

Learning theory from first principles

Lecture 2: Linear least-squares regression

Francis Bach

September 25, 2020

Class summary

- Guarantees in the fixed design settings (simple in closed-form as $\sigma^2 d/n$)
- Ridge regression: dimension independent bounds
- Guarantees in the random design setting
- Lower bound of performance

Announcements

- Ask questions! (chat or directly)
- Coding assignment for each lecture: reproduce the experiments and send the figures to the following address: `learning.theory.first.principles@gmail.com` (and not my personal email address).



Each student is expected to read class notes before the class.



1 Introduction

In this class, we introduce and analyze linear least-squares regression, a tool that can be traced back to Legendre (1805) and Gauss (1809)—see https://en.wikipedia.org/wiki/Least_squares#The_method for an interesting discussion and the claim that Gauss knew about it already in 1795.

Why should we study linear least-squares regression? Isn't there any progress since 1805? A few reasons:

- It already captures many of the concepts in learning theory, such as the bias-variance trade-off, as well as the dependence of generalization performance on the underlying dimension of the problem, or on dimension-less quantities.
- Because of its simplicity, many results can be easily derived without the need for complicated mathematics (simple linear algebra for the simplest results).
- Using features, it can be extended to arbitrary non-linear predictions (see kernel methods in Lecture 6).

In subsequent lectures, we will extend many of these results beyond least-squares.

2 Least-squares framework

- We recall the goal of supervised machine learning from Lecture 1: given some observations $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$, of inputs/outputs, features/variables (training data), given a new $x \in \mathcal{X}$, predict $y \in \mathcal{Y}$ (testing data) with a *regression* function f such that $y \approx f(x)$. We use the square loss $\ell(y, z) = (y - z)^2$, for which we know from the previous lecture, that the optimal predictor is $f^*(x) = \mathbb{E}(y|x)$.
- In this lecture, we consider empirical risk minimization. We choose a parameterized family of prediction functions $f_\theta : \mathcal{X} \rightarrow \mathcal{Y}$ for $\theta \in \Theta$ and minimize the empirical risk

$$\frac{1}{n} \sum_{i=1}^n (y_i - f_\theta(x_i))^2,$$

leading to the estimator $\hat{\theta} \in \arg \min_{\theta} \frac{1}{n} \sum_{i=1}^n (y_i - f_\theta(x_i))^2$. Note that in most cases, the Bayes predictor f^* does not belong to the class of functions $\{f_\theta, \theta \in \Theta\}$, that is, the model is said *misspecified*.

- Least-squares regression can be carried out with parameterizations of the function f_θ which may be non-linear in the parameter θ . In this lecture, we will consider only situations where $f_\theta(x)$ is linear in θ , which is thus assumed to live in a vector space, and which we take to be \mathbb{R}^d for simplicity.



Being linear in x or linear θ is different!

- While we assume linearity in the parameter θ , nothing forces $f_\theta(x)$ to be linear in the input x . In fact, even the concept of linearity may be meaningless if \mathcal{X} is not a vector space. Through Riesz

representation theorem, for any $x \in \mathcal{X}$, there exists a vector in \mathbb{R}^d , which we denote $\varphi(x)$, such that

$$f_\theta(x) = \varphi(x)^\top \theta.$$

The vector $\varphi(x) \in \mathbb{R}^d$ is typically called the *feature vector*, which we assume to be known (in other words, it is given to us and can be computed explicitly when needed). We thus consider minimizing

$$\hat{\mathcal{R}}(\theta) := \frac{1}{n} \sum_{i=1}^n (y_i - \varphi(x_i)^\top \theta)^2.$$

- When $\mathcal{X} \subset \mathbb{R}^d$, we can make the extra assumptions that f_θ is an affine function, which can be obtained through $\varphi(x) = \begin{pmatrix} x \\ 1 \end{pmatrix} \in \mathbb{R}^{d+1}$. Other classical assumptions are $\varphi(x)$ composed of monomials. We will see in Lecture 6 (kernel methods) that we can consider infinite-dimensional features.
- Matrix notation: the cost function above can be rewritten in matrix notations. Let $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$ be the vector of outputs (sometimes called the *response vector*), and $\Phi \in \mathbb{R}^{n \times d}$ the matrix of inputs, which rows are $\varphi(x_i)^\top$. It is called the *design matrix*. In these notations, the empirical risk is

$$\hat{\mathcal{R}}(\theta) = \frac{1}{n} \|y - \Phi\theta\|_2^2, \quad (1)$$

where $\|\alpha\|_2^2 = \sum_{j=1}^d \alpha_j^2$ is the squared ℓ_2 -norm of α .

⚠ It is sometimes tempting at first to avoid matrix notations. We strongly advise against it as it leads to long and error-prone formulas.

3 Ordinary least-squares (OLS) estimator

We make the assumption that the matrix $\Phi \in \mathbb{R}^{n \times d}$ has full column rank (i.e., the rank of Φ is d). In particular, the problem is said “over-determined”, and $d \leq n$. Equivalently, we assume that $\Phi^\top \Phi \in \mathbb{R}^{d \times d}$ is invertible.

Definition 1 When Φ has full column rank, the minimizer of Eq. (1) is called the ordinary least-squares (OLS) estimator.

3.1 Closed-form solution

Proposition 1 When Φ has full column rank, the OLS estimator exists and is unique. It is given by

$$\hat{\theta} = (\Phi^\top \Phi)^{-1} \Phi^\top y.$$

Denote the (non-centered)¹ empirical covariance matrix $\hat{\Sigma} := \frac{1}{n} \Phi^\top \Phi \in \mathbb{R}^{d \times d}$; we have $\hat{\theta} = \frac{1}{n} \hat{\Sigma}^{-1} \Phi^\top y$.

¹The “centered” covariance matrix would be $\frac{1}{n} \sum_{i=1}^n [\varphi(x_i) - \mu][\varphi(x_i) - \mu]^\top$ where $\mu = \frac{1}{n} \sum_{i=1}^n \varphi(x_i)$ is the empirical mean, while we consider $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \varphi(x_i) \varphi(x_i)^\top$.

Proof Since the function $\hat{\mathcal{R}}$ is coercive (i.e., going to infinity at infinity) and continuous, it admits at least a minimizer. Moreover, it is differentiable, so a minimizer $\hat{\theta}$ must satisfy $\nabla \hat{\mathcal{R}}(\hat{\theta}) = 0$. For all $\theta \in \mathbb{R}^d$, we have

$$\hat{\mathcal{R}}(\theta) = \frac{1}{n} \left(\|y\|_2^2 - 2\theta^\top \Phi^\top y + \theta^\top \Phi^\top \Phi \theta \right) \quad \text{and} \quad \nabla \hat{\mathcal{R}}(\theta) = \frac{2}{n} \left(\Phi^\top \Phi \theta - \Phi^\top y \right).$$

The condition $\nabla \hat{\mathcal{R}}(\hat{\theta}) = 0$ gives the so-called *normal equations*:

$$\Phi^\top \Phi \hat{\theta} = \Phi^\top y.$$

The normal equations have a unique solution $\hat{\theta} = (\Phi^\top \Phi)^{-1} \Phi^\top y$. This shows the uniqueness of the minimizer of $\hat{\mathcal{R}}$ as well as its closed-form expression. ■

Another way to show uniqueness of the minimizer is by showing that $\hat{\mathcal{R}}$ is strongly convex since $\nabla^2 \hat{\mathcal{R}}(\theta) = 2\hat{\Sigma}$ for all $\theta \in \mathbb{R}^d$ (convexity will be studied in Lecture 4).

⚠ For readers worried about carrying a factor of two in the gradients, we will use an additional factor 1/2 in lectures on optimization (e.g., Lecture 4).

