A and a vector b which is called standard form.

Lemma 4.5.4. *Let $\mathcal{X} = \left\{x \in \mathbb{R}^d, Ax = b, x \geq 0\right\}$ be non empty. Then $\mathcal{X}$ has at least one extreme point and we have the following equivalence*

- *x is an extreme point of $\mathcal{X}$*
- *the columns of A corresponding to non zero entries of x are independent.*

Proof. The existence of extreme points follow from the characterization. If $A = 0$, then $x = 0$ is an extreme point. Suppose that $A \neq 0$, for any $x \in \mathbb{R}^d$, denote by A_x the matrix which columns correspond to the non zero entries of x. For any $x, x_1, x_2 \in \mathcal{X}$, if $x = \frac{x_1 + x_2}{2}$, then $\text{supp}(x_1) \subset \text{supp}(x)$ and $\text{supp}(x_2) \subset \text{supp}(x)$ and $A_x(x_1 - x_2) = 0$ hence, if the columns of A_x are independent, $x_1 = x_2$ and x is an extreme point. On the other hand, if the columns of A_x are not independant, choosing $d \in \mathbb{R}^d$ such that $\text{supp}(d) = \text{supp}(x)$ and $Ad = 0$, one has, for sufficiently small alpha that $x + \alpha d \in \mathcal{X}$ and $x - \alpha d \in \mathcal{X}$ so that x is not an extreme point of $\mathcal{X}$. □

As a result, polyhedra have only finitely many extreme points. A polytope is a compact polyhedra. Krein-Millman theorem ensures that $\mathcal{X}$ is a polytope if and only if it is the convex hull of finitely many points.

Example 4.5.2. *The ℓ_1 ball used to define the $\ell - 1$ constrained least squares estimator:*

$$\hat{\theta}_K^{LS} \in \arg \min_{\|\theta\|_1 \leq 1} \|\mathbb{X}\theta - y\|_2^2$$

is a polytope which has $2d$ extreme points corresponding to plus or minus the elements of the canonical basis. Linear fuction over the ℓ_1 ball attains their optimum at one of these extreme points which have a support of size 1. This illustrates the sprasity promoting role of this constraint.

4.6 Conic programming

4.6.1 Conic hierarchy

Definition 4.6.1. *$\mathcal{K} \subset \mathbb{R}^d$ is a cone if it satisfies for any $x \in \mathcal{K}$ and $\alpha \geq 0$, $\alpha x \in \mathcal{K}$.*

Given a closed convex cone $\mathcal{K}$, one can define the corresponding conic program, for any $A \in \mathbb{R}^{m \times d}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^d$,

$$p^* = \inf_{x \in \mathbb{R}^d} \quad c^T x \qquad \text{s.t.} \qquad Ax = b,\, x \in \mathcal{K}. \tag{P}$$

This gives rise to the following classes of optimization problems.

Linear programs: Choosing $\mathcal{K} = \mathbb{R}_+^d$, we obtain a linear program in standard form. The problem of computing $\hat{\theta}_{CS}^{\ell_1}$ can be expressed as a linear program as

$$\begin{aligned} &\min_{\theta \in \mathbb{R}^d} \|\theta\|_1 \quad \text{s.t.} \quad \mathbb{X}\theta = y \\ = &\min_{\theta_+ \in \mathbb{R}^d, \theta_- \in \mathbb{R}^d} 1^T(\theta_+ + \theta_-) \quad \text{s.t.} \quad \mathbb{X}(\theta_+ - \theta_-) = y,\, \theta_+ \in \mathcal{K},\, \theta_- \in \mathcal{K}. \end{aligned}$$

which is a linear program (LP).

Second order cone: The second order cone in $\mathbb{R}^{d+1}$ is given by $\mathcal{K} = \left\{(x,t) \in \mathbb{R}^{d+1},\, \|x\|_2 \leq t\right\}$. This allows to express linear optimization over convex quadratic constraints such as balls or ellipses and their intersection. Such a problem is called a second order cone program (SOCP).

Semidefinite cone: The set of symmetric positive semidefinite is called the semidefinite cone. Given a symmetric matrix $C \in \mathbb{R}^{d \times d}$, a linear function $\mathcal{A}\colon \mathbb{R}^{d \times d} \to \mathbb{R}^m$ and $b \in \mathbb{R}^m$ a semidefinite program has the form

$$\min_{X \in \mathbb{R}^{d \times d}} \operatorname{tr}(C^T X) \quad \text{s.t.} \quad \mathcal{A}(X) = b,\, X^T = X,\, X \succcurlyeq 0.$$

Such programs are called semidefinite programs (SDP).

Hierarchy of conic programs These conic programs are standard optimization problems for which there exists efficient algorithms allowing to solve numerically efficiently moderate size programs of this type. The term hierarchy refers to the fact that linear programs can be expressed as second order cone programs and second order cone programs can be expressed as semidefinite programs.

4.6.2 Conic duality

Definition 4.6.2. *Let $\mathcal{K} \subset \mathbb{R}^d$ be a convex cone, the dual cone of $\mathcal{K}$ is denoted by*

$$\mathcal{K}^* = \left\{ y \in \mathbb{R}^d,\, x^T y \geq 0,\, \forall x \in \mathcal{K} \right\}$$

If $\mathcal{K} = \mathcal{K}^*$, we say that $\mathcal{K}$ is self dual

All the cones given in the previous section are self-dual. The Lagrangian of problem (P) is given for any $x \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$, $\nu \in \mathbb{R}^m$, by

$$\mathcal{L}(x, \mu) = c^T x + \mu^T (b - Ax) \tag{4.1}$$

The dual problem to (P) is obtained by minimizing the Lagrangian over $x \in \mathcal{K}$. If $c^T - A^T\mu \notin \mathcal{K}^*$, the infimum of the Lagrangian over $x \in \mathcal{K}$ is $-\infty$. On the other hand, if $c - A^T\mu \in \mathcal{K}^*$, then the minimizer of the Lagrangian is $\mu^T b$. Hence the dual problem has the form

$$d^* = \sup b^T \mu \qquad \text{s.t.} \qquad c - A^T \mu \in \mathcal{K}^*. \tag{D}$$

We have the following relation between primal (P) and dual problems (D).

Theorem 4.6.1. *It holds that $d^* \leq p^*$. Furthermore, if $\operatorname{rank}(A) = m$, and there exists $\bar{x}$ such that $A\bar{x} = b$ and $\bar{x}$ is in the interior of $\mathcal{K}$ and $p^* > -\infty$, then $p^* = d^*$ and the dual problem has a solution. In this case, x is primal optimal if and only if it is primal feasible and there exists a dual feasible μ such that*

$$x^T(c - A^T\mu) = 0 \qquad \textit{or} \qquad x^T c = b^T \mu.$$

Proof. If either the primal (P) or the dual problem (D) are not feasible, then the result is obvious as $p^* = +\infty$ or $d^* = -\infty$.

Assuming that both are feasible, for any x feasible for (P) and μ feasible for (D), we have

$$c^T x = c^T x + \mu^T(b - Ax) = \mathcal{L}(x, \mu) = \mu^T b + (c - A^T\mu)^T x \geq \mu^T b \tag{4.2}$$

where the first equality is from primal feasibility, and the last inequality is because $x \in \mathcal{K}$ and $c - A^T\mu \in \mathcal{K}^*$ so that the dot product is nonnegative. This implies that $p^* \geq d^*$

To obtain strong duality (not assuming dual feasibility), consider the sets

$$S_1 = \left\{ (u - x), b - Ax, c^T x + t) \in \mathbb{R}^{d+m+1},\, x \in \mathbb{R}^d,\, u \in \mathcal{K},\, t \geq 0 \right\} \qquad S_2 = \{(0, 0, s),\, s < p^*\}$$

It holds that both S_1 and S_2 are convex. Furthermore, they are disjoint since an element of the intersection would provide a primal feasible x with $c^T x < p^*$. Theorem 4.5.6 ensures that there exists $\alpha_1 \in \mathbb{R}^d$, $\alpha_2 \in \mathbb{R}^m$, $\alpha_3 \in \mathbb{R}$, not all equal to 0, and $\alpha_4 \in \mathbb{R}$ such that for all $x \in \mathbb{R}^d$, $u \in \mathcal{K}$, $t \geq 0$, $s < p^*$

$$\begin{aligned} \alpha_1^T(u - x) + \alpha_2^T(b - Ax) + \alpha_3(c^T x + t) &\geq \alpha_4 \\ \alpha_3 s &\leq \alpha_4 \end{aligned} \tag{4.3}$$

It must hold that $\alpha_3 \geq 0$ and $\alpha_1 \in \mathcal{K}^*$, otherwise, the left hand side of the first inequality is unbounded from below. From the second inequality, we obtain $\alpha_3 p^* \leq \alpha_4$. We are going to show that the strict feasibility condition ensures that $\alpha_3 > 0$. First note that if $x \in \operatorname{int}(\mathcal{K})$, then for any nonzero $\mu \in \mathcal{K}^*$, we have $x^T\mu > 0$. Second assuming that $\alpha_3 = 0$, choosing $\bar{x}$ as given in the hypothesis and $u = 0$, one has

$$-\alpha_1^T \bar{x} \geq \alpha_3 p^* = 0,$$

and since $\alpha_1^T \neq 0$ implies $\alpha_1^T \bar{x} > 0$ we have $\alpha_1 = 0$. Furthermore, we have $\alpha_2 \neq 0$ and $\alpha_2^T(b-Ax) \geq 0$, for all $x \in \mathbb{R}^d$. This is impossible as it would imply $\alpha_2^T A = 0$ with $\alpha_2 \neq 0$ which contradicts the rank assumption on A.

Finally, $\alpha_3 > 0$ and we obtain from (4.3), for any $x \in \mathcal{K}$

$$\frac{\alpha_2^T}{\alpha_3}(b - Ax) + c^T x \geq p^* + \frac{\alpha_1^T}{\alpha_3} x \geq p^*,$$

where the last inequality follows because $\alpha_1 \in \mathcal{K}^*$. This implies that $c - A^T\alpha_2/\alpha 3 \in \mathcal{K}^*$ as otherwise, the right hand side would be unbounded from below. Hence $\mu = \alpha_2/\alpha_3$ is dual feasible. Minimizing over x, we obtain that $b^T\mu \geq p^*$. Hence $d^* \geq p^*$ and by weak duality, $d^* = p^*$ and μ is dual optimal.

The last statement follows from the existence of a dual optimal μ and inequality (4.2). □

4.6.3 Interior point methods

Interior point methods were discovered in the 80's, Karmarkar polynomail time (and empirically efficient) algorithm for linear programing was based on interior point methods. There has been an important activity around interior point methods in the 90's. We refer to [10] for a detailed presentation. In this section, we will only briefly touch the topic and describe the main ideas on a simple problem.

4.6.4 Strong convexity

This notion will be important to develop algorithmic ideas to solve the optimization problems which we have seen.

Definition 4.6.3. *A function $f\colon \mathbb{R}^d \mapsto \mathbb{R}$ is μ strongly convex, if $f - \frac{\mu}{2}\|\cdot\|$ is convex. The following provide sufficient conditions:*

- *If f is differentiable, $f(y) \geq f(x) + (y-x)^T\nabla f(x) + \frac{\mu}{2}\|y-x\|_2^2$, for all x, y.*
- *If f is differentiable, $(\nabla f(x) - \nabla f(y))^T (x-y) \geq \mu\|y-x\|_2^2$ for all x, y.*
- *If f is twice differentiable, the matrix $\nabla^2 f(x) - \mu I$ is positive semidefinite for all x.*

Exercise 4.6.1. *Prove that the function $f\colon x \mapsto -\log(1-\|x\|^2)$ is strongly convex (when restricted to the unit Euclidean ball).*

Newton's method

Newton's method is famously used to solve equations of the form $g(x) = 0$. In the context of convex optimization, one actually solves $f'(x) = 0$. Application of this method to find a zero of the gradient operator of a strongly convex function $f\colon \mathbb{R}^d \mapsto \mathbb{R}$, can be implemented as follows: choose x_0 and iterate for $k \in \mathbb{N}$,

$$x_{k+1} = x_k - \alpha \left(\nabla^2 f(x_k)\right)^{-1} \nabla f(x_k). \tag{4.4}$$

Where α is a positive stepsize, determined algorithmically. Note that this equation is well defined since by strong convexity, the Hessian is always positive definite and invertible. One intuition about this method is that it minimizes the second order Taylor expansion: $f(y) \simeq f(x) + \nabla f(x)^T(y-x) + (y-x)^T\nabla^2 f(x)(y-x)$.

A detailed convergence rate analysis of Newton's method can be found in [10, 13, 20]. We prove a local quadratic convergence result which illustrate the fast asymptotic convergence of the method. A more refined analysis is more involved and requires to analyse backtracking line search procedures. We limit ourselves here to a local result stating that when initialized close to the optimum, Newton's method with unit step sizes is extremely fast.

Theorem 4.6.2. *Let f be μ-strongly convex, twice continuoulsy differentiable, with L-Lipschitz Hessian (operator norm) and $\bar{x}$ be the (unique) minimum of f. Newton's method with unit step size satisfy, for all $k \in \mathbb{N}$,*

$$\frac{L}{2\mu^2}\|\nabla f(x_k)\|_2 \leq \left(\frac{L}{2\mu^2}\|\nabla f(x_0)\|_2\right)^{2^k},$$

In particular, if $\|\nabla f(x_0)\|_2 < \frac{L}{2\mu^2}$, we obtain extremely fast convergence for Newton's method with unit step size.

