Bayesian statistics

Let us go back to the problems of statistical inference that we considered in Chapter 2. We have data $X$ valued on a measurable space $(E, \mathcal{E})$ and a model $\{ P_\theta : \theta \in \Theta \}$ for its distribution, see Definition 1.1 from Chapter 1. For the problems of estimation and testing, we can define a set $A$ of decisions: for scalar estimation, it is $A = \Theta \subset \mathbb{R}$, while for testing, we have binary decisions, so that $A = \{0, 1\}$.

### 4.1 Elements of decision theory

Given a (measurable) statistical procedure $\delta : E \to A$, we decide $\delta(X) \in A$. In order to assess a decision, we use a loss function $\ell : A \times \Theta \to \mathbb{R}$. This means that if the true parameter is $\theta \in \Theta$ and if we decide $a \in A$, we incur a loss $\ell(a, \theta) \in \mathbb{R}$.

**Definition 4.1** Consider a statistical experiment with data $X \in E$ and a set of parameters $\Theta$, a set $A$ of decisions and a loss function $\ell : A \times \Theta \to \mathbb{R}$. The risk of a statistical procedure $\delta : E \to A$ is given by

$$R(\delta, \theta) = \mathbb{E}_\theta[\ell(\delta(X), \theta)]$$

for any $\theta \in \Theta$.

For the problem of estimation of a scalar parameter, we have $\Theta = \mathbb{R} = A$ and $\ell(\theta', \theta) = (\theta' - \theta)^2$, so that the risk is, in this case, the quadratic risk introduced in Definition 2.1 from Chapter 2. Note that we could consider other losses, such as $\ell(\theta', \theta) = |\theta' - \theta|^p$ for some $p \geq 1$.

Consider now statistical testing with hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ where $\{ \Theta_0, \Theta_1 \}$ is a partition of $\Theta$. Introduce the loss given by

$$\ell(i, \theta) = 0 \quad \text{if} \quad \theta \in \Theta_i \quad \text{and} \quad \ell(i, \theta) = c_i \quad \text{if} \quad \theta \in \Theta_{1-i}$$

for $i \in \{0, 1\}$ and constants $c_0, c_1 > 0$. The risk writes in this case

$$R(\delta, \theta) = c_i \mathbb{P}_\theta[\delta(X) = i] \quad \text{when} \quad \theta \in \Theta_{1-i}$$

for $i \in \{0, 1\}$. The constants $c_0, c_1 > 0$ allow to tune the importance given to the Type I and Type II errors: the approach described here leads to an approach of statistical testing different from the one described in Section 2.3.
4.2 Bayesian risk

Let us go back to the coin flip problem considered in Chapter 2. We observe \( X_1, \ldots, X_n \) iid distributed as \( \text{Bernoulli}(\theta) \) for \( \theta \in (0, 1) \). The estimator we introduced back then was the empirical mean \( \overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i \), and we know from (2.3) that the quadratic risk is given by \( R(\overline{X}_n, \theta) = \theta(1 - \theta)/n \) for any \( \theta \in (0, 1) \). But what about the estimator \( \overline{X}_n = 0 \)? It is a rather stupid estimator, but if \( \theta \) is close to 0, it turns out to be a good estimator, and actually, it is easy to see that \( R(\overline{X}_n, \theta) < R(\overline{X}_n, \theta) \) whenever \( \theta < 1/(n + 1) \). This proves that \( \overline{X}_n = 0 \) is better than \( \overline{X}_n \), when assessed by the quadratic risk, for \( \theta \) small enough.\(^1\) This very simple example illustrates the fact that it is not possible to find an estimator with an optimal risk for all \( \theta \in \Theta \).

As illustrated above, given two statistical procedures \( \delta, \delta' \) for the same problem of statistical inference, we do not have in general that \( R(\delta, \theta) < R(\delta', \theta) \) uniformly for \( \theta \in \Theta \). What we can do instead is to consider an averaged risk: choose a distribution \( \mu \) on \( \Theta \) and use it to integrate the risk over \( \Theta \). This distribution is called the prior distribution or simply the prior.