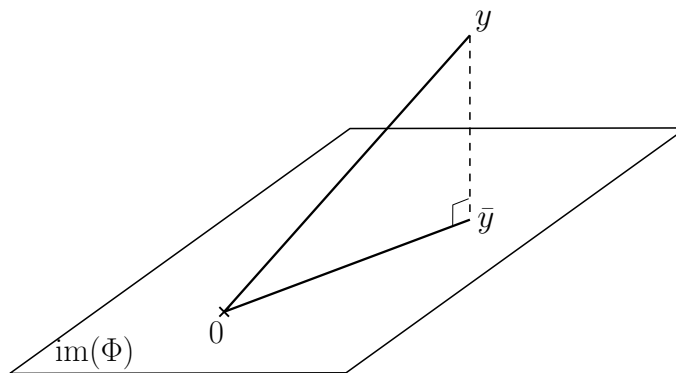
3.2 Geometric interpretation

Proposition 2 *The vector of predictions $\Phi \hat{\theta} = \Phi(\Phi^\top \Phi)^{-1} \Phi^\top y$ is the orthogonal projection of $y \in \mathbb{R}^n$ onto $\text{im}(\Phi) \subset \mathbb{R}^n$, the column space of Φ .*

Proof Let us show that $P := \Phi(\Phi^\top \Phi)^{-1} \Phi^\top \in \mathbb{R}^{n \times n}$ is the orthogonal projection on $\text{im}(\Phi)$. For any $a \in \mathbb{R}^d$, it holds $P\Phi a = \Phi(\Phi^\top \Phi)^{-1} \Phi^\top \Phi a = \Phi a$, so $Pu = u$ for all $u \in \text{im}(\Phi)$. Also, since $\text{im}(\Phi)^\perp = \text{null}(\Phi^\top)$, where $Pu' = 0$ for all $u' \in \text{im}(\Phi)^\perp$. These properties characterize the orthogonal projection on $\text{im}(\Phi)$. ■

Thus we can interpret the OLS estimation as doing the following (see below for an illustration):

1. compute \bar{y} the projection of y on the image of Φ ,
2. solve the linear system $\Phi \theta = \bar{y}$ which has a unique solution.



3.3 Numerical resolution

While the closed-form $\hat{\theta} = (\Phi^\top \Phi)^{-1} \Phi^\top y$ is convenient for analysis, inverting $\Phi^\top \Phi$ is sometimes unstable and has a large computational cost when d is large. The following methods are usually preferred.

QR factorization. The QR decomposition factorizes the matrix Φ as $\Phi = QR$ where $Q \in \mathbb{R}^{n \times d}$ has orthonormal columns and $R \in \mathbb{R}^{d \times d}$ is upper triangular (see [1]). Computing a QR decomposition is faster and more stable than inverting a matrix. One has

$$(\Phi^\top \Phi)\hat{\theta} = \Phi^\top y \Leftrightarrow R^\top Q^\top QR\hat{\theta} = R^\top Q^\top y \Leftrightarrow R^\top R\hat{\theta} = R^\top Q^\top y \Leftrightarrow R\hat{\theta} = Q^\top y.$$

It only remains to solve a triangular linear system which is easy. The overall running time complexity remains $O(d^3)$. The conjugate gradient algorithm can also be used.

Gradient descent. We can completely bypass the need of matrix inversion or factorization using gradient descent. It consists in minimizing approximately $\hat{\mathcal{R}}$ by taking an initial point $\theta_0 \in \mathbb{R}^d$ and iteratively going towards the minimizer by following the opposite of the gradient

$$\theta_{k+1} = \theta_k - \gamma \nabla \hat{\mathcal{R}}(\theta_k) \quad \text{for } k \geq 0,$$

where $\gamma > 0$ is the step-size. When these iterates converge, it is towards the OLS estimator since a fixed-point θ satisfies $\nabla \hat{\mathcal{R}}(\theta) = 0$. We will study such algorithms in Lecture 4, with running-time complexities going up to linear in d .

4 Statistical analysis of OLS

We now prove guarantees on the performance of the OLS estimator. There are two settings of analysis for least-squares:

- *Random design.* In this setting, both the input and the output are random. This is the classical setting of supervised machine learning, where the goal is *generalization* to unseen data (like in last lecture). Since it is bit more complicated, it will be done after the fixed design setting.
- *Fixed design.* In this setting, we assume that the input data (x_1, \dots, x_n) are *not* random and we are interested in obtaining a small prediction error *on those input points only*. Alternatively, this can be seen as a prediction problem where the input distribution $dp(x)$ is the empirical distribution of (x_1, \dots, x_n) .

Our goal is thus to minimize the fixed design risk (where thus Φ is deterministic):

$$\mathcal{R}(\theta) = \mathbb{E}_y \left[\frac{1}{n} \sum_{i=1}^n (y_i - \varphi(x_i)^\top \theta)^2 \right] = \mathbb{E}_y \left[\frac{1}{n} \|y - \Phi \theta\|_2^2 \right]. \quad (2)$$

This assumption allows a complete analysis with basic linear algebra. It is justified in some settings, e.g., when the input is a fixed grid, but is otherwise just a simplifying assumption. It can also be understood as learning the optimal vector $\Phi \theta_* \in \mathbb{R}^n$ of best predictions for a well-defined θ_* , instead of a function.

In the fixed design setting, no attempts are made to generalize to unseen input points x , and we want to estimate well a label vector y resampled from the same distribution than the observed y . The risk in Eq. (2) is often called the *in-sample prediction error*.

We will first consider below the fixed design setting, where the celebrated rate $\sigma^2 d/n$ will appear naturally.


Relationship to maximum likelihood estimation. If, in the fixed design setting, we make the stronger assumption that the noise is Gaussian with mean zero and variance σ^2 , i.e., $\varepsilon_i = y_i - \varphi(x_i)^\top \theta_* \sim \mathcal{N}(0, \sigma^2)$, then the least mean-squares estimator of θ_* coincides with the maximum likelihood estimator (where Φ is assumed fixed). Indeed, the density / likelihood of y is, using independence and the density of the normal distribution:

$$p(y|\theta, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left(- (y_i - \varphi(x_i)^\top \theta)^2 / (2\sigma^2) \right).$$

Taking the logarithm and removing constants, the maximum likelihood estimators $(\tilde{\theta}, \tilde{\sigma}^2)$ minimize

$$\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \varphi(x_i)^\top \theta)^2 + n \log(\sigma).$$

We immediately see that $\tilde{\theta} = \hat{\theta}$, that is, OLS corresponds to maximum likelihood.

-  While maximum likelihood under a Gaussian model provides an interesting interpretation, the Gaussian assumption is not needed for the forthcoming analysis.
- **Exercise:** what is $\tilde{\sigma}^2$ the maximum likelihood of σ^2 ? Solution: $\tilde{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \varphi(x_i)^\top \hat{\theta})^2$.

5 Fixed design setting

We now assume that Φ is deterministic, and as before, we assume that $\widehat{\Sigma} = \frac{1}{n}\Phi^\top\Phi$ is invertible.

Any kind of guarantee requires assumptions about how the data are generated. We assume that:

- there exists a vector $\theta_* \in \mathbb{R}^d$ such that the relationship between input and output is for $i \in \{1, \dots, n\}$

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

- ε_i are independent of expectation $\mathbb{E}[\varepsilon_i] = 0$ and variance $\mathbb{E}[\varepsilon_i^2] = \sigma^2$.

The vector $\varepsilon \in \mathbb{R}^n$ accounts for variabilities in the output which are due to unobserved factors or to noise. The ‘‘homoscedasticity’’ assumption above, where the noise variances are uniform, is made for simplicity (and allows for the later bound $\sigma^2 d/n$ bound to be an equality). Note that to prove upper-bounds in performance, we could also only assume that $\mathbb{E}[\varepsilon_i^2] \leq \sigma^2$ for each $i \in \{1, \dots, n\}$.

Denoting \mathcal{R}^* the minimum value of $\mathcal{R}(\theta) = \mathbb{E}_y \left[\frac{1}{n} \|y - \Phi\theta\|_2^2 \right]$ over \mathbb{R}^d , the following proposition shows that it is attained at θ_* , and that is is equal to σ^2 .

