Learning theory from first principles

Lecture 7: Sparse methods

Francis Bach

November 6, 2020

Class summary

-ℓ_0 penalty
-ℓ_1 penalty
-High-dimensional estimation

1 Introduction

In this course, we have seen the strong effect of the dimensionality of the input space $\mathcal{X}$ on the generalization performance of supervised learning methods, in two settings:

- When the target function f^* was only assumed to be Lipschitz-continuous on $\mathcal{X} = \mathbb{R}^d$, we saw that the excess risk for k-nearest-neigbors, Nadaraya-Watson estimation (Lecture 5), or positive kernel methods (Lecture 6), was scaling as $n^{-2/(d+2)}$.
- When the target function is linear in some features $\varphi(x) \in \mathbb{R}^d$, then the excess risk was scaling as d/n.

In these two situations, when d is large (of course much larger in the linear case), efficient learning is not possible in general.

In order to improve upon these rates, we study two techniques in this course. The first one is regularization, e.g., by the ℓ_2-norm, that allows to obtain dimension-independent bounds that cannot improve over the bounds above in the worst-case, but are typically adaptive to additional regularity (see Lectures 2 and 6).

In this lecture, we consider a another framework, namely *variable selection*, whose aim is to build predictors that depend only on a small number of variables. The key difficulty is that the identity of the selected variables is not known in advance. Note that this can be done by regularization techniques.

In practice, variable selection is used in mainly two ways:

- The original set of features is large.

- Given some input $x \in \mathcal{X}$, a large-dimensional feature vector $\varphi(x)$ is built where features are added that could potentially help predicting the response, but from which we expect only a small number to be relevant.

  ⚠ If no good predictor with small number of active variables exists, these methods are not supposed to work better.

In this lecture, we focus on linear methods, where we assume that we have a feature vector $\varphi(x) \in \mathbb{R}^d$, and we aim to minmize

$$\mathbb{E}[\ell(y, \varphi(x)^\top \theta)]$$

with respect to $\theta \in \mathbb{R}^d$, for some loss function $\ell : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}$. We will consider two variable selection techniques, namely the penalization by $\|\theta\|_0$ the number of non-zeros in θ (often called abusively the "ℓ_0-norm"), or the ℓ_1-norm.

Main focus on least-squares. These two types of penalties can be applied to all losses, but in this lecture, for simplicity we will mostly consider the square loss, and in most cases, the fixed design setting (see the classical set-up in Lecture 2), and assume that we have n observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, such that there exists $\theta_* \in \mathbb{R}^d$ for which for $i \in \{1, \dots, n\}$,

$$y_i = \varphi(x_i)^\top \theta_* + \varepsilon_i,$$

where x_i is assumed deterministic, and ε_i has zero mean and variance σ^2 (we also assume independence, and sometimes stronger regulariry, such as bounded almost surely, or Gaussian). The goal is then to find $\theta \in \mathbb{R}^d$, such that

$$\frac{1}{n}\|\Phi(\theta - \theta_*)\|_2^2 = (\theta - \theta_*)^\top \widehat{\Sigma}(\theta - \theta_*)$$

is as small as possible, where $\Phi \in \mathbb{R}^{n \times d}$ is the design matrix and $\widehat{\Sigma} = \frac{1}{n}\Phi^\top \Phi$ the non-centered empirical covariance matrix. We recall from Lecture 2 that for the ordinary least-squares estimator, this excess risk is less than $\sigma^2 d/n$. This is the best possible performance if we make no assumption on θ_*. In this lecture, we assume that θ_* is sparse, that is, only a few of its components are non-zero, or in other words, $\|\theta_*\|_0 = k$ is small compared to d.

1.1 Dedicated proof technique for constrained least-squares

In this lecture, we consider a more refined proof technique[1] that can extend to constrained versions of least-squares (while our technique in Lecture 2 heavily relies on having a closed form for the estimator, which is not possible in constrained or regularized cases except in few instances, such as ridge regression).

We denote by $\hat{\theta}$ a minimizer of $\frac{1}{n}\|y - \Phi\theta\|_2^2$ with the constraint that $\theta \in \Theta$. If $\theta_* \in \Theta$, then we have, by optimality of $\hat{\theta}$:

$$\|y - \Phi\hat{\theta}\|_2^2 \leqslant \|y - \Phi\theta_*\|_2^2.$$

[1] Taken from Philippe Rigollet's lecture notes, see `http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf`. See also [1] for an example of application.

By expanding with $y = \Phi\theta_* + \varepsilon$, we get $\|\varepsilon - \Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant \|\varepsilon\|_2^2$, leading to, by expanding the norms:

$$\|\varepsilon\|_2^2 - 2\varepsilon^\top \Phi(\hat{\theta} - \theta_*) + \|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant \|\varepsilon\|_2^2,$$

and thus

$$\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant 2\varepsilon^\top \Phi(\hat{\theta} - \theta_*).$$

We can write it as

$$\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant 2\|\Phi(\hat{\theta} - \theta_*)\|_2 \cdot \varepsilon^\top \Big(\frac{\Phi(\hat{\theta} - \theta_*)}{\|\Phi(\hat{\theta} - \theta_*)\|_2} \Big).$$

This reformulation is difficult to deal with because $\hat{\theta}$ appears on the right side of the equation. Like done for upper-bounding estimation errors in Lecture 3, we can maximize with respect to $\theta \in \Theta$, which leads to

$$\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant 2\|\Phi(\hat{\theta} - \theta_*)\|_2 \cdot \sup_{\theta \in \Theta} \varepsilon^\top \Big(\frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2} \Big),$$

and finally

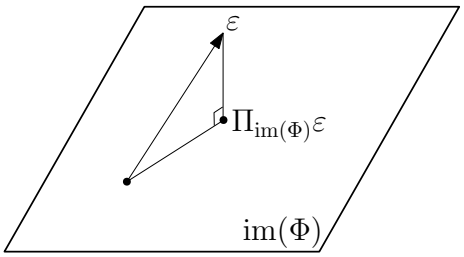
$$\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant 4 \sup_{\theta \in \Theta} \Big[\varepsilon^\top \Big(\frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2} \Big) \Big]^2. \tag{1}$$

This inequality is true almost surely, and we can take expectation (with respect to ε) to obtain bounds. Therefore, in this lecture, we will compute expectations of maxima of quadratic forms in ε.

For example, when $\Theta = \mathbb{R}^d$ (no constraints), we get, by taking $z = \frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2}$, with Π_Φ the orthogonal projector on the image space $\mathrm{im}(\Phi)$:

$$\mathbb{E}\big[\|\Phi(\hat{\theta} - \theta_*)\|_2^2\big] \leqslant 4\mathbb{E}\Big[\sup_{z \in \mathrm{im}(\Phi), \|z\|_2 = 1} \big[\varepsilon^\top z\big]^2 \Big].$$

By the simple geometric argument below,



we have

$$\sup_{z \in \mathrm{im}(\Phi), \|z\|_2 = 1} \big[\varepsilon^\top z\big]^2 = \|\Pi_\Phi \varepsilon\|^2,$$

leading to

$$\mathbb{E}\big[\|\Phi(\hat{\theta} - \theta_*)\|_2^2\big] \leqslant 4\mathbb{E}\Big[\|\Pi_\Phi \varepsilon\|^2\Big] = 4\sigma^2 \mathrm{rank}(\Phi).\Big] = 4\sigma^2 \mathrm{rank}(\Phi).$$

We thus, get up to a constant 4, the excess risk as $\sigma^2 d/n$, which is worse than the direct computation from Lecture 2, but allows extensions to more complex situations.

This reasoning also allows to get high probability bounds by adding assumptions on the noise ε. This also extends to penalized problems (see Section 2.2).