**Definition 4.2 (Bayesian risk)** The Bayesian risk of a procedure \( \delta \) associated to the prior \( \mu \) is given by

\[
R_B(\delta, \mu) = \int_{\Theta} R(\delta, \theta) \mu(d\theta) = \int_{\Theta} \mu(d\theta) \int_{E} \ell(\delta(x), \theta) P_{\theta}(dx).
\]

Note that \( R_B(\delta, \mu) \leq \sup_{\theta \in \Theta} R(\delta, \theta) \) which means that the Bayes risk is always smaller than the worst-case risk over \( \Theta \). We understand this risk as an average of the risk over \( \Theta \) “weighted” by the prior distribution \( \mu \). Given a prior \( \mu \), the Bayesian risk is a scalar value: we can compare procedures using it and even try to find a procedure that minimizes it.\(^2\)

**Example 4.1** For statistical testing with the loss given by (4.1), the Bayesian risk associated to a prior \( \mu \) writes

\[
R_B(\delta, \mu) = \sum_{i \in \{0, 1\}} c_i \int_{\Theta_{1-i}} P_{\theta}[\delta(X) = i] \mu(d\theta),
\]

which is a weighted combination of the Type I and Type II errors averaged by the prior \( \mu \).

Another interpretation of the Bayesian risk is of utmost importance in Bayesian statistics. Indeed, we could say that the parameter \( \theta \) is itself a random variable distributed as \( \mu \), that we denote \( T \) instead of \( \theta \), and that \( P_{\theta} \) is actually the distribution of \( X \) “conditionally“ on \( T = \theta \).\(^3\)

---

1: A longer story hides beneath this simple example: the Stein effect and the Stein estimator, which provably improves the sample average estimator using thresholding, see [stein1956inadmissibility, lehmann2006theory] for more details on this.

2: We will do it in Section 4.4, such an estimator is called a Bayesian estimator.

3: Of course the event \( T = \theta \) has zero probability if \( T \) is continuous with respect to the Lebesgue measure. We will explain clearly what such a conditional density is in Section 4.3 below.
Assuming that the joint distribution \( P_{T,X} \) of \((T, X)\) is given by
\[
P_{T,X}[B \times C] = \mathbb{P}[T \in B, X \in C] = \int_B \mu(d\theta) \int_C p_\theta(dx),
\]
we could write the Bayesian risk as an expectation with respect to \( P_{T,X} \), since
\[
R_B(\mu, \delta) = \int_{\Theta} \mu(d\theta) \int_E \ell(\delta(x), \theta)p_\theta(dx) = \mathbb{E}[\ell(\delta(X), T)].
\]

What we need to do now, in order to formalize this, is to explain what a conditional density is, and to explain some useful formulas, such as the Bayes formula for conditional densities.

### 4.3 Conditional densities and the Bayes formula

Let \( X \) and \( Y \) be random variables on the same probability space and valued in measurable sets \( \mathcal{X} \) and \( \mathcal{Y} \). Let \( \phi \) be a measurable function such that \( \phi(X) \) is integrable. Let us recall that we can define the conditional expectation \( \mathbb{E}[\phi(X)|Y] \) as the random variable \( r(Y) \) (for some measurable function \( r \), almost surely unique) such that
\[
\mathbb{E}[\phi(X)\varphi(Y)] = \mathbb{E}[r(Y)\varphi(Y)]
\]
for any measurable and bounded function \( \varphi \). The particular value \( r(y) \) for some \( y \in \mathcal{Y} \) is denoted \( \mathbb{E}[\phi(X)|Y = y] \). We know that
\[
\mathbb{E}[\phi(X)h(Y)|Y] = h(Y)\mathbb{E}[\phi(X)|Y]
\]
almost surely, for any measurable function \( h \) such that \( h(Y) \) and \( \phi(X)h(Y) \) are integrable, and we have also
\[
\mathbb{E}[\mathbb{E}[\phi(X)|Y]] = \mathbb{E}[\phi(X)].
\]
Finally, we have \( \mathbb{E}[\phi(X)|Y] = \mathbb{E}[\phi(X)] \) whenever \( X \) and \( Y \) are independent. Let us suppose now that the joint distribution \( P_{X,Y} \) of \((X, Y)\) has a density \( p(x, y) \) with respect to a product of dominating measures \( \nu_X \otimes \nu_Y \) on \( \mathcal{X} \times \mathcal{Y} \). We can define the marginal densities of \( X \) and \( Y \) as
\[
p_X(x) = \int_Y p(x, y)\nu_Y(dy) \quad \text{and} \quad p_Y(y) = \int_X p(x, y)\nu_X(dx),
\]
so that we have
\[
\mathbb{E}[\phi(X)] = \int_\mathcal{X} \phi(x)p_X(dx) = \int_\mathcal{X} \phi(x)p_X(x)\nu_X(dx) \\
\mathbb{E}[\varphi(Y)] = \int_\mathcal{Y} \varphi(y)p_Y(dy) = \int_\mathcal{Y} \varphi(y)p_Y(x)\nu_Y(dy)
\]
for any \( \phi \) and \( \varphi \) such that \( \phi(X) \) and \( \varphi(Y) \) are integrable. Let us introduce
\[
\mathcal{Y}_0 = \{ y \in \mathcal{Y} : p_Y(y) = 0 \}
\]
and remark that
\[
P_{X,Y}[\mathcal{X} \times \mathcal{Y}_0] = \int_{\mathcal{X} \times \mathcal{Y}_0} p(x,y) \nu_X(dx) \nu_Y(dy)
= \int_{\mathcal{Y}_0} \nu_Y(dy) \int_{\mathcal{X}} p(x,y) \nu_X(dx)
= \int_{\mathcal{Y}_0} p_Y(y) \nu_Y(dy) = 0.
\]