Proposition 3 (Risk decomposition) *Under the linear model and fixed design assumptions above, for any $\theta \in \mathbb{R}^d$, we have $\mathcal{R}^* = \sigma^2$ and*

$$\mathcal{R}(\theta) - \mathcal{R}^* = \|\theta - \theta_*\|_{\widehat{\Sigma}}^2,$$

where $\widehat{\Sigma} := \frac{1}{n}\Phi^\top\Phi$ is the input covariance matrix and $\|\theta\|_{\widehat{\Sigma}}^2 := \theta^\top\widehat{\Sigma}\theta$. If now $\hat{\theta}$ is a random variable (such as an estimator of θ_*), then

$$\mathbb{E}[\mathcal{R}(\hat{\theta})] - \mathcal{R}^* = \underbrace{\|\mathbb{E}[\hat{\theta}] - \theta_*\|_{\widehat{\Sigma}}^2}_{\text{Bias}} + \underbrace{\mathbb{E}\left[\|\hat{\theta} - \mathbb{E}[\hat{\theta}]\|_{\widehat{\Sigma}}^2\right]}_{\text{Variance}}.$$

Proof We have, using $y = \Phi\theta_* + \varepsilon$, with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\|\varepsilon\|_2^2] = n\sigma^2$:

$$\begin{aligned} \mathcal{R}(\theta) &= \mathbb{E}_y \left[\frac{1}{n} \|y - \Phi\theta\|_2^2 \right] = \mathbb{E}_y \left[\frac{1}{n} \|\Phi\theta_* + \varepsilon - \Phi\theta\|_2^2 \right] \\ &= \frac{1}{n} \mathbb{E}_y \left[\|\Phi(\theta_* - \theta)\|_2^2 + \|\varepsilon\|_2^2 + 2\Phi(\theta_* - \theta)^\top \varepsilon \right] \\ &= \sigma^2 + \frac{1}{n} (\theta - \theta_*)^\top \Phi^\top \Phi (\theta - \theta_*). \end{aligned}$$

Since $\widehat{\Sigma} = \frac{1}{n}\Phi^\top\Phi$ is invertible, this shows that θ_* is the unique global minimizer of $\mathcal{R}(\theta)$, and that the minimum value \mathcal{R}^* is equal to σ^2 . This shows the first claim.

Now if θ is random, we perform the usual bias/variance decomposition:

$$\begin{aligned} \mathbb{E}[\mathcal{R}(\hat{\theta})] - \mathcal{R}^* &= \mathbb{E} \left[\|\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta_*\|_{\widehat{\Sigma}}^2 \right] \\ &= \mathbb{E} \left[\|\hat{\theta} - \mathbb{E}[\hat{\theta}]\|_{\widehat{\Sigma}}^2 \right] + \mathbb{E} \left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^\top \widehat{\Sigma} (\mathbb{E}[\hat{\theta}] - \theta_*) \right] + \mathbb{E} \left[\|\mathbb{E}[\hat{\theta}] - \theta_*\|_{\widehat{\Sigma}}^2 \right] \\ &= \mathbb{E} \left[\|\hat{\theta} - \mathbb{E}[\hat{\theta}]\|_{\widehat{\Sigma}}^2 \right] + 0 + \|\mathbb{E}[\hat{\theta}] - \theta_*\|_{\widehat{\Sigma}}^2. \end{aligned}$$

(NB: this is also a simple application of $\mathbb{E}\|z - a\|_M^2 = \|\mathbb{E}z - a\|_M^2 + \mathbb{E}\|z - \mathbb{E}[z]\|_M^2$ to $a = \theta_*$, $M = \widehat{\Sigma}$ and $z = \hat{\theta}$). ■

- The quantity $\|\cdot\|_{\widehat{\Sigma}}$ is called the Mahalanobis distance norm (it is a “true” norm whenever $\widehat{\Sigma}$ is positive definite). It is the norm on the parameter space induced by the input data.

Statistical properties of the OLS estimator

We now analyze the properties of the OLS estimator.

Proposition 4 (Estimation properties of OLS) *The OLS estimator $\hat{\theta}$ has the following properties:*

1. *it is unbiased, that is, $\mathbb{E}[\hat{\theta}] = \theta_*$,*
2. *its variance is $\text{var}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta_*)(\hat{\theta} - \theta_*)^\top\right] = \frac{\sigma^2}{n}\widehat{\Sigma}^{-1}$; $\widehat{\Sigma}^{-1}$ is often called the precision matrix.*

Proof

1. Since $\mathbb{E}[y] = \Phi\theta_*$, we have directly $\mathbb{E}[\hat{\theta}] = (\Phi^\top\Phi)^{-1}\Phi^\top\Phi\theta_* = \theta_*$.
2. It follows that $\hat{\theta} - \theta_* = (\Phi^\top\Phi)^{-1}\Phi^\top(\Phi\theta_* + \varepsilon) - \theta_* = (\Phi^\top\Phi)^{-1}\Phi^\top\varepsilon$. Thus, using that $\mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2I$, we get

$$\text{var}(\hat{\theta}) = \mathbb{E}\left[(\Phi^\top\Phi)^{-1}\Phi^\top\varepsilon\varepsilon^\top\Phi(\Phi^\top\Phi)^{-1}\right] = \sigma^2(\Phi^\top\Phi)^{-1}(\Phi^\top\Phi)(\Phi^\top\Phi)^{-1} = \sigma^2(\Phi^\top\Phi)^{-1} = \frac{\sigma^2}{n}\widehat{\Sigma}^{-1}.$$

We can put back the expression of the variance in the risk. ■

Proposition 5 (Risk of OLS) *The excess risk of the OLS estimator is equal to*

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^* = \frac{\sigma^2 d}{n}. \quad (3)$$

Proof Note here that the expectation is over ε only as we are in the fixed design setting. Using the risk decomposition of Proposition 3 and the fact that $\mathbb{E}[\hat{\theta}] = \theta_*$, we have

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^* = \mathbb{E}\|\hat{\theta} - \theta_*\|_{\widehat{\Sigma}}^2.$$

We have: $\mathbb{E}\left[\mathcal{R}(\hat{\theta})\right] - \mathcal{R}^* = \text{tr}[\text{var}(\hat{\theta})\widehat{\Sigma}] = \frac{\sigma^2 d}{n} \text{tr}(I) = \frac{\sigma^2 d}{n}$.

We can also give a direct proof. Using the identity $\hat{\theta} - \theta_* = (\Phi^\top \Phi)^{-1} \Phi^\top \varepsilon$, we get

$$\begin{aligned} \mathbb{E}[\mathcal{R}(\hat{\theta})] - \mathcal{R}^* &= \mathbb{E}[\|(\Phi^\top \Phi)^{-1} \Phi^\top \varepsilon\|_{\hat{\Sigma}}^2] \\ &= \frac{1}{n} \mathbb{E}[\varepsilon^\top \Phi (\Phi^\top \Phi)^{-1} \Phi^\top \Phi (\Phi^\top \Phi)^{-1} \Phi^\top \varepsilon] \\ &= \frac{1}{n} \mathbb{E}[\varepsilon^\top \Phi (\Phi^\top \Phi)^{-1} \Phi^\top \varepsilon] \\ &= \frac{1}{n} \mathbb{E}[\varepsilon^\top P \varepsilon] = \frac{1}{n} \mathbb{E}[\text{tr}(P \varepsilon \varepsilon^\top)] = \frac{\sigma^2}{n} \text{tr}(P) = \frac{\sigma^2 d}{n}, \end{aligned}$$

where we used that $P = \Phi(\Phi^\top \Phi)^{-1} \Phi^\top$ is the orthogonal projection on $\text{im}(\Phi)$, which is d -dimensional. ■