1.2 Probabilistic and combinatorial lemmas

We start with two small probabilistic lemmas:

Lemma 1 *If $z \in \mathbb{R}^n$ is normally distributed with mean 0 and covariance matrix $\sigma^2 I$, then, if $s < \frac{1}{2\sigma^2}$, $\mathbb{E}\big[e^{s\|z\|_2^2}\big] = (1-2\sigma^2 s)^{-n/2}$.*

Proof We have, for $\sigma = 1$ (from which we can derive the result for all σ), and $s < 1/2$:

$$\begin{aligned}\mathbb{E}\big[e^{s\|z\|_2^2}\big] &= \mathbb{E}\big[e^{s\sum_{i=1}^n z_i^2}\big] = \prod_{i=1}^n \mathbb{E}\big[e^{sz_i^2}\big] = \frac{1}{(2\pi)^{n/2}} \prod_{i=1}^n \int_{-\infty}^{\infty} e^{(s-\frac{1}{2})z_i^2} dz_i \\ &= \frac{1}{(2\pi)^{n/2}} \prod_{i=1}^n \sqrt{2\pi}(1-2s)^{-1/2} = (1-2s)^{-n/2}.\end{aligned}$$

■

Lemma 2 *Let $u_1, \ldots, u_m$ be m random variables which are potentially dependent, and $s > 0, v > 0$ such that for each $i \in \{1, \ldots, m\}$, $\mathbb{E}\big[e^{su_i}\big] \leqslant v$. Then, $\mathbb{E}\big[\max\{u_1, \ldots, u_m\}\big] \leqslant \frac{1}{s}\log(mv)$.*

Proof Following the reasoning from Section 6.2 in Lecture 1, for any $s \in \mathbb{R}$,

$$\mathbb{E}\big[\max\{u_1, \ldots, u_m\}\big] \leqslant \frac{1}{s}\log\Big(\sum_{i=1}^m \mathbb{E}\big[e^{su_i}\big]\Big) \leqslant \frac{1}{s}\log(mv).$$

■

The previous two lemmas can be combined to upper-bound the expectation of squared norms of Gaussian random variables: if $z_1, \ldots, z_m \in \mathbb{R}^n$ are Gaussian random vectors which are potentially dependent, but for which the covariance matrix of z_i has eigenvalues less than σ^2, we have for $s = \frac{1}{4\sigma^2}$, and Lemma 1, $\mathbb{E}[e^{s\|z\|_2^2}] \leqslant 2^{n/2}$, and from Lemma 2,

$$\mathbb{E}[\max\{\|z_1\|_2^2, \ldots, \|z_m\|_2^2\}] \leqslant 4\sigma^2 \log(m2^{n/2}) = 2n\sigma^2\log(2) + 4\sigma^2\log(m),$$

which is to be compared to the expectation of each elements of the max, which is less than $\sigma^2 n$. We pay an additive factor proportion to $\sigma^2\log(m)$. This will be applied to $m \propto d^k$, leading to the extra term in $\sigma^2 k\log(d)$ for methods based on the ℓ_0-penalty.

The term in d^k comes from the following lemma.

Lemma 3 *Let $d > 0$ and $k \in \{1, \ldots, d\}$. Then $\log\binom{d}{k} \leqslant k(1 + \log\frac{d}{k})$.*

Proof By recursion on k, the inequality is trivial for $k = 1$, and if $\binom{d}{k-1} \leqslant \big(\frac{ed}{k-1}\big)^{k-1}$, then

$$\binom{d}{k} = \binom{d}{k-1}\frac{d-k}{k} \leqslant \big(\frac{ed}{k-1}\big)^{k-1}\frac{d}{k} \leqslant \big(\frac{ed}{k}\big)^{k-1}\big(1 + \frac{1}{k-1}\big)^{k-1}\frac{d}{k} \leqslant \big(\frac{ed}{k}\big)^{k-1} e\frac{d}{k} = \big(\frac{ed}{k}\big)^k,$$

where we use for $\alpha > 0$, $(1+\frac{1}{\alpha})^\alpha = \exp(\alpha\log(1+\alpha)) \leqslant \exp(1) = e$. ■

We now consider two types of variable selection frameworks, one based on ℓ_0-penalties, one based on ℓ_1-penalties.

2 Variable selection by ℓ_0 penalty

In this section, we assume that the target θ_* has k non-zero components, that is, $\|\theta_*\|_0 = k$. We denote by $A = \mathrm{supp}(\theta_*)$ the "support" of θ_*, that is, the subset of $\{1, \dots, d\}$ composed of j such that $(\theta_*)_j \neq 0$. We have $|A| = k$.

2.1 Assuming k is known

Price of adaptivity. If we knew the set A, then we could simply perform least-squares with the design matrix $\Phi_A \in \mathbb{R}^{n \times |A|}$, where Φ_B denotes the sub-matrix of Φ obtained by keeping only the columns from B, with an excess risk proportional to $\sigma^2 k/n$ (this is what we called the "oracle" in Section 4). Thus, as long as k is small compared to n, we can estimate θ_* correctly, regardless of the potentially large value of d.

However, we do not know A in advance, and we have to estimate it. We will see that this will lead to an extra factor of $\log \big(\frac{d}{k}\big) \leqslant \log d$, due to the potentially large number of models with k variables. We first start by assuming that the cardinality k is known in advance, and we consider Gaussian noise for simplicity (this extends to sub-Gaussian noise as well, see note below).

Proposition 1 (Model selection - known k) *Assume $y = \Phi\theta_* + \varepsilon$, with $\varepsilon \in \mathbb{R}^n$ a vector with independent Gaussian components of zero mean and variance σ^2, with $\|\theta_*\|_0 \leqslant k$, for $k \leqslant d/2$. Let $\hat{\theta}$ be the minimizer of $\|y - \Phi\theta\|_2^2$ with the constraint that $\|\theta\|_0 \leqslant k$. Then:*

$$\mathbb{E}\big[(\hat{\theta} - \theta_*)^\top \widehat{\Sigma}(\hat{\theta} - \theta_*)\big] = \mathbb{E}\big[\frac{1}{n}\|\Phi(\hat{\theta} - \theta_*)\|_2^2\big] \leqslant 32\sigma^2 \frac{k}{n}\Big(\log\Big(\frac{d}{k}\Big) + 1\Big).$$

Proof Starting from Eq. (1), we see that for any θ such that $\|\theta\|_0 \leqslant k$, we have $\|\theta - \theta_*\|_0 \leqslant 2k$, and thus we have, from Section 1.1:

$$\begin{aligned}
\|\Phi(\hat{\theta} - \theta_*)\|_2^2 &\leqslant 4 \sup_{\theta \in \mathbb{R}^d, \|\theta\|_0 \leqslant k} \Big[\varepsilon^\top \Big(\frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2}\Big)\Big]^2 \\
&\leqslant 4 \sup_{\theta \in \mathbb{R}^d, \|\theta - \theta_*\|_0 \leqslant 2k} \Big[\varepsilon^\top \Big(\frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2}\Big)\Big]^2 \text{ from the discussion above,} \\
&= 4 \sup_{B \subset \{1,\dots,n\},\ |B| \leqslant 2k} 4 \sup_{\mathrm{Supp}(\theta - \theta_*) = B} \Big[\varepsilon^\top \Big(\frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2}\Big)\Big]^2 \text{ by separating by supports,} \\
&\leqslant 4 \sup_{B \subset \{1,\dots,n\},\ |B| \leqslant 2k} \ \sup_{z \in \mathrm{im}(\Phi_B), \|z\|_2 = 1} \big[\varepsilon^\top z\big]^2 \\
&\leqslant 4 \sup_{B \subset \{1,\dots,n\},\ |B| \leqslant 2k} \|\Pi_{\Phi_B}\varepsilon\|^2 \leqslant 4 \sup_{B \subset \{1,\dots,n\},\ |B| = 2k} \|\Pi_{\Phi_B}\varepsilon\|^2,
\end{aligned}$$

because $\|\Pi_{\Phi_B}\varepsilon\|^2$ is non-decreasing in B.

The random variable $\|\Pi_{\Phi_B}\varepsilon\|^2$ has an expectation which is less than $2k$, given that there are $\binom{d}{2k} \leqslant \left(\frac{ed}{2k}\right)^{2k}$ sets B of cardinality $2k$ (bound from Lemma 3), we should expect, with concentration inequalities from Section 1.2, that we pay a price of $\log\left(\frac{ed}{2k}\right)^{2k} \approx k\log\frac{d}{k}$. We will make this reasoning formal.