Using (4.4)

Given any probability density \( q \) on \( \mathcal{X} \) with respect to \( \nu_X \), we can define
\[
p_X|Y(x|y) := \frac{p(x,y)}{p_Y(y)} \mathbf{1}_{\mathcal{Y}_0^c}(y) + q(x) \mathbf{1}_{\mathcal{Y}_0}(y), \tag{4.5}
\]
so that all the versions of \( p_X|Y \) associated to the choice of \( q \) are equal \( P_{X,Y} \)-almost surely. Moreover, we can check immediately that
\[
\int_{\mathcal{X}} p_X|Y(x|y) \nu_X(dx) = 1,
\]
so that it is a probability density with respect to \( \nu_X \) on \( \mathcal{X} \). Now, if we define
\[
r'(y) = \int_{\mathcal{X}} \phi(x) p_X|Y(x|y) \nu_X(dx),
\]
we can write, for any measurable and bounded \( \varphi \), that
\[
\mathbb{E}[r'(Y) \varphi(Y)] = \int_{\mathcal{Y}} r'(y) \varphi(y) p_Y(y) \nu_Y(dy)
= \int_{\mathcal{Y}_0^c} \varphi(y) p_Y(y) \nu_Y(dy) \int_{\mathcal{X}} \phi(x) \frac{p(x,y)}{p_Y(y)} \nu_X(dx)
= \int_{\mathcal{X} \times \mathcal{Y}} \phi(x) \varphi(y) p(x,y) \nu_X(dx) \nu_Y(dy)
= \mathbb{E}[^{\prime}\!\phi(X) \varphi(Y)].
\]

This corresponds to the definition of the conditional expectation, which is almost surely unique, so that we proved that \( r = r' \) almost surely.

Now, we know that we can compute a conditional expectation using the formula
\[
\mathbb{E}[\phi(X)|Y = y] = r(y) = \int_{\mathcal{X}} \phi(x) p_X|Y(x|y) \nu_X(dx).
\]

The density \( p_X|Y \) is called the conditional density of \( X \) knowing \( Y \).

We can define in the exact same way \( p_Y|X \), the conditional density of \( Y \) knowing \( X \), and by construction of \( p_X|Y \) and \( p_Y|X \), we have that the following equalities
\[
p(x,y) = p_{X|Y}(x|y)p_Y(y) = p_{Y|X}(y|x)p_X(x) \tag{4.6}
\]
holds \( P_{X,Y} \)-almost surely. From these equalities we can deduce that

\[
p_{X|Y}(x|y) = \frac{p(x,y)}{p_Y(y)} = \frac{p_{Y|X}(y|x)p_X(x)}{\int_X p(x',y)\nu(dx')}
\]

recalling that \( P_{X,Y}[X \times Y_0] = 0 \)

holds \( P_{X,Y} \)-almost surely, which leads, using again (4.6), to the Bayes formula

\[
p_{X|Y}(x|y) = \frac{p_{Y|X}(y|x)p_X(x)}{\int_X p_{Y|X}(y|x')p_X(x')\nu_X(dx')}, \tag{4.7}
\]

that holds, once again, \( P_{X,Y} \)-almost surely. This is a remarkable formula, since it allows to reverse the conditioning: we can write the conditional density of \( X \) knowing \( Y \) as a function of the conditional density of \( Y \) knowing \( X \). This formula is at the core of Bayesian statistics, as explained in the next Section.

