Chapter 8

Sparse methods

Chapter summary

- Model selection can be performed by adding a specific "sparsity-inducing" penalty on top of the empirical risk.
- ℓ_0-penalty: For fixed design linear regression, if the optimal predictor has k non-zeros, then we can replace the rate $\frac{\sigma^2 d}{n}$ by $\frac{\sigma^2 k \log d}{n}$ with an ℓ_0-penalty on the square loss (which is computationally hard).
- ℓ_1-penalty: With few assumptions, we can get a slow rate proportional to $k\sqrt{\frac{\log d}{n}}$ with an ℓ_1-penalty and efficient algorithms, while fast rates require strong assumptions on the design matrix in the fixed design setting. In the random design setting, fast rates can be obtained with invertible population covariance matrices.

8.1 Introduction

In previous chapters, 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 the set $\mathcal{X} = \mathbb{R}^d$, we saw that the excess risk for k-nearest-neigbors, Nadaraya-Watson estimation (Chapter 6), or positive kernel methods (Chapter 7), 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 for unregularized least-squares was scaling as d/n.

In these two situations, efficient learning is generally impossible when d is too large (of course, much larger in the linear case).

To improve upon these rates, we study two techniques in this book. The first one is regularization, e.g., by the ℓ_2-norm, that allows obtaining dimension-independent bounds that cannot improve over the bounds above in the worst-case but are typically adaptive to additional regularity (see Chapter 3 and Chapter 7).

In this chapter, we consider 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.

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

- The original set of features is already large (for example, in text or web data).
- Given some input $x \in \mathcal{X}$, a large-dimensional feature vector $\varphi(x)$ is built where features are added that could potentially help predict the response, but from which we expect only a small number to be relevant.

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

Linear variable selection. In this chapter, we focus on *linear* methods, where we assume that we have a feature vector $\varphi(x) \in \mathbb{R}^d$, and we aim to minimize

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

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. See extensions in Section 8.5.

Non-linear variable selection corresponds to selecting a subset of variables from the d available features $\varphi(x)_1, \dots, \varphi(x)_d$, but with a potentially non-linear model on top of them. This is considered in the context of neural networks in Chapter 9.

Main focus on least-squares. These two types of penalties can be applied to all losses, but in this chapter, for simplicity, we will primarily consider the square loss and, in most cases, the fixed design setting (see a thorough description of this setting in Section 3.5), 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 from x_i, and sometimes stronger regularity, 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 Chapter 3 that for the ordinary

least-squares estimator, the expectation of 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 chapter, 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.

The results presented in this section extend beyond the square loss, e.g., to the logistic loss, in a straightforward way for slow rates in $1/\sqrt{n}$ (see the end of Section 8.3.3), with significant additional work for fast rates in $O(1/n)$ (see the end of Section 8.3.4).

8.1.1 Dedicated proof technique for constrained least-squares

In this chapter, we consider a more refined proof technique[1] that can extend to constrained versions of least-squares (while our technique in Chapter 3 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 $\theta \in \Theta$, for some subset Θ of $\mathbb{R}^d$. If $\theta_* \in \Theta$, then we have, by optimality of $\hat{\theta}$:

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

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 $\hat{\theta}$ also appears on the right side of the equation. Like done for upper-bounding estimation errors in Chapter 4, we can maximize with respect to $\theta \in \Theta$ to get rid of this randomness, which leads to

$$\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant 2\|\Phi(\hat{\theta} - \theta_*)\|_2 \cdot \sup_{\theta \in \Theta} \varepsilon^\top \Big(\frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2} \Big), \tag{8.1}$$

where $\hat{\theta}$ has disappeared from the right-hand side. Finally, isolating $\|\Phi(\hat{\theta} - \theta_*)\|_2^2$, we get:

$$\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant 4 \sup_{\theta \in \Theta} \Big[\varepsilon^\top \Big(\frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2} \Big) \Big]^2. \tag{8.2}$$

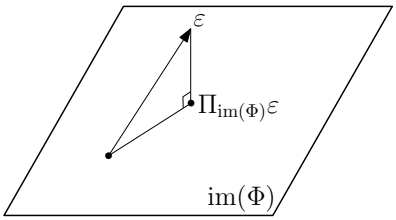
[1] Taken from Philippe Rigollet's lecture notes, see https://math.mit.edu/~rigollet/. See also Rigollet and Tsybakov (2007) for an example of application.

This inequality is true almost surely, and we can take expectation (with respect to ε) to obtain bounds. Therefore, in this chapter, 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 $\Pi_\Phi = \Pi_{\mathrm{im}(\Phi)}$ the orthogonal projector on the image space $\mathrm{im}(\Phi)$ (which has dimension $\mathrm{rank}(\Phi)$):

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

By a simple geometric argument (see below),



we have

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

leading to

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

We thus get a bound on the excess risk equal to $4\sigma^2 d/n$, which is (because of the constant 4) slightly worse than the direct computation from Chapter 3 (Prop. 3.5) but allows extensions to more complex situations.

This reasoning also allows getting high-probability bounds by adding assumptions on the noise ε. Finally, this also extends to penalized problems (see Section 8.2.2).

8.1.2 Probabilistic and combinatorial lemmas

In the proof technique above, we will need to bound expectations of maxima of squared norms of Gaussians, which we now consider. We start with two probabilistic lemmas.

Lemma 8.1 *If $z \in \mathbb{R}^n$ has a Gaussian distribution 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 $s < 1/2$ (using independence among the components of z):

$$\begin{aligned}\mathbb{E}\big[e^{s\|z\|_2^2}\big] &= \mathbb{E}\big[e^{s\sum_{i=1}^n z_i^2}\big] = \prod_{i=1}^n \mathbb{E}\big[e^{sz_i^2}\big] = \frac{1}{(2\pi)^{n/2}}\prod_{i=1}^n \int_{-\infty}^{\infty} e^{(s-\frac12)z_i^2} dz_i \\ &= \frac{1}{(2\pi)^{n/2}}\prod_{i=1}^n \sqrt{2\pi}(1-2s)^{-1/2} = (1-2s)^{-n/2}.\end{aligned}$$

■

Lemma 8.2 *Let $u_1, \dots, u_m$ be m random variables which are potentially dependent, and $s > 0$. Then, $\mathbb{E}\big[\max\{u_1, \dots, u_m\}\big] \leqslant \frac{1}{s} \log\big(\sum_{i=1}^m \mathbb{E}\big[e^{s u_i}\big]\big)$.*