Indeed, $\Pi_{\Phi_B}\varepsilon$ is normally distributed with isotropic covariance matrix of dimension $|B| \leqslant 2k$, and thus we have for $s\sigma^2 < 1/2$ small enough, from Lemma 1:

$$\mathbb{E}\big[e^{s\|\Pi_{\Phi_B}\varepsilon\|^2}\big] \leqslant (1-2\sigma^2 s)^{-k} = 2^{-k}.$$

Thus, with $s = 1/(4\sigma^2)$, we get, from Lemma 2:

$$\mathbb{E}\big[\|\Phi(\hat{\theta}-\theta_*)\|_2^2\big] \leqslant 16\sigma^2\log\left(\binom{d}{2k}2^k\right) \leqslant 16\sigma^2\log\left(\left(\frac{ed}{2k}\right)^{2k}2^k\right) = 16\sigma^2\left(2k\log\left(\frac{d}{k}\right)+(2-\log 2)k\right).$$

This leads to the desired result. ■

We can make the following observations:

- The result extends beyond Gaussian noise, that is, for all sub-Gaussian ε_i, for which $\mathbb{E}[e^{s\varepsilon_i}] \leqslant e^{s^2\tau^2}$ for all $s > 0$ (for some $\tau > 0$), or, equivalently $\mathbb{P}(|\varepsilon_i| > t) = O(e^{-ct^2})$ for some $c > 0$.
- The result extends if the minimisation is only done approximately.
- This result is not improvable by any algorithm (polynomial time or not), see, e.g., [2, Theorem 2.3].

Algorithms. In terms of algorithms, essentially all subsets of size k have to be looked at for exact minimization, with a cost proportional to $O(d^k)$, which starts to be a problem when k gets large. There are however two simple algorithms that come with guarantees when such fast rates are available for ℓ_1-regularization (see Section 3.3).

- Greedy algorithm: starting from the empty set, variables are added one by one that maximizing the resulting cost reduction. This is often referred to as orthogonal matching pursuit.
- Iterative sorting: Starting from $\theta_0 = 0$, the iterative algorithm goes as follows at iteration t; the upper bound (based on the L-smoothness of the quadratic loss, with $L = \lambda_{\max}(\frac{1}{n}\Phi^\top\Phi)$, see Lecture 4):

  $$\frac{1}{n}\|y-\Phi\theta_{t-1}\|_2^2 - \frac{2}{n}(y-\Phi\theta_{t-1})^\top\Phi(\theta-\theta_{t-1}) + \lambda_{\max}(\frac{1}{n}\Phi^\top\Phi)\|\theta-\theta_{t-1}\|_2^2$$

  on the cost function $\frac{1}{n}\|y-\Phi\theta\|_2^2$ is built and minimized with respect to $\|\theta\|_0 \leqslant k$ to obtain θ_t, which is done (check as an exercise) by computing the unconstrained minimizer $\theta_{t-1} + \dfrac{1}{\lambda_{\max}(\frac{1}{n}\Phi^\top\Phi)}\dfrac{1}{n}\Phi^\top(y-\Phi\theta_{t-1})$, and selecting the k largest components.

2.2 Estimating k (♦♦)

In practice, regardless of the computational cost, one also needs to estimate k. A classical idea to consider penalized maximum likelihood and minimize

$$\frac{1}{n}\|y-\Phi\theta\|_2^2+\lambda\|\theta\|_0. \tag{2}$$

This is known to be a hard problem to solve, which essentially requires to look at all 2^d subsets. For a well chosen λ, this (almost) leads to the same performance as if k were known.

Proposition 2 (Model selection - ℓ_0-penalty) *Assume $y=\Phi\theta_*+\varepsilon$, with $\varepsilon\in\mathbb{R}^n$ a vector of with independent Gaussian components of zero mean and variance σ^2, with $\|\theta_*\|_0\leqslant k$. Let $\hat{\theta}$ be the minimizer of Eq. (2). Then, for $\lambda=\frac{2\sigma^2}{n}(3+2\log d)$, we have:*

$$\mathbb{E}\big[\frac{1}{n}\|\Phi(\hat{\theta}-\theta_*)\|_2^2\big]\leqslant\frac{16\sigma^2k}{n}(3+2\log d)+\frac{5\sigma^2}{n}.$$

Proof We follow the same proof technique than in Section 1.1, but now for regularized problems. We have by optimality of $\hat{\theta}$:

$$\|y-\Phi\hat{\theta}\|_2^2+n\lambda\|\hat{\theta}\|_0\leqslant\|y-\Phi\theta_*\|_2^2+n\lambda\|\theta_*\|_0,$$

which leads to, using the inequality $2ab\leqslant 2a^2+\frac{1}{2}b^2$:

$$\begin{aligned}\|\Phi(\hat{\theta}-\theta_*)\|_2^2 &\leqslant 2\|\Phi(\hat{\theta}-\theta_*)\|_2\cdot\varepsilon^\top\Big(\frac{\Phi(\hat{\theta}-\theta_*)}{\|\Phi(\hat{\theta}-\theta_*)\|_2}\Big)+n\lambda\|\theta_*\|_0-n\lambda\|\hat{\theta}\|_0\\ &\leqslant 2\Big(\varepsilon^\top\Big(\frac{\Phi(\hat{\theta}-\theta_*)}{\|\Phi(\hat{\theta}-\theta_*)\|_2}\Big)\Big)^2+\frac{1}{2}\|\Phi(\hat{\theta}-\theta_*)\|_2^2+n\lambda\|\theta_*\|_0-n\lambda\|\hat{\theta}\|_0,\end{aligned}$$

leading to, by taking the supremum over $\theta\in\mathbb{R}^d$:

$$\|\Phi(\hat{\theta}-\theta_*)\|_2^2\leqslant\sup_{\theta\in\mathbb{R}^d}\Big\{4\Big(\varepsilon^\top\Big(\frac{\Phi(\theta-\theta_*)}{\|\Phi(\theta-\theta_*)\|_2}\Big)\Big)^2+2n\lambda\|\theta_*\|_0-2n\lambda\|\theta\|_0\Big\}.$$

We then take the supremum by layers, as $\sup_{\theta\in\mathbb{R}^d}=\sup_{k\in\{1,\dots,d\}}\sup_{|B|=k}\sup_{\mathrm{supp}(\theta)=B}$, that is, and using the same derivations as for Prop. 1:

$$\begin{aligned}\mathbb{E}\big[\|\Phi(\hat{\theta}-\theta_*)\|_2^2\big] &\leqslant \mathbb{E}\Big[\sup_{k\in\{1,\dots,d\}}\sup_{|B|=k}\sup_{\mathrm{supp}(\theta)=B}\Big\{4\Big(\varepsilon^\top\Big(\frac{\Phi(\theta-\theta_*)}{\|\Phi(\theta-\theta_*)\|_2}\Big)\Big)^2+2n\lambda\|\theta_*\|_0-2n\lambda k\Big\}\Big]^2\\ &\leqslant 4\mathbb{E}\Big[\sup_{k\in\{1,\dots,d\}}\sup_{|B|=k}\Big\{\|\Pi_{\Phi_{A\cup B}}\varepsilon\|^2+\frac{n\lambda}{2}\|\theta_*\|_0-\frac{n\lambda}{2}k\Big\}\Big]^2.\end{aligned}$$

We thus get with the same reasoning as in Section 2.1 (based on the probabilistic lemmas from Section 1.2):

$$\begin{aligned}
\mathbb{E}\big[\|\Phi(\hat{\theta}-\theta_*)\|_2^2\big] &\leqslant 16\sigma^2 \log\Big(\sum_{k=1}^{d}\binom{d}{2k}2^{2k}\exp(\frac{n}{2}\frac{\lambda}{\sigma^2}\|\theta_*\|_0 - \frac{n}{2}\frac{\lambda}{\sigma^2}k)\Big)\\
&\leqslant 8n\lambda\|\theta_*\|_0 + 16\sigma^2 \log\Big(\sum_{k=1}^{d}\binom{d}{2k}2^{2k}\exp(-\frac{n}{2}\frac{\lambda}{\sigma^2}k)\Big)\\
&\leqslant 8n\lambda\|\theta_*\|_0 + 16\sigma^2 \log\Big(\sum_{k=1}^{d}(\frac{ed}{2k})^{2k}2^{2k}\exp(-\frac{n}{2}\frac{\lambda}{\sigma^2}k)\Big)\\
&\leqslant 8n\lambda\|\theta_*\|_0 + 16\sigma^2 \log\Big(\sum_{k=1}^{d}\big(\exp(k(2\log(d)+2-\frac{n}{2}\frac{\lambda}{\sigma^2})\Big).
\end{aligned}$$

We thus simply impose that $2\log(d)+2-\frac{n}{2}\frac{\lambda}{\sigma^2}\leqslant -\log 2$, to get

$$\mathbb{E}\big[\|\Phi(\hat{\theta}-\theta_*)\|_2^2\big] \leqslant 8n\lambda\|\theta_*\|_0 + 16\sigma^2\log(2).$$

We can thus choose: $\lambda = \frac{2\sigma^2}{n}(3+2\log d) \geqslant \frac{2\sigma^2}{n}(2+\log 2)$, and get the desired result. ■

We can make the following observations:

- The penalty proportional to $\|\theta\|_0 \log d$ is often referred to as the "BIC penalty".
- Note that we need to know σ^2 in advance, which can be a problem in practice. See [3] for more details and alternative formulations.
- The three most important aspects are that: (1) the bound does not require any assumption on the design matrix Φ, (2) that we observer a positive high-dimensional phenomenon, where d only appears as $\frac{\log d}{n}$, but (3) only exponential-time algorithms are possible for solving the problem with guarantees (see algorithms below).
- **Exercise (♦)**: With a penalty proportional to $\|\theta\|_0 \log\frac{d}{\theta_0}$, show the same bound than for d known.