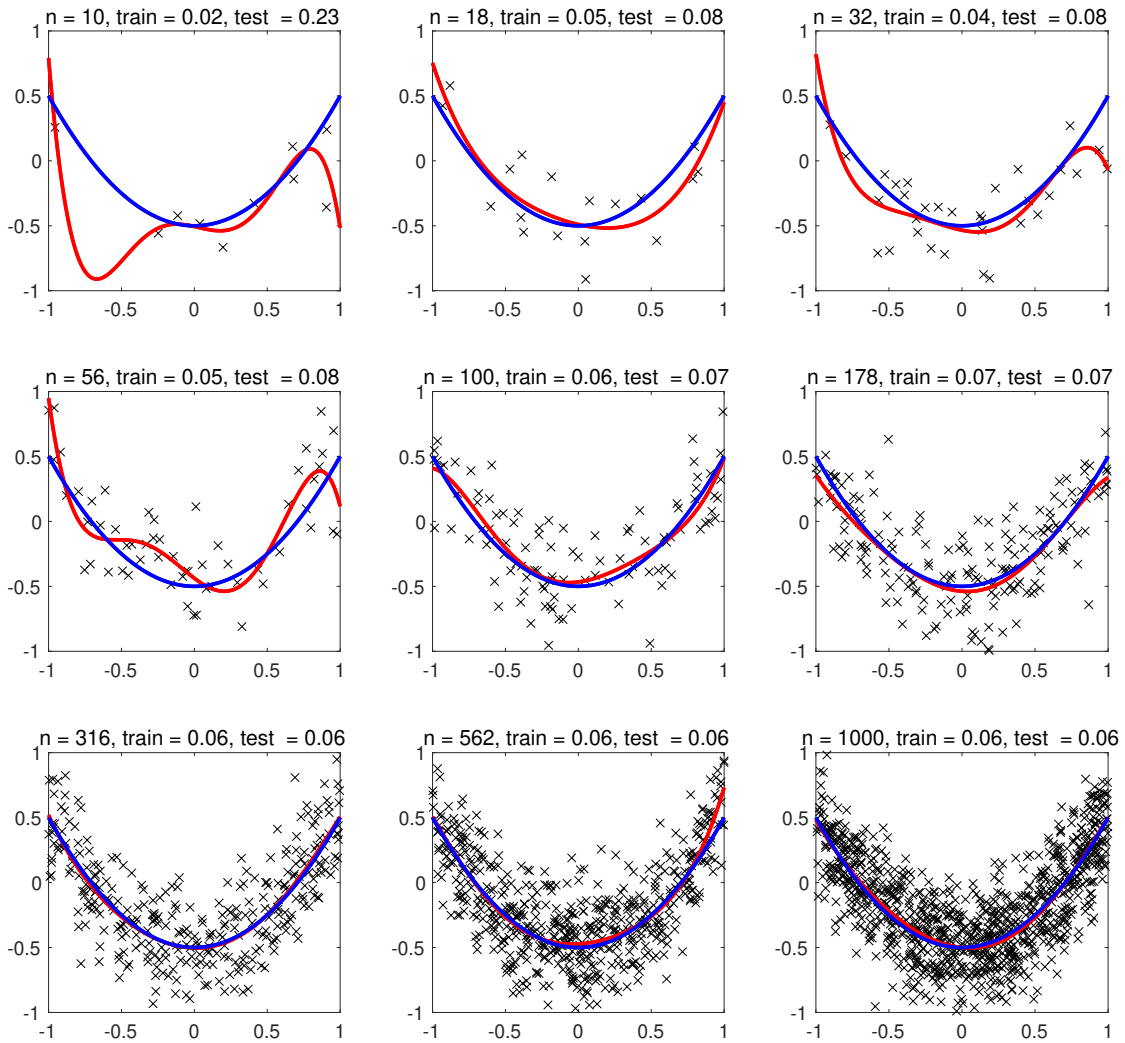
- **⚠** In the fixed design setting, the expectation over ε appears twice: (1) in the definition of the risk of some θ in Eq. (2), and when taking expectation over the data in Eq. (3).
- **Exercise:** what is the expected empirical risk $\mathbb{E}[\hat{\mathcal{R}}(\hat{\theta})]$? Solution: $\mathbb{E}[\hat{\mathcal{R}}(\hat{\theta})] = \frac{n-d}{n} \sigma^2$. In particular, when $n > d$, an unbiased estimator of the noise variance σ^2 is given by $\frac{\|Y - \Phi \hat{\theta}\|_2^2}{n-d}$.
- Above, we have an expression of the expected training error, which is equal to $\frac{n-d}{n} \sigma^2 = \sigma^2 - \frac{d}{n} \sigma^2$, while the expected testing error is $\sigma^2 + \frac{d}{n} \sigma^2$. We thus see that in context of least-squares, the training error underestimates (in expectation) the testing error by a factor of $2\sigma^2 d/n$, which characterizes the amount of overfitting. This difference can be used to perform model selection (see https://en.wikipedia.org/wiki/Mallows\%27s_Cp).

Discussion

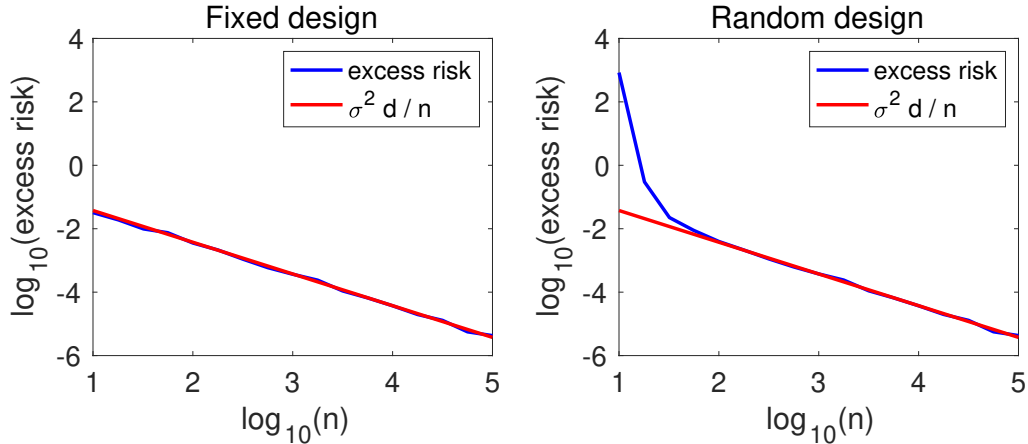
- In the fixed design setting, OLS thus leads to unbiased estimation, with an excess risk of $\sigma^2 d/n$.
- On the positive side, the math is very simple, and as we will show below, the obtained convergence rate is optimal.
- On the negative side, for the excess risk being small compared to σ^2 , we need d/n to be small, which seems to exclude high-dimensional problems where d is closed to n (let alone problems where $d > n$ or d much larger than n). Regularization (ridge or ℓ_1) will come to the rescue (see below).
- This is only for the fixed design setting. We consider the random design setting below, which is a bit more involved mathematically, mostly because of the presence of $\hat{\Sigma}^{-1}$ which does not cancel anymore (leading to a term $\hat{\Sigma}^{-1} \Sigma$).

Experiments

To illustrate the $\sigma^2 d/n$ bound, we consider polynomial regression in one dimension, with $x \in \mathbb{R}$, $\varphi(x) = (1, x, x^2, \dots, x^k)^\top \in \mathbb{R}^{k+1}$, so $d = k + 1$. The inputs are sampled from the uniform distribution in $[-1, 1]$, while the optimal regression function is a degree 2 polynomial (blue curve below). Gaussian noise is added to generate the outputs (black crosses below). The ordinary least-squares estimator is plotted in red, for various values of n , from $n = 10$ to $n = 1000$, for $k = 5$.



We can now plot the expected excess risk as a function of n , estimated by 32 replications of the experiment, together with the bound. In the right plot, we consider the random design setting (generalization error), while in the left plot we consider the fixed design setting (in-sample error). Notice the closeness of the bound for all n for the fixed design (as predicted by our bounds), while this is only true for n large enough in the random design setting.



