Chapter 3

Linear least-squares regression

Chapter summary

– Ordinary least-squares estimator: least-squares regression with linearly parameterized predictors leads to a linear system of size d (the number of predictors).

– Guarantees in the fixed design setting with no regularization: when the inputs are assumed deterministic and $d < n$, the excess risk is equal to $\sigma^2 d/n$.

– Ridge regression: with ℓ_2-regularization, excess risk bounds become dimension independent and allow high-dimensional feature vectors where $d > n$.

– Guarantees in the random design setting: although they are harder to show, they have a similar form.

– Lower bound of performance: under well-specification, the rate $\sigma^2 d/n$ is unimprovable.

3.1 Introduction

In this chapter, we introduce and analyze linear least-squares regression, a tool that can be traced back to Legendre (1805) and Gauss (1809).[1]

Why should we study linear least-squares regression? Has there not been 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 with no regularization or on dimension-less quantities when regularization is added.

- Because of its simplicity, many results can be easily derived without the need for

[1] see https://en.wikipedia.org/wiki/Least_squares for an interesting discussion and the claim that Gauss knew about it already in 1795.

complicated mathematics, both in terms of algorithms and statistical analysis (simple linear algebra for the simplest results in the fixed design setting).

- Using non-linear features, it can be extended to arbitrary non-linear predictions (see kernel methods in Chapter 7).

In subsequent chapters, we will extend many of these results beyond least-squares with the proper additional mathematical tools.

3.2 Least-squares framework

We recall the goal of supervised machine learning from Chapter 2: 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 assume that $\mathcal{Y} = \mathbb{R}$ and we use the square loss $\ell(y, z) = (y - z)^2$, for which we know from the previous chapter that the optimal predictor is $f^*(x) = \mathbb{E}[y|x]$.

In this chapter, we consider empirical risk minimization. We choose a parameterized family of prediction functions (often referred to as "models") $f_\theta : \mathcal{X} \to \mathcal{Y} = \mathbb{R}$ for some parameter $\theta \in \Theta$ and minimize the empirical risk

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

leading to the estimator $\hat{\theta} \in \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2$. 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 (such as for neural networks in Chapter 9). In this chapter, 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 in 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. If $f_\theta(x)$ is linear in $\theta \in \mathbb{R}^d$, then it has to be a linear combination of the form $f_\theta(x) = \sum_{i=1}^{d} \alpha_i(x)\theta_i$, where $\alpha_i : \mathcal{X} \to \mathbb{R}$, $i = 1, \dots, d$, are d functions. By concatenating them in a vector $\varphi(x) \in \mathbb{R}^d$ where $\varphi(x)_i = \alpha_i(x)$, we get the representation

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

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 the empirical risk

$$\hat{\mathcal{R}}(\theta) := \frac{1}{n}\sum_{i=1}^{n}(y_i - \varphi(x_i)^\top \theta)^2. \tag{3.1}$$

When $\mathcal{X} \subset \mathbb{R}^d$, we can make the extra assumptions that f_θ is an affine function, which could be obtained through $\varphi(x) = \binom{x}{1} = (x^\top, 1)^\top \in \mathbb{R}^{d+1}$. Other classical assumptions are $\varphi(x)$ composed of monomials (so that prediction functions are polynomials). We will see in Chapter 7 (kernel methods) that we can consider infinite-dimensional features.

Matrix notation. The cost function above in Eq. (3.1) 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, whose rows are $\varphi(x_i)^\top$. It is called the *design* matrix or *data* matrix. In these notations, the empirical risk is

$$\hat{\mathcal{R}}(\theta) = \frac{1}{n}\|y - \Phi\theta\|_2^2, \tag{3.2}$$

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 lengthy and error-prone formulas.

3.3 Ordinary least-squares (OLS) estimator

We assume 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 to be "over-determined," and we must have $d \leqslant n$, that is, more observations than feature dimension. Equivalently, we assume that $\Phi^\top\Phi \in \mathbb{R}^{d\times d}$ is invertible.

Definition 3.1 (OLS) *When Φ has full column rank, the minimizer of Eq. (3.2) is unique and called the* ordinary least-squares *(OLS) estimator.*

3.3.1 Closed-form solution

Since the objective function is quadratic, the gradient will be linear, and zeroing it will lead to a closed-form solution.

Proposition 3.1 *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[2]) empirical covariance *matrix by $\widehat{\Sigma} := \frac{1}{n}\Phi^\top\Phi \in \mathbb{R}^{d\times d}$; we have $\hat{\theta} = \frac{1}{n}\widehat{\Sigma}^{-1}\Phi^\top y$.*

[2] The "centered" covariance matrix would be $\frac{1}{n}\sum_{i=1}^n[\varphi(x_i)-\mu][\varphi(x_i)-\mu]^\top$ where $\mu = \frac{1}{n}\sum_{i=1}^n \varphi(x_i) \in \mathbb{R}^d$ is the empirical mean, while we consider $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \varphi(x_i)\varphi(x_i)^\top$.

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 $\hat{\mathcal{R}}'(\hat{\theta}) = 0$ where $\hat{\mathcal{R}}'(\theta) \in \mathbb{R}^d$ is the gradient of $\hat{\mathcal{R}}$ at θ. For all $\theta \in \mathbb{R}^d$, we have, by expanding the square and computing the gradient:

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

The condition $\hat{\mathcal{R}}'(\hat{\theta}) = 0$ 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 the uniqueness of the minimizer is by showing that $\hat{\mathcal{R}}$ is strongly convex since $\hat{\mathcal{R}}''(\theta) = 2\widehat{\Sigma}$ is invertible for all $\theta \in \mathbb{R}^d$ (convexity will be studied in Chapter 5).

⚠ For readers worried about carrying a factor of two in the gradients, we will use an additional factor 1/2 in chapters on optimization (e.g., Chapter 5).

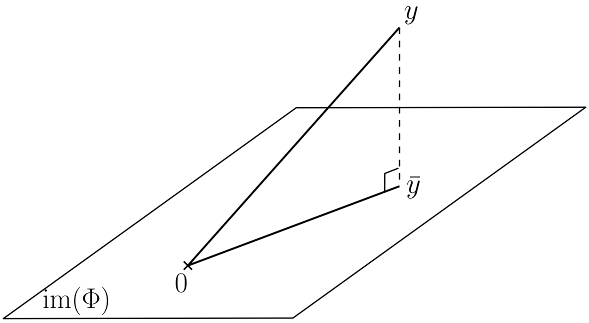
3.3.2 Geometric interpretation

Proposition 3.2 *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 $\mathrm{im}(\Phi) \subset \mathbb{R}^n$, 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 $\mathrm{im}(\Phi)$. 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 $u \in \mathrm{im}(\Phi)$. Also, since $\mathrm{im}(\Phi)^\perp = \mathrm{null}(\Phi^\top)$, $Pu' = 0$ for all $u' \in \mathrm{im}(\Phi)^\perp$. These properties characterize the orthogonal projection on $\mathrm{im}(\Phi)$. ■

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.