Algorithms. We can extend the two algorithms from Section 2.1 for the penalized case:

- Forward-backward algorithm to minimize a function of a set B: Starting from the empty set $B=\varnothing$, at every step of the algorithm, one tries both a forward algorithm (adding a node to B) and a backward algorithm (removing a node from B), and only perform a step if it decreases the overall cost function.
- Iterative hard thresholding: compared to the constrained case, we minimize

$$\frac{1}{n}\|y-\Phi\theta_{t-1}\|_2^2 - \frac{2}{n}(y-\Phi\theta_{t-1})^\top\Phi(\theta-\theta_{t-1}) + \lambda_{\max}(\frac{1}{n}\Phi^\top\Phi)\|\theta-\theta_{t-1}\|_2^2 + \lambda\|\theta\|_0,$$

  which can also be computed in closed form (by iterative hard thresholding). That is, with $\theta_t = \theta_{t-1} + \frac{1}{\lambda_{\max}(\Phi^\top\Phi)}\Phi^\top(y-\Phi\theta_{t-1})$, all components $(\theta_t)_j$ such that $|(\theta_t)_j|^2 \geqslant \frac{\lambda}{\frac{1}{n}\lambda_{\max}(\Phi^\top\Phi)}$, are left unchanged and all others are set to zero (left as an exercise).

This is referred to as iterative hard thresholding (while for the ℓ_1-norm, this will be iterative soft thresholding, because, a component is either kept intact or set exactly to zero, leading to a discontinuous behavior.

3 High-dimensional estimation through ℓ_1-regularization

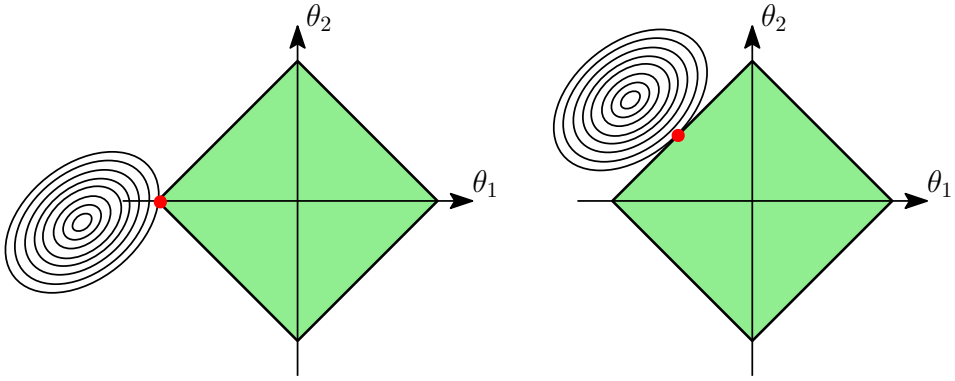
We now consider a computationally efficient alternative to ℓ_0 penalties, namely using ℓ_1 penalties, by minimizing, for the square loss:

$$\frac{1}{2n}\|y - \Phi\theta\|_2^2 + \lambda\|\theta\|_1. \tag{3}$$

This is a convex optimization problem on which algorithms from Lecture 4 can be applied (see instances below). It is often referred to as the "Lasso"problem, for "least absolute shrinkage and selection operator".

3.1 Intuition and algorithms

Sparsity-inducing effect. As opposed to the sqaured ℓ_2-norm used in ridge regression, the ℓ_1-norm is non differentiable, and its non-differentiability is not limited to $\theta = 0$, but in many other points. To see this, we can look at the ℓ_1-ball and its different geometry compared to the ℓ_2-ball. This is directly relevant to situations where we constrain the value of the norm instead of penalizing by it.

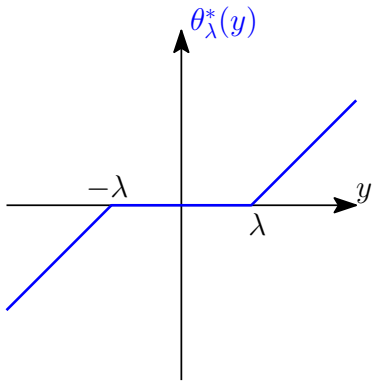


As shown above, where we represent the level set of a potential loss function, the solution of the minimization of the loss subject to the ℓ_1-constraint (in green), is obtained when level sets are "tangent" to the constraint set. In right part, this is obtained in a point away from the axes, but on the left part, this is achieved at one of the corners of the ℓ_1-ball, which are points where one of the components of θ is equal to zero. Such corners are attractive and thus typically lead to sparse solutions.

One-dimensional problem. Another classical way to understand the sparsity-inducing effect is to consider the one-dimensional problem:

$$\min_{\theta\in\mathbb{R}} F(\theta) = \frac{1}{2}(y-\theta)^2 + \lambda|\theta|.$$

Since F is strongly-convex, it has a unique minimizer $\theta^*_\lambda(y)$. For $\lambda = 0$ (no regularization), we have $\theta^*_0(y) = y$, while for $\lambda > 0$, by computing left and right derivatives at zero (to be done as an exercise), one can check that $\theta^*_\lambda(y) = 0$ if $|y| \leqslant \lambda$, and $\theta^*_\lambda(y) = y - \lambda$ for $y > \lambda$, and $\theta^*_\lambda(y) = y + \lambda$ for $y < -\lambda$, which can be put all together as $\theta^*_\lambda(y) = \max\{|y| - \lambda, 0\} \operatorname{sign}(y)$, which is depicted below. This referred to as iterative soft thresholding (this will be useful for proximal methods below).



Note that the minimizer is either sent to zero, or shrunk towards zero.

Optimization algorithms. We can adapt algorithms from Lecture 4 to the problem in Eq. (3).

- Iterative hard-thresholding: We can apply proximal methods to the objective function of the form $F(\theta) + \lambda\|\theta\|_1$ for $F(\theta) = \frac{1}{2n}\|y - \Phi\theta\|_2^2$, for which $F'(\theta) = \frac{1}{2n}\Phi^\top(y - \Phi\theta)$. The plain (non-accelerated) proximal method recursion is

$$\theta_t = \arg\min_{\theta \in \mathbb{R}^d} F(\theta_{t-1}) + F'(\theta_{t-1})^\top(\theta - \theta_{t-1}) + \frac{L}{2}\|\theta - \theta_{t-1}\|_2^2 + \lambda\|\theta\|_1,$$

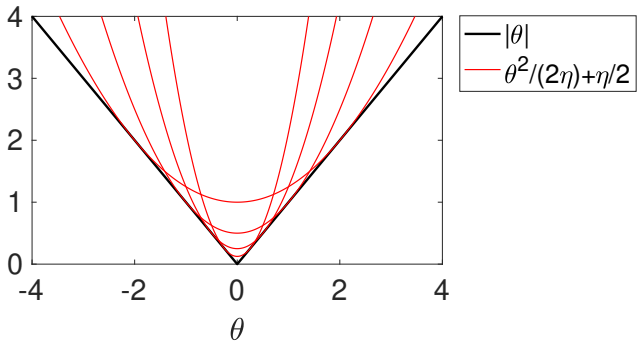
  with $L = \lambda_{\max}(\frac{1}{n}\Phi^\top\Phi)$. This leads to $(\theta_t)_j = \max\{|(\eta_t)_j| - \lambda, 0\} \operatorname{sign}((\eta_t)_j)$, for $\eta_t = \theta_{t-1} - \frac{1}{L}F'(\theta_{t-1})$. This simple algorithm can also be accelerated. The convergence rate then depends on invertibility of $\frac{1}{n}\Phi^\top\Phi$.