Proof. Fix $k \in \mathbb{N}$. From the Newton iterate, we have $\nabla^2 f(x_k)(x_{k+1} - x_k) = -\nabla f(x_k)$. Hence integrating along the segment $[x_{k+1}, x_k]$, we have

$$\begin{aligned}\nabla f(x_{k+1}) &= \nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)(x_{k+1} - x_k) \\ &= \int_{t=0}^{1} \left(\nabla^2 f(x_k + t(x_{k+1} - x_k)) - \nabla^2 f(x_k)\right)(x_{k+1} - x_k)dt\end{aligned}$$

Using the Lipschitz assumption, we obtain

$$\|\nabla f(x_{k+1})\|_2 \leq \frac{L}{2}\|x_{k+1} - x_k\|_2^2 = \frac{L}{2}\|\nabla^2 f(x_k)^{-1}\nabla f(x_k)\|_2^2 \leq \frac{L}{2\mu^2}\|\nabla f(x_k)\|_2^2$$

By a simple recursion, we have

$$\frac{L}{2\mu^2}\|\nabla f(x_{k+1})\|_2 \leq \left(\frac{L}{2\mu^2}\|\nabla f(x_k)\|_2\right)^2 \leq \left(\frac{L}{2\mu^2}\|\nabla f(x_0)\|_2\right)^{2^k}$$

□

Interior point method

We refer to [10] for a more detailed overview. We illustrate the idea of interior point methods for the follwing toy problem, for a given $a \in \mathbb{R}^d$, $b \in \mathbb{R}$, and $f\colon \mathbb{R}^d \mapsto \mathbb{R}$, convex differentiable

$$f^* = \min_{x \in \mathbb{R}^d} f(x) \qquad \text{s.t.} \qquad \|x\|_2 \leq 1,\ a^T x \leq b \tag{4.5}$$

We only use this problem to illustrate the main idea of interior point methods. The main idea of interior points methods is to replace this problem by an unconstrained problem using a barier function, for any $t \geq 0$,

$$\min_{x \in \mathbb{R}^d} tf(x) - \log(1 - \|x\|_2^2) - \log(b - a^T x) \tag{4.6}$$

Note that we need to restrict the domain of definition of the objetctive, since the logarithms explode on the boundary of the feasible set. By example 4.6.1, the objective in (4.6) is 2 strongly convex. Denoting by x_t the minimal value of (4.6) for a given $t \geq 0$, this defines the notion of *central path*, a quick argument shows that

$$f(x_t) \underset{t \to \infty}{\to} f^*$$

and furthermore for each t, x_t can be computed efficiently using Newton's method. This provides an algorithm to solve problem (4.6). A detailled complexity analysis of these types of methods is found for example in [46]. Let us mention that the optimality conditions for (4.6), ensure that

$$t\nabla f(x_t) + 2x_t \frac{1}{1 - \|x_t\|_2} + a\frac{1}{b - a^T x_t}$$

so that x_t minimizes also

$$x \mapsto tf(x) + \frac{1}{1 - \|x_t\|_2} \left(\|x\|_2^2 - 1 \right) + \frac{1}{b - a^T x_t} \left(a^T x - b \right)$$

This entails that for any feasible x, we have

$$tf(x_t) - 2 \leq tf(x) + \frac{1}{1 - \|x_t\|_2} \left(\|x\|_2^2 - 1 \right) + \frac{1}{b - a^T x_t} \left(a^T x - b \right) \leq f(x),$$

so that $f(x_t) \leq f^* + \frac{2}{t}$ and an ϵ suboptimal solution for (4.5) can be found by choosing $t = 2/\epsilon$.

General purpose solvers

One of the most important topics in Optimization during the 90's was interior point methods. These developments led to theoretical and practical results which materialize in the existence of efficient numerical solvers for the classes of conic problems which were discussed in this section.

4.6.5 Polynomial time LP solvers over $\mathbb{Q}$

Algorithms to solve the LP problem date back to Fourier, Kantorovitch and Dantzig who proposed the simplex method still used in many numerical solvers.

Theorem 4.6.3 (Khachiyan,Karmarkar)**.** *Given inputs $A \in \mathbb{Q}^{n \times d}$, $b \in \mathbb{Q}^n$ and $c \in \mathbb{Q}^d$ consider the problem of computing*

$$\rho = \inf_{x \in \mathbb{Q}^d} c^T x \quad \text{s.t. } Ax \leq b. \tag{4.7}$$

This problem is in $\mathcal{P}$.

Proof sketch. We only sketch the main ideas, a full detailed proof is very tedious. We refer to Schiver's book [57] for more details.

- First if the infimum is not attained, either the original problem or its dual are unfeasible and there polynomial time certificates for this can be found in polynomial time.
- If the problem attains its optimum, then it must attain its optimum at one of the vertices of the polyhedra described by the linear inequalities. There are only finitely many of them.
- There are only polynomially many candidate optimal values for ρ. This is because we have finitely many candidate solutions and the size of the input allows to estimate size of largest common denominators and condition numbers of A.
- Local search methods such as ellipsoid method (for Khachiyan's algorithm) or interior point methods (for Karmarkar's algorithm) converge exponentially fast to ρ (see interior point methods).
- Carefully controling the magnitude of accumulated errors allong the local search path and the degree of approximation required to dicriminate between any two candidate optimal values allow to conclude.

□

Historically, the ellipsoid method was the first polynomial time algorithm for linear programming, it has been studied by various authors in the 70's including Shor, Yudin and Nemirovski. It was proved to be polynomial time by Khachiyan [35] but is quite inefficient in practice. Karmarkar proposed the first polynomial time algorithm which was efficient empirically, based on interior point methods [34].

Corollary 4.6.1. *Assuming the model 3.10 holds and $\theta^* \in \mathbb{Q}^d$, $\hat{\theta}^{\ell_1}_{CS}$ in (3.12) is computable exactly using a number of operations which is at most polynomial in n, d and the number of bits required to encode $\mathbb{X}$ and $\mathbb{X}\theta^*$.*

Remark 4.6.1. *Such a result cannot hold for second order cone programs and semidefinite programs. This is because the solution of such programs may not be in $\mathbb{Q}$ eventhough the data is in $\mathbb{Q}$. For example*

$$\begin{aligned} & \min_x \|x\|_2 \quad \text{s.t.} \quad x_1 \geq 1, x_2 \geq 2 \\ = \quad & \min_{x,t} t \quad \text{s.t.} \quad x_1 \geq 1, x_2 \geq 2, t \geq \|x\|_2 \end{aligned}$$

is a second order cone program which value is attained only for $x_1 = x_2 = 1$ and $t = \sqrt{2}$. Hence the solution of this program cannot be found over $\mathbb{Q}$ and one must switch to computation over $\mathbb{R}$, in particular, the program cannot be solved exactly by finite precision numerical methods. As we have seen, computation over $\mathbb{R}$ has different formulations and connections with practice on physical computers is sometimes a bit far fetched. Hence when one talks about polynomial time solvability of general convex program, this is not in the classical Church-Turing thesis sense but in a different sense such as: polynomial time approximation to a any fixed precision, or polynomial time computation over real machines (which do not exist in the physical world).

Another remark of the same kind goes as follows, the matrix

$$\begin{pmatrix} 1 & y \\ y & x \end{pmatrix}$$

being semidefinite positive implies that $x \geq y^2$ and pilling up k such equalities allows to express numbers of the order 2^{2^k} which bit representation size is exponential in k. Hence such a number cannot be approximated in time polynomial in k using standard numerical integer encoding.

Remark 4.6.2. *In the context of linear programing (LP), since the number of candidate solution is finite (extreme points of the undelying polyhedra), and we have explicit description of these points (lemma 4.5.4), one could try to build an algorithm for finding an optimal extreme point. This is the basis for the Simplex method proposed by Dantzig in 1947 and still used in many numerical softwares. We do not describe it here, but mention that it is an efficient method in practice. However there do not exist polynomial time worst case bounds for these types of algorithm. There exist polynomial time bounds for average instances of linear programs and the simplex method is one of the candidate polynomial time algorithm to solve linear programing over the reals. It also motivate many questions about the geometry of polyhedra such as the Hirsh conjecture.*

Exercises

Exercise 4.6.2. *Prove that Lemma 4.5.1, for any convex set $\mathcal{X} \subset \mathbb{R}^d$ we have*

- *The closure of $\mathcal{X}$ is convex.*
- *The interior of $\mathcal{X}$ is convex.*
- *For any $u \in \text{int}(\mathcal{X})$ and $v \in \text{cl}(\mathcal{X})$, $[u, v) \subset \text{int}(\mathcal{X})$.*
- *If the interior of $\mathcal{X}$ is non empty, then $\text{cl}(\mathcal{X}) = \text{cl}(\text{int}(\mathcal{X}))$.*
- *The interior of $\mathcal{X}$ is empty if and only if it is contained in a lower dimensional affine subspace.*

Exercise 4.6.3. *Prove Theorem 4.5.1, et $f \colon \mathbb{R}^d \to \mathbb{R}$:*

1. *If f is continuously differentiable, then f is convex if and only if or any $x, y \in \mathbb{R}^d$, $f(y) \geq f(x) + \nabla f(x)^T (y - x)$.*
2. *If f is continuously differentiable, then f is convex if and only if or any $x, y \in \mathbb{R}^d$, $(\nabla f(x) - \nabla f(y))^T (y - x) \geq 0$.*
3. *If f is twice continuously differentiable, then f is convex if and only if or any $x \in \mathbb{R}^d$, $\nabla^2 f(x)$ is positive semidefinite.*

Exercise 4.6.4. *Prove Theroem 4.5.6, let $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$ be two disjoint convex sets, then there exists a vector $v \in \mathbb{R}^d$, $v \neq 0$ and a number $c \in \mathbb{R}$ such that $x^T v \geq c$ for all $x \in \mathcal{X}$ and $y^T v \leq c$ for all $y \in \mathcal{Y}$.*

Exercise 4.6.5. *Prove that the different conditions for strong convexity are indeed equivalent to $f - \mu \| \cdot \|_2^2$:*

- *If f is differentiable, $f(y) \geq f(x) + (y - x)^T \nabla f(x) + \frac{\mu}{2} \|y - x\|_2^2$, for all x, y.*
- *If f is differentiable, $(\nabla f(x) - \nabla f(y))^T (y - x) \geq \mu \|y - x\|_2^2$ for all x, y.*
- *If f is twice differentiable, the matrix $\nabla^2 f(x) - \mu I$ is positive semidefinite for all x.*

Exercise 4.6.6. *Prove that the function $f \colon x \mapsto -\log(1 - \|x\|^2)$ is strongly convex (when restricted to the unit Euclidean ball).*

Exercise 4.6.7. *Let $\mathcal{S}_d^+$ denote the cone of positive semidefinite matrices in $\mathbb{R}^{d \times d}$. We consider the function $h \colon S \mapsto \log(\det(S))$ over $\mathcal{S}_d^{++}$ the cone of positive definite matrices.*

- *Compute the gradient of $\det$ over $\mathcal{S}_d^{++}$ (Hint: use the relation between S^{-1}, $\det(S)$ and C the adjugate matrix of S).*
- *Compute the gradient of h.*
- *Show that h is convex.*
- *Explain how h could be used as a barrier function for interior point methods in semi-definite programming.*

Chapter 5

First order methods

The preceeding chapter was the occasion to describe one of the a fundamental difference between statistical estimation problems which can or cannot be solved in polynomial time. Efficient numerical solvers exist for $\mathcal{NP}$-hard problems and their use in high dimensional statistics is explored [14]. Nontheless, we will focus on algorithms which are less computationally demanding, and better scale in very large dimensions.

We have seen that large families of convex optimization problems can be solved via generic purpose solvers which have efficient implementations. These solvers have the following properties, in dimension d:

- The cost of a single iteration is of the order of d^3
- They lead to fast converging sequences allowing to obtain very accurate solutions.

In very large dimensions d^3 may be too big to be considered as reasonable and we need cheaper algorithms. This observation motivated the rise of first order methods as efficient alternatives in high dimensional statistics and signal processing. These methods have a long history in applied mathematics and the recent trends in data analysis bolstered new developments.

Sources for this chapter include the classic book of Rockafellar [55], the book of Nesterov [45] as well as elements presented in Sébastien Bubeck's book [21]. Good references on this topic include the surveys [26, 6] which is very close to the statistical matters presented in these notes.

5.1 Gradient descent

In this section f denotes a continuously differentiable function. The gradient descent algorithm can be described as follows, choose $x_0 \in \mathbb{R}^d$ and iterate for $k \in \mathbb{N}$:

$$x_{k+1} = x_k - s_k \nabla f(x_k) \tag{5.1}$$

Each iteration costs a call to the gradient with a vector addtion which. A vector addition costs of the order of d operations. Hence it is much cheaper than the d^3 operations required to run interior point methods. For example, if one is given a computational budget of the order of d^2, then one can implement d steps of gradient descent while Newton step simply cannot be considered. We review basic theoretical results known for the gradient method for convex optimization.

5.1.1 Dynamical systems intuition

The minimizing properties of gradient descent are easily seen in continuous time.

Proposition 5.1.1. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be twice differentiable with compact sublevel sets. Consider the differential equation, for $x_0 \in \mathbb{R}^p$,*

$$\dot{x}(t) = -\nabla f(x(t)) \tag{5.2}$$
$$x(0) = x_0. \tag{5.3}$$

Then, there exists a solution to the initial value problem defined for all $t > 0$.

- *$\int_0^{+\infty} \|\nabla f(x(t))\|_2^2 dt < +\infty$ and $\lim_{t\to\infty} \|\nabla f(x(t))\| = 0$.*
- *Any accumulation point $\bar{x}$ of the trajectory satisfies $\nabla f(\bar{x}) = 0$.*
- *If in addition f is convex, set $f^* = \inf_{x\in\mathbb{R}^p} f(x)$ and assume that it is attained at x^*, we have for any $t \in \mathbb{R}$, $t > 0$,*

$$f(x(t)) - f^* \leq \frac{\|x_0 - x^*\|_2^2}{2t}.$$

And $x(t) \underset{t\to\infty}{\to} \bar{x}$ where $\bar{x}$ is a global minimizer of f.