3.3.3 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, that is, $Q^\top Q = I$, and $R \in \mathbb{R}^{d\times d}$ is upper triangular (see Golub and Loan, 1996). Computing a QR decomposition is faster and more stable than inverting a matrix. We then have $\Phi^\top\Phi = R^\top Q^\top QR = R^\top R$, and R is thus the Cholesky factor of $\Phi^\top\Phi$. One then 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 (see Golub and Loan, 1996, for details).

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

$$\theta_t = \theta_{t-1} - \gamma\hat{\mathcal{R}}'(\theta_{t-1}) \quad \text{for } t \geqslant 1,$$

where $\gamma > 0$ is the step-size. When these iterates converge, it is towards the OLS estimator since a fixed-point θ satisfies $\hat{\mathcal{R}}'(\theta) = 0$. We will study such algorithms in Chapter 5, with running-time complexities going down to linear in d.

3.4 Statistical analysis of OLS

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

- *Random design.* In this setting, both the inputs and the outputs are random. This is the classical setting of supervised machine learning, where the goal is *generalization* to unseen data (as in the last chapter). Since obtaining guarantees is mathematically 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 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):

$$\mathcal{R}(\theta) = \mathbb{E}_y\Big[\frac{1}{n}\sum_{i=1}^n (y_i - \varphi(x_i)^\top\theta)^2\Big] = \mathbb{E}_y\Big[\frac{1}{n}\|y - \Phi\theta\|_2^2\Big]. \tag{3.3}$$

This assumption allows a complete analysis with basic linear algebra. It is justified in some settings, e.g., when the inputs are equally spaced along 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 instead of a function from $\mathcal{X}$ to $\mathbb{R}$.

In the fixed design setting, no attempts are made to generalize to unseen input points $x \in \mathcal{X}$, and we want to estimate well a label vector y resampled from the same distribution as the observed y. The risk in Eq. (3.3) is often called the *in-sample prediction error*, and the task can be seen as "denoising" the labels.

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:

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

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

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

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.

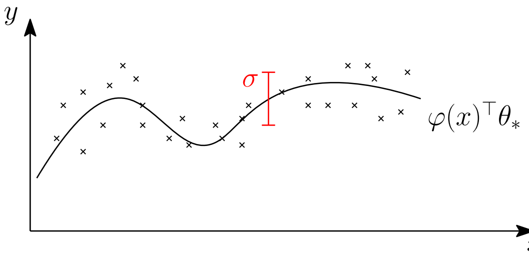
Exercise 3.1 *In the Gaussian model above, compute $\tilde{\sigma}^2$ the maximum likelihood estimator of σ^2.*

3.5 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 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 $i \in \{1, \dots, n\}$
$$y_i = \varphi(x_i)^\top \theta_* + \varepsilon_i. \tag{3.4}$$
- All noise variables ε_i, $i \in \{1, \dots, n\}$, are independent with 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 due to unobserved factors or noise. The "homoscedasticity" assumption above, where the noise variances are uniform, is made for simplicity (and allows for the later bound $\sigma^2 d/n$ to be an equality). Note that to prove upper-bounds in performance, we could also only assume that $\mathbb{E}[\varepsilon_i^2] \leqslant \sigma^2$ for each $i \in \{1, \dots, n\}$. The noise variance σ^2 is the expected squared error between the observations y_i and the model $\varphi(x_i)^\top \theta_*$.



⚠ In Eq. (3.4), we assume the model is *well-specified*, that is, the target function is a linear function of $\varphi(x)$. In general, an additional approximation error is incurred because of a misspecified model (see Chapter 4).

Denoting by $\mathcal{R}^*$ 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 it is equal to σ^2.

Proposition 3.3 (Risk decomposition for OLS - fixed design) *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 $\hat{\theta}$ is now a random variable (such as an estimator of θ_*), then

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

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

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

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*:

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

(note that this is also a simple application of the law of total variance for vectors, that is, $\mathbb{E}\big[\|z-a\|_M^2\big] = \|\mathbb{E}[z]-a\|_M^2 + \mathbb{E}\big[\|z - \mathbb{E}[z]\|_M^2\big]$ to $a = \theta_*$, $M = \widehat\Sigma$ and $z = \hat\theta$). ■

Note that 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.

3.5.1 Statistical properties of the OLS estimator

We can now analyze the properties of the OLS estimator, which has a closed form $\hat\theta = (\Phi^\top\Phi)^{-1}\Phi^\top y = \widehat\Sigma^{-1}(\frac{1}{n}\Phi^\top y)$, with the model $y = \Phi\theta_* + \varepsilon$. The only randomness comes from ε, and we thus need to compute the expectation of linear and quadratic forms in ε.

Proposition 3.4 (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 $\mathrm{var}(\hat\theta) = \mathbb{E}\big[(\hat\theta - \theta_*)(\hat\theta - \theta_*)^\top\big] = \frac{\sigma^2}{n}\widehat\Sigma^{-1}$, where $\widehat\Sigma^{-1}$ 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 $\mathbb{E}[\varepsilon\varepsilon^\top] = \sigma^2 I$, we get

$$\mathrm{var}(\hat\theta) = \mathbb{E}\big[(\Phi^\top\Phi)^{-1}\Phi^\top\varepsilon\varepsilon^\top\Phi(\Phi^\top\Phi)^{-1}\big] = \sigma^2(\Phi^\top\Phi)^{-1}(\Phi^\top\Phi)(\Phi^\top\Phi)^{-1} = \sigma^2(\Phi^\top\Phi)^{-1},$$

which leads to the desired result $\frac{\sigma^2}{n}\widehat\Sigma^{-1}$.