- Coordinate descent: Although the ℓ_1-norm is a non-differentiable function, coordinate descent can be applied (because the ℓ_1-norm is "separable"). At each iteration, we select a coordinate to update (at random or by cycling), and optimize with respect to this coordinate, which is a one-dimensional problem which can be solved in closed form. The convergence properties are similar to proximal methods [4].

η-trick. The non-differentiability of the ℓ_1-norm may also be treated through the simple identity:

$$|\theta_j| = \inf_{\eta_j > 0} \frac{\theta_j^2}{2\eta_j} + \frac{\eta_j}{2},$$

where the minimizer is attained at $\eta_j = |\theta_j|$. See below.



This leads to the reformulation of Eq. (3) as

$$\inf_{\theta\in\mathbb{R}^d} \frac{1}{2n}\|y-\Phi\theta\|_2^2 + \lambda\|\theta\|_1 = \inf_{\eta\in\mathbb{R}_+^d} \inf_{\theta\in\mathbb{R}^d} \frac{1}{2n}\|y-\Phi\theta\|_2^2 + \frac{\lambda}{2}\sum_{j=1}^d \frac{\theta_j^2}{2\eta_j} + \frac{\lambda}{2}\sum_{j=1}^d \eta_j,$$

and alternating optimization algorithms can be used: (a) minimizing with respect to η when θ is fixed can be done in closed form as $\eta_j = |\theta_j|$, while minimizing with respect to θ when η is fixed is a quadratic optimization problem which can be solved by a linear system. See more details in `https://francisbach.com/the-\%ce\%b7-trick-or-the-effectiveness-of-reweighted-least-squares/`.

Optimality conditions (♦). In order to study the estimator defined by Eq. (3), it is often necessary to characterize when a certain θ is optimal or not, that is, to derive optimality conditions.

Since the objective function $H(\theta) = F(\theta) + \lambda\|\theta\|_1$ is not differentiable, we need other tools than having the gradient equal to zero. The gradient looks only at d directions (along the coordinate axis), while, in the non-smooth context, we need to look at all directions, that is, for all $\Delta \in \mathbb{R}^d$, we need that the directional derivative

$$\partial H(\theta,\Delta) = \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\big[H(\theta+\varepsilon\Delta) - H(\theta)\big],$$

is non-negative. That is, we need to go up in all directions. When H is differentiable at θ, then $\partial H(\theta,\Delta) = H'(\theta)^\top\Delta$, and the positivity for all Δ is equivalent to $H'(\theta) = 0$.

For $H(\theta) = F(\theta) + \lambda\|\theta\|_1$, we have:

$$\partial H(\theta,\Delta) \quad = \quad F'(\theta)^\top\Delta + \lambda \sum_{j,\ \theta_j\neq 0} \operatorname{sign}(\theta_j)\Delta_j + \lambda \sum_{j,\ \theta_j=0} |\Delta_j|.$$

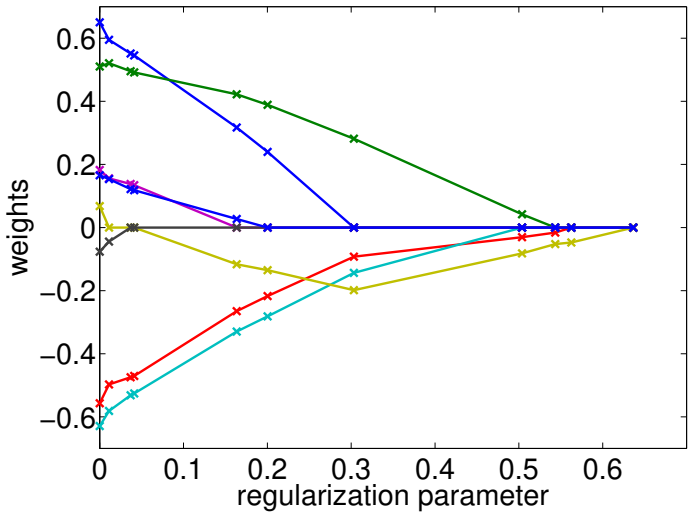
It is separable in Δ_j, $j = 1,\dots,d$, and it is non-negative for all j, if and only if, all components that depend on Δ_j are non-negative.

When $\theta_j \neq 0$, then this requires $F'(\theta)_j + \lambda\operatorname{sign}(\theta_j) = 0$, while when $\theta_j = 0$, then we need $F'(\theta)_j\Delta_j + \lambda|\Delta_j| \geqslant 0$ for all Δ_j, which is equivalent to $|F'(\theta)_j| \leqslant \lambda$. This leads to the set of conditions:

$$\begin{cases} F'(\theta)_j + \lambda\operatorname{sign}(\theta_j), & \forall j \in \{1,\dots,d\} \text{ such that } \theta_j \neq 0, \\ |F'(\theta)_j| \leqslant \lambda, & \forall j \in \{1,\dots,d\} \text{ such that } \theta_j = 0. \end{cases}$$

See [2] for more details.

Homotopy method (♦♦). We assume for simplicity that $\Phi^\top \Phi$ is invertible so that the minimizer $\theta(\lambda)$ is unique, Given a certain sign pattern for θ, optimality conditions are all convex in λ and thus define an interval in λ where the sign is constant. Given the sign, then the solution $\theta(\lambda)$ is affine in λ, leading to a piecewise affine function in λ (see an example below). This is leads to the regularization path below.



If we know the break points in λ and the associated signs, then we can compute all solutions for all λ. This is the source of the homotopy algorithm for Eq. (3), which starts with large λ and builds the path of solutions by computing break points one by one. See more details in [5].

3.2 Slow rates

We first consider analysis based on simple tools and with no assumptions on the design matrix Φ. We will see that we can deal with high-dimensional inference problems where d can be large, but it will be rates in $1/\sqrt{n}$ and not $1/n$, hence the denomination "slow".

We study the penalization by a general norm $\Omega : \mathbb{R}^d \to \mathbb{R}$ with dual norm Ω^*. We thus denote by $\hat{\theta}$ a minimizer of

$$\frac{1}{2n}\|y - \Phi\theta\|_2^2 + \lambda\Omega(\theta). \tag{4}$$

We first start by a lemma characterizing the excess risk in two situations: (a) where λ is large enough, and (b) in the general case.

Lemma 4 *Let $\hat{\theta}$ be a minimizer of Eq. (4).*

(a) If $\Omega^*(\Phi^\top \varepsilon) \leqslant \frac{n\lambda}{2}$, then we have $\Omega(\hat{\theta}) \leqslant 3\Omega(\theta_*)$ and $\frac{1}{n}\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant 3\lambda\Omega(\theta_*)$.

(b) In all cases, $\frac{1}{n}\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant \frac{8}{n}\|\varepsilon\|_2^2 + 4\lambda\Omega(\theta_*)$.

Proof We have, like in previous proofs, by optimality of $\hat{\theta}$ for Eq. (4):

$$\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant 2\varepsilon^\top \Phi(\hat{\theta} - \theta_*) + 2n\lambda\Omega(\theta_*) - 2n\lambda\Omega(\hat{\theta}).$$

Then, with the dual norm $\Omega^*(z) = \sup_{\Omega(\theta)\leqslant 1} z^\top \theta$, assuming that $\Omega^*(\Phi^\top \varepsilon) \leqslant \frac{n\lambda}{2}$, and using the triangle inequality:

$$\begin{aligned}\|\Phi(\hat\theta - \theta_*)\|_2^2 &\leqslant 2\Omega^*(\Phi^\top\varepsilon)\Omega(\hat\theta - \theta_*) + 2n\lambda\Omega(\theta_*) - 2n\lambda\Omega(\hat\theta)\\ &\leqslant n\lambda\Omega(\hat\theta - \theta_*) + 2n\lambda\Omega(\theta_*) - 2n\lambda\Omega(\hat\theta)\\ &\leqslant n\lambda\Omega(\hat\theta) + n\lambda\Omega(\theta_*) + 2n\lambda\Omega(\theta_*) - 2n\lambda\Omega(\hat\theta) \leqslant 3n\lambda\Omega(\theta_*) - n\lambda\Omega(\hat\theta).\end{aligned}$$

This implies that $\Omega(\hat\theta) \leqslant 3\Omega(\theta_*)$ and $\frac{1}{n}\|\Phi(\hat\theta - \theta_*)\|_2^2 \leqslant 3\lambda\Omega(\theta_*)$.