Proof. First note that ∇f is continuous and locally Lipschitz so that there is a unique maximal solution to the initial value problem (Cauchy-Lipschitz). By differentiation, we obtain, for any t in the interval of definition of the solution,

$$\frac{d}{dt}\left(f(x(t))\right) = \dot{x}(t)^T \nabla f(x(t)) = -\|\nabla f(x(t))\|_2^2.$$

We deduce that $f(x(t)) \leq f\left(x_0\right)$ and by compacity the trajectory remains bounded and is defined for all $t > 0$ (sortie de tout compact). Integrating between 0 and $T > 0$, we obtain $f(x(T)) = f(x(0)) - \int_0^T \|\nabla f(x(t))\|_2^2 dt$. The function f is decreasing allong the trajectory and bounded below by compacity. This proves the first point. Since f has bounded level sets, the trajectory remains bounded so that both ∇f and x can be considered to be Lipschtiz. Square integrable Lipschitz function converge to 0 and this proves that the gradient goes to 0.

The second point is a direct consequence.

For the third point, using convexity of f, we obain

$$\frac{d}{dt}\|x(t) - x^*\|_2^2 = -2\left\langle \nabla f(x(t)), x(t) - x^*\right\rangle \leq 2\left(f(x^*) - f(x(t))\right) < 0.$$

Integrading between 0 and $t > 0$, we obtain

$$t(f(x(t)) - f^*) \leq \int_0^t f(x(s)) - f^* ds \leq \frac{1}{2}\|x(0) - x^*\|_2^2$$

where the first inequality follows because f is decreasing allong the trajectory. For the convergence of $x(t)$, we have

$$\frac{d}{dt}\|x(t) - x^*\|_2^2 \leq 2\left(f(x^*) - f(x(t))\right) < 0,$$

so that $\|x(t) - x^*\|_2^2$ is decreasing and the trajectory remains bounded. Since $x(t)$ has at least one accumulation point which attains the minimum of f, $x(t)$ must converge to this point (Opial's Lemma). □

5.1.2 Convergence of gradient descent

We start with the following Lemma which proof is left as an exercise.

Lemma 5.1.1. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be continuously differentiable with L-Lipschitz gradient ($L > 0$), then for any $x, y \in \mathbb{R}^p$,*

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \frac{L}{2}\|y - x\|_2^2.$$

Proof. Using the fundamental theorem of calculus, we have, for any x, y

$$\begin{aligned} f(y) - f(x) &= \int_{t\in[0,1]} \langle \nabla f((1-t)x + ty), y - x\rangle \, dt \\ &= \int_{t\in[0,1]} \langle \nabla f((1-t)x + ty) - \nabla f(x) + \nabla f(x), y - x\rangle \, dt \\ &= \langle \nabla f(x), y - x\rangle + \int_{t\in[0,1]} \langle \nabla f((1-t)x + t(y)) - \nabla f(x), y - x\rangle \, dt. \end{aligned}$$

We deduce that

$$\begin{aligned} |f(y) - f(x) - \langle \nabla f(x), y - x\rangle| &= \left| \int_{t\in[0,1]} \langle \nabla f((1-t)x + ty) - \nabla f(x), y - x\rangle \, dt \right| \\ &\leq \int_{t\in[0,1]} |\langle \nabla f((1-t)x + ty) - \nabla f(x), y - x\rangle| \, dt \\ &\leq \int_{t\in[0,1]} \|f((1-t)x + ty) - \nabla f(x)\| \times \|y - x\| dt \\ &\leq \int_{t\in[0,1]} tL \times \|y - x\|^2 dt \\ &= \frac{L}{2}\|y - x\|^2, \end{aligned}$$

which proves the result. □

The gradient descent algorithm can be seen as an explicit discretisation of the differential equation (5.3). It preserves the same qualitative properties as seen in the following proposition.

Proposition 5.1.2. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be continuously differentiable with L-Lipschitz gradient and such that $\inf_{x\in\mathbb{R}^p} f(x) > -\infty$. Consider the algorithm, for $x_0 \in \mathbb{R}^p$ and*

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k). \tag{5.4}$$

Then

- $\lim_{k\to\infty} \|\nabla f(x_k)\| = 0$, *(any accumulation point $\bar{x}$ of the trajectory satisfies $\nabla f(\bar{x}) = 0$).*
- *If in addition f is convex, set $f^* = \inf_{x\in\mathbb{R}^p} f(x)$ and assume that it is attained at x^*, we have for any $k \in \mathbb{N}$, $k > 0$,*

$$f(x_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2k}.$$

Furthermore x_k converges to $\bar{x}$ a global minimum of f

- *If in addition f is μ-strongly convex, then we have for any $k \in \mathbb{N}$*

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)(f(x_k) - f^*).$$

Proof. The ideas are the same, first, the descent Lemma ensures that for any $k \in \mathbb{N}$

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 \\ &= f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|_2^2. \end{aligned} \tag{5.5}$$

Note that that f is decreasing allong the iterates of the algorithm. We have

$$x_{k+1} = \arg\min_{y \in \mathbb{R}^p} f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|_2^2, \tag{5.6}$$

so that for all $y \in \mathbb{R}^d$,

$$f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|_2^2 \tag{5.7}$$

$$= \quad f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + \frac{L}{2}\|y - x_{k+1}\|_2^2. \tag{5.8}$$

We obtain

$$\begin{aligned} & f(x^*) + \frac{L}{2}\|x^* - x_k\|_2^2 \\ \geq\, & f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle + \frac{L}{2}\|x^* - x_k\|_2^2 && \text{convexity} \\ =\, & f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + \frac{L}{2}\|x_{k+1} - x^*\|_2^2 && (5.8) \\ \geq\, & f(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x^*\|_2^2, && (5.5) \end{aligned}$$

By summing up, we obtain for any $K \in \mathbb{N}$, $K \geq 1$,

$$\frac{L}{2}\|x^* - x_0\|_2^2 \geq \sum_{k=1}^{K} f(x_k) - f^* \geq K(f(x_K) - f^*).$$

For the last point we have by strong convexity for any $x \in \mathbb{R}^d$,

$$f(x^*) \geq f(x) + \langle \nabla f(x), x^* - x \rangle + \frac{\mu}{2}\|x^* - x\|_2^2 \geq f(x) - \frac{1}{2\mu}\|\nabla f(x)\|_2^2$$

$$\|\nabla f(x)\|_2^2 \geq 2\mu(f(x) - f^*)$$

We have for all $k \in \mathbb{N}$,

$$f(x_{k+1}) - f^* \leq f(x_k) - f^* - \frac{1}{2L}\|\nabla f(x_k)\|_2^2 \leq (f(x_k) - f^*)\left(1 - \frac{\mu}{L}\right).$$

□

5.2 Recap on nonsmooth analysis

The following content is treated in greater generality in [55]. In what follows f denotes a lower semi-continuous convex function on $\mathbb{R}^p$ which is finite at least at one point. Lower semi-continuity refers to the fact that the epigraph is closed:

$$\mathrm{epi}_f = \left\{(x, z) \in \mathbb{R}^{p+1},\, z \geq f(x)\right\}.$$

which is expressed equivalently as for any $x \in \mathbb{R}^p$

$$\lim \inf_{y \to x} f(y) \geq f(x).$$

The function f is allowed to take value $+\infty$, we denote its domain by

$$\mathrm{dom}_f = \{x \in \mathbb{R}^p,\, f(x) < +\infty\},$$

which is a convex set.

Exercise 5.2.1. *Show that a convex function is continuous on the interior of its domain.*

5.2.1 Notion of subgradient

Definition 5.2.1. *For any $x \in \mathrm{dom}_f$, the subgradient of f denotes the set*

$$\partial f(x) = \{v \in \mathbb{R}^p,\, f(y) \geq f(x) + \langle v, y - x\rangle,\, \forall y \in \mathbb{R}^p\}.$$

For $x \notin \mathrm{dom}_f$, $\partial f(x)$ is set to be empty.

We deduce from the definition the generalization of Fermat rule

Theorem 5.2.1. *$x^* \in \arg\min_x f(x)$ if and only if $0 \in \partial f(x^*)$.*

Proposition 5.2.1. *For any $x \in \mathbb{R}^p$, $\partial f(x)$ is a closed convex set. Furthermore, at any $x \in \mathrm{int}(\mathrm{dom}_f)$, $\partial f(x)$ is non empty and bounded*

Proof. Closedness and convexity follow from the definition. Take $x \in \mathbb{R}^p$ and assume that x is in the interior of the domain of f this means that f is finite around x. The set epi_f is convex in $\mathbb{R}^{p+1}$, and $(x, f(x))$ belongs to the boundary of epi_f. Consider a suporting hyperplane of epi_f at $(x, f(x))$ as given by Theorem 4.5.4, this provides a vector $v \in \mathbb{R}^p$ and a number $a \in \mathbb{R}$ such that for all $y \in \mathrm{dom}_f$

$$az + v^T y \geq af(x) + v^T x, \qquad \forall z \geq f(y).$$

If $a = 0$ then v is different from 0 and this provides a supporting hyperplane to dom_f at x which contradicts the fact that f is finite around x. Hence $a \neq 0$. It must holda that $a > 0$ and $\frac{-v}{a}$ provides a subgradient for f. Boundedness follows because for any $v \in \partial f(x)$,

$$f\left(x + v\frac{1}{\|v\|_2^{3/2}}\right) \geq f(x) + \|v\|_2^{1/2},$$

if the set of such v was unbounded, the left hand side should remain finite while the right hand side should diverge to $+\infty$. □

Exercise 5.2.2. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be a convex function, show that ∂f is sequentialy closed in the sence that, for any $\bar{x}$*

$$\left\{v \in \mathbb{R}^p,\, \exists\, (x_k, v_k)_{k\in\mathbb{N}},\, x_k \to \bar{x},\, v_k \to v,\, v_k \in \partial f(x_k),\, f(x_k) \to f(\bar{x})\right\} \subset \partial f(\bar{x})$$

Exercise 5.2.3. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$, show that f is L-Lipschitz if and only if $\sup_{x\in\mathbb{R}^p,\, v\in\partial f(x)} \|v\|_2 \leq L$.*

Theorem 5.2.2. *Let f be convex and lower semicontinuous and finite at least at one point, then f is the supremum of all its affine minorants: for any $x \in \mathbb{R}^p$*

$$f(x) = \sup_{r\in\mathbb{R}, v\in\mathbb{R}^p} r + v^T x \qquad \text{s.t.} \qquad f(y) \geq r + v^T y,\, \forall y \in \mathbb{R}^p.$$

Proof. epi_f is a closed set in $\mathbb{R}^{p+1}$. Reducing the dimension if necessary and restricting to affine subspaces, we may consider that $\text{int}(\text{dom}_f) \neq \emptyset$. Fix $(x, \mu) \notin \text{epi}_f$, this means that $\mu < \min\{f(x), +\infty\}$. From the separating hyperplane theorem, there exists, $v \in \mathbb{R}^p$, $\beta \in \mathbb{R}$ and $a \in \mathbb{R}$ such that

$$\begin{aligned} &v^T y + \beta z - a \leq 0 \qquad \forall y \in \text{dom}_f,\, z \geq f(y) \\ &v^T x + \beta \mu - a > 0. \end{aligned}$$

If $\beta = 0$, this means that $x \notin \text{dom}_f$. Consider $\bar{x} \in \text{int}(\text{dom}_f)$ and $\tilde{v} \in \partial f(\bar{x})$ (non empty by Proposition 5.2.1), for any $\lambda \geq 0$ and any $y \in \text{dom}_f$

$$\lambda(v^T y - a) + \tilde{v}^T(y - \bar{x}) + f(\bar{x}) \leq f(y),$$

So that we have a family of affine minorants of f parametrized by $\lambda \geq 0$. Furthermore, $\lambda(v^T x - a) + \tilde{v}^T(x - \bar{x}) + f(\bar{x})$ can be chosen arbitrarily big as $\lambda \to \infty$ and the supremum is $+\infty$.

Assume that $\beta \neq 0$, then $\beta < 0$ and we have for any $y \in \text{dom}_f$

$$\frac{1}{-\beta}(v^T y - a) \leq f(y)$$

and furthermore $\frac{1}{-\beta}(v^T x - a) > \mu$. We obtain

$$\begin{aligned} &\frac{1}{-\beta}(v^T y - a) \leq f(y), \qquad \forall y \in \text{dom}_f \\ &\mu < \frac{1}{-\beta}(v^T x - a) \leq \min\{f(x), +\infty\} \end{aligned}$$

since μ is arbitrary, if $f(x)$ is finite, the supremum over all affine lower bounds is $f(x)$, if it is not finite, the supremum is $+\infty$. □

The following is due to Moreau and Rockafellar

Theorem 5.2.3. *For any $x \in \text{int}(\text{dom}_f)$ and any $h \in \mathbb{R}^p$,*

$$D_h f(x) = \sup_{v \in \partial f(x)} \langle v, h \rangle,$$

where D_h denotes the directional derivative of f,

$$D_h f(x) = \lim_{t > 0, t \to 0} \frac{f(x + th) - f(x)}{t}.$$

Proof. By convexity of f, $t \mapsto (f(x + th) - f(x))/t$ is an increasing function of $t > 0$. Indeed, for any $s > t$,

$$f(x + th) = f\left(\left(1 - \frac{t}{s}\right) x + \frac{t}{s}(x + sh)\right) \leq \left(1 - \frac{t}{s}\right) f(x) + \frac{t}{s} f(x + sh),$$

so that

$$\frac{f(x + th) - f(x)}{t} \leq \frac{f(x + sh) - f(x)}{s},$$

Using the Definition 5.2.1, we have for any $v \in \partial f(x)$, any $h \in \mathbb{R}^p$ and $t > 0$,

$$\frac{f(x + th) - f(x)}{t} \geq \langle v, h \rangle$$

which shows by letting $t \to 0$ and taking the supremum on v that

$$D_h f(x) \geq \sup_{v \in \partial f(x)} \langle v, h \rangle .$$

Hence $D_h f(x)$ is well defined for all h and we have one inequality. The function $g\colon h \mapsto D_h f(x)$ is convex, has full domain and is positively homogeneous. By Theorem 5.2.2, we have for any h

$$D_h f(x) = \sup_{r,v} \quad r + v^T h \qquad \text{s.t.} \qquad D_{h'}(x) \geq r + v^T h',\ \forall h' \in \mathbb{R}^p.$$

From positive homogeneity of g, the constraint enforce that for any $t > 0$, $D_{th'} f(x) = t D_{h'} \geq r + t v^T h'$, $\forall h' \in \mathbb{R}^p$ and $D_{h'} f(x) \geq v^T h'$, letting $t \to \infty$, so that r may be chosen to be 0. We deduce that

$$D_h f(x) = \sup_{v} \quad v^T h \qquad \text{s.t.} \qquad D_{h'} f(x) \geq v^T h',\ \forall h' \in \mathbb{R}^p.$$

We notice that if $D_{h'} f(x) \geq v^T h'$, $\forall h' \in \mathbb{R}^p$, then $f(x + h') - f(x) \geq v^T h'$ for all $h' \in \mathbb{R}^p$ so that v is a subgradient of f at x. We obtain

$$\begin{aligned} D_h f(x) &= \sup_{v} \quad v^T h \qquad \text{s.t.} \qquad D_{h'} f(x) \geq v^T h',\ \forall h' \in \mathbb{R}^p \\ &\leq \sup_{v \in \partial f(x)} \quad v^T h. \end{aligned}$$

□

We deduce from this result that f is differentiable at $x \in \operatorname{int}(\operatorname{dom}_f)$ if and only if $\partial f(x) = \{\nabla f(x)\}$.