■

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

Proposition 3.5 (Risk of OLS) *The excess risk of the OLS estimator is equal to*

$$\mathbb{E}\big[\mathcal{R}(\hat\theta)\big] - \mathcal{R}^* = \frac{\sigma^2 d}{n}. \tag{3.5}$$

Proof Note here that the expectation is over ε only as we are in the fixed design setting. Using the risk decomposition of Proposition 3.3 and the fact that $\mathbb{E}[\hat\theta] = \theta_*$, we have

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

We have: $\mathbb{E}\big[\mathcal{R}(\hat\theta)\big] - \mathcal{R}^* = \mathrm{tr}[\mathrm{var}(\hat\theta)\widehat\Sigma] = \mathrm{tr}[\frac{\sigma^2}{n}\widehat\Sigma^{-1}\widehat\Sigma] = \frac{\sigma^2}{n}\,\mathrm{tr}(I) = \frac{\sigma^2 d}{n}$.

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

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

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

We can make the following observations:

- ⚠ In the fixed design setting, the expectation over ε appears twice: (1) in the definition of the risk of some θ in Eq. (3.3), and when taking an expectation over the data in Eq. (3.5).

  Exercise 3.2 *Show that the expected empirical risk* $\mathbb{E}[\hat{\mathcal{R}}(\hat\theta)]$ *is equal to* $\mathbb{E}[\hat{\mathcal{R}}(\hat\theta)] = \frac{n-d}{n}\sigma^2$. *In particular, when* $n > d$*, deduce that an unbiased estimator of the noise variance* σ^2 *is given by* $\frac{\|Y-\Phi\hat\theta\|_2^2}{n-d}$.

- In the exercise 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 the 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.[3]

- 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 elementary, and as we will show in Section 3.7, 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 close to n (let alone problems where $d > n$ or d much larger than n). Regularization (ridge in this chapter or with the ℓ_1-norm in Chapter 8) will come to the rescue.

- This is only for the fixed design setting. We consider the random design setting below, which is a bit more involved mathematically, primarily because of the presence of $\widehat\Sigma^{-1}$ which does not cancel anymore, leading to the term $\widehat\Sigma^{-1}\Sigma$, where Σ is the population covariance matrix.

Exercise 3.3 *(general noise) We consider the fixed design regression model* $y = \Phi\theta_* + \varepsilon$

[3] See https://en.wikipedia.org/wiki/Mallows's_Cp.

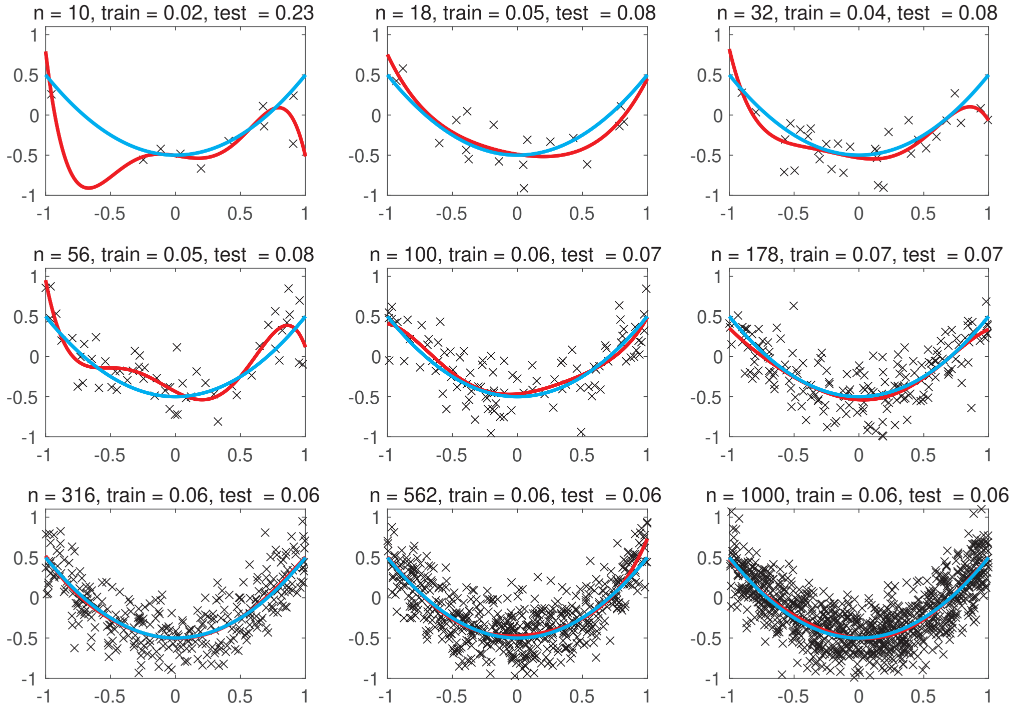


Figure 3.1: Polynomial regression with a varying number of observations. Blue: Optimal prediction, red: estimated prediction by ordinary least-squares with degree 5 polynomials.

with ε with zero mean and covariance matrix equal to C (not anymore $\sigma^2 I$). Show that the expected excess risk of the OLS estimator is equal to $\frac{1}{n}\operatorname{tr}\left[\Phi(\Phi^\top\Phi)^{-1}\Phi^\top C\right]$.

Exercise 3.4 *(♦) (multivariate regression) We consider $\mathcal{Y} = \mathbb{R}^k$ and the multivariate regression model $y = \theta_*^\top\varphi(x) + \varepsilon \in \mathbb{R}^k$, where $\theta_* \in \mathbb{R}^{d\times k}$, and ε has zero-mean with covariance matrix $C \in \mathbb{R}^{k\times k}$. In the fixed regression setting with design matrix $\Phi \in \mathbb{R}^{n\times d}$ and $Y \in \mathbb{R}^{n\times k}$* the matrix *of responses, derive the OLS estimator and its excess risk.*

3.5.2 Experiments

To illustrate the bound $\sigma^2 d/n$, 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 $f(x) = x^2 - \frac{1}{2}$ (blue curve in Figure 3.1). Gaussian noise with standard deviation $\frac{1}{4}$ is added to generate the outputs (black crosses). 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 in Figure 3.2 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

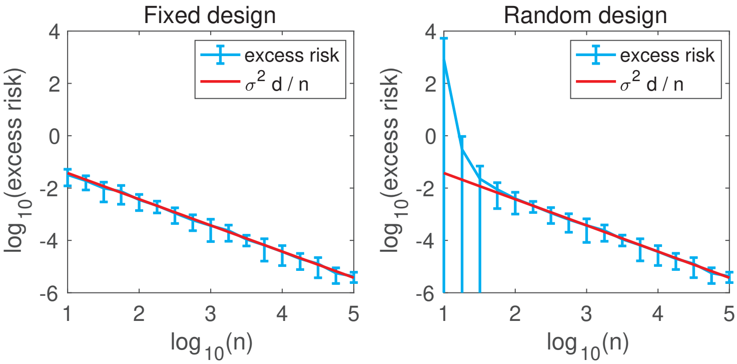


Figure 3.2: Convergence rate for polynomial regression with error bars (obtained from 32 replications by adding/subtracting standard deviations), plotted in logarithmic scale, with fixed design (left plot) and random design (right plot). The large error bars for small n in the right plot are due to the lower error bar being negative before taking the logarithm.

the random design setting (generalization error, considered in Section 3.8), 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 valid for n large enough in the random design setting.