Proof Following the reasoning from Section 1.2.4 in Chapter 1, for any $s \in \mathbb{R}$,

$$\mathbb{E}\big[\max\{u_1, \dots, u_m\}\big] \leqslant \frac{1}{s} \log\big(\mathbb{E}\big[e^{s \max\{u_1, \dots, u_m\}}\big]\big) = \frac{1}{s} \log\big(\mathbb{E}\big[\max\{e^{s u_1}, \dots, e^{s u_m}\}\big]\big),$$

which is thus less than $\frac{1}{s} \log\big(\sum_{i=1}^m \mathbb{E}\big[e^{s u_i}\big]\big)$. ■

The previous two lemmas can be combined to upper-bound the expectation of squared norms of Gaussian random variables: if $z_1, \dots, z_m \in \mathbb{R}^n$ are centered (that is, zero-mean) Gaussian random vectors which are potentially dependent, but for which the covariance matrix of z_i has eigenvalues less than σ^2, we can first use the rotational invariance of Gaussian densities, to assume without loss of generality that the Gaussians have diagonal covariance matrices with components $\sigma_{ij}^2 \leqslant \sigma^2$ (for $i \in \{1, \dots, m\}$ and $j \in \{1, \dots, n\}$), then, we have from Lemma 8.1,

$$\mathbb{E}[e^{s\|z_i\|_2^2}] = \prod_{j=1}^n \mathbb{E}[e^{s z_{ij}^2}] \leqslant \prod_{j=1}^n (1 - 2\sigma_{ij}^2 s)^{-1/2} \leqslant (1 - 2\sigma^2 s)^{-n/2}.$$

Thus, for $s = \frac{1}{4\sigma^2}$, $\mathbb{E}[e^{s\|z_i\|_2^2}] \leqslant 2^{n/2}$ for all $i \in \{1, \dots, m\}$, and from Lemma 8.2,

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

which is to be compared to the expectation of each argument of the max, which is less than $\sigma^2 n$. We pay an additive factor proportional to $\sigma^2 \log(m)$. This will be applied to $m \propto d^k$, leading to the additional 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 8.3 *Let $d > 0$ and $k \in \{1, \dots, d\}$. 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

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

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

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

Exercise 8.1 (Concentration of chi-squared variables) *We consider n independent standard normal variables $z_1, \dots, z_n$ and the variables $y = z_1^2 + \cdots + z_n^2$. Using Lemma 8.1, show that for any $\varepsilon > 0$, $\mathbb{P}(y \geqslant n(1 + \varepsilon)) \leqslant \big(\frac{1+\varepsilon}{\exp(\varepsilon)}\big)^{n/2}$, and for any $\varepsilon \in (0, 1)$, $\mathbb{P}(y \leqslant n(1 - \varepsilon)) \leqslant \big(\frac{1-\varepsilon}{\exp(-\varepsilon)}\big)^{n/2}$.*

8.2 Variable selection by the ℓ_0-penalty

In this section, we assume that the target vector 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 $|A| = k$.

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 call the "oracle" in Section 8.4). 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 $\log\binom{d}{k} \leqslant \log d$ due to the potentially large number of models with k variables.

8.2.1 Assuming k is known

We 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 8.1 (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, the (fixed design) excess risk is upper-bounded as:*

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

Proof For any θ such that $\|\theta\|_0 \leqslant k$, we have $\|\theta - \theta_*\|_0 \leqslant 2k$. Thus, we have, using the bounding technique from Section 8.1.1:

$$\begin{aligned}
\|\Phi(\hat\theta - \theta_*)\|_2^2 &\leqslant 4 \sup_{\theta \in \mathbb{R}^d, \|\theta\|_0 \leqslant k} \Big[\varepsilon^\top \Big(\frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2}\Big)\Big]^2 \text{ from Eq. (8.2)}, \\
&\leqslant 4 \sup_{\theta \in \mathbb{R}^d, \|\theta - \theta_*\|_0 \leqslant 2k} \Big[\varepsilon^\top \Big(\frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2}\Big)\Big]^2 \text{ from the discussion above}, \\
&= 4 \sup_{B \subset \{1,\dots,d\},\ |B| \leqslant 2k} \ \sup_{\mathrm{supp}(\theta - \theta_*) \subset B} \Big[\varepsilon^\top \Big(\frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2}\Big)\Big]^2
\end{aligned}$$

by separating by supports. Thus, using the same argument as in Section 8.1.1,

$$\begin{aligned}
\|\Phi(\hat\theta - \theta_*)\|_2^2 &\leqslant 4 \sup_{B \subset \{1,\dots,d\},\ |B| \leqslant 2k} \ \sup_{z \in \mathrm{im}(\Phi_B), \|z\|_2 = 1} \big[\varepsilon^\top z\big]^2 \\
&\leqslant 4 \sup_{B \subset \{1,\dots,d\},\ |B| \leqslant 2k} \|\Pi_{\Phi_B}\varepsilon\|^2 \leqslant 4 \sup_{B \subset \{1,\dots,d\},\ |B| = 2k} \|\Pi_{\Phi_B}\varepsilon\|^2,
\end{aligned}$$

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 8.3), we should expect, with concentration inequalities from Section 8.1.2, that we pay a price of $\log\left[\left(\frac{ed}{2k}\right)^{2k}\right] \approx k\log\frac{d}{k}$. 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 $s\sigma^2 < 1/2$, from Lemma 8.1:

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

Therefore, with $s = 1/(4\sigma^2)$, for which $(1-2\sigma^2 s)^{-k} = 2^k$, we get, from Lemma 8.2:

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

This leads to the desired result. ■

We can make the following observations:

- The term $k\log(d/k)$ comes from the logarithm of the number m of subsets of $\{1,\dots,d\}$ of size $2k$, which is a result of the expectation of the maximum of m squared norms of Gaussians.
- The assumption that $k < d/2$ is not a real issue, as when $k \geqslant d/2$, then the classical bound $\sigma^2 d/n$ is of the same order as $\sigma^2 k\log(d/k)/n$.
- 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 minimization of the empirical risk is only done approximately.
- This result is not improvable by any algorithm (polynomial time or not), see, e.g., Giraud (2014, Theorem 2.3) and Chapter 15.