6 Ridge least-squares regression

- When d/n approaches 1, we are essentially memorizing the observations y_i . Also when $d > n$, then $\Phi^\top \Phi$ is not invertible and the normal equations admit a linear subspace of solutions. These behaviors of OLS in high dimension (d large) are often undesirable.
- Several solutions exist to fix these issues. The most common is to regularize the least-squares objective, either by adding an ℓ_1 -penalty $\|\theta\|_1$ (leading to, “Lasso” regression, see Lecture 7) or $\|\theta\|_2^2$ (leading to *ridge* regression, this lecture and also Lecture 6) to the empirical risk.

Definition 2 (Ridge least-squares regression) For a regularization parameter $\lambda > 0$, we define the ridge least-squares estimator $\hat{\theta}_\lambda$ as the minimizer of

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \|y - \Phi\theta\|_2^2 + \lambda \|\theta\|_2^2.$$

Proposition 6 We recall that $\hat{\Sigma} = \frac{1}{n} \Phi^\top \Phi$. We have

$$\hat{\theta}_\lambda = \frac{1}{n} (\hat{\Sigma} + \lambda I)^{-1} \Phi^\top y.$$

Proof Left as an exercise (similar to the proof of Proposition 1). ■

As for the OLS, we can analyze the statistical properties of this estimator under the linear model and fixed design assumptions. See Lecture 6 for an analysis for random design and potentially infinite-dimensional features.

Proposition 7 Under the linear model assumption, the ridge least-squares estimator $\hat{\theta}_\lambda = \frac{1}{n} \Sigma_\lambda^{-1} \Phi^\top Y$ has the following excess risk

$$\mathbb{E} \left[\mathcal{R}(\hat{\theta}_\lambda) \right] - \mathcal{R}^* = \lambda^2 \theta_*^\top (\hat{\Sigma} + \lambda I)^{-2} \hat{\Sigma} \theta_* + \frac{\sigma^2}{n} \text{tr} \left[\hat{\Sigma}^2 (\hat{\Sigma} + \lambda I)^{-2} \right].$$

Proof We use the risk decomposition of Proposition 3 into a bias B and variance V terms. Since we have $\mathbb{E}[\hat{\theta}_\lambda] = \frac{1}{n}(\widehat{\Sigma} + \lambda I)^{-1}\Phi^\top \Phi \theta_* = (\widehat{\Sigma} + \lambda I)^{-1}\widehat{\Sigma}\theta_* = \theta_* - \lambda(\widehat{\Sigma} + \lambda I)^{-1}\theta_*$, it follows

$$\begin{aligned} B &= \|\mathbb{E}[\hat{\theta}_\lambda] - \theta_*\|_{\widehat{\Sigma}}^2 \\ &= \lambda^2 \theta_*^\top (\widehat{\Sigma} + \lambda I)^{-2} \widehat{\Sigma} \theta_*. \end{aligned}$$

For the variance term, using the fact that $\mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2 I$, we have

$$\begin{aligned} V &= \mathbb{E}\left[\|\hat{\theta}_\lambda - \mathbb{E}[\hat{\theta}_\lambda]\|_{\widehat{\Sigma}}^2\right] = \mathbb{E}\left[\left\|\frac{1}{n}(\widehat{\Sigma} + \lambda I)^{-1}\Phi^\top \varepsilon\right\|_{\widehat{\Sigma}}^2\right] = \mathbb{E}\left[\frac{1}{n^2} \text{tr}\left(\varepsilon^\top \Phi (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{-1} \Phi^\top \varepsilon\right)\right] \\ &= \mathbb{E}\left[\frac{1}{n^2} \text{tr}\left(\Phi^\top \varepsilon \varepsilon^\top \Phi (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{-1}\right)\right] = \frac{\sigma^2}{n} \text{tr}\left(\widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{-1}\right). \end{aligned}$$

The proposition follows by summing the bias and variance terms. ■

We can make the following observations:

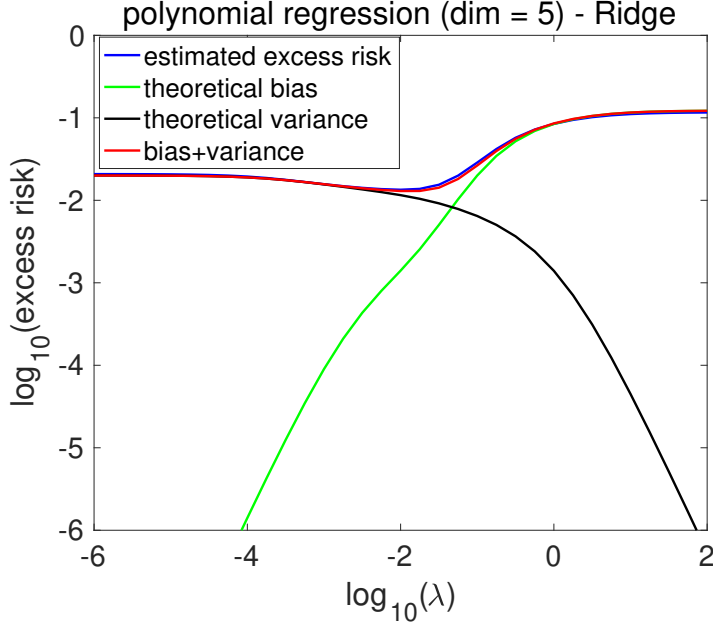
- The result above is also a bias / variance decomposition with the bias term equal to $B = \lambda^2 \theta_*^\top (\widehat{\Sigma} + \lambda I)^{-2} \widehat{\Sigma} \theta_*$, and the variance term equal to $V = \frac{\sigma^2}{n} \text{tr}[\widehat{\Sigma}^2 (\widehat{\Sigma} + \lambda I)^{-2}]$.
- The bias term is increasing in λ and equal to zero for $\lambda = 0$ if $\widehat{\Sigma}$ is invertible, while when λ goes to infinity, the bias goes to $\theta_*^\top \widehat{\Sigma} \theta_*$. It is independent of n and plays the role of the approximation error in the risk decomposition.
- The variance term is decreasing in λ , and equal to $\sigma^2 d/n$ for $\lambda = 0$ if $\widehat{\Sigma}$ is invertible, and converging to zero when λ goes to infinity. It depends on n and plays the role of the estimation error in the risk decomposition.

The quantity $\text{tr}[\widehat{\Sigma}^2 (\widehat{\Sigma} + \lambda I)^{-2}]$ is often called the “degrees of freedom”, and is often considered as an implicit number of parameters. It can be expressed as $\sum_{j=1}^d \frac{\lambda_j^2}{(\lambda_j + \lambda)^2}$, where (λ_j) are the eigenvalues of $\widehat{\Sigma}$. This quantity will be very important in the analysis of kernel methods in Lecture 6.

- Observe how this converges to the OLS estimator (when it is defined) as $\lambda \rightarrow 0$.
- In most cases, $\lambda = 0$ is not the optimal choice, that is biased estimation (with controlled bias) is preferable to unbiased estimation.

Experiments

With the same polynomial regression set-up as above, with $k = 10$, we can plot the various quantities above as a function of λ . We can see the monotonicity of bias and variance with respect to λ as well as the presence of an optimal choice of λ .



Choice of λ

Based on the expression for the risk, we can tune the parameter λ to obtain a potentially better bound than with the OLS (which corresponds to $\lambda = 0$ and the excess risk $\sigma^2 d/n$).