3.6 Ridge least-squares regression

Least-squares in high dimensions. When d/n approaches 1, we are essentially memorizing the observations y_i (that is, for example, when $d = n$ and Φ is a square invertible matrix, $\theta = \Phi^{-1}y$ leads to $y = \Phi\theta$, that is, ordinary least-squares will lead to a perfect fit, which is typically not good for generalization to unseen data, see more details in Chapter 12). Also, when $d > n$, $\Phi^\top\Phi$ is not invertible, and the normal equations admit a linear subspace of solutions. These behaviors of OLS in high dimensions (d large) are often undesirable.

Two main classes of solutions exist to fix these issues: dimension reduction and regularization. Dimension reduction aims to replace the feature vector $\varphi(x)$ with another feature of lower dimension, with a classical method being principal component analysis, presented in Section 3.9. Regularization adds a term to the least-squares objective, typically either an ℓ_1-penalty $\|\theta\|_1$ (leading to *"Lasso"* regression, see Chapter 8) or $\|\theta\|_2^2$ (leading to *ridge* regression, as done in this chapter and also in Chapter 7).

Definition 3.2 (Ridge least-squares regression) *For a regularization parameter $\lambda > 0$, we define the ridge least-squares estimator $\hat{\theta}_\lambda$ as the minimizer of*

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

The ridge regression estimator can be obtained in closed form, and we do not require anymore $\Phi^\top \Phi$ to be invertible.

Proposition 3.6 *We recall that $\widehat{\Sigma} = \frac{1}{n}\Phi^\top \Phi \in \mathbb{R}^{d\times d}$. We have $\hat{\theta}_\lambda = \dfrac{1}{n}(\widehat{\Sigma} + \lambda I)^{-1}\Phi^\top y$.*

Proof As for the proof of Proposition 3.1, we can compute the gradient of the objective function, which is equal to $\frac{2}{n}\big(\Phi^\top \Phi\theta - \Phi^\top y\big) + 2\lambda\theta$. Setting it to zero leads to the estimator. Note that when $\lambda > 0$, the linear system always has a unique solution regardless of the invertibility of $\widehat{\Sigma}$. ■

Exercise 3.5 *Using the matrix inversion lemma (Section 1.1.3), show that the estimator above can be written $\hat{\theta}_\lambda = (\Phi^\top\Phi + n\lambda I)^{-1}\Phi^\top y = \Phi^\top(\Phi\Phi^\top + n\lambda I)^{-1} y$. What could be the computational benefits?*

As for the OLS estimator, we can analyze its statistical properties under the linear model and fixed design assumptions. See Chapter 7 for an analysis of random design and potentially infinite-dimensional features.

Proposition 3.7 *Under the linear model assumption (and for the fixed design setting), the ridge least-squares estimator $\hat{\theta}_\lambda = \frac{1}{n}(\widehat{\Sigma} + \lambda I)^{-1}\Phi^\top y$ has the following excess risk*

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

Proof We use the risk decomposition of Proposition 3.3 into a bias term B and a variance term V. 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

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

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

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

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 $V = \frac{\sigma^2}{n}\operatorname{tr}\big[\widehat{\Sigma}^2(\widehat{\Sigma}+\lambda I)^{-2}\big]$. They are plotted in Figure 3.3.

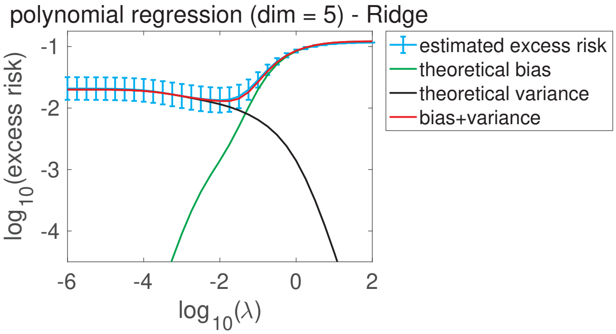


Figure 3.3: Polynomial regression (same set-up as Figure 3.2, with $n = 300$), with $k = 5$: bias/variance trade-offs for ridge regression as a function of λ. We can see the monotonicity of bias and variance with respect to λ and the presence of an optimal choice of λ.

The bias/variance decomposition can be related to the decomposition in approximation error and estimation error presented in Section 2.3.2 and further developed in Chapter 4. The bias term is the part of the excess risk due to the regularization term constraining the proper estimation of the model. It plays the role of the approximation error, while the variance term characterizes the effect of the noise and plays the role of the estimation error.

- 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 $\operatorname{tr}\big[\widehat{\Sigma}^2(\widehat{\Sigma}+\lambda I)^{-2}\big]$ is 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 $(\lambda_j)_{j\in\{1,\dots,d\}}$ are the eigenvalues of $\widehat{\Sigma}$. This quantity will be very important in analyzing kernel methods in Chapter 7. Since the function $\mu \mapsto \mu^2/(\mu+\lambda^2)$ in increasing from 0 to 1, close to zero if $\mu \ll \lambda$, close to one if $\mu \gg \lambda$, the degrees of freedom provide a soft count of the number of eigenvalues that are larger than λ.
- Observe how this converges to the OLS estimator (when defined) as $\lambda \to 0$.
- In most cases, $\lambda = 0$ is not the optimal choice; that is, biased estimation (with controlled bias) is preferable to unbiased estimation. In other words, the mean square error is minimized for a biased estimator.