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 is a problem when k gets large. There are, however, two simple algorithms that only come with guarantees when such fast rates are available for ℓ_1-regularization (see Section 8.3.4, and Zhang, 2009).

- **Greedy algorithm:** Starting from the empty set, variables are added one by one, maximizing the resulting cost reduction. This is often referred to as orthogonal matching pursuit (Pati et al., 1993).
- **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 Chapter 5):

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

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. This is done (proof left as an exercise) by computing the unconstrained minimizer $\theta_{t-1} + \frac{1}{L}\frac{1}{n}\Phi^\top(y - \Phi\theta_{t-1})$, and selecting the k largest components.

8.2.2 Estimating k (♦)

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

$$\frac{1}{n}\|y - \Phi\theta\|_2^2 + \lambda\|\theta\|_0. \tag{8.3}$$

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

Proposition 8.2 (Model selection - ℓ_0-penalty) *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 $\|\theta_*\|_0 \leqslant k$. Let $\hat{\theta}$ be a minimizer of Eq. (8.3). Then, for $\lambda = \frac{8\sigma^2}{n}\log(2d)$, we have:*

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

Proof We follow the same proof technique than in Section 8.1.1, 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 $2ab \leqslant 2a^2 + \frac{1}{2}b^2$, and the same arguments that led to Eq. (8.1):

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

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 $\sup_{\theta\in\mathbb{R}^d} = \sup_{k'\in\{1,\dots,d\}} \sup_{|B|=k'} \sup_{\mathrm{supp}(\theta)\subset B}$, that is, using the same derivations as for Prop. 8.1 (A is the support of θ_*):

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

We thus get with the same reasoning as in Section 8.2.1 (based on the probabilistic lemmas from Section 8.1.2), using $s=\frac{1}{4\sigma^2}$ within Lemma 8.2:

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

To find a good regularization parameter, we can then approximately minimize the bound above with respect to λ. We obtain a good balance of the two terms by having $-\log d=\log(2)-\frac{n\lambda}{8\sigma^2}$, that is, $\lambda=\frac{8\sigma^2}{n}\log(2d)$, for which we get:

$$\mathbb{E}\big[\|\Phi(\hat\theta-\theta_*)\|_2^2\big]\ \leqslant\ (2n\lambda+16\log(2)\sigma^2)\|\theta_*\|_0+16\sigma^2\leqslant 16\sigma^2\big((\log(d)+2)\|\theta_*\|_0+1\big),$$

and get the desired result. ■

We can make the following observations:

- Penalties on the number of parameters on top of the empirical risk can be obtained from various perspectives, for square loss depending on whether the noise variance is known or more generally for other losses. For example, the Bayesian information criterion (BIC) penalizes by penalty proportional to $\|\theta\|_0\log n$ (often a smaller penalty than proposed here).
- Note that we need to know σ^2 in advance to compute λ, which can be a problem in practice. See Giraud et al. (2012) 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) we observe 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).

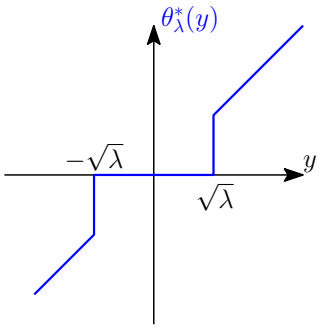
Exercise 8.2 *(♦) With a penalty proportional to $\|\theta\|_0\log\frac{d}{\|\theta\|_0}$, show the same bound than for k known.*

Algorithms. We can extend the two algorithms from Section 8.2.1 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. See an analysis by Zhang (2011).
- **Iterative hard-thresholding:** compared to the constrained case, we minimize

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

with $L = \lambda_{\max}(\frac{1}{n}\Phi^\top\Phi)$, which can also be computed in closed form (by iterative hard thresholding). That is, with $\theta_t = \theta_{t-1} + \frac{1}{nL}\Phi^\top(y - \Phi\theta_{t-1})$, all components $(\theta_t)_j$ such that $|(\theta_t)_j|^2 \geqslant \frac{\lambda}{L}$, are left unchanged and all others are set to zero. Indeed, for one-dimensional problems, the minimizer of $|\theta - y|^2 + \lambda 1_{\theta\neq 0}$ is $\theta_\lambda^*(y) = 0$ if $|y|^2 \leqslant \lambda$ and $\theta_\lambda^*(y) = y$ otherwise (see below).



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. See an analysis by Blumensath and Davies (2009).

8.3 Variable selection by 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:

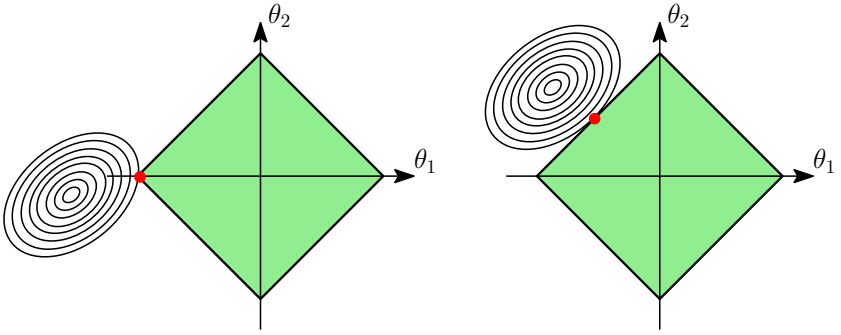
$$\frac{1}{2n}\|y - \Phi\theta\|_2^2 + \lambda\|\theta\|_1. \tag{8.4}$$

This is a convex optimization problem on which algorithms from Chapter 5 can be applied (see instances below). It is often called the "Lasso" problem, for "least absolute shrinkage and selection operator" (Tibshirani, 1996).

We first present algorithms dedicated to solving the optimization problem, then present "slow rate" analyses leading to excess risks in $O(1/\sqrt{n})$, first in the random design case with Lipschitz-continuous losses, and then in the fixed design case with the square loss. We then present "fast rate" analyses leading to excess risks in $O(1/n)$.

8.3.1 Intuition and algorithms

Sparsity-inducing effect. Unlike the squared 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 occurs 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 it.



As shown above, where we represent the level set of a potential loss function, the solution of minimizing the loss subject to the ℓ_1-constraint (in green) is obtained when level sets are "tangent" to the constraint set. In the right part, this is obtained at 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", that is, minimizers tend to be precisely at these corners, and this exactly leads to sparse solutions.