Proposition 8 *With the choice $\lambda^* = \frac{\sigma \sqrt{\text{tr}(\widehat{\Sigma})}}{\|\theta_*\|_2 \sqrt{n}}$, we have*

$$\mathbb{E} [\mathcal{R}(\hat{\theta}_{\lambda^*})] - \mathcal{R}^* \leq \frac{\sigma \sqrt{\text{tr}(\widehat{\Sigma})} \|\theta_*\|_2}{\sqrt{n}}.$$

Proof We have, using the fact that the eigenvalues of $(\widehat{\Sigma} + \lambda I)^{-2} \lambda \widehat{\Sigma}$ are less than $1/2$ (which is a simple consequence of $(\mu + \lambda)^{-2} \mu \lambda \leq 1/2 \Leftrightarrow (\mu + \lambda)^2 \geq 2\lambda\mu$ for all eigenvalues μ of $\widehat{\Sigma}$):

$$B = \lambda^2 \theta_*^\top (\widehat{\Sigma} + \lambda I)^{-2} \widehat{\Sigma} \theta_* = \lambda \theta_*^\top (\widehat{\Sigma} + \lambda I)^{-2} \lambda \widehat{\Sigma} \theta_* \leq \frac{\lambda}{2} \|\theta_*\|_2^2.$$

Similarly, we have

$$V = \frac{\sigma^2}{n} \text{tr} [\widehat{\Sigma}^2 (\widehat{\Sigma} + \lambda I)^{-2}] = \frac{\sigma^2}{\lambda n} \text{tr} [\widehat{\Sigma} \lambda \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{-2}] \leq \frac{\sigma^2 \text{tr} \widehat{\Sigma}}{2\lambda n}.$$

Plugging in λ^* (which was chosen to minimize the upper bound on $B + V$) gives the result. ■

- Observe that if we write $R = \max_{i \in \{1, \dots, n\}} \|\varphi(x_i)\|_2$, then we have

$$\text{tr}(\widehat{\Sigma}) = \sum_{j \geq 1} \widehat{\Sigma}_{jj} = \frac{1}{n} \sum_{i=1}^n \sum_{j \geq 1} \varphi(x_i)_j^2 = \frac{1}{n} \sum_{i=1}^n \|\varphi(x_i)\|_2^2 \leq R^2.$$

Thus in the excess risk bound, the dimension d plays no role and it could even be infinite (given that R and $\|\theta_*\|_2$ remain finite). This type of bounds are called *dimension-free* bounds.



The number of parameters is not the only way to measure the generalization capabilities of a learning method.

- Comparing this bound with that of the OLS estimator, we see that it converges slower to 0 as a function of n (from n^{-1} to $n^{-1/2}$) but it has a milder dependence on the noise (from σ^2 to σ). The presence of a “fast” rate in $O(n^{-1})$ with a potentially large constant, and of “slow” rate $O(n^{-1/2})$ with a smaller constant will appear several time in this course.

⚠ Depending on n and the constant, the “fast” rate result is not always the best.

- The value of λ^* involves quantities which we typically do not know in practice (such as σ and $\|\theta_*\|_2$). This is still useful to highlight the existence of some λ with good predictions (which can be found by cross-validation).
- Note here that the choice of $\lambda^* = \frac{\sigma\sqrt{\text{tr}(\hat{\Sigma})}}{\|\theta_*\|_2\sqrt{n}}$ is optimizing the *upper-bound* $\frac{\lambda}{2}\|\theta_*\|_2^2 + \frac{\sigma^2\text{tr}\hat{\Sigma}}{2\lambda n}$, and is thus typically not optimal for the true expected risk.
- ⚠ Check homogeneity!

Choosing λ in practice. The regularization λ is an example of a *hyper-parameter*. This term refers broadly to any quantity that influences the behavior of a machine learning algorithm and that is left to choose by the practitioner. While theory often offers guidelines and qualitative understanding on how to best chose the hyper-parameters, their precise numerical value depends on quantities which are often difficult to know or even guess. In practice, we typically resort to validation and cross-validation.

- **Exercise:** Compute the expected risk of the estimator obtained by regularizing by $\theta^\top \Lambda \theta$, where $\Lambda \in \mathbb{R}^{d \times d}$ is a positive definite matrix.

7 Lower-bound (◆)

In order to show a lower bound in the fixed design setting, we will consider only Gaussian noise, that is, ε has a joint Gaussian distribution with mean zero and covariance matrix $\sigma^2 I$ (adding an extra assumption can only make the lower bound smaller). We follow the elegant and simple proof technique outlined by [2].

The only uncertainty in the model is the location of θ_* . In order to make the dependence on θ_* explicit, we denote by $\mathcal{R}_{\theta_*}(\theta)$ the risk (in the previous lecture, we were using the notation \mathcal{R}_{dp} to make the dependence on the distribution explicit), which is equal to

$$\mathcal{R}_{\theta_*}(\theta) = \|\theta - \theta_*\|_{\Sigma}^2.$$

Our goal is to lower bound

$$\sup_{\theta_* \in \mathbb{R}^d} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \mathcal{R}_{\theta_*}(\mathcal{A}(\Phi \theta_* + \varepsilon)),$$

over all functions \mathcal{A} from \mathbb{R}^n to \mathbb{R}^d (these functions are allowed to depend on the observed deterministic quantities such as Φ). Indeed, algorithms take $y = \Phi \theta_* + \varepsilon \in \mathbb{R}^n$ as an input and outputs a vector of parameters in \mathbb{R}^d .

The main idea, which is classical in the Bayesian analysis of learning algorithms, is to lower bound the supremum by the expectation with respect to some probability on θ_* , called the prior distribution in Bayesian statistics. That is, we have, for any algorithm / estimator \mathcal{A} :

$$\sup_{\theta_* \in \mathbb{R}^d} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \mathcal{R}_{\theta_*}(\mathcal{A}(\Phi \theta_* + \varepsilon)) \geq \mathbb{E}_{\theta_* \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda n} I)} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \mathcal{R}_{\theta_*}(\mathcal{A}(\Phi \theta_* + \varepsilon)).$$

Here, we choose the normal distribution with mean 0 and covariance matrix $\frac{\sigma^2}{\lambda n} I$ as a prior distribution.

Using the expression of the excess risk (and ignoring the additive constant $\sigma^2 = \mathcal{R}^*$), we thus get the lower bound

$$\mathbb{E}_{\theta_* \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda n} I)} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \|\mathcal{A}(\Phi \theta_* + \varepsilon) - \theta_*\|_{\Sigma}^2,$$

which we need to minimize with respect to \mathcal{A} . By making θ_* random, we now have a joint Gaussian distribution for (θ_*, ε) . The joint distribution of $(\theta_*, y) = (\theta_*, \Phi \theta_* + \varepsilon)$ is also Gaussian with mean zero and covariance matrix

$$\begin{pmatrix} \frac{\sigma^2}{\lambda n} I & \frac{\sigma^2}{\lambda n} \Phi^\top \\ \frac{\sigma^2}{\lambda n} \Phi & \frac{\sigma^2}{\lambda n} \Phi \Phi^\top + \sigma^2 I \end{pmatrix} = \frac{\sigma^2}{\lambda n} \begin{pmatrix} I & \Phi^\top \\ \Phi & \Phi \Phi^\top + n \lambda I \end{pmatrix}.$$

We need to perform a similar operation as for computing the Bayes predictor in Lecture 1. This will be done by conditioning on y , by writing

$$\begin{aligned} \mathbb{E}_{\theta_* \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda n} I)} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \|\mathcal{A}(\Phi \theta_* + \varepsilon) - \theta_*\|_{\Sigma}^2 &= \mathbb{E}_{(\theta_*, y)} \|\mathcal{A}(y) - \theta_*\|_{\Sigma}^2 \\ &= \int_{\mathbb{R}^n} \left(\int_{\mathbb{R}^d} \|\mathcal{A}(y) - \theta_*\|_{\Sigma}^2 dp(\theta_* | y) \right) dp(y). \end{aligned}$$

Thus, for each y , the optimal $\mathcal{A}(y)$ has to minimize $\int_{\mathbb{R}^d} \|\mathcal{A}(y) - \theta_*\|_{\Sigma}^2 dp(\theta_* | y)$, which is exactly the posterior mean of θ_* given y . Indeed, the vector that minimizes the expected squared deviation is the expectation (exactly like when we computed the Bayes predictor for regression), here applied to the distribution $dp(\theta_* | y)$.