5.2.2 Legendre transform

Definition 5.2.2. *Given f convex, the Fenchel-Legendre transform of f is given as follows*

$$f^*\colon z \mapsto \sup_{y \in \mathbb{R}^p} z^T y - f(y)$$

Theorem 5.2.4. *For any f convex, f^* is convex and for any $x, z \in \mathbb{R}^p$*

$$f(x) + f^*(z) \geq z^T x$$

and the preceeding inequality holds if and only if $z \in \partial f(x)$. This is called Fenchel-Young's inequality. Furthermore, if f is lower semicontinuous if and only if $(f^*)^* = f$.

Proof. Convexity follows because f^* is the pointwise supremum of affine functions which are convex and convexity is preserved by pointwise suprema. If we have equality, this means that x attains the minimum of the convex function $y \mapsto f(y) - y^T z$ and we must have zero in the subdifferential of this function at x.

From Fenchel-Young's inequality, we have that $f(x) \geq z^T x - f^*(z)$ for all z so that taking the supremum over z, we obtain $f(x) \geq (f^*)^*(x)$ to get equality, we use Theorem 5.2.2. For any $x \in \mathbb{R}^p$

$$\begin{aligned} (f^*)^*(x) &= \sup_{v \in \mathbb{R}^p} v^T x - f^*(v) \\ &= \sup_{v \in \mathbb{R}^p} v^T x - \sup_{y \in \mathbb{R}^p} v^T y - f(y) \\ &= \sup_{v \in \mathbb{R}^p} v^T x + \inf_{y \in \mathbb{R}^p} f(y) - v^T y \\ &= \sup_{v \in \mathbb{R}^p} v^T x + \sup_{r \in \mathbb{R}} r, \qquad \text{s.t.} \qquad f(y) - v^T y \geq r, \quad \forall y \in \mathbb{R}^p \\ &= \sup_{v, r \in \mathbb{R}^p} v^T x + r, \qquad \text{s.t.} \qquad f(y) \geq r + v^T y, \quad \forall y \in \mathbb{R}^p \end{aligned}$$

Hence f^{**} is the supremum of all affine lower bounds of f. As such it is always lower-semicontinuous since its graph is an intersection of closed sets which is closed. Furthermore, when f is lower-semicontinuous, we obtain $f^{**} = f$. □

Example 5.2.1. *Let $f\colon x \mapsto \max_i x_i$, compute the subgradient of this function.*

5.3 Subgradient descent

Subgradient descent generalizes gradient descent to nonsmooth functions.

Proposition 5.3.1. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be a convex function which attains its infimum and has full domain. Consider the algorithm, for $x_0 \in \mathbb{R}^p$, a sequence of positive numbers $\alpha_k > 0$, $k \in \mathbb{N}$, iterate*

$$x_{k+1} = x_k - \alpha_k v_k \tag{5.9}$$
$$v_k \in \partial f(x_k). \tag{5.10}$$

Then for any global minimizer x^*, setting, $y_k = \sum_{i=0}^k \alpha_i x_i / \left(\sum_{i=0}^k \alpha_i\right)$

$$\min_{i=1,\dots,k} f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2 + \sum_{i=0}^k \alpha_i^2 \|v_i\|_2^2}{2\sum_{i=0}^k \alpha_i}$$
$$f(y_k) - f^* \leq \frac{\|x_0 - x^*\|^2 + \sum_{i=0}^k \alpha_i^2 \|v_i\|_2^2}{2\sum_{i=0}^k \alpha_i}.$$

Proof. We have for any $k \in \mathbb{N}$

$$\begin{aligned}\frac{1}{2}\|x_{k+1} - x^*\|_2^2 &= \frac{1}{2}\|x_k - \alpha_k v_k - x^*\|_2^2 \\ &= \frac{1}{2}\|x_k - x^*\|_2^2 + \alpha_k v_k^T(x^* - x_k) + \frac{\alpha_k^2}{2}\|v_k\|_2^2 \\ &\leq \frac{1}{2}\|x_k - x^*\|_2^2 + \alpha_k(f(x^*) - f(x_k)) + \frac{\alpha_k^2}{2}\|v_k\|_2^2.\end{aligned}$$

By summing up, we obtain

$$\frac{\sum_{i=0}^k \alpha_i (f(x_i) - f^*)}{\sum_{i=0}^k \alpha_i} \leq \frac{\|x_0 - x_*\|^2 + \sum_{i=0}^k \alpha_i^2 \|v_i\|_2^2}{2\sum_{i=0}^k \alpha_i}$$

and the result follows from convexity of f. □

Corollary 5.3.1. *If f is L-Lipschitz, we have the following convergence result for subgradient method.*

- *If $\alpha_k = \alpha$ is constant, we have*

$$\min_{i=1,\dots,k} f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2}{2(k+1)\alpha} + \frac{L^2\alpha}{2}.$$

- *In particular, choosing $\alpha_i = \frac{\|x_0 - x^*\|/L}{\sqrt{k+1}}$, we have*

$$\min_{i=1,\dots,k} f(x_k) - f^* \leq \frac{\|x_0 - x^*\| L}{\sqrt{k+1}}.$$

- *Choosing $\alpha_k = \|x_0 - x^*\|/(L\sqrt{k})$ for all k, we obtain for all k*

$$\min_{i=1,\dots,k} f(x_k) - f^* = O\left(\frac{\|x_0 - x^*\|_2 L(1+\log(k))}{\sqrt{k}}\right).$$

Remark 5.3.1. *We have for any $k \in \mathbb{N}$*

$$\frac{1}{2}\|x_{k+1} - x^*\|_2^2 \leq \frac{1}{2}\|x_k - x^*\|_2^2 + \frac{\alpha_k^2}{2}\|v_k\|_2^2.$$

If f is Lipschitz, choosing $\alpha_k = 1/(1+k)^{\frac{1}{2}+\epsilon}$ with $\epsilon > 0$ small, for all k, we have that $(x_k)_{k\in\mathbb{N}}$ converges. Indeed, for any x^* solution, for any $k \in \mathbb{N}$, set $u_k = \|x_k - x^*\|_2$, we have,

$$S_k = \sum_{i=1}^{k} (u_i - u_{i-1})_+$$

converges. Setting $R^k = \sum_{i=1}^{k}(u_i - u_{i-1})_-$, we have $u_k = S_k + R_k + u_0$ and since $u_k \geq 0$ R_k also converges and finally u_k converges.

5.4 Composite optimization

The subgradient method is slow in practice. Furthermore, convergence depends a lot on step size tuning. In favorable situations there exists better suited algorithms. Good introduction to the topic of proximal algorithms with connection to statistics and signal processing are found in [26, 6].

5.4.1 Motivation

The Lasso estimator is given by

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2n}\|\mathbb{X}\theta - Y\|^2 + \lambda\|\theta\|_1.$$

This is the solution of a nonsmooth convex optimization problem. The subgradient method can be used to solve this problem as it can be used to solve any continuous convex optimization problem for which subgradients are available. However, this method is slow and hard to tune in practice. It turns out that the objective function has additional structure which can be leveraged to devise more powerful and easier to implements algorithms. Indeed the objective function is of the form $f + g$ where f is a smooth (quadratic) convex function and g is the ℓ_1 norm, a nonsmooth convex function. Objective functions falling in this classe are sometimes called "composite objectives". Under additional restriction on g (easily computable proximity operator), there exists numerical algorithm which efficiency is comparable to the that of gradient descent for smooth optimization.

5.4.2 Proximity operator

The construction of the following object is due to Jean-Jacques Moreau [40].

Definition 5.4.1. *Given a closed convex function, $f\colon \mathbb{R}^d \mapsto \mathbb{R}$, the proximity operator of f is defined as follows*

$$\mathrm{prox}_f\colon z \mapsto \arg\min_{y\in\mathbb{R}^d} f(y) + \frac{1}{2}\|y - z\|_2^2.$$

By strong convexity, the minimum is attained and is strict.

Note that we have $x = \mathrm{prox}_f(z)$ if and only if $z = \partial f(x) + x$ and the proximity operator is someties denoted $(\partial f + I)^{-1}$.

Exercise 5.4.1. *Describe the* prox *applications for the following functions:*

- *A constant*
- *A linear function*
- *The indicator of a closed convex set C:*

$$\delta\colon x \mapsto \begin{cases} 0 \text{ if } x \in C \\ +\infty \text{ otherwise} \end{cases}$$

- *The function $x \mapsto \frac{1}{2}\|x\|_2^2$*
- *The function $x \mapsto \|x\|_2$*
- *The function $x \mapsto \|x\|_1$*

Exercise 5.4.2. *Let f and g be convex. Show that $\partial(f+g)(x) \supset \partial f(x) + \partial g(x)$ for every x such that $\partial f(x)$ and $\partial g(x)$ are non empty.*

Lemma 5.4.1. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be convex continuously differentiable with L-Lipschitz gradient and g be convex lower semicontinuous. Fix any $x \in \mathbb{R}^p$ and set*

$$y = \operatorname{prox}_{g/L}\left(x - \frac{1}{L}\nabla f(x)\right).$$

Then, for any $z \in \mathbb{R}^d$,

$$f(z) + g(z) + \frac{L}{2}\|x - z\|_2^2 \geq f(y) + g(y) + \frac{L}{2}\|y - z\|_2^2.$$

Proof. First, the descent Lemma ensures that for any $k \in \mathbb{N}$

$$f(y) \leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 \tag{5.11}$$

We have

$$y = \arg\min_{y \in \mathbb{R}^p} f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 + g(y), \tag{5.12}$$

so that by strong convexity, for all $z \in \mathbb{R}^d$,

$$\begin{aligned} & f(x) + \langle \nabla f(x), z - x\rangle + \frac{L}{2}\|z - x\|_2^2 + g(z) \\ \geq\ & f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 + g(y) + \frac{L}{2}\|z - y\|_2^2. \end{aligned} \tag{5.13}$$

Combining (5.11) and (5.13), we obtain

$$\begin{aligned} & f(z) + g(z) + \frac{L}{2}\|z - x\|_2^2 \\ \geq\ & f(x) + \langle \nabla f(x), z - x\rangle + \frac{L}{2}\|z - x\|_2^2 + g(z) && \text{convexity} \\ \geq\ & f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|_2^2 + g(y) + \frac{L}{2}\|y - z\|_2^2 && (5.13) \\ \geq\ & f(y) + g(y) + \frac{L}{2}\|y - z\|_2^2, && (5.11) \end{aligned}$$

□

Proposition 5.4.1. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be convex continuously differentiable with L-Lipschitz gradient and g be convex lower semicontinuous such that $\rho = \inf_{x\in\mathbb{R}^p} f(x) + g(x) > -\infty$ is attained at x^*. Consider the algorithm, for $x_0 \in \mathbb{R}^p$ and*

$$x_{k+1} = \operatorname{prox}_{g/L}\left(x_k - \frac{1}{L}\nabla f(x_k)\right). \tag{5.14}$$

Then x_k converges to a global minimum and we have for any $k \in \mathbb{N}$, $k > 0$,

$$f(x_k) + g(x_k) - \rho \leq \frac{L\|x_0 - x^*\|_2^2}{2k}.$$

If in addition $f + g$ is μ-strongly convex, we have in addition

$$\|x_{k+1} - x^*\|_2^2 \leq \frac{L}{L+\mu}\|x_k - x^*\|_2^2.$$

Proof. From Lemma 5.4.1 with $x = x_k = z$ and $y = x_{k+1}$ we read that $f + g$ is decreasing along the sequence. From Lemma 5.4.1 with $x = x_k$ z^* and $y = x_{k+1}$ we read that for any $k \in \mathbb{N}$

$$f(x^*) + g(x^*) + \frac{L}{2}\|x_k - x^*\|_2^2 \geq f(x_{k+1}) + g(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x^*\|_2^2.$$

By summing up, we obtain for any $K \in \mathbb{N}$, $K \geq 1$,

$$\frac{L}{2}\|x^* - x_0\|_2^2 \geq \sum_{k=1}^{K} f(x_k) + g(x_k) - \rho \geq K(f(x_K) + g(x_K) - \rho).$$

We also have that $\|x_k - x^*\|_2^2$ is decreasing and the convergence follows (this is Opial's Lemma). For the last statement, Lemma 5.4.1 with $y = x_{k+1}$, $z = x^*$ and $x = x_k$ combined with μ-strong convexity gives

$$\begin{aligned} f(x^*) + g(x^*) + \frac{L}{2}\|x_k - x^*\|_2^2 &\geq f(x_{k+1}) + g(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x^*\|_2^2 \\ &\geq f(x^*) + g(x^*) + \frac{L+\mu}{2}\|x_{k+1} - x^*\|_2^2, \end{aligned}$$

which is the desired result. □

5.5 Acceleration

We have obtained $1/k$ convergence rates for the gradient algorithm and the proximal gradient algorithm. Could we do better?