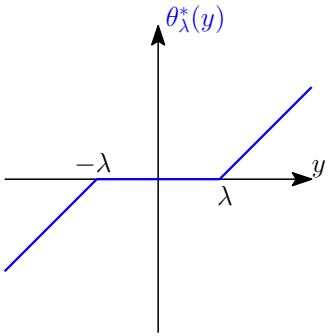
The ℓ_1-norm is also often introduced as the "convex relaxation" of the ℓ_0-penalty. Indeed, the ℓ_1-norm is the convex envelope (the largest convex function which is a lower-bound) of the ℓ_0-penalty on the set $[-1,1]^d$ (proof left as an exercise). While this provides some intuition about the ℓ_1-norm and its potential generalization to other sparse situations, this does not directly justify its good behavior on sparse problems.

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 $\theta^*_\lambda(y) = \max\{|y| - \lambda, 0\}\,\mathrm{sign}(y)$, which is depicted below. This is 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 toward zero.

Optimization algorithms. We can adapt algorithms from Chapter 5 to the problem in Eq. (8.4).

- **Iterative soft-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 the gradient is $F'(\theta) = -\frac{1}{n}\Phi^\top(y - \Phi\theta)$. 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 $(\theta_t)_j = \max\{|(\eta_t)_j| - \lambda, 0\}\,\mathrm{sign}((\eta_t)_j)$, 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 the invertibility of $\frac{1}{n}\Phi^\top\Phi$ (if invertible, we get an exponential convergence rate in t, with only $O(1/t)$ otherwise).

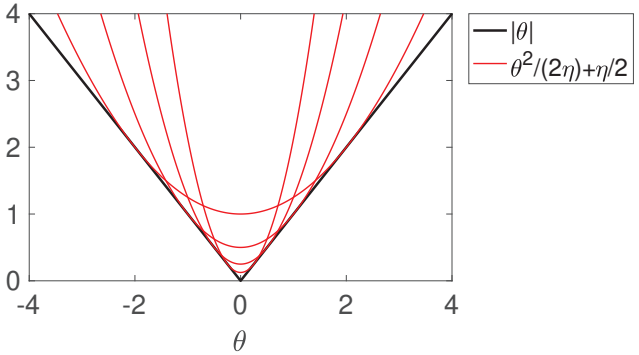
- **Coordinate descent:** Although the ℓ_1-norm is a non-differentiable function, coordinate descent can be applied (because the ℓ_1-norm is "separable"). At each iteration, we select a coordinate to update (at random or by cycling) and optimize with respect to this coordinate, which is a one-dimensional problem that can be solved in closed form. The convergence properties are similar to proximal methods (Fercoq and Richtárik, 2015).

  Exercise 8.3 *Provide a closed-form expression for the iteration of the coordinate descent algorithm described above.*

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 an example in one dimension, with $|\theta|$ and several quadratic upper bounds.



This leads to the reformulation of Eq. (8.4) as

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

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.[2]

Optimality conditions (♦). To study the estimator defined by Eq. (8.4), 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 require 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:

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

[2]See more details in https://www.di.ens.fr/~fbach/ltfp/etatrick.html and by Bach et al. (2012a, Section 5).

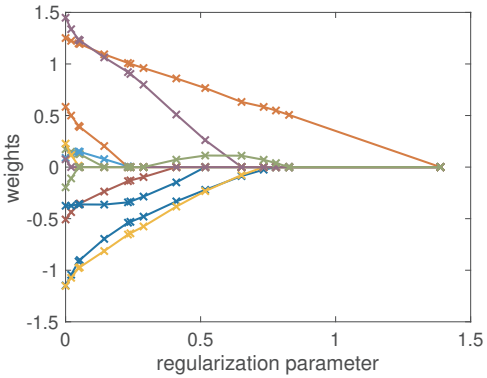


Figure 8.1: Regularization path for a Lasso problem in dimension $d = 32$ and $n = 32$ input observations sampled from a standard Gaussian distribution with 4 non-zero weights equal to -1 or $+1$, and outputs generated with additive Gaussian noise with unit variance. The random seed was chosen so that at least one weight comes in and out in the regularization path.

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:

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

See Giraud (2014) for more details. Note that we could have also used subgradients to derive these optimality conditions (derivations left as an exercise).

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 of a regularization path in Figure 8.1).

If we know the breakpoints in λ and the associated signs, we can compute all solutions for all λ. This is the source of the homotopy algorithm for Eq. (8.4), which starts with large λ and builds the path of solutions by computing break points one by one. See more details by Osborne et al. (2000); Mairal and Yu (2012).

8.3.2 Slow rates - random design

In this section, we consider Lipschitz-continuous loss functions and, thus, an empirical risk of the form

$$\widehat{\mathcal{R}}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \varphi(x_i)^\top \theta),$$

with ℓ having Lipschitz constant G with respect to the second variable. We assume that the expected risk $\mathcal{R}(\theta) = \mathbb{E}[\ell(y, \varphi(x)^\top\theta)]$ is minimized at a certain $\theta_* \in \mathbb{R}^d$ and, for simplicity, we consider the estimator $\hat{\theta}_D$ obtained by minimizing $\widehat{\mathcal{R}}(\theta)$ with the constraint that $\|\theta\|_1 \leqslant D$, where we will use tools from Section 4.5.4 (we could also consider penalized formulation using Section 4.5.5). We assume that $\|\varphi(x)\|_\infty \leqslant R$ almost surely.