### 4.4 Posterior distribution and Bayes estimator

Let us go back to the setting introduced in Section 4.2. We have data \( X \) and a statistical model \( \{ P_{\theta} : \theta \in \Theta \} \). We consider a prior distribution \( \mu \) on \( \Theta \). We assume that \( \mu \) has a density \( p(\cdot) \) with respect to a measure \( \lambda \) on \( \Theta \), namely \( \mu(d\theta) = p(\theta)\lambda(d\theta) \) and that \( P_{\theta} \) has a density that we will denote as \( p(\cdot|\theta) \) with respect to a measure \( \nu \) on \( E \). We want to apply Bayesian reasoning: the density \( p(\cdot|\theta) \) of the data is understood as a conditional density of \( X \) “knowing the parameter \( \theta \”).

**The posterior distribution.** In order to formalize this, we introduce a random variable \( T \) distributed as \( \mu \), and we apply (4.6) in order to express the joint density of \((X, T)\) through the product of the conditional density of \(X|T\) and the density of \(T\):

\[
p_{X,T}(x,\theta) = p_{X|T}(x|\theta)p_T(\theta) = p(x|\theta)p(\theta). \tag{4.8}
\]

We can only proceed like this to express \( p_{X,T} \), since what we are given is the prior density \( p(\cdot) \) and the model, namely the density \( p(\cdot|\theta) \). We know that the marginal density of \( X \) can be computed as

\[
p_X(x) = \int_{\Theta} p_{X,T}(x,\theta)\lambda(d\theta) = \int_{\Theta} p(x|\theta)p(\theta)\lambda(d\theta).
\]

Now, using the Bayes formula (4.7), we can reverse the conditioning, and define what we call the posterior density

\[
p(\theta|x) := p_{T|X}(\theta|x) = \frac{p_{X,T}(x,\theta)}{p_X(x)} = \frac{p(x|\theta)p(\theta)}{\int_{\Theta} p(x|\theta')p(\theta')\lambda(d\theta')}.
\]

This formula expresses the conditional density of the parameter \( \theta \) knowing the data \( x \) (more formally the conditional density of \( T \) knowing \( X \))
through the model (the conditional density of $X$ knowing $T$) and the prior (the density of $T$) that are both known and chosen beforehand. Let us wrap-up what we constructed in the following definition.

**Definition 4.3** Consider a prior $\mu(d\theta) = p(\theta)\lambda(d\theta)$ and a model $P_x(dx) = p(x|\theta)\nu(dx)$ for $\theta \in \Theta$, and the corresponding joint distribution $P(dx, d\theta) = p(x|\theta)p(\theta)\nu(dx)\lambda(d\theta)$. The posterior distribution is the distribution with density

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int_\Theta p(x|\theta')p(\theta')\lambda(d\theta')}$$

with respect to $\lambda$. It is well-defined and unique for $P$--almost all $(x, \theta)$.

The Bayesian reasoning is therefore as follows: choose a prior density $p(\theta)$ and a model $p(x|\theta)$ for the data $x$ knowing the parameter $\theta$. Then, compute the posterior distribution $p(\theta|x)$ of $\theta$ knowing the data $x$. A nice aspect of this approach is that we can quantify uncertainty right out of the box, since instead of an estimator $\hat{\theta}_n$ (which is, given data, a single value), we obtain a full posterior distribution $p(\theta|x)$, which takes into account the data $x$ that we observed. However, such a reasoning is of course only possible when we know how to choose a prior, and when we are able to compute exactly or to approximate efficiently the posterior.

**The Bayes estimator.** Let us consider the estimation problem where $A = \Theta \subset \mathbb{R}$ and use the Bayes risk to assess the error of an estimator $\hat{\theta} : E \rightarrow \Theta$. Arguably, an optimal Bayesian estimator should minimize the Bayes risk, and a beautiful aspect of the Bayesian approach is that such a minimizer can be defined precisely. Indeed, we can rewrite the Bayes risk as follows:

$$R_B(\hat{\theta}, \mu) = \int_{\Theta} \int_E \ell(\hat{\theta}(x), \theta)p(x|\theta)p(\theta)\nu(dx)\lambda(d\theta)$$

$$= \int_E p_X(x)\nu(dx) \int_{\Theta} \ell(\hat{\theta}(x), \theta)p(\theta|x)\lambda(dx).$$