5.5.1 A lower bound

This is taken from Bubeck's book [21] and originally due to Nesterov [45]. Such results first appeared in the litterature in Newmirovski and Yudin [43].

Definition 5.5.1. *A first order method to minimize a smooth convex function f when initiated at $x_0 = 0$, produces a sequence of points $(x_i)_{i\in\mathbb{N}}$ such that for any $k \in \mathbb{N}$,*

$$x_{k+1} \in \operatorname{span}\left(\nabla f(x_0), \ldots, \nabla f(x_k)\right).$$

Theorem 5.5.1. *Let $k \leq (d-1)/2$, $L > 0$. There exists a convex function f with L-Lipschitz gradient over $\mathbb{R}^d$, such that for any first order method satisfying definition* (5.5.1)*,*

$$\min_{1 \leq s \leq k} f(x_s) - f(x^*) \geq \frac{3L}{32} \frac{\|x_0 - x^*\|^2}{(k+1)^2}.$$

Proof. In this proof for $h : \mathbb{R}^d \to \mathbb{R}$ we denote $h^* = \inf_{x \in \mathbb{R}^d} h(x)$. For $k \leq d$ let $A_k \in \mathbb{R}^{d \times d}$ be the symmetric and tridiagonal matrix defined by

$$(A_k)_{i,j} = \begin{cases} 2, & i = j, i \leq k \\ -1, & j \in \{i-1, i+1\}, i \leq k, j \neq k+1 \\ 0, & \text{otherwise.} \end{cases}$$

We verify that $0 \preceq A_k \preceq 4I$ since

$$x^\top A_k x = 2\sum_{i=1}^{k} x(i)^2 - 2\sum_{i=1}^{k-1} x(i)x(i+1) = x(1)^2 + x(k)^2 + \sum_{i=1}^{k-1} (x(i) - x(i+1))^2 \leq 4\sum_{i=1}^{k} x(i)^2.$$

We consider now the following convex function:

$$f(x) = \frac{L}{8} x^\top A_{2k+1} x - \frac{L}{4} x^\top e_1.$$

For any $s = 1, \ldots, k$, x_s must lie in the linear span of $e_1, \ldots, e_{s-1}$ (because of our assumption on the black-box procedure). In particular for $s \leq k$ we necessarily have $x_s(i) = 0$ for $i = s, \ldots, n$, which implies $x_s^\top A_{2k+1} x_s = x_s^\top A_k x_s$. In other words, if we denote

$$f_k(x) = \frac{L}{8} x^\top A_k x - \frac{L}{4} x^\top e_1,$$

We proved that, for all $s \leq k$

$$f(x_s) - f^* = f_k(x_s) - f_{2k+1}^* \geq f_k^* - f_{2k+1}^*.$$

Thus it simply remains to compute the minimizer x_k^* of f_k, its norm, and the corresponding function value f_k^*.

The point x_k^* is the unique solution in the span of $e_1, \ldots, e_k$ of $A_k x = e_1$. One can verify (Exercise) that it is defined by $x_k^*(i) = 1 - \frac{i}{k+1}$ for $i = 1, \ldots, k$. Thus we have:

$$f_k^* = \frac{L}{8}(x_k^*)^\top A_k x_k^* - \frac{L}{4}(x_k^*)^\top e_1 = -\frac{L}{8}(x_k^*)^\top e_1 = -\frac{L}{8}\left(1 - \frac{1}{k+1}\right).$$

Furthermore note that

$$\|x_k^*\|^2 = \sum_{i=1}^{k} \left(1 - \frac{i}{k+1}\right)^2 = \sum_{i=1}^{k} \left(\frac{i}{k+1}\right)^2 \leq \frac{k+1}{3}.$$

Thus one obtains:

$$f_k^* - f_{2k+1}^* = \frac{L}{4}\left(\frac{1}{k+1} - \frac{1}{2k+2}\right) \geq \frac{3L}{32} \frac{\|x_{2k+1}^*\|^2}{(k+1)^2},$$

□

5.5.2 Accelerated algorithm

The previous lower bound shows that there is a gap between the convergence speed of gradient descent for smooth convex functions and the and the lower bound. It remained an open question if the gap was due to gradient descent or if it was due to the fact that the lower bound is loose until Nesterov published in 1983 an algorithm which achieves $1/k^2$ rate [47]. We extend bellow the original proof of Nesterov. An extension to the proximal setting has been developped by Beck and Teboulle in [9].

Theorem 5.5.2. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be convex continuously differentiable with L-Lipschitz gradient $\inf_{x\in\mathbb{R}^p} f(x) > -\infty$. Consider the algorithm, for $x_{-1} \in \mathbb{R}^p$, set $y_0 = x_{-1}$, $t_1 = 1$ and for $k \in \mathbb{N}$,*

$$\begin{aligned} x_k &= y_k - \frac{1}{L}\nabla f(y_k) \\ t_{k+1} &= \frac{1+\sqrt{1+4t_k^2}}{2} \\ y_{k+1} &= x_k + \left(\frac{t_k - 1}{t_{k+1}}\right)(x_k - x_{k-1}). \end{aligned} \tag{5.15}$$

Then for any $k \in \mathbb{N}$

$$f(x_k) - f^* \leq \frac{4L\|x_0 - x^*\|_2^2}{(k+2)^2}.$$

Proof. We introduce the following notation which is taken from the original proof, for any $k \in \mathbb{N}$,

$$p_k := (t_k - 1)(x_{k-1} - x_k) \qquad \text{so that} \qquad y_{k+1} = x_k - \frac{p_k}{t_{k+1}}$$

First, we have for any $k \geq 1$

$$t_k \geq \frac{1+\sqrt{4t_{k-1}^2+1}}{2} \geq t_{k-1} + \frac{1}{2} \geq t_0 + \frac{k}{2} = 1 + \frac{k}{2}. \tag{5.16}$$

$$(t_{k+1}^2 - t_{k+1}) = t_k^2. \tag{5.17}$$

The main argument of the proof is the following. The sequence $\{z_k\}_{k\in\mathbb{N}}$ defined as

$$z_k := \frac{2t_k^2}{L}(f(x_k) - f^*) + \|p_k - x_k + x^*\|^2, \tag{5.18}$$

is non-increasing and $z_0 \leq 2||x_0 - x^*||^2$. The result can be deduced by combining (5.16) and (5.18).

We have a series of three inequalities.

$$\begin{aligned} p_{k+1} - x_{k+1} &= p_k - x_k + \frac{t_{k+1}}{L}\nabla f(y_{k+1}) \\ p_{k+1} - x_{k+1} &= (t_{k+1} - 1)(x_k - x_{k+1}) - x_{k+1} \\ &= (t_{k+1} - 1)x_k - t_{k+1}x_{k+1} \\ &= (t_{k+1} - 1)x_k - t_{k+1}\left(y_{k+1} - \frac{1}{L}\nabla f(y_{k+1})\right) \\ &= (t_{k+1} - 1)x_k - t_{k+1}x_k - (t_k - 1)(x_k - x_{k-1}) - \frac{t_{k+1}}{L}\nabla f(y_{k+1}) \\ &= p_k - x_k + \frac{t_{k+1}}{L}\nabla f(y_{k+1}) \end{aligned}$$

This implies

$$\begin{aligned}
\|p_{k+1} - x_{k+1} + x^*\|_2^2 &= \|p_k - x_k + \frac{t_{k+1}}{L}\nabla f(y_{k+1}) + x^*\|_2^2 \\
&= \|p_k - x_k + x^*\|_2^2 + 2\left\langle p_k - x_k + x^*, \frac{t_{k+1}}{L}\nabla f(y_{k+1})\right\rangle \\
&\quad + \frac{t_{k+1}^2}{L^2}\|\nabla f(y_{k+1})\|_2^2 \\
y_{k+1} &= x_k - \frac{p_k}{t_{k+1}} \\
\left\langle p_k - x_k + x^*, \frac{t_{k+1}}{L}\nabla f(y_{k+1})\right\rangle &= \left\langle p_k - y_{k+1} - \frac{p_k}{t_{k+1}} + x^*, \frac{t_{k+1}}{L}\nabla f(y_{k+1})\right\rangle \\
&= \frac{(t_{k+1}-1)}{L}\left\langle p_k, \nabla f(y_{k+1})\right\rangle + \frac{t_{k+1}}{L}\left\langle x^* - y_{k+1}, \nabla f(y_{k+1})\right\rangle \\
\|p_{k+1} - x_{k+1} + x^*\|_2^2 &= \|p_k - x_k + x^*\|_2^2 + 2\frac{(t_{k+1}-1)}{L}\left\langle p_k, \nabla f(y_{k+1})\right\rangle \\
&\quad + 2\frac{t_{k+1}}{L}\left\langle x^* - y_{k+1}, \nabla f(y_{k+1})\right\rangle + \frac{t_{k+1}^2}{L^2}\|\nabla f(y_{k+1})\|_2^2
\end{aligned}$$

From the Lipschitz gradient assumption, we obtain

$$\begin{aligned}
f(x_{k+1}) - f^* &\leq f(y_{k+1}) - f^* - \frac{1}{2L}\|\nabla f(y_{k+1})\|_2^2 \leq \langle \nabla f(y_{k+1}), y_{k+1} - x^*\rangle - \frac{1}{2L}\|\nabla f(y_{k+1})\|_2^2 \\
\frac{1}{2L}\|\nabla f(y_{k+1})\|_2^2 &\leq f(y_{k+1}) - f(x_{k+1}) \leq f(x_k) - f(x_{k+1}) - \frac{1}{t_{k+1}}\left\langle p_k, \nabla f(y_{k+1})\right\rangle
\end{aligned}$$

Using the last three identities, we obtain

$$\begin{aligned}
&\|p_{k+1} - x_{k+1} + x^*\|_2^2 - \|p_k - x_k + x^*\|_2^2 \\
&= 2\frac{(t_{k+1}-1)}{L}\left\langle p_k, \nabla f(y_{k+1})\right\rangle + 2\frac{t_{k+1}}{L}\left\langle x^* - y_{k+1}, \nabla f(y_{k+1})\right\rangle + \frac{t_{k+1}^2}{L^2}\|\nabla f(y_{k+1})\|_2^2 \\
&\leq 2t_{k+1}\frac{(t_{k+1}-1)}{L}\left(f(x_k) - f(x_{k+1}) - \frac{1}{2L}\|\nabla f(y_{k+1})\|_2^2\right) \\
&\quad + 2\frac{t_{k+1}}{L}\left(f^* - f(x_{k+1}) - \frac{1}{2L}\|\nabla f(y_{k+1})\|_2^2\right) + \frac{t_{k+1}^2}{L^2}\|\nabla f(y_{k+1})\|_2^2 \\
&= 2t_{k+1}\frac{(t_{k+1}-1)}{L}\left(f(x_k) - f^* + f^* - f(x_{k+1})\right) + 2\frac{t_{k+1}}{L}\left(f^* - f(x_{k+1})\right) \\
&= 2\frac{t_k^2}{L}\left(f(x_k) - f^*\right) - 2\frac{t_{k+1}^2}{L}\left(f(x_{k+1} - f^*)\right)
\end{aligned}$$

where we used (5.16) for the last step. This proves that the sequence $(z_k)_{k\in\mathbb{N}}$ is non increasing. It remains to compute z_0,

$$z_0 = \frac{2}{L}(f(x_0) - f^*) + \|x^* - x_0\|^2 \leq 2\|x_0 - x^*\|_2^2.$$

Putting things together

$$f(x_k) - f^* \leq \frac{Lz_0}{2t_k^2} \leq \frac{4L\|x_0 - x^*\|_2^2}{(k+2)^2}.$$

□

5.6 Non convex problems

Most algorithm described in this chapter have extensions to nonconvex problems. In this setting, the only hope is to find first order critical points instead of global minima. The notion of subgradient in this case has to be treated with a lot of care. A reference on the topic is [54].

Exercises

Exercise 5.6.1. *Show that if $f\colon \mathbb{R}^d \mapsto \mathbb{R}$ is $\mathcal{C}^2$ then any accumulation point of the system $\dot{x} = -\nabla f(x)$ is a critical point of f.*

Exercise 5.6.2. *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be a convex function,*

- *Show that ∂f is sequencialy closed in the sence that, for any $\bar{x}$*

$$\left\{v \in \mathbb{R}^p,\, \exists\, (x_k, v_k)_{k\in\mathbb{N}}\,,\, x_k \to \bar{x},\, v_k \to v,\, f(x_k) \to f(\bar{x})\right\} \subset \partial f(\bar{x})$$

- *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$, show that f is L-Lipschitz if and only if $\sup_{x\in\mathbb{R}^p,\, v\in\partial f(x)} \|v\|_2 \leq L$.*

Exercise 5.6.3.

- *Let $f\colon \mathbb{R} \mapsto \mathbb{R}$ be convex (with full domain), show that for any $s < t < u$,*

$$\frac{f(t) - f(s)}{t - s} \leq \frac{f(u) - f(s)}{u - s} \leq \frac{f(u) - f(t)}{u - t}.$$

- *Deduce that f is continuous on $\mathbb{R}$.*

Exercise 5.6.4.

- *Let $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ be convex (with full domain), show that for any $x \in \mathbb{R}^d$ and $h \in \mathbb{R}^d*$, with $\|h\|_1 < 1$, we have*

$$f(x + h) \leq (1 - \|h\|_1) f(x) + \|h\|_1 \max_{i=1,\dots,d} f(x \pm e_i)$$

  where e_i are elements of the canonical basis.

- *Deduce that f is continuous at x.*

- *What can you say about an extended valued convex function which domain has nonempty interior?*

Exercise 5.6.5. *Let $\|\cdot\|$ be a norm, it is then convex. Its dual norm is defined by*

$$\|z\|_* = \sup z^T x \qquad \textit{such that} \qquad \|x\| \leq 1.$$

Consider the function $f\colon x \mapsto \|x\|$, compute the Legendre transform of f.