Since the joint distribution of (θ_*, y) is Gaussian with known parameters, we could use classical results about conditioning for Gaussian vectors², but we can also use the property that for Gaussian variables, the posterior mean given y is equal to the posterior mode given y , that is, it can be obtained by maximizing the log-likelihood $\log p(\theta_*, y)$ with respect to θ_* . Up to constants and using independence of ε and θ_* , this log-likelihood is

$$-\frac{1}{2\sigma^2}\|\varepsilon\|^2 - \frac{\lambda n}{2\sigma^2}\|\theta_*\|_2^2 = -\frac{1}{2\sigma^2}\|y - \Phi\theta_*\|^2 - \frac{\lambda n}{2\sigma^2}\|\theta_*\|_2^2,$$

which is exactly (up to a sign and a constant) the ridge regression cost function. Thus, we have: $\mathcal{A}^*(y) = (\Phi^\top \Phi + n\lambda I)^{-1} \Phi^\top y$, which is exactly the ridge regression estimator $\hat{\theta}_\lambda$, and we can compute the corresponding optimal risk, to get:

$$\begin{aligned} & \inf_A \sup_{\theta_* \in \mathbb{R}^d} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \mathcal{R}_{\theta_*}(\mathcal{A}(\Phi\theta_* + \varepsilon)) - \mathcal{R}^* \\ & \geq \inf_A \mathbb{E}_{\theta_* \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda n} I)} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \mathcal{R}_{\theta_*}(\mathcal{A}(\Phi\theta_* + \varepsilon)) - \mathcal{R}^* \\ & = \mathbb{E}_{\theta_* \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda n} I)} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \mathcal{R}_{\theta_*}(\mathcal{A}^*(\Phi\theta_* + \varepsilon)) - \mathcal{R}^* \\ & = \mathbb{E}_{\theta_* \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda n} I)} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \|\mathcal{A}^*(\Phi\theta_* + \varepsilon) - \theta_*\|_{\hat{\Sigma}}^2 \\ & = \mathbb{E}_{\theta_* \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda n} I)} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \|(\Phi^\top \Phi + n\lambda I)^{-1} \Phi^\top (\Phi\theta_* + \varepsilon) - \theta_*\|_{\hat{\Sigma}}^2 \\ & = \mathbb{E}_{\theta_* \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda n} I)} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \|(\Phi^\top \Phi + n\lambda I)^{-1} \Phi^\top \varepsilon - n\lambda (\Phi^\top \Phi + n\lambda I)^{-1} \theta_*\|_{\hat{\Sigma}}^2 \\ & = \mathbb{E}_{\theta_* \sim \mathcal{N}(0, \frac{\sigma^2}{\lambda n} I)} \| -n\lambda (\Phi^\top \Phi + n\lambda I)^{-1} \theta_* \|_{\hat{\Sigma}}^2 + \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \|(\Phi^\top \Phi + n\lambda I)^{-1} \Phi^\top \varepsilon\|_{\hat{\Sigma}}^2 \text{ by independence,} \\ & = \frac{\sigma^2}{n\lambda} (n\lambda)^2 \text{tr} [(\hat{\Sigma} + \lambda I)^{-2} \hat{\Sigma}] + \frac{\sigma^2}{n} \text{tr} [(\hat{\Sigma} + \lambda I)^{-2} \hat{\Sigma}^2] \\ & = \frac{\sigma^2}{n} \text{tr} [(\hat{\Sigma} + \lambda I)^{-1} \hat{\Sigma}] \end{aligned}$$

This risk tends to d when λ tends to zero. This such show that

$$\inf_A \sup_{\theta_* \in \mathbb{R}^d} \mathbb{E}_{\varepsilon \sim \mathcal{N}(0, \sigma^2 I)} \mathcal{R}_{\theta_*}(\mathcal{A}(\Phi\theta_* + \varepsilon)) \geq \frac{\sigma^2 d}{n}.$$

This gives us a lower-bound on performance, which exactly matches the upper-bound obtained by OLS. In the general non least-squares case, such results are significantly harder to show. See Lectures 3 and 6.

²See https://en.wikipedia.org/wiki/Multivariate_normal_distribution#Conditional_distributions.

8 Random design analysis

In this section, we consider the regular random design setting, that is, both x and y are considered random, and each pair (x_i, y_i) is assumed independent and identically distributed from a distribution $dp(x, y)$. Our goal is to show that the bound on the the excess risk that we have shown for the fixed design setting, namely $\sigma^2 d/n$, is still valid. We will make the following assumptions regarding the joint distribution $dp(x, y)$, transposed from the fixed design setting to the random design setting:

- there exists a vector $\theta_* \in \mathbb{R}^d$ such that the relationship between input and output is

$$y = \varphi(x)^\top \theta_* + \varepsilon.$$

- ε is independent from x , and $\mathbb{E}[\varepsilon] = 0$ and with variance $\mathbb{E}[\varepsilon^2] = \sigma^2$.

With the assumption above, $\mathbb{E}(y|x) = \varphi(x)^\top \theta_*$, and thus, we perform empirical risk minimization where our class of functions includes the Bayes predictor, a situation that is often referred to as the *well-specified* setting. The risk also has a simple expression:

Proposition 9 *Under the linear model above, for any $\theta \in \mathbb{R}^d$, the excess risk is equal to:*

$$\mathcal{R}(\theta) - \mathcal{R}^* = \|\theta - \theta_*\|_\Sigma^2$$

where $\Sigma := \mathbb{E}[\varphi(x)\varphi(x)^\top]$ is the (non-centered) covariance matrix, and $\mathcal{R}^* = \sigma^2$.

Proof We have:

$$\begin{aligned} \mathcal{R}(\theta) &= \mathbb{E}[(y - \theta^\top \varphi(x))^2] \\ &= \mathbb{E}[(\varphi(x)^\top \theta_* + \varepsilon - \theta^\top \varphi(x))^2] \\ &= \mathbb{E}[(\varphi(x)^\top \theta_* - \theta^\top \varphi(x))^2] + \mathbb{E}[\varepsilon^2] \\ &= (\theta - \theta_*)^\top \Sigma (\theta - \theta_*) + \sigma^2, \end{aligned}$$

which leads to the desired result. ■

Note that the only difference with the fixed design setting is the replacement of $\hat{\Sigma}$ by Σ . We can now express the risk of the OLS estimator.

Proposition 10 *Under the linear model above, assuming $\hat{\Sigma}$ is invertible, the expected excess risk of the OLS estimator is equal to*

$$\frac{\sigma^2}{n} \mathbb{E}[\text{tr}(\Sigma \hat{\Sigma}^{-1})].$$

Proof Since the OLS estimator is equal to $\hat{\theta} = \frac{1}{n} \hat{\Sigma}^{-1} \Phi^\top y = \frac{1}{n} \hat{\Sigma}^{-1} \Phi^\top (\Phi \theta_* + \varepsilon) = \theta_* + \frac{1}{n} \hat{\Sigma}^{-1} \Phi^\top \varepsilon$, we have:

$$\begin{aligned} \mathbb{E}[\mathcal{R}(\hat{\theta})] - \mathcal{R}^* &= \mathbb{E}\left[\left(\frac{1}{n} \hat{\Sigma}^{-1} \Phi^\top \varepsilon\right)^\top \Sigma \left(\frac{1}{n} \hat{\Sigma}^{-1} \Phi^\top \varepsilon\right)\right] \\ &= \mathbb{E}\left[\text{tr}\left(\Sigma \left(\frac{1}{n} \hat{\Sigma}^{-1} \Phi^\top \varepsilon\right) \left(\frac{1}{n} \hat{\Sigma}^{-1} \Phi^\top \varepsilon\right)^\top\right)\right] = \frac{1}{n^2} \mathbb{E}\left[\text{tr}\left(\Sigma \hat{\Sigma}^{-1} \Phi^\top \varepsilon \varepsilon^\top \Phi \hat{\Sigma}^{-1}\right)\right] \\ &= \frac{1}{n^2} \mathbb{E}\left[\text{tr}\left(\Sigma \hat{\Sigma}^{-1} \Phi^\top \mathbb{E}[\varepsilon \varepsilon^\top] \Phi \hat{\Sigma}^{-1}\right)\right] = \frac{\sigma^2}{n^2} \text{tr}\left(\Sigma \hat{\Sigma}^{-1} \Phi^\top \Phi \hat{\Sigma}^{-1}\right) = \frac{\sigma^2}{n} \text{tr}(\Sigma \hat{\Sigma}^{-1}). \end{aligned}$$

■

Thus, to compute the expected risk of the OLS estimator, we need to compute $\mathbb{E}[\text{tr}(\Sigma\widehat{\Sigma}^{-1})]$. One difficulty here is the potential non-invertibility of $\widehat{\Sigma}$. Under simple assumptions (e.g., $\varphi(x)$ has a density on \mathbb{R}^d), as soon as $n > d$, $\widehat{\Sigma}$ is almost surely invertible, however its smallest eigenvalue can be very small. Extra assumptions are then needed to control it (see, e.g., [2, Section 3]).