From Section 4.5.4, we get that,

$$\mathbb{E}\big[\mathcal{R}(\hat{\theta}_D)\big] \leqslant \inf_{\|\theta\|_1 \leqslant D} \mathcal{R}(\theta) + 4G\mathrm{R}_n(\mathcal{F}_D),$$

where $\mathrm{R}_n(\mathcal{F}_D)$ is the Rademacher complexity of the set of linear predictors with weight vectors bounded by D in ℓ_1-norm, which we can compute as:

$$\mathrm{R}_n(\mathcal{F}_D) = \mathbb{E}\Big[\sup_{\|\theta\|_1 \leqslant D} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^{\mathrm{R}} \varphi(x_i)^\top\theta\Big] = D\mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^{\mathrm{R}}\varphi(x_i)\Big\|_\infty\Big],$$

where $\varepsilon_i^{\mathrm{R}} \in \{-1, 1\}$ are Rademacher random variables. We can now compute a bound on the expectation, first conditioned on the data. Indeed, $\varepsilon_i^{\mathrm{R}}\varphi(x_i)$ has conditional zero mean and is bounded in absolute value by R. It is thus sub-Gaussian with constant R (see Section 1.2.1, which implies that $\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^{\mathrm{R}}\varphi(x_i)$ is sub-Gaussian with constant R^2/n^2). We can then use Prop. 1.4, to get that the maximum of the $2d$ sub-Gaussian variables is less than $\big(2R^2\log(2d)/n^2\big)^{1/2}$. This leads to:

$$\mathbb{E}\big[\mathcal{R}(\hat{\theta}_D)\big] \leqslant \inf_{\|\theta\|_1 \leqslant D} \mathcal{R}(\theta) + \frac{GRD\sqrt{\log(2d)}}{\sqrt{n}}.$$

When D is large enough, e.g., $D = \|\theta_*\|_1$, then we get an excess risk bounded by $GRD\sqrt{\log(2d)}/\sqrt{n}$. If θ_* has only k non-zero, its ℓ_1-norm will typically grow as $O(k)$, and we see a high-dimensional phenomenon with a bound proportional to $k\sqrt{\log d}/\sqrt{n}$, where d can be much larger than n, as long as $k^2\log(d)/n$ is small. This is a "slow" rate because of the dependence in n, which is $O(1/\sqrt{n})$ rather than in $O(1/n)$.

8.3.3 Slow rates - fixed design (square loss)

We now look at the fixed design setting with the square loss. We first consider an analysis based on simple tools 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 with 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^* defined as $\Omega^*(z) = \sup_{\Omega(\theta)\leqslant 1} z^\top \theta$ (see Exercise 8.4 below for classical examples). We thus denote by $\hat{\theta}$ a minimizer of

$$\frac{1}{2n}\|y - \Phi\theta\|_2^2 + \lambda\Omega(\theta). \tag{8.5}$$

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

Lemma 8.4 *Let $\hat{\theta}$ be a minimizer of Eq. (8.5).*

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

Proof We have, like in Section 8.1.1, by optimality of $\hat{\theta}$ for Eq. (8.5):

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

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, using the identity $2ab \leqslant \frac{1}{2}a^2 + 2b^2$,

$$\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant \frac{1}{2}\|\Phi(\hat{\theta} - \theta_*)\|_2^2 + 2\|\varepsilon\|_2^2 + 2n\lambda\Omega(\theta_*),$$

which leads to the desired bound. ■

Exercise 8.4 *For $p \in [1, \infty]$, show that the dual of the ℓ_p-norm is the ℓ_q-norm, for $\frac{1}{p} + \frac{1}{q} = 1$.*

We can now use the lemma above to compute the excess risk of the Lasso, for which $\Omega = \|\cdot\|_1$ and $\Omega^*(\Phi^\top\varepsilon) = \|\Phi^\top\varepsilon\|_\infty$.[3] The key is to note that since $\|\Phi^\top\varepsilon\|_\infty$ is a maximum of $2d$ zero-mean terms that scale as $\sqrt{n}$, according to Section 1.2.4, its maximum scales as $\sqrt{n\log(d)}$, and we will apply the lemma above when λ is larger than $\sqrt{\log(d)/n}$. We denote by $\|\widehat{\Sigma}\|_\infty$ the largest element of the matrix $\widehat{\Sigma}$ in absolute value.

[3] Developments similar to Prop. 4.7 in Section 4.5.5 for general norms could also be carried out.

Proposition 8.3 (Lasso - slow 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. (8.4). Then, for $\lambda = \frac{2\sigma}{\sqrt{n}}\sqrt{2\|\widehat{\Sigma}\|_\infty}\sqrt{\log(d) + \log\frac{1}{\delta}}$, we have, with probability greater than $1-\delta$:*

$$\frac{1}{n}\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant 3\|\theta_*\|_1 \cdot \frac{2\sigma}{\sqrt{n}}\sqrt{2\|\widehat{\Sigma}\|_\infty}\sqrt{\log(d) + \log\frac{1}{\delta}}.$$

Proof For each j, the random variable $(\Phi^\top\varepsilon)_j$ is Gaussian with mean zero and variance $n\sigma^2\widehat{\Sigma}_{jj}$. Thus, we get from the union bound and from the fact that for a standard Gaussian variable z, $\mathbb{P}(|z| \geqslant t) \leqslant \exp(-t^2/2)$:[4]

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

Thus, with probability greater than $1-\delta$, we can apply the first part of Lemma 8.4, and therefore the error is less than $3\lambda\|\theta^*\|_1$. For a result in expectation, see the exercise below. ■

Check homogeneity!

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, so that $\|\theta_*\|_1$ does not grow with d). Note that the proposed regularization parameter depends on the unknown noise variance. A simple trick known as the "square root Lasso" allows to avoid that dependence on σ (see Giraud, 2014, Section 5.4), by minimizing $\frac{1}{\sqrt{n}}\|y - \Phi\theta\|_2 + \lambda\|\theta\|_1$.

The proposition above suggests a regularization parameter λ proportional to $1/\sqrt{n}$, which does enable estimation in high-dimensional situations but can also add a significant bias because all non-zero components of $\hat{\theta}$ are shrunk towards zero. See Section 8.5 for methods to alleviate this effect.

Exercise 8.5 *(♦) With the same assumptions as Prop. 8.3, and with the choice of regularization parameter $\lambda = 4\sigma\sqrt{\frac{\log(dn)}{n}}\sqrt{\|\widehat{\Sigma}\|_\infty}$, show the following bound in expectation:* $\mathbb{E}\big[\frac{1}{n}\|\Phi(\hat{\theta} - \theta_*)\|_2^2\big] \leqslant 32\sigma\sqrt{\frac{\log(dn)}{n}}\sqrt{\|\widehat{\Sigma}\|_\infty}\|\theta_*\|_1 + \frac{32}{n}\sigma^2$.

Beyond square loss. The slow rates proportional to $\|\theta_*\|_1\sqrt{(\log d)/n}$ for regularization by the ℓ_1-norm can also be achieved for Lipschitz-continuous losses (such as the logistic loss and the hinge loss), as shown in Prop. 4.7 in Section 4.5.5.