Exercise 5.6.6. *Let $f_i\colon \mathbb{R}^d \mapsto \mathbb{R}$ be convex and differentiable on $\mathbb{R}^d$ for $i = 1 \dots n$. Set $F\colon x \mapsto \max_i f_i(x)$. Show that*

$$\partial F(x) = \operatorname{conv}\left(\{\nabla f_i(x),\, f_i(x) = F(x)\}\right).$$

How does this result extend to non differentiable convex functions?

Exercise 5.6.7. *Let $f\colon \mathbb{R}^d \mapsto \mathbb{R}$ be convex with full domain. Show that f is upper bounded if and only if f is constant.*

Exercise 5.6.8. *Describe the* prox *applications for the following functions: a constant, a linear function, the indicator of a closed convex set C:*

$$\delta\colon x \mapsto \begin{cases} 0 \text{ if } x \in C \\ +\infty \text{ otherwise} \end{cases}$$

The function $x \mapsto \frac{1}{2}\|x\|_2^2$, the function $x \mapsto \|x\|_2$, the function $x \mapsto \|x\|_1$

Exercise 5.6.9. *Let f and g be convex. Show that $\partial(f + g)(x) \supset \partial f(x) + \partial g(x)$ for every x such that $\partial f(x)$ and $\partial g(x)$ are non empty. What do you think about the reverse inclusion?*

Chapter 6

Stochastic approximation

This chapter is dedicated to stochastic approximation for large sums. Stochastic approximation has a long history starting with Robbins-Monro algorithm [53] with the ODE method [38] from Ljung and latter extensions, see [11] for a complete exposition and [16]. The idea of using stochastic approximation in large scale setting gained significance interest in the machine learning litterature see for example [17]. We provide example of non asymptotic convergence rate analyses for stochastic subgradient and stochastic proximal gradient for finite sums.

6.1 Motivation, large n

6.1.1 Lasso estimator

The Lasso estimator is given as follows:

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2n}\|\mathbb{X}\theta - Y\|^2 + \lambda\|\theta\|_1.$$

We have seen that the optimization problem has a favorable structure which allow to devise efficient algorithms. Another way to write the same optimization problem is to consider

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}(x_i^T\theta - y_i)^2 + \lambda\|\theta\|_1,$$

which actually exhibits an additional sum structure. In this chapter we will be considering optimization problems of the form

$$\min_{x\in\mathbb{R}^d} \quad F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x) + g(x). \tag{6.1}$$

where f_i and g are convex lower semicontinuous convex functions.

6.1.2 Stochastic approximation

To solve problem (6.1), one may use first order methods such as the ones described in the previous chapter. Computing a subgradient in this case require to compute subgradient of f_i, $i = 1, \dots, n$ and average them. The computational cost is of the order of n subgradient computation and n vector operations. When n is very large, or even infinite, this could be prohibitive. Intuitively, if there is redundancy in the elements of the sum, one should be able to take advantage of it. For example, suppose that $f_i = f$, for all i, then blindly computing the gradient of F has a cost of the order $n \times d$, while only d operations (computing one gradient would suffice).

More generally, one could rewrite the objective function in (6.1) in the following form:

$$F\colon x \mapsto \mathbb{E}\left[f_I(x)\right] + g(x),$$

where I denotes a uniform random variable over $\{1, \dots, n\}$. Stochastic approximation, or stochastic optimization algorithm allow to handle such objectives. The main algorithmic step is as follows:

- For any $x \in \mathbb{R}^d$,
- Sample i uniformly at random in $\{1, \dots, n\}$.
- Perform an algorithmic step using only the value of $f_i(x)$ and $\nabla f_i(x)$ or eventually $v \in \partial f_i(x)$

The simple example can be extended to more general random variables I and under proper integrability and domination conditions, one can invert gradient (or subgradient) and expectation, assuming $g = 0$ for simplicity

- If for each value of I, f_I is continuously differentiable, we then have for any $x \in \mathbb{R}^d$,

$$\mathbb{E}\left[\nabla f_I(x)\right] = \nabla \mathbb{E}\left[f_I(x)\right] = \nabla F(x)$$

- Assume that f_I is convex for all realizations of I. Assume that we have access to a random variable v_I such that $v_I \in \partial f_I(x)$ almost surely, then the expectation is convex and

$$\mathbb{E}\left[v_I\right] \in \partial \mathbb{E}\left[f_I(x)\right] = \partial F(x).$$

Hence the process of using a single element of the sum in an algorithm can be seen as performing optimization based on noisy unbiased estimates of the gradient, or subgradient, of the objective. This intuition is described more formaly in the coming section.

6.2 Prototype stochastic approximation algorithm

This section describes Robbins-Monro algorithm for stochastic approximation. Consider a Lipschitz map $h\colon \mathbb{R}^p \mapsto \mathbb{R}^p$, the goal is to find a zero of h. The operator only has access to unbiased noisy estimates of h. The Robins-Monro algorithm is described as follows, $(X_k)_{k\in\mathbb{N}}$ is a sequence of random variables such that for any $k \in \mathbb{N}$

$$X_{k+1} = X_k + \alpha_k \left(h(X_k) + M_{k+1}\right) \tag{6.2}$$

where

- $(\alpha_k)_{k\in\mathbb{N}}$ is a sequence of positive step sizes satisfying

$$\sum_{i=1}^{n} \alpha_k = +\infty$$
$$\sum_{i=1}^{n} \alpha_k^2 < +\infty$$

- $(M_k)_{k\in\mathbb{N}}$ is a martingale difference sequence with respect to the increasing family of σ-fields

$$\mathcal{F}_k = \sigma(X_m, M_m, m \leq k) = \sigma(X_0, M_1, \dots, M_k).$$

  This means that $\mathbb{E}\left[M_{k+1}|\mathcal{F}_k\right] = 0$, for all $k \in \mathbb{N}$.

- In addition, we assume that there exists a positive constant C such that

$$\sup_{k\in\mathbb{N}} \mathbb{E}\left[\|M_{k+1}\|_2^2 | \mathcal{F}_k\right] \leq C.$$

The intuition here is that our hypotheses on the step size ensure that the quantity $\sum_{k=0}^{+\infty} \mathbb{E}\left[\alpha_k^2 \|M_{k+1}\|^2 | \mathcal{F}_k\right]$ is finite and hence the zero mean martingale $\sum_{k=0}^{K} \alpha_k M_{k+1}$ has square summable increments and converges to a square integrable random varible M in $\mathbb{R}^p$ both almost surely and in L^2 (see for example [29, Section 5.4]).

The long term behaviour of such recursions is at the heart of the field of stochastic approximation. The fact that the step sizes tends to 0 and that the sum of perturbation stabilizes suggests that in the limit one obtains trajectories of a continuous time differential equation. This is formalized in the next section.

6.3 The ODE approach

For optimization we may choose $h = -\nabla F(x)$ assuming that F has Lispchitz gradient. We consider Robbins-Monro algorithm in this setting. This idea dates back to Ljung [38], see also [11] for an advanced presentation. An accessible exposition of the following result is found in [16],

Theorem 6.3.1. *Conditioning on boundedness of $\{X_k\}_{k\in\mathbb{N}}$, almost surely, the (random) set of accumulation point of the sequence is compact connected and invariant by the flow generated by the continuous time limit:*

$$\dot{x} = h(x).$$

This theorems means that for any $\bar{x}$ accumulation point of the algorithm, the unique solution $x\colon t \mapsto \mathbb{R}^p$ of the continuous time ODE satisfying $x(0) = \bar{x}$ remains bounded for all $t \in \mathbb{R}$. This allows to conclude in the convex case.

Corollary 6.3.1. *If F is convex, differentiable and attains its minimum, setting $h = -\nabla F$, conditioning on the event that $\sup_{k\in\mathbb{N}} \|X_k\|$ is finite, almost surely, all the accumulation points of X_k are critical points of F.*

Proof. Fix $\bar{x} \in \mathbb{R}^p$ such that $\nabla F(\bar{x}) \neq 0$, this means that $F(\bar{x}) - F^* > 0$. Consider the solution to

$$\dot{x} = \nabla F(x),$$

starting at $\bar{x}$, we have

$$\begin{aligned} \frac{\partial}{\partial t} F(x(t)) &= \|\nabla F(x(t))\|_2^2 \geq 0 \\ \frac{\partial}{\partial t} \|x(t) - x^*\|_2^2 &= \langle \nabla F(x(t)), x(t) - x^* \rangle \geq F(x(t)) - F^* \geq F(\bar{x}) - F^* > 0. \end{aligned}$$

We deduce that F is increasing along the trajectory and diverges, hence the solution escapes any compact set which means that $\bar{x}$ does not belong to a compact invariant set. □

The power of the ODE approach lies in the fact that it allows to treat much more complicated situations beyond convexity and differentiability.

6.4 Rates for convex optimization

In the context of convex optimization problems of the form described in the introduction of this chapter, one can obtain precise covergence rate estimates using elementary arguments.

6.4.1 Stochastic subgradient descent

Proposition 6.4.1. *Consider the problem*

$$\min_{x\in\mathbb{R}^d} F(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x),$$

where each f_i is convex and L-Lipschitz. Choose $x_0 \in \mathbb{R}$ and a sequence of random variables $(i_k)_{k\in\mathbb{N}}$ independently identically distributed uniformly on $\{1,\dots,n\}$ and a sequence of positive step sizes $(\alpha_k)_{k\in\mathbb{N}}$. Consider the recursion

$$x_{k+1} = x_k - \alpha_k v_k \tag{6.3}$$
$$v_k \in \partial f_{i_k}(x_k) \tag{6.4}$$

Then for all $K \in \mathbb{N}$, $K \geq 1$

$$\mathbb{E}\left[F(\bar{x}_K) - F^*\right] \leq \frac{L\|x_0 - x^*\|_2^2 + L^2\sum_{k=0}^K \alpha_k^2}{2\sum_{k=0}^K \alpha_k}$$

where $\bar{x}_K = \frac{\sum_{k=0}^K \alpha_k x_k}{\sum_{k=0}^K \alpha_k}$.

Proof. We fix $k \in \mathbb{N}$ and condition on $i_1,\dots,i_k$ so that x_k and x_{k+1} are fixed. We have for any $k \in \mathbb{N}$

$$\begin{aligned}\frac{1}{2}\|x_{k+1} - x^*\|_2^2 &= \frac{1}{2}\|x_k - \alpha_k v_k - x^*\|_2^2 \\ &= \frac{1}{2}\|x_k - x^*\|_2^2 + \alpha_k v_k^T(x^* - x_k) + \frac{\alpha_k^2}{2}\|v_k\|_2^2 \\ &\leq \frac{1}{2}\|x_k - x^*\|_2^2 + \alpha_k(f_{i_k}(x^*) - f_{i_k}(x_k)) + \frac{\alpha_k^2}{2}L^2.\end{aligned}$$

Conditioning on x_k and taking expectation with respect to i_k,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{2}\|x_{k+1} - x^*\|_2^2 | x_k\right] &\leq \mathbb{E}\left[\frac{1}{2}\|x_k - x^*\|_2^2 | x_k\right] + \frac{\alpha_k^2 L^2}{2} + \mathbb{E}\left[\alpha_k(f_{i_k}(x^*) - f_{i_k}(x_k)) | x_k\right] \\ &= \frac{1}{2}\|x_k - x^*\|_2^2 + \frac{\alpha_k^2 L^2}{2} + \alpha_k(F(x^*) - F(x_k)).\end{aligned}$$

Taking expectation with respect to x_k, using tower property of conditional expectation, we have

$$\mathbb{E}\left[\frac{1}{2}\|x_{k+1} - x^*\|_2^2\right] \leq \mathbb{E}\left[\frac{1}{2}\|x_k - x^*\|_2^2\right] + \frac{\alpha_k^2 L^2}{2} + \alpha_k\mathbb{E}\left[(F(x^*) - F(x_k))\right].$$

By summing up, we obtain, for all $K \in \mathbb{N}$, $K \geq 1$

$$\frac{\sum_{k=0}^K \alpha_k \mathbb{E}\left[F(x_k) - F^*\right]}{\sum_{i=0}^k \alpha_i} \leq \frac{\|x_0 - x_*\|^2 + L^2\sum_{k=0}^K \alpha_k^2}{2\sum_{k=0}^K \alpha_i}$$

and the result follows from convexity of f. □

Corollary 6.4.1. *Under the hypotheses of Proposition 6.4.1, we have the following*

- *If $\alpha_k = \alpha$ is constant, we have*

$$\mathbb{E}\left[F(\bar{x}_k) - F^*\right] \leq \frac{\|x_0 - x^*\|^2}{2(k+1)\alpha} + \frac{L^2\alpha}{2}.$$

- *In particular, choosing* $\alpha_i = \frac{\|x_0 - x^*\|/L}{\sqrt{k+1}}$*, we have*

$$\mathbb{E}\left[F(\bar{x}_k) - F^*\right] \leq \frac{\|x_0 - x^*\|L}{\sqrt{k+1}}.$$

- *Choosing* $\alpha_k = \|x_0 - x^*\|/(L\sqrt{k})$ *for all* k*, we obtain for all* k

$$\mathbb{E}\left[F(\bar{x}_k) - F^*\right] = O\left(\frac{\|x_0 - x^*\|_2 L(1+\log(k))}{\sqrt{k}}\right).$$

6.4.2 Stochastic proximal gradient descent

This method is sometimes called FOBOS in the litterature. I could not find a reference for the following result.