Choice of λ. Based on the expression for the risk, we can tune the regularization 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 3.8 (choice of regularization parameter) *With the choice of regularization parameter* $\lambda^* = \dfrac{\sigma \operatorname{tr}(\widehat{\Sigma})^{1/2}}{\|\theta_*\|_2\sqrt{n}}$, *we have*

$$\mathbb{E}\big[\mathcal{R}(\hat{\theta}_{\lambda^*})\big] - \mathcal{R}^* \leqslant \frac{\sigma \operatorname{tr}(\widehat{\Sigma})^{1/2}\|\theta_*\|_2}{\sqrt{n}}.$$

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 $(\mu+\lambda)^{-2}\mu\lambda \leqslant 1/2 \Leftrightarrow (\mu+\lambda)^2 \geqslant 2\lambda\mu$ for all eigenvalues μ of $\widehat{\Sigma}$):

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

Similarly, we have $V = \dfrac{\sigma^2}{n}\operatorname{tr}\big[\widehat{\Sigma}^2(\widehat{\Sigma}+\lambda I)^{-2}\big] = \dfrac{\sigma^2}{\lambda n}\operatorname{tr}\big[\widehat{\Sigma}\lambda\widehat{\Sigma}(\widehat{\Sigma}+\lambda I)^{-2}\big] \leqslant \dfrac{\sigma^2\operatorname{tr}\widehat{\Sigma}}{2\lambda n}$. This leads to

$$\mathbb{E}\big[\mathcal{R}(\hat{\theta}_{\lambda^*})\big] - \mathcal{R}^* \leqslant \frac{\lambda}{2}\|\theta_*\|_2^2 + \frac{\sigma^2\operatorname{tr}\widehat{\Sigma}}{2\lambda n}. \tag{3.6}$$

Plugging in λ^* (which was chosen to minimize the upper bound on $B+V$) gives the result.[4] ■

We can make the following observations:

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

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

  Thus, the dimension d plays no explicit role in the excess risk bound and could even be infinite (given that R and $\|\theta_*\|_2$ remain finite). This type of bounds is called *dimension-free* bounds (see more details in Chapter 7).

  The number of parameters is not the only way to measure the generalization capabilities of a learning method, hence the need for explicit constants that depend on problem parameters.

- 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 a "slow" rate $O(n^{-1/2})$ with a smaller constant will appear several times in this book.

  Depending on n and the constants, the "fast" rate result is not always the best.

[4] We have used the property that for any vector u, any symmetric matrix M, and any symmetric positive semi-definite matrix A, $u^\top M u \leqslant \|u\|_2^2 \cdot \lambda_{\max}(M)$ and $\operatorname{tr}(AM) \leqslant \operatorname{tr}(A)\cdot\lambda_{\max}(M)$.

- The value of λ^* involves quantities that 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, as presented in Section 2.1).
- Note here that the choice of $\lambda^* = \frac{\sigma\sqrt{\operatorname{tr}(\widehat{\Sigma})}}{\|\theta_*\|_2\sqrt{n}}$ is optimizing the *upper-bound* $\frac{\lambda}{2}\|\theta_*\|_2^2 + \frac{\sigma^2 \operatorname{tr}\widehat{\Sigma}}{2\lambda n}$, and is thus typically not optimal for the true expected risk.
- We can check the homogeneity of the various formulas by a basic dimensional analysis. We use the bracket notation to denote the unit. Then $[\lambda] \times [\theta]^2 = [y^2] = [\sigma^2]$ since $\lambda\|\theta\|_2^2$ appears in the same objective function as y^2 (or σ^2). Moreover, we have $[y] = [\sigma] = [\varphi][\theta]$, leading to $[\lambda] = [\varphi]^2$. The value of λ suggested above has the dimension $\frac{[\varphi]\times[\sigma]}{[\theta]}$, which is indeed equal to $[\varphi]^2$. Similarly, we can check that the bias and variance terms have the correct dimensions.

Choosing λ in practice. The regularization λ is an example of a *hyper-parameter*. This term broadly refers 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 best choosing the hyper-parameters, their precise numerical value depends on quantities that are often difficult to know or even guess. In practice, we typically resort to validation and cross-validation.

Exercise 3.6 *Compute the expected risk of the estimators obtained by regularizing by $\theta^\top \Lambda \theta$ instead of $\lambda\|\theta\|_2^2$, where $\Lambda \in \mathbb{R}^{d\times d}$ is a positive definite matrix.*

Exercise 3.7 *(♦) We consider the leave-one-out estimator $\theta_\lambda^{-i} \in \mathbb{R}^d$ obtained, for each $i \in \{1, \dots, n\}$, by minimizing $\frac{1}{n}\sum_{j\neq i}(y_j - \theta^\top\varphi(x_j))^2 + \lambda\|\theta\|_2^2$. Given the matrix $H = \Phi(\Phi^\top\Phi + n\lambda I)^{-1}\Phi^\top \in \mathbb{R}^{n\times n}$, and its diagonal $h = \operatorname{diag}(H) \in \mathbb{R}^n$, show that*

$$\frac{1}{n}\sum_{i=1}^n (y_i - \varphi(x_i)^\top \theta_\lambda^{-i})^2 = \frac{1}{n}\|(I - \operatorname{Diag}(h))^{-1}(I-H)^\top y\|_2^2.$$

3.7 Lower-bound (♦)

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 Mourtada (2019).

The only unknown in the model is the location of θ_*. To make the dependence on θ_* explicit, we denote by $\mathcal{R}_{\theta_*}(\theta) - \mathcal{R}^*$ the excess risk (in the previous chapter, we were using the notation $\mathcal{R}_p$ to make the dependence on the distribution p explicit), which is equal to

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

Our goal is to lower bound

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

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 output 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 $\mathcal{A}$:

$$\sup_{\theta_*\in\mathbb{R}^d} \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)}\mathcal{R}_{\theta_*}(\mathcal{A}(\Phi\theta_*+\varepsilon)) \ \geqslant\ \mathbb{E}_{\theta_*\sim\mathcal{N}(0,\frac{\sigma^2}{\lambda n}I)}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)}\mathcal{R}_{\theta_*}(\mathcal{A}(\Phi\theta_*+\varepsilon)). \quad (3.7)$$

Here, we choose the normal distribution with mean 0 and covariance matrix $\frac{\sigma^2}{\lambda n}I$ as a prior distribution since this will lead to closed-form computations.