We also have a general bound through:

$$\|\Phi(\hat\theta - \theta_*)\|_2^2 \leqslant 2\|\varepsilon\|_2\|\Phi(\hat\theta - \theta_*)\|_2 + 2n\lambda\Omega(\theta_*),$$

which leads to[2], without any constraint on λ:

$$\|\Phi(\hat\theta - \theta_*)\|_2 \leqslant 2\|\varepsilon\|_2 + \sqrt{2n\lambda\Omega(\theta_*)},$$

which leads to the desired bound. ■

We can now use the lemma above to compute the excess risk of the Lasso, for which $\Omega^*(\Phi^\top\varepsilon) = \|\Phi^\top\varepsilon\|_\infty$. The key is to note that since $\|\Phi^\top\varepsilon\|_\infty$ is a maximum of $2d$ terms that scales as $\sqrt{n}$, it maximum scales as $\sqrt{n\log(d)}$ and we will apply the lemma above when λ is larger than $\sqrt{\frac{\log d}{n}}$.

Proposition 3 (Lasso - slow rate) *Assume $y = \Phi\theta_* + \varepsilon$, with $\varepsilon \in \mathbb{R}^n$ a vector of with independent Gaussian components of zero mean and variance σ^2. Let $\hat\theta$ be the minimizer of Eq. (3). Then, for $\lambda = 4\sigma\sqrt{\frac{\log(dn)}{n}}\sqrt{\|\widehat\Sigma\|_\infty}$, we have:*

$$\mathbb{E}\big[\frac{1}{n}\|\Phi(\hat\theta - \theta_*)\|_2^2\big] \leqslant 40\sigma\sqrt{\frac{\log(dn)}{n}}\sqrt{\|\widehat\Sigma\|_\infty}\|\theta_*\|_1 + \frac{6\sqrt{2}}{n}\sigma^2.$$

Proof For each j, the random variable $(\Phi^\top\varepsilon)_j$ is Gaussian with mean zero and variance $\sigma^2\widehat\Sigma_{jj}$. Thus, we get from the union bound and from the fact that for a standard Gaussian variable z, $\mathbb{P}(|z| \geqslant t) \leqslant 2\exp(-t^2/2)$:

$$\mathbb{P}\big(\|\Phi^\top\varepsilon\|_\infty > \frac{n\lambda}{2}\big) \leqslant \sum_{j=1}^d \mathbb{P}\big(|\Phi^\top\varepsilon|_j > \frac{n\lambda}{2}\big) \leqslant 2\sum_{j=1}^d \exp\big(-\frac{n\lambda^2}{8\widehat\Sigma_{jj}}\big) \leqslant 2d\exp\big(-\frac{n\lambda^2}{8\sigma^2\|\widehat\Sigma\|_\infty}\big) = \delta.$$

Thus, with probability greater than $1-\delta$, we can apply the first part of Lemma 4, and thus the error is less than $3\lambda\|\theta^*\|_1$. This would be the end of the proof if a high-probability result was desired. For a result in expectation, we need also the second part.

Overall, we get, denoting $\mathcal{A}$ the event $\mathcal{A} = \big\{\Omega^*(\Phi^\top\varepsilon) \leqslant \frac{n\lambda}{2}\big\}$, and the previous lemma:

$$\begin{aligned}\mathbb{E}\big[\|\Phi(\hat\theta - \theta_*)\|_2^2\big] &= \mathbb{E}\big[1_{\mathcal{A}}\|\Phi(\hat\theta - \theta_*)\|_2^2\big] + \mathbb{E}\big[1_{\mathcal{A}^c}\|\Phi(\hat\theta - \theta_*)\|_2^2\big]\\ &\leqslant 3n\lambda\|\theta_*\|_1 + \mathbb{P}(\mathcal{A}^c)^{1/2}\Big(\mathbb{E}\big[(2\|\varepsilon\|_2 + \sqrt{2n\lambda\|\theta_*\|_1})^4\big]\Big)^{1/2}\\ &\leqslant 3n\lambda\|\theta_*\|_1 + 2\mathbb{P}(\mathcal{A}^c)^{1/2}\Big(4\Big(\mathbb{E}\big[\|\varepsilon\|_2^4\big]\Big)^{1/2} + 2n\lambda\|\theta_*\|_1\Big).\end{aligned}$$

[2]Using the lemma: if $a \geqslant 0$ is such that $a^2 \leqslant ab + c$ for some $b, c \geqslant 0$, then $a \leqslant b + \sqrt{c}$. Note that we could also use the identity $2ab \leqslant 2a^2 + \frac{1}{2}b^2$.

With Gaussian noise, we have: $\sqrt{\mathbb{E}\big[\|\varepsilon\|_2^4\big]} \leqslant 3n\sigma^2$, leading to:

$$\frac{1}{n}\mathbb{E}\big[\|\Phi(\hat{\theta}-\theta_*)\|_2^2\big] \leqslant 3\lambda\|\theta_*\|_1 + 2\sqrt{2d}\exp\big(-\frac{n\lambda^2}{16\sigma^2\|\widehat{\Sigma}\|_\infty}\big)(3\sigma^2+2\lambda\|\theta_*\|_1).$$

With $\frac{n\lambda^2}{16\sigma^2\|\widehat{\Sigma}\|_\infty} = \log(dn)$, we get

$$\begin{aligned}\frac{1}{n}\mathbb{E}\big[\|\Phi(\hat{\theta}-\theta_*)\|_2^2\big] &\leqslant 3\lambda\|\theta_*\|_1 + \frac{2\sqrt{2}}{n}(3\sigma^2+2\lambda\|\theta_*\|_1) \\ &\leqslant 40\sigma\sqrt{\frac{\log(dn)}{n}}\sqrt{\|\widehat{\Sigma}\|_\infty}\|\theta_*\|_1 + \frac{6\sqrt{2}}{n}\sigma^2.\end{aligned}$$

Check homogeneity!

■

We can make the following observations:

- We already observe some high-dimensional phenomenon with the term $\sqrt{\frac{\log d}{n}}$, where n can be much larger than d (if of course we assume that the optimal predictor θ_* is sparse).
- **Exercise (♦)**: Using Rademacher complexities from Lecture 3, show a similar slow rate for ℓ_1-constrained optimization with Lipschitz-continuous losses.

3.3 Fast rates (♦)

We now consider conditions to obtain a fast rate with leading term proportional to $\sigma^2 k\log dn$, which is the same as for ℓ_0-penalty, but with tractable algorithms. This will come with extra (very) strong conditions on the design matrix Φ.

We start with a simple (but crucial) lemma, characterizing the solution of Eq. (3) in terms of the support A of θ_*.