Proposition 6.4.2. *Consider the problem*

$$\min_{x\in\mathbb{R}^d} F(x) \;:=\; \frac{1}{n}\sum_{i=1}^{n} f_i(x) + g(x)$$

where each f_i *is convex with* L*-Lipschitz gradient and* g *is convex. Choose* $x_0 \in \mathbb{R}$ *and a sequence of random variables* $(i_k)_{k\in\mathbb{N}}$ *independently identically distributed uniformly on* $\{1,\dots,n\}$ *and a sequence of positive step sizes* $(\alpha_k)_{k\in\mathbb{N}}$*. Consider the recursion*

$$x_{k+1} = \operatorname{prox}_{\alpha_k g/L}\left(x_k - \alpha_k/L\nabla f_{i_k}(x_k)\right). \tag{6.5}$$

Assume the following

- $0 < \alpha_k \leq 1$*, for all* $k \in \mathbb{N}$*.*
- f_i *and* g *are* G*-Lipschitz for all* $i = 1,\dots,n$*;*

Then for all $K \in \mathbb{N}$*,* $K \geq 1$

$$\mathbb{E}\left[F(\bar{x}_K) - F^*\right] \leq \frac{L\|x_0 - x^*\|_2^2 + \frac{2G^2}{L}\sum_{k=0}^{K}\alpha_k^2}{2\sum_{k=0}^{K}\alpha_k}$$

where $\bar{x}_K = \frac{\sum_{k=0}^{K}\alpha_k x_k}{\sum_{k=0}^{K}\alpha_k}$*.*

Proof. We fix $k \in \mathbb{N}$ and condition on $i_1,\dots,i_k$ so that x_k and x_{k+1} are deterministic. Note that the prox iteration gives

$$\frac{\alpha_k}{L}\partial g(x_{k+1}) + x_{k+1} = x_k - \frac{\alpha_k}{L}\nabla f_{i_k}(x_k)$$
$$\|x_{k+1} - x_k\|_2 \leq 2G\frac{\alpha_k}{L}$$

Fix $k \in \mathbb{N}$, applying Lemma 5.4.1 with $x = x_k$, $z = x^*$ and $y = x_{k+1}$, using the fact that f_{i_k} has L/α_k Lipschitz gradient,

$$\begin{aligned}
&f_{i_k}(x^*) + g(x^*) + \frac{L}{2\alpha_k}\|x^* - x_k\|_2^2 - \frac{L}{2\alpha_k}\|x_{k+1} - x^*\|_2^2 \\
&\geq f_{i_k}(x_{k+1}) + g(x_{k+1}) \\
&\geq f_{i_k}(x_k) + g(x_k) - 2G\|x_{k+1} - x_k\|_2 \\
&\geq f_{i_k}(x_k) + g(x_k) - 4G^2\frac{\alpha_k}{L}
\end{aligned}$$

And

$$\frac{\alpha_k}{L}(f_{i_k}(x_k) + g(x_k) - F^*) \leq \frac{1}{2}\|x^* - x_k\|_2^2 - \frac{1}{2}\|x_{k+1} - x^*\|_2^2 + 4G^2\frac{\alpha_k^2}{L^2}$$

We have, considering tower expectation, with respect to i_k first and the remaining randomness in a second step

$$\begin{aligned}
\mathbb{E}\left[\frac{\alpha_k}{L}(f_{i_k}(x_k) + g(x_k) - F^*)\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\alpha_k}{L}(f_{i_k}(x_k) + g(x_k) - F^*)|x_k\right]\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\alpha_k}{L}(F(x_k) - F^*)|x_k\right]\right] \\
&= \mathbb{E}\left[\frac{\alpha_k}{L}(F(x_k) - F^*)\right] \\
&\leq \mathbb{E}\left[\frac{1}{2}\|x^* - x_k\|_2^2\right] - \mathbb{E}\left[\frac{1}{2}\|x_{k+1} - x^*\|_2^2\right] + 4G^2\frac{\alpha_k^2}{L^2}
\end{aligned}$$

By summing, we obtain, for any $K \in \mathbb{N}$

$$\frac{\mathbb{E}\left[\sum_{k=0}^K \alpha_k(F(x_k) - F^*)\right]}{\sum_{k=0}^K \alpha_k} \leq \frac{L\|x_0 - x^*\|_2^2 + \frac{2G^2}{L}\sum_{k=0}^K \alpha_k^2}{2\sum_{k=0}^K \alpha_k}$$

and the result follows by Jensen's inequality. □

Corollary 6.4.2. *Under the hypotheses of Proposition 7.4.1.*

- *If $\alpha_k = \alpha$ is constant, we have for all $k \geq 1$*

$$F(\bar{x}_k) - F^* \leq \frac{L\|x_0 - x^*\|^2}{2(k+1)\alpha} + \frac{G^2\alpha}{L}.$$

- *In particular, choosing $\alpha_i = \frac{1}{\sqrt{2k+2}}$, for $i = 1 \ldots, k$, for some $k \in \mathbb{N}$, we have*

$$F(\bar{x}_k) - F^* \leq \frac{L\|x_0 - x^*\|_2^2 + \frac{G^2}{L}}{\sqrt{2k+2}}.$$

- *Choosing $\alpha_k = 1/\sqrt{2k+2}$ for all k, we obtain for all k*

$$F(x_k) - F^* = O\left(\frac{L\|x_0 - x^*\|_2^2 + \frac{G^2}{L}\log(k)}{\sqrt{2k+2}}\right).$$

6.5 Minimizing the population risk

The methods which we have seen can be used to minimize functions of the form

$$x \mapsto \mathbb{E}_Z\left[f(x, Z)\right]$$

where x denotes some model parameters and Z denotes a random variable describing our population. In this case, Z could the input output pair (X, Y) of a regression problem, for which we try to minimize the exepected prediction error over a certain parametric regression function class $\mathcal{F}$.

$$R(f) = \mathbb{E}\left[(f(X) - Y)^2\right] = \int_{\mathcal{X}\times\mathcal{Y}} (f(x) - y))^2 P(dx, dy).$$

This can be done by replacing the finite sum by an expectation and sampling of independant indices by iis samples of the random variable Z. The results are exactly the same.

Such a procedure are usually called "single pass" procedure: given a dataset $(x_i, y_i)_{i=1}^n$ for a regression problem, performing one pass of a stochastic algorithm, looking at each data point only once amount to perform n step of the same stochastic algorithm on the population risk.

This illustrates a strong relation between stochastic optimization and statistics. We have seen that in the linear regression setting, there is no hope to obtain estimators with statistical rates much faster than $1/n$ in terms of mean squared error. Similarly, the rates which we obtained for stochastic algorithms are of the order of $1/\sqrt{k}$. This is also optimal in a precise sense. These algorithms provide estimator with statistical efficiency of the order of $1/\sqrt{n}$.

The gap stands because we considered regression problems with squared loss, a very special structure, while here the convex functions which we considered are arbitrary. For strongly convex functions, stochastic optimization algorithms may show faster convergence rate of the order $1/k$.

Chapter 7

Block coordinate methods

Block decomposition methods appeared as alternatives to solve optimization problems involving large number of dimensions. The idea is to reduce the complexity of a single iteration by updating only a few coordinates at a time. The use of such methods was advocated by Nesterov [44], extensions such as [50] appeared in the continuity of these works. The survey [64] is a good entry point to the litterature.

7.1 Motivation, large d

The Lasso estimator is given as follows:

$$\hat{\theta}^{\ell_1} \in \arg\min_{\theta\in\mathbb{R}^d} \frac{1}{2n}\|\mathbb{X}\theta - Y\|^2 + \lambda\|\theta\|_1.$$

We have seen that the optimization problem has a favorable structure which allow to devise efficient algorithms. The cost of each iteration is depends on the dimension (here d^2) which for some problems may be limiting. A possible alternative is to update coordinates independantly, reducing the cost of each iteration.

In general, this approach is not convergent for nonsmooth functions (can you see why?), however, the Lasso problem, despite being nonsmooth, fits coordinate descent methods because the nonsmooth part is separable. We shall see two variations of such algorithms, deterministic and random, with convergence rate estimates in both cases. A good introduction to the topic cand be found in [64] and a pioneering work in optimization is described in [44]. The litterature on the subject has completely exploded in the past years.

7.2 Description of the algorithm

We consider optimization problems of the form

$$\min_{x\in\mathbb{R}^p} F(x) = f(x) + \sum_{i=1}^{p} g_i(x_i),$$

where $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ has L-Lipschitz gradient and $g_i\colon \mathbb{R} \mapsto \mathbb{R}$ are convex lower semicontinuous univariate functions. We denote by $e_1, \dots, e_p$ the elements of the canonical basis. Block coordinate descent algorithms are given a sequence of coordinate indices $(i_k)_{k\in\mathbb{N}}$, and, starting at $x_0 \in \mathbb{R}^p$ updates coordinates one by one at each iteration. For example

$$\begin{aligned} x_{k+1} &= \arg\min_{y=x_k+te_{i_k}} f(x_k) + \langle \nabla f(x_k), y - x_k\rangle + \frac{L}{2}\|y - x_k\|_2^2 + g_{i_k}(y) \\ x_{k+1} &= \arg\min_{y=x_k+te_{i_k}} f(y) + g_{i_k}(y). \end{aligned}$$

The first option corresponds to a block proximal gradient algorithm, the second option corresponds to exact block minimization. Block coordinate descent algorithms are usualy analysed under coercvity assumptions:

Assumption 7.2.1. *The sublevelset $\{y \in \mathbb{R}^p,\ F(y) \leq F(x_0)\}$ is compact, for any $y \in \mathbb{R}^p$ such that $F(y) \leq F(x_0)$, $\|y - x^*\|_2 \leq R$.*

7.3 Convergence rate analysis using random blocks

7.3.1 Smooth setting

The following technical Lemma is classical.

Lemma 7.3.1. *Let $(A_k)_{k\in\mathbb{N}}$ be a sequence of positive real numbers and $\gamma > 0$ be such that*

$$A_k - A_{k+1} \geq \gamma A_k^2$$

then for all $k \in \mathbb{N}$, $k \geq 1$, $A_k \leq (\gamma k)^{-1}$.

Proof. We have for all $k \in \mathbb{N}$,

$$\frac{1}{A_{k+1}} - \frac{1}{A_k} = \frac{A_k - A_{k+1}}{A_k A_{k+1}} \geq \gamma \frac{A_k^2}{A_{k+1}A_k} = \gamma \frac{A_k}{A_{k+1}} \geq \gamma.$$

Hence for all $k \in \mathbb{N}$,

$$\frac{1}{A_k} \geq \frac{1}{A_0} + \gamma k \geq k\gamma.$$

□

Proposition 7.3.1. *Consider the problem*

$$\min_{x\in\mathbb{R}^p} f(x)$$

where $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ is convex differentiable with L-Lipschitz gradient. Choose $x_0 \in \mathbb{R}$ and a sequence of random variables $(i_k)_{k\in\mathbb{N}}$ independently identically distributed uniformly on $\{1, \dots, p\}$ and a sequence of positive step sizes. Consider the recursion

$$x_{k+1} = x_k - \frac{1}{L}\nabla_{i_k} f(x_k) \tag{7.1}$$

Then for all $k \in \mathbb{N}$, $k \geq 1$

$$\mathbb{E}\left[f(x_k) - f^*\right] \leq \frac{2pLR^2}{k}.$$

Proof. Fix, $k \in \mathbb{N}$, and condition on x_k and $i_0, \dots, i_k$ so that x_{k+1} is deterministic. We remark that $t \mapsto f(x_k + te_{i_k})$ is convex with L-Lipschitz gradient. Applying Lemma 7.3.1 with $x = z = x_k$, $y = x_{k+1}$,

$$f(x_k) \geq f(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 = f(x_{k+1}) + \frac{1}{2L}\|\nabla_i f(x_k)\|_2^2,$$

and in particular f is decreasing along the sequence. Taking expectation with respect to i_k, we obtain

$$\mathbb{E}\left[f(x_{k+1})|x_k\right] \leq f(x_k) - \frac{1}{2pL}\|\nabla f(x_k)\|_2^2 \tag{7.2}$$

From convexity, we have $f^* \geq f(x) - \|\nabla f(x)\|\|x - x^*\|$ and using the fact that f is decreasing along the sequence, $\|x_k - x^*\|$ remains bounded. We have

$$\|\nabla f(x)\|_2^2 \geq \frac{(f(x_k) - f^*)^2}{R^2}$$

and

$$\mathbb{E}\left[f(x_{k+1})|x_k\right] - f^* \leq f(x_k) - f^* - \frac{1}{2pL}\|\nabla f(x_k)\|_2^2$$
$$f(x_k) - f^* - \frac{(f(x_k) - f^*)^2}{2pLR^2}$$

Taking expectation with respect to x_k and using the fact that $\mathbb{E}\left[Z^2\right] \geq \mathbb{E}\left[Z\right]^2$, we obtain

$$\mathbb{E}\left[f(x_{k+1}) - f^*\right] \leq \mathbb{E}\left[f(x_k) - f^*\right] - \frac{\mathbb{E}\left[f(x_k) - f^*\right]^2}{2pLR^2}.$$

Applying Lemma 7.3.1, we obtain for all $k \in \mathbb{N}$, $k \geq 1$,

$$\mathbb{E}\left[f(x_k) - f^*\right] \leq \frac{2pLR^2}{k}.$$

□

7.3.2 Extension to the nonsmooth setting

The following is a simplification of the arguments given in [50].

Proposition 7.3.2. *Consider the problem*

$$\min_{x \in \mathbb{R}^d} F(x) \;:=\; f(x) + \sum_{i=1}^{p} g_i(x)$$

where $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ is convex differentiable with L-Lipschitz gradient, each $g_i : \mathbb{R}^p \mapsto \mathbb{R}$ is convex and lower semicontinuous and only depends on coordinate i. Choose $x_0 \in \mathbb{R}$ and a sequence of random variables $(i_k)_{k\in\mathbb{N}}$ independently identically distributed uniformly on $\{1, \ldots, p\}$ and a sequence of positive step sizes. Consider the recursion

$$x_{k+1} = \arg\min_y f(x_k) + \langle \nabla_{i_k} f(x_k), y - x_k\rangle + \frac{L}{2}\|y - x_k\|_2^2 + g_{i_k}(y) \tag{7.3}$$
$$= \operatorname{prox}_{g_{i_k}/L}\left(x_k - \frac{1}{L}\nabla_{i_k} f(x_k)\right). \tag{7.4}$$

Set $C = \max\left\{LR^2, F(x_0) - F^*\right\}$, where R is given in Assumption 7.2.1, we have, for all $k \geq 1$,

$$\mathbb{E}\left[F(x_k) - F^*\right] \leq \frac{2pC}{k}.$$

Proof. Fix, $k \in \mathbb{N}$, and condition on x_k and $i_0, \ldots, i_k$ so that x_{k+1} is deterministic. We remark that $t \mapsto f(x_k + te_{i_k})$ is convex with L-Lipschitz gradient. Noting that the iteration actually solves a univariate problem, applying Lemma 7.3.1 with $x = z = x_k$, $y = x_{k+1}$,

$$f(x_k) + g_{i_k}(x_k) \geq f(x_{k+1}) + g_{i_k}(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2.$$

By assumption, g_i depends only on coordinate i so that, $g_i(x_k) = g_i(x_{k+1})$ for $i \neq i_k$.

$$F(x_k) \geq F(x_{k+1}) + \frac{L}{2}\|x_{k+1} - x_k\|_2^2,$$

So that F is non increasing along the sequence and for any $k \in \mathbb{N}$, $\|x_k - x^*\|_2^2 \leq R^2$, almost surely.