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

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

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 an operation similar to computing the Bayes predictor in Chapter 2. This will be done by conditioning on y by writing

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

Thus, for each y, the optimal $\mathcal{A}(y)$ has to minimize $\int_{\mathbb{R}^d}\|\mathcal{A}(y)-\theta_*\|^2_{\widehat{\Sigma}}dp(\theta_*|y)$, 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 vectors (see Section 1.1.3). Still, 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_{\mathcal{A}} \sup_{\theta_*\in\mathbb{R}^d} \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)}\big[\mathcal{R}_{\theta_*}(\mathcal{A}(\Phi\theta_* + \varepsilon))\big] - \mathcal{R}^* \\
\geqslant\ & \inf_{\mathcal{A}} \mathbb{E}_{\theta_*\sim\mathcal{N}(0,\frac{\sigma^2}{\lambda n}I)}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)}\big[\mathcal{R}_{\theta_*}(\mathcal{A}(\Phi\theta_* + \varepsilon))\big] - \mathcal{R}^* \text{ using Eq. (3.7)}, \\
=\ & \mathbb{E}_{\theta_*\sim\mathcal{N}(0,\frac{\sigma^2}{\lambda n}I)}\ \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)}\big[\mathcal{R}_{\theta_*}(\mathcal{A}^*(\Phi\theta_* + \varepsilon))\big] - \mathcal{R}^* \text{ using the reasoning above}, \\
=\ & \mathbb{E}_{\theta_*\sim\mathcal{N}(0,\frac{\sigma^2}{\lambda n}I)}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)}\big[\|\mathcal{A}^*(\Phi\theta_* + \varepsilon) - \theta_*\|_{\widehat\Sigma}^2\big] \text{ using the expression of the risk}, \\
=\ & \mathbb{E}_{\theta_*\sim\mathcal{N}(0,\frac{\sigma^2}{\lambda n}I)}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)}\big[\|(\Phi^\top\Phi + n\lambda I)^{-1}\Phi^\top(\Phi\theta_* + \varepsilon) - \theta_*\|_{\widehat\Sigma}^2\big] \\
& \qquad\qquad\qquad \text{using the closed-form expression of the OLS estimator}, \\
=\ & \mathbb{E}_{\theta_*\sim\mathcal{N}(0,\frac{\sigma^2}{\lambda n}I)}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)}\big[\|(\Phi^\top\Phi + n\lambda I)^{-1}\Phi^\top\varepsilon - n\lambda(\Phi^\top\Phi + n\lambda I)^{-1}\theta_*\|_{\widehat\Sigma}^2\big] \\
=\ & \mathbb{E}_{\theta_*\sim\mathcal{N}(0,\frac{\sigma^2}{\lambda n}I)}\big[\|-n\lambda(\Phi^\top\Phi + n\lambda I)^{-1}\theta_*\|_{\widehat\Sigma}^2\big] + \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)}\big[\|(\Phi^\top\Phi + n\lambda I)^{-1}\Phi^\top\varepsilon\|_{\widehat\Sigma}^2\big] \\
& \qquad\qquad\qquad \text{by independence}, \\
=\ & \frac{\sigma^2}{n\lambda}(n\lambda)^2\frac{1}{n^2}\operatorname{tr}\big[(\widehat\Sigma + \lambda I)^{-2}\widehat\Sigma\big] + \frac{\sigma^2}{n}\operatorname{tr}\big[(\widehat\Sigma + \lambda I)^{-2}\widehat\Sigma^2\big] \\
=\ & \frac{\sigma^2}{n}\operatorname{tr}\big[(\widehat\Sigma + \lambda I)^{-1}\widehat\Sigma\big].
\end{aligned}$$

When Φ (and thus $\widehat\Sigma$) has full rank, this last expression tends to $\frac{\sigma^2}{n}\operatorname{tr}(I) = \frac{\sigma^2 d}{n}$ when λ tends to zero (otherwise, it tends to $\frac{\sigma^2}{n}\operatorname{rank}(\Phi)$). This such shows that

$$\inf_{\mathcal{A}} \sup_{\theta_*\in\mathbb{R}^d} \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,\sigma^2 I)}\big[\mathcal{R}_{\theta_*}(\mathcal{A}(\Phi\theta_* + \varepsilon))\big] - \mathcal{R}^* \geqslant \frac{\sigma^2 d}{n}.$$

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 more general lower bounds in Chapter 15.

3.8 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 probability distribution p on $\mathcal{X} \times \mathbb{R}$. We aim to show that the bound on the excess risk 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 p, 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 for all i,
$$y_i = \varphi(x_i)^\top \theta_* + \varepsilon_i.$$
- The noise distribution of $\varepsilon_i \in \mathbb{R}$ is independent from x_i, and $\mathbb{E}[\varepsilon_i] = 0$ and with variance $\mathbb{E}[\varepsilon_i^2] = \sigma^2$ (and is the same for all i, as observations are i.i.d.).

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

Proposition 3.9 (Excess risk for random design least-squares regression) *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, for a pair (x_0, y_0) sampled from the same distribution as all (x_i, y_i), $i = 1, \dots, n$, with ε_0 the corresponding noise variable:

$$\begin{aligned} \mathcal{R}(\theta) &= \mathbb{E}\big[(y_0 - \theta^\top \varphi(x_0))^2\big] = \mathbb{E}\big[(\varphi(x_0)^\top \theta_* + \varepsilon_0 - \theta^\top \varphi(x_0))^2\big] \\ &= \mathbb{E}\big[(\varphi(x_0)^\top \theta_* - \theta^\top \varphi(x_0))^2\big] + \mathbb{E}\big[\varepsilon_0^2\big] = (\theta - \theta_*)^\top \Sigma (\theta - \theta_*) + \sigma^2, \end{aligned}$$

which leads to the desired result. ■

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

Proposition 3.10 *Under the linear model above, assuming $\widehat{\Sigma}$ is invertible, the expected excess risk of the OLS estimator is equal to*

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

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

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

■

Thus, to compute the expected risk of the OLS estimator, we need to compute $\mathbb{E}\big[\operatorname{tr}(\Sigma\widehat{\Sigma}^{-1})\big]$. 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. Additional assumptions are then needed to control it (see, e.g., Mourtada, 2019, Section 3).

Exercise 3.8 *Show that for the random design setting with the same assumptions as Prop. 3.10, the expected risk of the ridge regression estimator is*

$$\mathbb{E}\big[\mathcal{R}(\hat{\theta}_\lambda) - \mathcal{R}^*\big] = \lambda^2 \mathbb{E}\Big[\theta_*^\top(\widehat{\Sigma} + \lambda I)^{-1}\Sigma(\widehat{\Sigma} + \lambda I)^{-1}\theta_*\Big] + \frac{\sigma^2}{n}\mathbb{E}\Big[\operatorname{tr}\big[(\widehat{\Sigma} + \lambda I)^{-2}\widehat{\Sigma}\Sigma\big]\Big].$$

3.8.1 Gaussian designs

Suppose we assume that $\varphi(x)$ is normally distributed with mean 0 and covariance matrix Σ. In that case, 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 $\mathbb{E}\big[\operatorname{tr}(\Sigma\widehat{\Sigma}^{-1})\big] = n\mathbb{E}\big[\operatorname{tr}(Z^\top Z)^{-1}\big]$.