Lemma 5 *Let $\hat{\theta}$ be a minimizer of Eq. (4). If $\Delta = \hat{\theta}-\theta_*$, then $\|\Delta_{A^c}\|_1 \leqslant 3\|\Delta_A\|_1$ and $\|\Phi\Delta\|_2^2 \leqslant 3n\lambda\|\Delta_A\|_1$.*

Proof We have, like in previous proofs, with $\Delta = \hat{\theta}-\theta_*$, and A the support of θ_*:

$$\|\Phi\Delta\|_2^2 \leqslant 2\varepsilon^\top\Phi\Delta + 2n\lambda\|\theta_*\|_1 - 2n\lambda\|\hat{\theta}\|_1.$$

Then, assuming that $\|\Phi^\top\varepsilon\|_\infty \leqslant \frac{n\lambda}{2}$,

$$\begin{aligned}\|\Phi\Delta\|_2^2 &\leqslant 2\|\Phi^\top\varepsilon\|_\infty\|\Delta\|_1 + 2n\lambda\|\theta_*\|_1 - 2n\lambda\|\hat{\theta}\|_1 \\ \|\Phi\Delta\|_2^2 &\leqslant n\lambda\|\Delta\|_1 + 2n\lambda\|\theta_*\|_1 - 2n\lambda\|\hat{\theta}\|_1.\end{aligned}$$

We then use, by using the decomposability of the ℓ_1-norm and the triangle inequality:

$$\|\theta_*\|_1 - \|\hat{\theta}\|_1 = \|(\theta_*)_A\|_1 - \|\theta_*+\Delta\|_1 = \|(\theta_*)_A\|_1 - \|(\theta_*+\Delta)_A\|_1 - \|\Delta_{A^c}\|_1 \leqslant \|\Delta_A\|_1 - \|\Delta_{A^c}\|_1,$$

to get

$$\begin{aligned}\|\Phi\Delta\|_2^2 &\leqslant n\lambda\|\Delta\|_1 + 2n\lambda(\|\theta_*\|_1 - \|\hat{\theta}\|_1) \leqslant n\lambda\|\Delta\|_1 + 2n\lambda(\|\Delta_A\|_1 - \|\Delta_{A^c}\|_1) \\ &\leqslant n\lambda(\|\Delta_A\|_1 + \|\Delta_{A^c}\|_1) + 2n\lambda(\|\Delta_A\|_1 - \|\Delta_{A^c}\|_1) = 3n\lambda\|\Delta_A\|_1 - n\lambda\|\Delta_{A^c}\|_1.\end{aligned}$$

This leads to $\|\Delta_{A^c}\|_1 \leqslant 3\|\Delta_A\|_1$ and the other desired inequality. ■

We can now add an extra assumption that will make the proof go through, namely

$$\frac{1}{n}\|\Phi\Delta\|_2^2 \geqslant \kappa\|\Delta_A\|_2^2 \tag{5}$$

for all Δ that satisfies the condition $\|\Delta_{A^c}\|_1 \leqslant 3\|\Delta_A\|_1$. This is called the "restrictive eigenvalue property", because is the smallest eigenvalue of $\frac{1}{n}\Phi^\top\Phi$ is great than κ, the condition is satisfied (but this is only possible if $n \geqslant d$). This leads to the following proposition.

Proposition 4 (Lasso - fast rate) *Assume $y = \Phi\theta_* + \varepsilon$, with $\varepsilon \in \mathbb{R}^n$ a vector with independent Gaussian components of zero mean and variance σ^2. Let $\hat{\theta}$ be the minimizer of Eq. (3). Then, for $\lambda = 4\sigma\sqrt{\frac{\log(dn)}{n}}\sqrt{\|\widehat{\Sigma}\|_\infty}$, we have, if Eq. (5) is satisfied:*

$$\mathbb{E}\big[\frac{1}{n}\|\Phi(\hat{\theta} - \theta_*)\|_2^2\big] \leqslant \frac{144|A|\sigma^2\|\widehat{\Sigma}\|_\infty}{\kappa}\frac{\log(dn)}{n} + \frac{6\sqrt{2}}{n}\sigma^2 + \frac{8}{n}\|\theta_*\|_1\sigma\sqrt{\frac{\log(dn)}{n}}\sqrt{\|\widehat{\Sigma}\|_\infty}.$$

Proof (♦) We have, when λ is large enough, and by application of Lemma 5, and using Eq. (5):

$$\|\Delta_A\|_1 \leqslant |A|^{1/2}\|\Delta_A\|_2 \leqslant \frac{|A|^{1/2}}{\sqrt{n\kappa}}\|\Phi\Delta\|_2 \leqslant \frac{|A|^{1/2}}{\sqrt{n\kappa}}\sqrt{3n\lambda\|\Delta_A\|_1},$$

which leads to $\|\Delta_A\|_1 \leqslant \frac{3|A|\lambda}{\kappa}$. We then get $\frac{1}{n}\|\Phi\Delta\|_2^2 \leqslant \frac{9|A|\lambda^2}{\kappa}$, and we can reuse the same reasoning as for the slow rate, to get

$$\begin{aligned}\mathbb{E}\big[\frac{1}{n}\|\Phi\Delta\|_2^2\big] &\leqslant \frac{9|A|\lambda^2}{\kappa} + \frac{2\sqrt{2}}{n}(3\sigma^2 + 2\lambda\|\theta_*\|_1) \\ &\leqslant \frac{144|A|\sigma^2\|\widehat{\Sigma}\|_\infty}{\kappa}\frac{\log(dn)}{n} + \frac{6\sqrt{2}}{n}\sigma^2 + \frac{8}{n}\|\theta_*\|_1\sigma\sqrt{\frac{\log(dn)}{n}}\sqrt{\|\widehat{\Sigma}\|_\infty}.\end{aligned}$$

■

The dominant part of the rate is proportional to $\sigma^2 k\frac{\log d}{n}$, which is a fast rate, but depends crucially on a very strong assumption.

3.4 Zoo of conditions (♦♦)

Conditions to obtain fast rates are plentyful: they all assume that there is low-correlation among predictors, which is rarely the case in practice (in particular, if there is two features which are equal, they are never satisfied).

Restricted eigenvalue property (REP). The most direct condition is the so-called restricted eigenvalue property (REP), which is exactly Eq. (5), with the supremum taken over the unknown set A of cardinality less than k:

$$\inf_{|A|\leqslant k} \inf_{\|\Delta_{A^c}\|_1 \leqslant 3\|\Delta_A\|_1} \frac{\|\Phi\Delta\|_2^2}{n\|\Delta_A\|_2^2} \geqslant \kappa > 0.$$

Mutual incoherence condition. A simpler one to check, but weaker, is the mutual incoherence condition:

$$\sup_{i\neq j} |\widehat{\Sigma}_{ij}| \leqslant \frac{\min_{j\in\{1,\dots,d\}} \widehat{\Sigma}_{jj}}{14k}, \tag{6}$$

which states that all cross-correlation coefficient are small (pure decorrelation would set them to zero).

This is weaker than the REP condition above. Indeed, by expanding, we have:

$$\|\Phi\Delta\|_2^2 = \|\Phi_A\Delta_A + \Phi_{A^c}\Delta_{A^c}\|_2^2 = \|\Phi_A\Delta_A\|_2^2 + 2\Delta_A^\top\Phi_A^\top\Phi_{A^c}\Delta_{A^c} + \|\Phi_{A^c}\Delta_{A^c}\|_2^2 \geqslant \|\Phi_A\Delta_A\|_2^2 + 2\Delta_A^\top\Phi_A^\top\Phi_{A^c}\Delta_{A^c}.$$

Moreover, we have:

$$\Delta_A^\top\widehat{\Sigma}_{AA}\Delta_A = \Delta_A^\top \operatorname{Diag}(\operatorname{diag}(\widehat{\Sigma}_{AA}))\Delta_A + \Delta_A^\top(\widehat{\Sigma}_{AA} - \operatorname{Diag}(\operatorname{diag}(\widehat{\Sigma}_{AA}))\Delta_A \geqslant \min_{j\in\{1,\dots,d\}} \widehat{\Sigma}_{jj}\big(\|\Delta_A\|_2^2 - \frac{1}{14k}\|\Delta_A\|_1^2\big),$$

and

$$|\Delta_A^\top\Phi_A^\top\Phi_{A^c}\Delta_{A^c}| \leqslant \frac{\min_{j\in\{1,\dots,d\}} \widehat{\Sigma}_{jj}}{14k}\|\Delta_{A^c}\|_1\|\Delta_A\|_1 \leqslant \frac{3\min_{j\in\{1,\dots,d\}} \widehat{\Sigma}_{jj}}{14k}\|\Delta_A\|_1^2.$$

This leads to $\frac{1}{n}\|\Phi\Delta\|_2^2 \geqslant \min_{j\in\{1,\dots,d\}} \widehat{\Sigma}_{jj}\Big(\|\Delta_A\|_2^2 - \frac{7}{14k}\|\Delta_A\|_1^2\Big) \geqslant \min_{j\in\{1,\dots,d\}} \widehat{\Sigma}_{jj}\Big(\|\Delta_A\|_2^2 - \frac{7k}{14k}\|\Delta_A\|_2^2\Big)$, thus leading to $\kappa = \min_{j\in\{1,\dots,d\}} \widehat{\Sigma}_{jj}/2$ for the REP condition.

Restricted isometry property. One of the earlier conditions was the restricted isometry property: all eigenvalues of submatrices of $\widehat{\Sigma}$ of size less than $2k$, are between $1-\delta$ and $1+\delta$ for δ small enough. See [2, 6] for details.