8.1 Gaussian designs

If we assume that $\varphi(x)$ is normally distributed with mean 0 and covariance matrix Σ , then we can directly compute the desired expectation, by first considering $z = \Sigma^{-1/2}\varphi(x)$, which has a standard normal distribution (that is, with mean zero and identity covariance matrix), with the corresponding normalized design matrix $Z \in \mathbb{R}^{n \times d}$, and compute $\mathbb{E}[\text{tr}(Z^\top Z)^{-1}]$.

Note that $\mathbb{E}[Z^\top Z] = nI$, and by convexity of the function $M \mapsto \text{tr}(M^{-1})$ on the cone of positive definite matrices, and using Jensen's inequality, we see that $\mathbb{E}[\text{tr}(Z^\top Z)^{-1}] \geq \frac{d}{n}$ (here we have not used the Gaussian assumption). However, this bound is in the incorrect direction (this happens a lot with Jensen's inequality).

It turns out that for Gaussians, the matrix $(Z^\top Z)^{-1}$ has a specific distribution, called the inverse Wishart distribution³, with an expectation that can be computed exactly as $\mathbb{E}[(Z^\top Z)^{-1}] = \frac{1}{n-d+1}I$. Thus, we have: $\mathbb{E}[\text{tr}(Z^\top Z)^{-1}] = \frac{d}{n-d+1}$ if $n > d + 1$, thus leading to the expected excess risk of

$$\frac{\sigma^2 d}{n-d-1} = \frac{\sigma^2 d}{n} \frac{1}{1-(d+1)/n}.$$

See [3] for further details. Note here that for Gaussian designs, the expected risk is exactly equal to the expression above, and that this will only be upper-bounds later in this course.

We see that we have an explicit non-asymptotic bound on the risk, which is equivalent to $\sigma^2 d/n$ when n goes to infinity.

8.2 General designs (◆◆)

In this last more technical section, we highlight how the Gaussian assumption can be avoided. The main idea is to show that with high probability, the lowest eigenvalue of $\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}$ is larger than some $1-t$, for some $t \in (0, 1)$. Since the excess risk is $\frac{\sigma^2}{n} \text{tr}(\Sigma\widehat{\Sigma}^{-1})$, this immediately shows that with high probability, the excess risk is less than $\frac{\sigma^2 d}{n} \frac{1}{1-t}$.

In order to obtain such results, more refined concentration inequalities are needed, such as described in [4], [5], [6], and [7]. The sharpest known results for least-squares regression can be found in [2].

Matrix concentration inequality. We will use the matrix Bernstein bound, adapted from [4, Theorem 1.4]. Note the similarity with the regular Bernstein bound for scalars when $d = 1$.

³See https://en.wikipedia.org/wiki/Inverse-Wishart_distribution.

Proposition 11 (matrix Bernstein bound) Given n independent symmetric matrices $M_i \in \mathbb{R}^{d \times d}$, such that for all $i \in \{1, \dots, n\}$, $\mathbb{E}[M_i] = 0$, $\lambda_{\max}(M_i) \leq b$ almost surely. Then for all $t \geq 0$,

$$\mathbb{P}\left(\lambda_{\max}\left(\frac{1}{n} \sum_{i=1}^n M_i\right) \leq t\right) \leq d \cdot \exp\left(-\frac{nt^2/2}{\tau^2 + bt/3}\right),$$

for $\tau^2 = \lambda_{\max}\left(\frac{1}{n} \sum_{i=1}^n M_i^2\right)$.

Application to re-scaled covariance matrices. Given $\Sigma = \mathbb{E}[\varphi(x)\varphi(x)^\top] \in \mathbb{R}^{d \times d}$, we consider the random vector $z = \Sigma^{-1/2}\varphi(x) \in \mathbb{R}^d$, which is such that $\mathbb{E}[zz^\top] = I$ and $\mathbb{E}[\|z\|_2^2] = d$.

We make the extra assumption that $\lambda_{\max}\left(\mathbb{E}\left[\|z\|_2^2 zz^\top\right]\right) \leq \rho d$, for ρ a constant. A sufficient condition is that almost surely $\|z\|_2^2 \leq \rho d$. For a Gaussian distribution with zero mean, one can check as an exercise that $\rho = (1 + 2/d)$.

We consider the random symmetric matrix $M_i = I - z_i z_i^\top$, which is such that $\mathbb{E}M_i = 0$, $\lambda_{\max}(M_i) \leq 1$ almost surely, and $\mathbb{E}[M_i^2] = \mathbb{E}\left[\|z_i\|_2^2 z_i z_i^\top\right] - I$ with largest eigenvalue less than ρd . We thus have for any $t \geq 0$,

$$\mathbb{P}\left(\lambda_{\max}\left(I - \frac{1}{n} Z^\top Z\right) \geq t\right) \leq d \cdot \exp\left(-\frac{nt^2/2}{\rho d + t/3}\right).$$

Thus, if t is such that $\frac{nt^2}{2\rho d + t/3} \geq \log \frac{d}{\delta}$, then, with probability greater than $1 - \delta$, we have $I - \Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} \preceq tI$, that is

$$\Sigma^{-1/2} \widehat{\Sigma} \Sigma^{-1/2} \succeq (1 - t)I,$$

and thus the risk is less than $\frac{\sigma^2 d}{n} \frac{1}{1-t}$, which is the desired result. We have used the order between symmetric matrices, defined as $A \succeq B \Leftrightarrow B \preceq A \Leftrightarrow A - B$ positive semi-definite.

This is possible when $t \geq \sqrt{\frac{2\rho d}{n} \log \frac{d}{\delta}} + \frac{2}{3n} \log \frac{d}{\delta}$. The bound is non-vacuous only when $t < 1$, that is, it is sufficient to impose $\frac{2}{3n} \log \frac{d}{\delta} < 1/2$ and $\sqrt{\frac{2\rho d}{n} \log \frac{d}{\delta}} < 1/2$, which is equivalent to $n \geq \frac{4}{3} \log \frac{d}{\delta}$, and $n \geq 8\rho d \log \frac{d}{\delta}$. Without surprise, the bound is non vacuous only for $n \geq d$.

Acknowledgements

These class notes have been adapted from the notes of many colleagues I have the pleasure to work with, in particular L ena ic Chizat, Pierre Gaillard, Alessandro Rudi and Simon Lacoste-Julien. Special thanks to Jaouad Mourtada for his help on lower bounds and random design analysis.

References

- [1] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.
- [2] Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. Technical Report 1912.10754, arXiv, 2019.

- [3] Leo Breiman and David Freedman. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78(381):131–136, 1983.
- [4] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.
- [5] Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on Learning Theory*, 2012.
- [6] Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms, with applications to ordinary least squares and restricted eigenvalue properties. Technical Report 1312.2903, arXiv, 2013.
- [7] Guillaume Lecué and Shahar Mendelson. Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22(3):1520–1534, 2016.