[4] We have for $t \geqslant 0$: $e^{t^2/2}\mathbb{P}(|z| \geqslant t) = \frac{2}{\sqrt{2\pi}}\int_t^{+\infty} e^{t^2/2 - s^2/2}dt \leqslant \frac{2}{\sqrt{2\pi}}\int_t^{+\infty} e^{-(s-t)^2/2}dt = 1$.

8.3.4 Fast rates (♦)

We now consider conditions to obtain a fast rate with a leading term proportional to $\sigma^2 \frac{k \log d}{n}$, 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. (8.4) in terms of the support A of θ_*.

Lemma 8.5 *Let $\hat{\theta}$ be a minimizer of Eq. (8.5). Assume $\|\Phi^\top \varepsilon\|_\infty \leqslant \frac{n\lambda}{2}$. 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 (e.g., Lemma 8.4), 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 that there exists a constant $\kappa > 0$ such that

$$\frac{1}{n}\|\Phi\Delta\|_2^2 \geqslant \kappa\|\Delta_A\|_2^2 \tag{8.6}$$

for all Δ that satisfies the condition $\|\Delta_{A^c}\|_1 \leqslant 3\|\Delta_A\|_1$. This is called the "restrictive eigenvalue property" because if the smallest eigenvalue of $\frac{1}{n}\Phi^\top\Phi$ is greater than κ, the condition is satisfied (but this is only possible if $n \geqslant d$). The relevance of this assumption is discussed in Section 8.3.5.

This leads to the following proposition.

Proposition 8.4 (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. (8.4). Then, for $\lambda = \frac{2\sigma}{\sqrt{n}}\sqrt{2\|\widehat{\Sigma}\|_\infty}\sqrt{\log(2d) + \log\frac{1}{\delta}}$, we have, if Eq. (8.6) is satisfied, and with probability greater than $1 - \delta$:*

$$\mathbb{E}\Big[\frac{1}{n}\|\Phi(\hat{\theta} - \theta_*)\|_2^2\Big] \leqslant \frac{72|A|\sigma^2}{n}\frac{\|\widehat{\Sigma}\|_\infty}{\kappa}\Big(\log(2d) + \log\frac{1}{\delta}\Big).$$

Proof (♦) We have, when λ is large enough, and by application of Lemma 8.5, and using Eq. (8.6):

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

which leads to $\|\Delta_A\|_1 \leqslant \dfrac{3|A|\lambda}{\kappa}$. We then get $\frac{1}{n}\|\Phi\Delta\|_2^2 \leqslant \frac{9|A|\lambda^2}{\kappa}$, which leads to the desired result.

■

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. Such results can be extended beyond the square loss using the notion of self-concordance (see, e.g., Ostrovskii and Bach, 2021b, and references therein).

Exercise 8.6 *(♦♦) With the same assumptions as Prop. 8.4, with the choice of regularization parameter* $\lambda = 4\sigma\sqrt{\frac{\log(dn)}{n}}\sqrt{\|\widehat{\Sigma}\|_\infty}$*, show that we have the bound in expectation* $\mathbb{E}\big[\frac{1}{n}\|\Phi(\hat{\theta}-\theta_*)\|_2^2\big] \leqslant \frac{144|A|\sigma^2\|\widehat{\Sigma}\|_\infty}{\kappa}\frac{\log(dn)}{n} + \frac{24}{n}\sigma^2 + \frac{32}{dn^2}\|\theta_*\|_1\sigma\sqrt{\frac{\log(dn)}{n}}\sqrt{\|\widehat{\Sigma}\|_\infty}$.

8.3.5 Zoo of conditions (♦♦)

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

Restricted eigenvalue property (REP). The most direct condition is the so-called restricted eigenvalue property (REP), which is exactly Eq. (8.6), with the supremum taken over the unknown set A of cardinality less than k:

$$\inf_{|A|\leqslant k}\ \inf_{\|\Delta_{A^c}\|_1\leqslant 3\|\Delta_A\|_1}\frac{\|\Phi\Delta\|_2^2}{n\|\Delta_A\|_2^2} \geqslant \kappa > 0. \tag{8.7}$$

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

$$\sup_{i\neq j}|\widehat{\Sigma}_{ij}| \leqslant \frac{\min_{j\in\{1,\dots,d\}}\widehat{\Sigma}_{jj}}{14k}, \tag{8.8}$$

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

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

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

Moreover, we have:

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

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 $\frac{1}{n}\|\Phi\Delta\|_2^2 \geqslant \min_{j\in\{1,\dots,d\}}\widehat{\Sigma}_{jj}\big(\|\Delta_A\|_2^2 - \frac{7}{14k}\|\Delta_A\|_1^2\big)$, which is greater than $\min_{j\in\{1,\dots,d\}}\widehat{\Sigma}_{jj}\big(\|\Delta_A\|_2^2 - \frac{7k}{14k}\|\Delta_A\|_2^2\big) = \kappa\|\Delta_A\|_2^2$, with $\kappa = \min_{j\in\{1,\dots,d\}}\widehat{\Sigma}_{jj}/2$, thus leading to the REP condition in Eq. (8.7).

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 Giraud (2014); Wainwright (2019) 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). For our results to be non-trivial, we need that $k\frac{\log d}{n}$ to be small but not too small. In this paragraph, we show, without proof, that when sampling from Gaussian distributions, the assumptions above are satisfied. This is a first step towards a random design assumption.

Proposition 8.5 (Wainwright, 2019, 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 $\kappa = \frac{1}{16}\lambda_{\min}(\Sigma)$ as soon as $k\frac{\log d}{n} \leqslant \frac{1}{3200}\frac{\lambda_{\min}(\Sigma)}{\|\Sigma\|_\infty}$.*

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

Exercise 8.7 *(♦♦♦) 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 $k^2\dfrac{\log d}{n}$, the mutual incoherence property in Eq. (8.8) 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 Giraud (2014); Wainwright (2019) for details.

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

8.3.6 Random design (♦)

In this section, we study the Lasso in the random design setting instead of the fixed design setting. For slow rates in $1/\sqrt{n}$, we can directly use Section 4.5.5 to get the exact same slow rate as for fixed design. In this section, we will only consider fast rates.