Gaussian designs (♦). It is not obvious that the conditions above are non-trivial (that is, there may exist no matrix with good sizes d and n for k large enough). In order for our results to be non-trivial, we need that $k\frac{\log d}{n}$ is small but not too small. We show in this paragraph that when sampling from Gaussian distributions, then assumptions above are satisfied. This is a first step towards a random design assumption.

Theorem 1 ([6], Theorem 7.16) *If sampling $\varphi(x)$ from a Gaussian with mean zero and covariance matrix Σ, then with probability greater than $1 - \frac{e^{-n/32}}{1-e^{-n/32}}$, the REP property is satisfied with $\kappa = \frac{c_1}{2}\lambda_{\min}(\Sigma)$ as soon as $k\frac{\log d}{n} \leqslant \frac{c_1}{8c_2}\frac{\lambda_{\min}(\Sigma)}{\|\Sigma\|_\infty}$, with $c_1 = 1/8$ and $c_2 = 50$.*

The theorem above is hard to prove, the following exercise proposes to prove a weaker result, showing that the guarantees for the maximal cardinality k of the support has to be smaller.

- **Exercise (♦♦♦)**: If sampling $\varphi(x)$ from a Gaussian with mean zero and covariance matrix identity, then with large probability, for n greater than a constant times $k^2 \frac{\log d}{n}$, then mutual incoherence property in Eq. (6) is satisfied.

Model selection and irrepresentable condition (♦). Given that the Lasso aims at performing variable selection, it is natural to study its capacity to find the support of θ_*, that is, the set of non-zero variables. It turns out that it also depends on some conditions on the design matrix, which are stronger than the REP conditions, and called the " irrepresentable condition", and also valid for Gaussian random matrices with similar scalings between n, d and k. See [2, 6] for details.

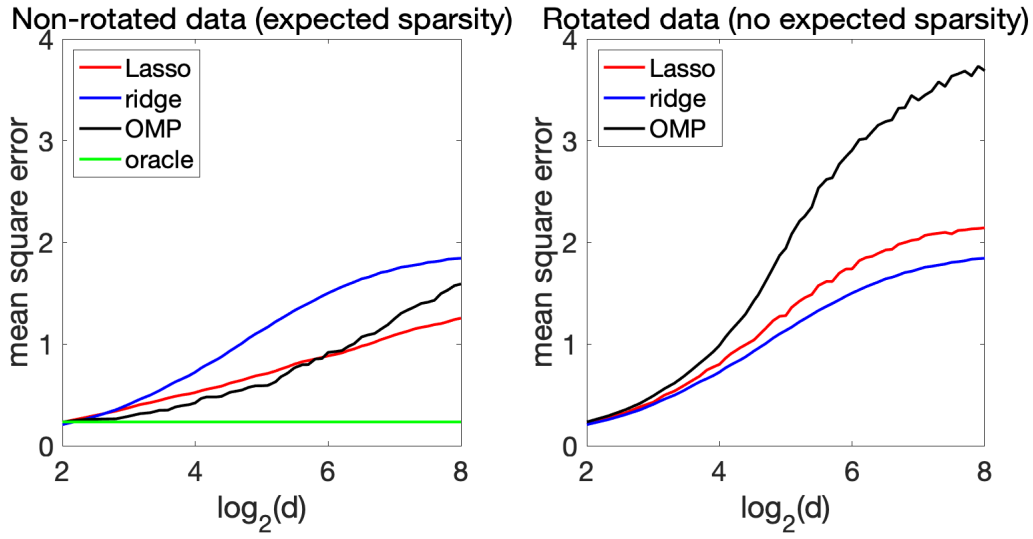
Algorithmic and theoretical tools are similar to compressed sensing, where the design matrix represents a set of measurements, which can be chosen by the user/theoretician. In this context, sampling from i.i.d. Gaussians make sense. For machine learning and statistics, the design matrix is the data, and comes as it is, often with strong correlations.

4 Experiments

In this section, we perform a simple experiment on Gaussian design matrices, where all entries in $\Phi \in \mathbb{R}^{n \times d}$ are sampled independently from a standard Gaussian distribution. Then θ_* is taken to be zero except on k components where it is randomly equal to -1 or 1. We consider $\sigma = \sqrt{k}$ (to have a signal to noise ratio that remains constant when k varies). We perform 128 replications. For each method and each value of its hyperparameter, we averaged the test risk over the 128 replications and report the minimum value (with respect to the hyperparameter). We compare the following three methods:

- Ridge regression: penalty by $\lambda \|\theta\|_2^2$.
- Lasso regression: penalty by $\lambda \|\theta\|_1$.
- Orthogonal matching pursuit (greedy forward method), with hyperparameter k (the number of included variables).

We compare two situations: (1) non-rotated data (exactly the model above), and (2) rotated data, where we replace Φ by ΦR and θ_* by $R^\top \theta_*$, where R is a rotation matrix. For the rotated data, we do not expect sparse solutions, and hence sparse methods are not expected to work better than ridge regression (and OMP performs significantly because once the support is chosen, there is no regularization). Note that the two curves for ridge regression are exactly the same (as expected from rotation invariance of the ℓ_2-norm). The oracle performance corresponds to the estimator where the true support is given.



Sparse methods make assumptions regarding the best predictor. Like all assumptions, when this assumed prior knowledge is not correct, the method does not perform better.

5 Extensions

Sparse methods are more general than the ℓ_1-norm, and can be extended in a number of ways:

- Group penalties: in many cases, $\{1, \dots, d\}$ is partitioned in to m subsets $A_1, \dots, A_m$, and the goal is to consider "group sparsity", that is, if we select one variable within a group A_j, the entire group should be selected. Such behavior can be obtained using the penalty $\sum_{i=1}^m \|\theta_{A_i}\|_2$ or $\sum_{i=1}^m \|\theta_{A_i}\|_\infty$. See, e.g., [2] for details.
- Structured sparsity: it is also possible to favor other specific patterns for the selected variables, such as blocks, trees, etc. See [7] for details.
- Nuclear norm: when learning on matrices, a natural form of sparsity is for a matrix to have low rank. This can be achieved by penalizing by the sum of singular values of a matrix, which is a norm called the nuclear norm or the trace norm. See [8] and references therein.
- Multiple kernel learning: the group penalty can be extended when the groups have an infinite dimension and ℓ_2-norms are replaced by RKHS norms defined in Lecture 6. This becomes a tool to learn the kernel matrix from data. See [9] for details.
- Elastic net: often, when both effects of the ℓ_1-norm (sparsity) and of the squared ℓ_2-norm (strong-convexity) are desired, we can sum the two, which is referred to as the "elastic net" penalty.
- Concave penalization and debiasing: in order to obtain a sparsity-inducing effect, the penalty in the ℓ_1-norm has to be quite large, such as in $1/\sqrt{n}$, which often creates a strong bias in the estimation once the support is selected. There are several ways on debiasing the Lasso, an elegant one being to

use a "concave" penalty. That is, we use $\sum_{i=1}^d a(|\theta_i|)$ where a is a concave increasing function on $\mathbb{R}^+$ such as $a(u) = u^\alpha$ for $\alpha \in (0,1)$. This leads to a non-convex optimization problem, where iterative weighted ℓ_1-minimization provides natural algorithms (see [10] and references therein).

Acknowledgements

These class notes have been adapted from the notes of many colleagues I have the pleasure to work with, in particular Lénaïc Chizat, Pierre Gaillard, Alessandro Rudi and Simon Lacoste-Julien. The notes from Philippe Rigollet have also been a very precious help.

References

[1] Ph. Rigollet and Alexander B. Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16(3):260–280, 2007.

[2] Christophe Giraud. *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, 2014.

[3] Christophe Giraud, Sylvie Huet, and Nicolas Verzelen. High-dimensional regression with unknown variance. *Statistical Science*, 27(4):500–518, 2012.

[4] Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

[5] Michael R. Osborne, Brett Presnell, and Berwin A. Turlach. On the Lasso and its dual. *Journal of Computational and Graphical statistics*, 9(2):319–337, 2000.

[6] Martin J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[7] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012.

[8] Francis Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9(Jun):1019–1048, 2008.

[9] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

[10] Julien Mairal, Francis Bach, Jean Ponce, et al. Sparse modeling for image and vision processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014.