We write $g = \sum_{i=1}^p g_i$. From the definition of the proximity operator, we have

$$f(x_{k+1}) + g_{i_k}(x_{k+1}) \leq f(x_k) + \langle \nabla_{i_k} f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + g_{i_k}(x_{k+1})$$
$$F(x_{k+1}) \leq f(x_k) + \langle \nabla_{i_k} f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + g_{i_k}(x_{k+1}) + \sum_{i \neq i_k} g_i(x_k).$$

Since each g_i only depends on coordinate i, prox_g can be computed coordinate by coordinate. Taking expectation with respect to i_k and setting $z_k = \operatorname{prox}_{g/L}\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$, we obtain

$$\mathbb{E}\left[F(x_{k+1})|x_k\right] \leq \frac{1}{p}\left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_k)\right) + \frac{p-1}{p}F(x_k). \tag{7.5}$$

By definition of the proximity operator, we have for any $y \in \mathbb{R}^p$,

$$\begin{aligned} &f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_k) \\ &\leq f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \frac{L}{2}\|y - x_k\|_2^2 + g(y) \\ &\leq F(y) + \frac{L}{2}\|y - x_k\|_2^2 \end{aligned}$$

In particular, for any $\alpha \in [0, 1]$,

$$\begin{aligned} &f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_k) \\ &\leq F(\alpha x^* + (1-\alpha)x_k) + \frac{\alpha^2 L}{2}\|x^* - x_k\|_2^2 \\ &\leq \alpha F(x^*) + (1-\alpha)F(x_k) + \frac{\alpha^2 C}{2} \end{aligned}$$

The minimum is attained for $\alpha = (F(x_k) - F^*)/C \leq 1$ so that

$$\begin{aligned} &f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_k) - F^* \\ &\leq \left(1 - \frac{F(x_k) - F^*}{2C}\right)(F(x_k) - F^*) \end{aligned}$$

Pluging this in (7.5), we obtain

$$\begin{aligned} \mathbb{E}\left[F(x_{k+1})|x_k\right] - F^* &\leq \frac{F(x_k) - F^*}{p}\left(1 - \frac{F(x_k) - F^*}{2C}\right) \\ &\quad + \frac{p-1}{p}(F(x_k) - F^*) \\ &= (F(x_k) - F^*)\left(1 - \frac{F(x_k) - F^*}{2pC}\right) \end{aligned} \tag{7.6}$$

Taking expectation with respect to x_k and using the fact that $\mathbb{E}[Z^2] \geq \mathbb{E}[Z]^2$, we have

$$\mathbb{E}\left[F(x_{k+1}) - F^*\right] \leq \mathbb{E}[F(x_k) - F^*] - \frac{1}{2pC}\mathbb{E}[F(x_k) - F^*]^2. \tag{7.7}$$

The result follows from Lemma 7.3.1. □

7.4 Convergence rates using deterministic blocks

Deterministic block selection may lead to similar theoretical guaranties, there is some computational overhead, but as we should see, this is affordable for Lasso instances. A broader discussion on this aspect is found in [48].

Proposition 7.4.1. *Consider the problem*

$$\min_{x\in\mathbb{R}^d} f(x)$$

where $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ is convex differentiable with L-Lipschitz gradient. Choose $x_0 \in \mathbb{R}$, and consider the recursion

$$x_{k+1} = x_k - \frac{1}{L}\nabla_{i_k} f(x_k) \tag{7.8}$$

where i_k is the largest block of $\nabla f(x_k)$ in Euclidean norm. Then for all $k \in \mathbb{N}$, $k \geq 1$

$$f(x_k) - f^* \leq \frac{2pLR^2}{k}.$$

Proof. The proof is essentially the same as in Proposition 7.3.1, we have for any $k \in \mathbb{N}$,

$$\|\nabla f(x_k)\|_2^2 \leq p\|\nabla_i f(x_k)\|_2^2$$

and one obtain using the same arguments as in (7.2)

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2pL}\|\nabla f(x_k)\|_2^2 \tag{7.9}$$

and the rest of the analysis is the same. □

Using similar ideas, one obtains the same behaviour for deterministic block proximal gradient algorithm.

Proposition 7.4.2. *Consider the problem*

$$\min_{x\in\mathbb{R}^d} F(x) \;:=\; f(x) + \sum_{i=1}^{p} g_i(x)$$

where $f\colon \mathbb{R}^p \mapsto \mathbb{R}$ is convex differentiable with L-Lipschitz gradient, each $g_i : \mathbb{R}^p \mapsto \mathbb{R}$ is convex and lower semicontinuous and only depends on coordinate i. Choose $x_0 \in \mathbb{R}$ and consider the recursion

$$x_{k+1} = \arg\min_y f(x_k) + \langle \nabla_{i_k} f(x_k), y - x_k\rangle + \frac{L}{2}\|y - x_k\|_2^2 + g_{i_k}(y) \tag{7.10}$$

$$= \operatorname{prox}_{g_{i_k}/L}\left(x_k - \frac{1}{L}\nabla_i f(x_k)\right). \tag{7.11}$$

where i_k is given by

$$\arg\min_i \left\{ \langle y - x_k, \nabla_i f(x_k)\rangle + \frac{L}{2}\|y - x_k\|_2^2 + g(y) - g_i(x_k),\; y = \operatorname{prox}_{g_i/L}\left(x_k - \frac{1}{L}\nabla_i f(x_k)\right)\right\}$$

Set $C = \max\left\{LR^2, F(x_0) - F^*\right\}$, where R is given in Assumption 7.2.1, we have, for all $k \geq 1$,

$$F(x_k) - F^* \leq \frac{2pC}{k}.$$

Proof. Using the same arguments as in the proof of Proposition 7.3.2, we obtain for any $k \in \mathbb{N}$,

$$\begin{aligned} F(x_{k+1}) &\leq f(x_k) + \langle \nabla_{i_k} f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + g_{i_k}(x_{k+1}) + \sum_{i \neq i_k} g_i(x_k) \\ &= F(x_k) + \langle \nabla_{i_k} f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 + g_{i_k}(x_{k+1}) - g_{i_k}(x_k). \end{aligned}$$

Since each g_i only depends on coordinate i, prox_g can be computed coordinate by coordinate. Setting $z_k = \operatorname{prox}_{g/L}\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$, and $g = \sum_i g_i$, we deduce from the definition of i_k,

$$\begin{aligned} F(x_{k+1}) &\leq F(x_k) + \frac{1}{p}\left(\langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_{k+1}) - g(x_k)\right) \\ &= F(x_k) + \frac{1}{p}\left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_k) - g(x_k) - f(x_k)\right) \\ &= \frac{p-1}{p}F(x_k) + \frac{1}{p}\left(f(x_k) + \langle \nabla f(x_k), z_k - x_k \rangle + \frac{L}{2}\|z_k - x_k\|_2^2 + g(z_k)\right) \end{aligned}$$

This is similar to (7.5) and the result follows from the same arguments. □

7.5 Comments on complexity for quadratic problems

The Lasso problem is a special case for block descent methods since the objective is quadratic. This leads to the following remark

- Computing the gradient of the Lasso problem costs a matrix vector product which complexity is of the order of d^2.
- Given $\theta \in \mathbb{R}^d$ and $\beta = \mathbb{X}^T(\mathbb{X}\theta - Y)$, choosing $\tilde{\theta}$ differing from θ in at most one coordinate, computing $\mathbb{X}^T(\mathbb{X}\tilde{\theta} - Y)$ given β costs only of the order of d operations by only considering the corresponding column of $\mathbb{X}^T\mathbb{X}$.

As a consequence the cost of performing one iteration of full proximal gradient for the Lasso problem is roughly equivalent to the cost of performing d iterations of random block proximal gradient.

Given the value of the gradient, the added complexity of computing the deterministic block is of the order of d as it requires only one path through the coordinates of the gradient and the current estimate θ. Hence the deterministic rule has similar complexity per iteratation as the random block rules.

Chapter 8

Further reading

We provides a non exhaustive list of themes and references which are connex to the matter treated in these notes.

- Learning theory [63, 19, 17].
- Compressed sensing [22, 28, 23].
- Conditions for consistency of ℓ_1 norm minimization [7, 22, 62, 25].
- Model selection consistency of Lasso [65, 60, 61].
- Stochastic approximation [11, 49, 16, 41].
- Variance reduction in stochastic approximation for finite sums [56, 32, 27].
- Dual methods in learning [58, 5, 31].
- First order methods for nonconvex problems [4, 2**?** , 30].
- Polynomial optimization [36, 37].
- Tradeoffs in large scale learning and lower bounds [17, 12, 3, 24].

References

Bibliography

[1] Scott Aaronson. Np-complete problems and physical reality. *ACM Sigact News*, 36(1):30–52, 2005.

[2] Pierre-Antoine Absil, Robert Mahony, and Benjamin Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.

[3] Alekh Agarwal, Martin J Wainwright, Peter L Bartlett, and Pradeep K Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*, pages 1–9, 2009.

[4] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.

[5] Francis Bach. Duality between subgradient and conditional gradient methods. *SIAM Journal on Optimization*, 25(1):115–129, 2015.

[6] Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

[7] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3): 253–263, 2008.

[8] Alexander Barvinok. *A course in convexity*, volume 54. American Mathematical Soc., 2002.

[9] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[10] Ahron Ben-Tal and Arkadi Nemirovski. *Lectures on modern convex optimization: analysis, algorithms, and engineering applications*, volume 2. Siam, 2001.

[11] Michel Benaïm. Dynamics of stochastic approximation algorithms. In *Seminaire de probabilites XXXIII*, pages 1–68. Springer, 1999.

[12] Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *Conference on Learning Theory*, pages 1046–1066, 2013.

[13] Dimitri P Bertsekas. *Nonlinear programming*. Athena scientific Belmont, 1999.

[14] Dimitris Bertsimas, Angela King, Rahul Mazumder, et al. Best subset selection via a modern optimization lens. *The annals of statistics*, 44(2):813–852, 2016.

[15] Lenore Blum, Mike Shub, Steve Smale, et al. On a theory of computation and complexity over the real numbers: *np*-completeness, recursive functions and universal machines. *Bulletin (New Series) of the American Mathematical Society*, 21(1):1–46, 1989.

[16] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[17] Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Advances in neural information processing systems*, pages 161–168, 2008.

[18] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[19] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[20] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[21] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[22] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

[23] Emmanuel J Candès and Michael B Wakin. An introduction to compressive sampling. *IEEE signal processing magazine*, 25(2):21–30, 2008.

[24] Venkat Chandrasekaran and Michael I Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, page 201302293, 2013.

[25] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12(6):805–849, 2012.

[26] Patrick L Combettes and Jean-Christophe Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[27] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654, 2014.

[28] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4): 1289–1306, 2006.

[29] Rick Durrett. *Probability: theory and examples*. Cambridge university press, 2010.

[30] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[31] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

[32] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323, 2013.

[33] Sudeep Kamath. Concentration of measure. Nexus of Information and Computation Theories Tutorial Week at CIRM, 2016. URL `http://www.youtube.com/watch?v=mpbWQbkl8_g#t=20m15s`.

[34] Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311. ACM, 1984.

[35] Leonid G Khachiyan. Polynomial algorithms in linear programming. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 20(1):51–68, 1980.

[36] Jean B Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on optimization*, 11(3):796–817, 2001.

[37] Jean-Bernard Lasserre. *Moments, positive polynomials and their applications*, volume 1. World Scientific, 2010.

[38] Lennart Ljung. Analysis of recursive stochastic algorithms. *IEEE transactions on automatic control*, 22(4):551–575, 1977.

[39] David G Luenberger. *Optimization by vector space methods*. John Wiley & Sons, 1997.

[40] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93(2):273–299, 1965.

[41] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[42] Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

[43] Arkadii Semenovich Nemirovsky and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

[44] Yu Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[45] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

[46] Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*, volume 13. Siam, 1994.

[47] Yurii E Nesterov. A method for solving the convex programming problem with convergence rate o (1/kˆ 2). In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547, 1983.

[48] Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate descent converges faster with the gauss-southwell rule than random selection. In *International Conference on Machine Learning*, pages 1632–1641, 2015.

[49] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[50] Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2): 1–38, 2014.

[51] Phillippe Rigollet and Jan-Christian Hütter. High dimensional statistics. *Lecture notes (MIT)*, 2017.

[52] Omar Rivasplata. Subgaussian random variables: An expository note. Unpublished notes, 2012.

[53] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[54] R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 1998.

[55] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 1970.

[56] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

[57] Alexander Schrijver. *Theory of linear and integer programming*. John Wiley & Sons, 1986.

[58] Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(Feb):567–599, 2013.

[59] Steve Smale. Mathematical problems for the next century. *The mathematical intelligencer*, 20(2):7–15, 1998.

[60] Samuel Vaiter, Gabriel Peyré, Charles Dossal, and Jalal Fadili. Robust sparse analysis regularization. *IEEE Transactions on information theory*, 59(4):2001–2016, 2013.

[61] Samuel Vaiter, Gabriel Peyré, and Jalal Fadili. Model consistency of partly smooth regularizers. *IEEE Transactions on Information Theory*, 64(3):1725–1737, 2018.

[62] Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[63] Vladimir Naumovich Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[64] Stephen J Wright. Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34, 2015.

[65] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.