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

It turns out that for Gaussians, the matrix $(Z^\top Z)^{-1}$ has a specific distribution, called the inverse Wishart distribution[5], with an expectation that can be computed exactly as $\mathbb{E}[(Z^\top Z)^{-1}] = \frac{1}{n-d-1}I$. Thus, we have: $\mathbb{E}\big[\operatorname{tr}(Z^\top Z)^{-1}\big] = \frac{d}{n-d-1}$ 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 Breiman and Freedman (1983) for further details. Note here that for Gaussian designs, the expected risk is precisely equal to the expression above and that later in this book, we will only consider upper bounds. See also a further analysis in Section 12.2.3.

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

3.8.2 General designs (♦♦)

This last more technical section highlights how the Gaussian assumption can be avoided. The main idea is to show that with high probability, the lowest eigenvalue of $\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2}$

[5] See https://en.wikipedia.org/wiki/Inverse-Wishart_distribution.

is larger than some $1-t$, for some $t \in (0,1)$. Since the excess risk is $\frac{\sigma^2}{n}\operatorname{tr}(\Sigma\widehat{\Sigma}^{-1})$, this immediately shows that with high probability, the excess risk is less than $\frac{\sigma^2 d}{n}\frac{1}{1-t}$.

To obtain such results, more refined concentration inequalities are needed, such as described by Tropp (2012), Hsu et al. (2012), Oliveira (2013), and Lecué and Mendelson (2016). See complementary results by Mourtada (2019).

Matrix concentration inequality. We will use the matrix Bernstein bound, adapted from (Tropp, 2012, Theorem 1.4), already discussed in Section 1.2.6 and recalled here.

Proposition 3.11 (Matrix Bernstein bound) *Given n independent symmetric matrices $M_i \in \mathbb{R}^{d\times d}$, such that for all $i \in \{1,\dots,n\}$, $\mathbb{E}[M_i]=0$, $\lambda_{\max}(M_i) \leqslant b$ almost surely, for all $t \geqslant 0$, we have:*

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

for $\tau^2 = \lambda_{\max}\Big(\frac{1}{n}\sum_{i=1}^n \mathbb{E}[M_i^2]\Big)$.

Application to re-scaled covariance matrices. We can now prove the following proposition that will give the desired high-probability bound for the excess risk with one extra assumption. Below, we will use the order between symmetric matrices, defined as $A \succcurlyeq B \Leftrightarrow B \preccurlyeq A \Leftrightarrow A-B$ positive semi-definite.

Proposition 3.12 *Given $\Sigma = \mathbb{E}[\varphi(x)\varphi(x)^\top] \in \mathbb{R}^{d\times d}$, and i.i.d. observations $\varphi(x_1),\dots,\varphi(x_n) \in \mathbb{R}^d$, assume that, for some $\rho>0$,*

$$\mathbb{E}\Big[\varphi(x)^\top\Sigma^{-1}\varphi(x)\varphi(x)\varphi(x)^\top\Big] \preccurlyeq \rho d\Sigma. \tag{3.8}$$

For $\delta \in (0,1)$, if $n \geqslant 5\rho d\log\frac{d}{\delta}$, then with probability greater than $1-\delta$,

$$\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} \succcurlyeq \frac{1}{4}I. \tag{3.9}$$

Before giving the proof, note that from the discussion earlier, the bound in Eq. (3.9) leads to an excess risk less than $\frac{\sigma^2 d}{n}\frac{1}{1-t} = 4\frac{\sigma^2 d}{n}$ for $t=3/4$. Moreover, without surprise, the bound is non-vacuous only for $n \geqslant d$ (and, in fact, because of the constraint on n, more than a constant times $d\log d$). The extra assumption in Eq. (3.8) can be interpreted as follows. 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 $\mathbb{E}[\|z\|_2^2]=d$. The assumption in Eq. (3.8) is then equivalent to

$$\lambda_{\max}\Big(\mathbb{E}\Big[\|z\|^2 zz^\top\Big]\Big) \leqslant \rho d.$$

A sufficient condition is that almost surely $\|z\|_2^2 \leqslant \rho d$, that is, $\varphi(x)^\top\Sigma^{-1}\varphi(x) \leqslant \rho d$. Moreover, for a Gaussian distribution with zero mean for z, one can check as an exercise that $\rho = (1+2/d)$. Similar results will be obtained for ridge regression in Chapter 7.

Proof We consider the random symmetric matrix $M_i = I - z_i z_i^\top$, which is such that $\mathbb{E}[M_i] = 0$, $\lambda_{\max}(M_i) \leqslant 1$ almost surely, and $\mathbb{E}[M_i^2] = \mathbb{E}\big[\|z_i\|^2 z_i z_i^\top\big] - I$ with largest eigenvalue less than ρd. We thus have for any $t \geqslant 0$, using Prop. 3.11:

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

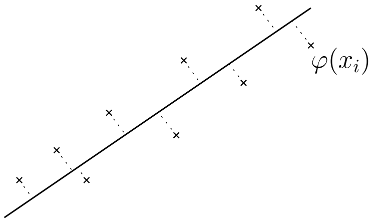
Thus, if t is such that $\frac{nt^2}{2\rho d + 2t/3} \geqslant \log\frac{d}{\delta}$, then, with probability greater than $1-\delta$, we have $I - \Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} \preccurlyeq tI$, that is, the desired result $\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} \succcurlyeq (1-t)I$.

For $t = 3/4$, the condition becomes $n \geqslant (32\rho d/9 + 8/3)\log\frac{d}{\delta}$, which is implied by $n \geqslant 5\rho d \log\frac{d}{\delta}$ since we always $\rho \geqslant 1$. ■

3.9 Principal component analysis (♦)

Unsupervised dimension reduction is an effective way of reducing the number of features, either for computational efficiency (by storing and manipulating smaller feature vectors) or to avoid overfitting in a way complementary to ridge regularization. In this section, we present principal component analysis (PCA), which corresponds to looking for a low-dimensional subspace that contains approximately all feature vectors.

We consider n feature vectors $\varphi(x_1), \dots, \varphi(x_n) \in \mathbb{R}^d$, with the corresponding design matrix $\Phi \in \mathbb{R}^{n\times d}$. PCA aims at finding a subspace of dimension k such that all feature vectors are close to their orthogonal projections onto that subspace (see an illustration below for $d = 2$ and $k = 1$, where the goal is to minimize the sum of squares of all dotted segments).