We now consider the well-specified Lasso case, where the expected risk is equal to $\mathcal{R}(\theta) = \frac{\sigma^2}{2} + \frac{1}{2}(\theta-\theta^*)^\top \Sigma(\theta-\theta^*)$. We assume that $\lambda_{\min}(\Sigma) \geqslant \mu \geqslant 0$, that is, the *expected* risk is μ-strongly convex (and not the empirical risk).

We assume that $y_i = \varphi(x_i)^\top \theta^* + \varepsilon_i$, and denote $\Phi \in \mathbb{R}^{n\times d}$ the design matrix, as well as $\varepsilon \in \mathbb{R}^n$ the vector of noises, which we assume independent and sub-Gaussian. Therefore we have

$$\widehat{\mathcal{R}}(\theta) = \frac{1}{2n}\|\Phi(\theta-\theta^*)-\varepsilon\|_2^2 = \frac{1}{2}(\theta-\theta^*)^\top \widehat{\Sigma}(\theta-\theta^*) - (\theta-\theta^*)^\top\Big(\frac{1}{n}\Phi^\top\varepsilon\Big) + \frac{1}{2n}\|\varepsilon\|_2^2, \quad (8.9)$$

where $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \varphi(x_i)\varphi(x_i)^\top = \frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{d\times d}$ is the empirical non-centered covariance matrix.

We will need that $\big\|\frac{1}{n}\Phi^\top\varepsilon\big\|_\infty = \big\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\varphi(x_i)\big\|_\infty$ is small enough, as well as the error in the covariance matrix $\big\|\widehat{\Sigma}-\Sigma\big\|_\infty$. Assuming that ε is sub-Gaussian with constant σ^2, and that $\|\varphi(x)\|_\infty \leqslant R$ almost surely, we get that, using results from Section 1.2.1,

$$\mathbb{P}\Big(\big\|\frac{1}{n}\Phi^\top\varepsilon\big\|_\infty \geqslant \frac{\sigma R t}{\sqrt{n}}\Big) \leqslant 2d\exp(-t^2/2) \text{ and } \mathbb{P}\Big(\big\|\widehat{\Sigma}-\Sigma\big\|_\infty \geqslant \frac{R^2 t}{\sqrt{n}}\Big) \leqslant 2d(d+1)/2\exp(-t^2/2).$$

Thus, the probability that at least one is satisfied is less than $d(d+3)\exp(-t^2/2) \leqslant 4d^2\exp(-t^2/2)$.

We now assume that $\big\|\frac{1}{n}\Phi^\top\varepsilon\big\|_\infty \leqslant \frac{\sigma R t}{\sqrt{n}}$ and $\big\|\widehat{\Sigma}-\Sigma\big\|_\infty \leqslant \frac{R^2 t}{\sqrt{n}}$, which happens with probability at least $1-4d^2\exp(-t^2/2)$. From Lemma 8.5, we know that if $\lambda \geqslant 2\big\|\frac{1}{n}\Phi^\top\varepsilon\big\|_\infty$, then we have, with $\hat{\Delta} = \hat{\theta}_\lambda - \theta^*$, and A the support of θ^*:

$$\|\hat{\Delta}_{A^c}\|_1 \leqslant 3\|\hat{\Delta}_A\|_1 \text{ and } \|\hat{\theta}_\lambda\|_1 \leqslant 3\|\theta^*\|_1.$$

Let $v = \mathcal{R}(\hat{\theta}_\lambda) - \mathcal{R}(\theta^*)$. We have:

$$\begin{aligned}
v &\leqslant \mathcal{R}(\hat{\theta}_\lambda) - \mathcal{R}(\theta^*) - \widehat{\mathcal{R}}_\lambda(\hat{\theta}_\lambda) + \widehat{\mathcal{R}}_\lambda(\theta^*) \text{ since } \hat{\theta}_\lambda \text{ minimizes } \widehat{\mathcal{R}}_\lambda, \\
&= \mathcal{R}(\hat{\theta}_\lambda) - \mathcal{R}(\theta^*) - \widehat{\mathcal{R}}(\hat{\theta}_\lambda) + \widehat{\mathcal{R}}(\theta^*) + \lambda\|\theta^*\|_1 - \lambda\|\hat{\theta}_\lambda\|_1 \text{ by definition of } \widehat{\mathcal{R}}_\lambda, \\
&= \frac{1}{2}\hat{\Delta}^\top(H - \hat{H})\hat{\Delta} + \hat{\Delta}^\top\Big(\frac{1}{n}\Phi^\top\varepsilon\Big) + \lambda\|\theta^*\|_1 - \lambda\|\hat{\theta}_\lambda\|_1 \text{ using Eq. (8.9)} , \\
&\leqslant \frac{1}{2}\big\|\widehat{\Sigma} - \Sigma\big\|_\infty \cdot \|\hat{\Delta}\|_1^2 + \big\|\frac{1}{n}\Phi^\top\varepsilon\big\|_\infty \cdot \|\hat{\Delta}\|_1 + \lambda\|\hat{\Delta}\|_1 \text{ using norm inequalities}, \\
&\leqslant \frac{\sigma R t}{\sqrt{n}} \cdot \|\hat{\Delta}\|_1 + \frac{R^2 t}{2\sqrt{n}} \cdot \|\hat{\Delta}\|_1^2 + \lambda\|\hat{\Delta}\|_1 \text{ using our assumptions.}
\end{aligned}$$

Moreover, we have, since $\lambda_{\min}(\Sigma) \geqslant \mu$, $v = \mathcal{R}(\hat{\theta}_\lambda) - \mathcal{R}(\theta^*) \geqslant \frac{\mu}{2}\|\hat{\Delta}\|_2^2 \geqslant \frac{\mu}{2|A|}\|\hat{\Delta}_A\|_1^2$, leading to $\|\hat{\Delta}\|_1 \leqslant 4\|\hat{\Delta}_A\|_1 \leqslant 4\sqrt{\frac{2|A|v}{\mu}}$. We also have $\|\hat{\Delta}\|_1 \leqslant \|\theta^*\|_1 + \|\hat{\theta}_\lambda\|_1 \leqslant \|\theta^*\|_1 + 3\|\theta^*\|_1 \leqslant 4\|\theta^*\|_1$. We thus get, with $\lambda = \frac{2\sigma R t}{\sqrt{n}}$, two inequalities:

$$v \leqslant \frac{3\sigma R t}{\sqrt{n}} \cdot \|\hat{\Delta}\|_1 + \frac{R^2 t}{2\sqrt{n}} \cdot \|\hat{\Delta}\|_1^2 \text{ and } \|\hat{\Delta}\|_1 \leqslant 4\sqrt{\frac{2|A|v}{\mu}}. \tag{8.10}$$