In the formulation presented below, we consider a *linear* subspace (which contains 0), but it is common in practice to look for the optimal *affine* subspace (that may not contain 0), which can be done by first centering the data, that is, subtracting the mean from all feature vectors.

Formulation as an eigenvalue problem. We can parameterize the subspace (non-uniquely) by an orthonormal basis $V \in \mathbb{R}^{d\times k}$ such that $V^\top V = I$. Then each feature vector $\varphi(x_i)$, $i = 1, \dots, n$, has projection $VV^\top\varphi(x_i)$, and thus the design matrix of all projected vectors is $\Phi V V^\top$, and the optimal V is found by minimizing:

$$\begin{aligned}\|\Phi - \Phi V V^\top\|_{\mathrm{F}}^2 &= \operatorname{tr}\big[(\Phi - \Phi V V^\top)^\top (\Phi - \Phi V V^\top)\big] \\ &= \operatorname{tr}\big[\Phi^\top \Phi\big] + \operatorname{tr}\big[V V^\top \Phi^\top \Phi V V^\top\big] - 2 \operatorname{tr}\big[\Phi^\top \Phi V V^\top\big] \\ &= \operatorname{tr}\big[\Phi^\top \Phi\big] - \operatorname{tr}\big[V^\top \Phi^\top \Phi V\big].\end{aligned}$$

Thus, minimizing $\|\Phi - \Phi V V^\top\|_{\mathrm{F}}^2$ is equivalent to maximizing $\operatorname{tr}\big[V^\top \Phi^\top \Phi V\big]$ with respect to an orthonormal matrix $V \in \mathbb{R}^{d\times k}$. Given an eigenvalue decomposition of the non-centered empirical covariance matrix $\widehat{\Sigma} = \frac{1}{n}\Phi^\top \Phi = U \operatorname{Diag}(\lambda) U^\top$, with $U \in \mathbb{R}^{d\times d}$ orthogonal and λ a vector with non-increasing components, an optimal V is obtained by taking the first k columns of U, that is, a basis of the principal subspace of dimension k. Such a basis can be computed by various algorithms from numerical algebra (Golub and Loan, 1996). See Exercise 3.9 for a simple alternative optimization algorithm.

Exercise 3.9 *Given $\Phi \in \mathbb{R}^{n\times d}$, we consider minimizing $\|\Phi - AD\|_{\mathrm{F}}^2$ with respect to $D \in \mathbb{R}^{k\times d}$ and $A \in \mathbb{R}^{n\times k}$. Show that the optimal solution is such that AD is the data matrix after performing principal component analysis. Show that an alternating minimization algorithm that iteratively minimizes $\|\Phi - AD\|_{\mathrm{F}}^2$ with respect to A and then D, converges to the global optimum for almost all initializations D, and compute the corresponding updates.*

Exercise 3.10 (K-means clustering) *Given $\Phi \in \mathbb{R}^{n\times d}$, we consider minimizing $\|\Phi - AD\|_{\mathrm{F}}^2$ with respect to $D \in \mathbb{R}^{k\times d}$ and $A \in \{0,1\}^{n\times k}$ such that each row of A sums to one. Compute the updates of an alternative optimization algorithm that minimizes $\|\Phi - AD\|_{\mathrm{F}}^2$.*

PCA and least-squares regression (♦♦). While regularization is a common way to avoid overfitting for least-squares (as shown in Section 3.6), performing PCA and then (unregularized) ordinary least-squares provides an alternative with similar behavior. That is, we now consider the feature vector $\Phi V \in \mathbb{R}^{n\times k}$, and minimize $\|y - \Phi V \eta\|_2^2$ with respect to $\eta \in \mathbb{R}^k$, with solution $\eta = (V^\top \Phi^\top \Phi V)^{-1} V^\top \Phi^\top y$, leading to the prediction vector $\Phi V \eta = \Phi V (V^\top \Phi^\top \Phi V)^{-1} V^\top \Phi^\top y \in \mathbb{R}^n$.

If we assume the linear model $y = \Phi\theta_* + \varepsilon$ like in Section 3.6, we have:

$$\begin{aligned}\frac{1}{n}\mathbb{E}_\varepsilon\big[\|\Phi V \eta - \Phi\theta_*\|_2^2\big] &= \frac{\sigma^2 k}{n} + \frac{1}{n}\big\|\Phi V (V^\top \Phi^\top \Phi V)^{-1} V^\top \Phi^\top \Phi \theta_* - \Phi\theta_*\big\|_2^2 \\ &= \frac{\sigma^2 k}{n} + \theta_*^\top \widehat{\Sigma}\theta_* - \theta_*^\top \widehat{\Sigma} V (V^\top \widehat{\Sigma} V)^{-1} \widehat{\Sigma}\theta_* \\ &= \frac{\sigma^2 k}{n} + \theta_*^\top (I - V V^\top)\widehat{\Sigma}(I - V V^\top)\theta_*,\end{aligned}$$

using the fact that the columns of V are eigenvector of $\widehat{\Sigma}$. We can now use the upper-bound

$$\leqslant \frac{\sigma^2 k}{n} + \|\theta_*\|_2^2 \lambda_{k+1},$$

where λ_{k+1} is the $(k+1)$-th largest eigenvalue of $\widehat{\Sigma}$, which is less than $1/(k+1)$ times $\mathrm{tr}[\widehat{\Sigma}]$ (the sum of all eigenvalues). Thus, the excess risk of OLS after PCA is less than $\frac{\sigma^2 k}{n} + \|\theta_*\|_2^2 \frac{\mathrm{tr}[\widehat{\Sigma}]}{k}$, which is similar to Eq. (3.6). A good value of k is then the closest integer to $\|\theta_*\|_2 \cdot \mathrm{tr}[\widehat{\Sigma}]\sqrt{n}/\sigma$, leading to, up to constants, the same excess risk than for ridge regression.

3.10 Conclusion

In this chapter, we have considered the simplest machine learning set-up, that is, square loss and prediction functions linearly parameterized by a finite-dimensional parameter. This simplest setup led to estimation algorithms based on numerical linear algebra (solving linear systems) and a statistical analysis based on simple probabilistic arguments (mostly variance computations). In particular, we highlighted the importance of regularization, which allows good predictive performance with high-dimensional features through dimension-free bounds.

Going beyond the square loss will require iterative algorithms based on optimization (presented in Chapter 5), and a more refined statistical analysis with deeper probabilistic tools (presented in Chapter 4).