If $1 \geqslant \frac{32R^2t}{\sqrt{n}}\frac{|A|}{\mu}$, then the last term in the first inequality in Eq. (8.10) is less than $\frac{v}{2}$, and we get $\frac{v}{2} \leqslant \frac{3\sigma R t}{\sqrt{n}} 4\sqrt{\frac{2|A|v}{\mu}}$, that is, $\sqrt{v} \leqslant \frac{24\sigma R t}{\sqrt{n}}\sqrt{\frac{2|A|}{\mu}}$. This leads to, with $\lambda = \frac{2\sigma R}{\sqrt{n}}t = \frac{2\sigma R}{\sqrt{n}}\sqrt{2\log\frac{4d^2}{\delta}}$, with probability greater than $1 - \delta$,

$$\mathcal{R}(\hat{\theta}_\lambda) - \mathcal{R}(\theta^*) \leqslant 2304 \cdot \frac{R^2}{\mu}\frac{\sigma^2|A|}{n}\log\frac{4d^2}{\delta}.$$

Exercise 8.8 *With the notations above, show that if $\mu = 0$, from Eq. (8.10) we can recover the slow rate* $\mathcal{R}(\hat{\theta}_\lambda) - \mathcal{R}(\theta^*) \leqslant \frac{4R\|\theta^*\|_1}{\sqrt{n}}(3\sigma + 2R\|\theta^*\|_1)\sqrt{2\log\frac{4d^2}{\delta}}$.

8.4 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, with $n = 64$, and varying d. Then θ_* is taken to be zero except on $k = 4$ components where it is randomly equal to -1 or 1. We consider $\sigma = \sqrt{k}$ (to have a constant signal-to-noise ratio 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 reported the minimum value (with respect to the hyperparameter). We compare the following three methods in Figure 8.2:

- Ridge regression: penalty by $\lambda\|\theta\|_2^2$.

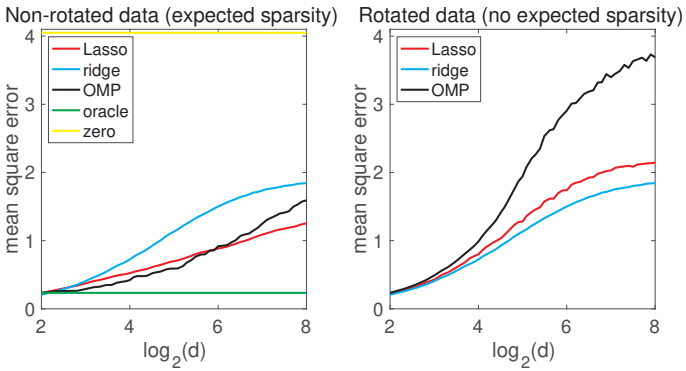


Figure 8.2: Comparison of estimators on least-squares regression: problem with sparse optimal predictor (left), and non-sparse optimal predictor (right).

- 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 random rotation matrix. For the rotated data, we do not expect sparse solutions. Hence, sparse methods are not expected to work better than ridge regression (and OMP performs significantly worse 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.

8.5 Extensions

Sparse methods are more general than the ℓ_1-norm and can be extended in several ways:

- **Group penalties:** in many cases, $\{1, \dots, d\}$ is partitioned into 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 $\sum_{i=1}^m \|\theta_{A_i}\|_\infty$. This is particularly used when the output y is multi-dimensional (such as in multivariate regression or multi-category classification) to select variables relevant to all outputs. See, e.g., Giraud (2014) for details.

Exercise 8.9 *Assuming that the design matrix Φ is orthogonal, compute the min-*

imizer of $\frac{1}{2n}\|y - \Phi\theta\|_2^2 + \lambda \sum_{i=1}^m \|\theta_{A_i}\|_2$.

- **Structured sparsity:** It is also possible to favor other specific patterns for the selected variables, such as blocks, trees, etc., when such prior knowledge is needed. See Bach et al. (2012b) for details.

  Exercise 8.10 *We consider the d (overlapping) sets* $A_i = \{1, \dots, i\}$*, and the norm* $\sum_{i=1}^d \|\theta_{A_i}\|_2$*. Show that penalization with this norm will tend to select patterns of non-zeros of the form* $\{i+1, \dots, d\}$.

- **Nuclear norm:** When learning on matrices, a natural form of sparsity is for a matrix to have a low rank. This can be achieved by penalizing the sum of singular values of a matrix, a norm called the nuclear norm or the trace norm. See Bach (2008) and references therein.

  Exercise 8.11 *Compute the minimizer of* $\frac{1}{2n}\|Y - \Theta\|_F^2 + \lambda\|\Theta\|_*$*, where* $\|M\|_F$ *is the Frobenius norm and* $\|M\|_*$ *the nuclear norm.*

- **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 Chapter 7. This becomes a tool for learning the kernel matrix from data. See Bach et al. (2012a) for details.

- **Elastic net:** often, when both effects of the ℓ_1-norm (sparsity) and the squared ℓ_2-norm (strong-convexity) are desired, we can sum the two, which is referred to as the "elastic net" penalty. This leads to a strongly-convex optimization problem, which is numerically better behaved.

- **Concave penalization and debiasing:** 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 of 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 Mairal et al., 2014, and references therein).

8.6 Conclusion

In this chapter, we have considered sparse methods based on the penalization by the ℓ_0 or ℓ_1 penalties of the weight vector of a linear model. For the square loss, ℓ_0-penalties led to an excess risk proportional to $\sigma^2 k \log(d)/n$, with a price of adaptivity of $\log(d)$, with few conditions on the problem, but no provably computationally efficient procedures. On the contrary, ℓ_1-norm penalization can be solved efficiently with appropriate convex optimization algorithms (such as proximal methods) but only obtained a slow rate proportional to $\sqrt{\log(d)/n}$, exhibiting a high-dimensional phenomenon, but a worse dependence in n.

Fast rates can only be obtained with stronger assumptions on the covariance matrix of the features.

This chapter was limited to linear models, and in the next chapter on neural networks, we will see how models that are non-linear in their parameters can lead to non-linear variable selection, still exhibiting a high-dimensional phenomenon but at the expense of a harder optimization.