What is remarkable with this rewriting is that in order to minimize $R_B(\hat{\theta}, \mu)$, we need to minimize, for any fixed $x \in E$, the quantity

$$\int_{\Theta} \ell(\hat{\theta}(x), \theta)p(\theta|x)\lambda(d\theta) = \mathbb{E}[\ell(\hat{\theta}(X), T)|X = x],$$

namely the expectation of the loss with respect to the posterior distribution given by Definition 4.3. This leads to the following definition of a Bayes estimator.
4 Bayesian statistics

Definition 4.4 Given a prior \( \mu(d\theta) = p(\theta)\lambda(d\theta) \), a model \( P_\theta(dx) = p(x|\theta)\nu(dx) \) and a loss \( \ell \), any estimator \( \hat{\theta}(X) \) defined as

\[
\hat{\theta}(x) \in \arg\min_{\theta \in \Theta} \int_{\Theta} \ell(t, \theta)p(\theta|x)\lambda(d\theta) = \arg\min_{\theta \in \Theta} \mathbb{E}[\ell(t, T)|X = x],
\]

namely a minimizer of the expectation of the loss with respect to the posterior distribution, is called a Bayes or a Bayesian estimator.

For the square loss \( \ell(\theta', \theta) = (\theta' - \theta)^2 \), the Bayes estimator is given by the expectation of the posterior distribution. Indeed, it is easy to see that

\[
\hat{\theta}(x) = \arg\min_{t \in \mathbb{R}} \int_{\Theta} (t - \theta)^2 p(\theta|x)\lambda(d\theta) = \int_{\Theta} \theta p(\theta|x)\lambda(d\theta). \tag{4.9}
\]

If \( \ell(\theta', \theta) = |\theta' - \theta| \), we can see that

\[
\hat{\theta}(x) = \arg\min_{t \in \mathbb{R}} \int_{\Theta} |t - \theta| p(\theta|x)\lambda(d\theta) = F_x^{-1}(1/2),
\]

where \( F_x^{-1}(1/2) \) is the median of the posterior distribution. Here, the notation \( F_x^{-1} \) stands for the generalized inverse of the distribution function \( F_x(\theta) = \int_{-\infty}^{\theta} p(\theta'|x)\lambda(d\theta') \) of the posterior distribution.\(^6\)

Recipe

On some examples, we can compute explicitly the posterior distribution. Given the data density \( p(x|\theta) \) and the prior density \( p(\theta) \), the joint density of the data and the prior is \( p(x|\theta)p(\theta) \) and we know from Definition 4.3 that the posterior density is proportional to the joint density, namely

\[
p(\theta|x) = \text{constant}(x)p(x|\theta)p(\theta),
\]

where \( \text{constant}(x) = 1/\int_{\Theta} p(x|\theta)p(\theta)\lambda(d\theta) \). So, using the fact that the integral of the posterior density with respect to \( \theta \) equals 1, we can try to identify directly the posterior distribution by having a careful look at \( p(x|\theta)p(\theta) \), together with some coffee, and looking for a density with respect to \( \theta \).

4.5 Examples

Let us give some standard examples of priors and data distributions where can apply this recipe.

\(^6\): This comes from the fact that if \( X \) is an integrable real random variable with distribution function \( F_x \), a minimizer of \( t \mapsto \mathbb{E}|X - t| \) is given by the median \( F_x^{-1}(1/2) \) of \( X \).
4.5.1 How to choose a restaurant? (Bayesian coin flip)

The first example is, once again, a coin flip, but this time with a Bayesian flavor. Consider the data distribution $X \sim \text{Binomial}(n, \theta)$, namely the model with density

$$p(x|\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} 1_{\{0,...,n\}}(x)$$

with respect to the counting measure $\nu$ on $\mathbb{N}$ and the \textit{flat prior} on $\theta$ with distribution $\text{Uniform}(0, 1)$ on $\theta$, namely a density $p(\theta) = 1_{[0,1]}(\theta)$ with respect to the Lebesgue measure $\lambda$.

This model can be useful to help us choose a restaurant: given a restaurant, $\theta$ is the unknown probability that a customer is happy (1) or unhappy (0) with it and $n$ is the number of customers who gave their (binary) opinion. We have no prior information on $\theta$, so we consider the flat prior. For each restaurant, we observe the percentage of happy customers (rescaled to a 0 to 5 stars rating in Figure 4.1) and the number $n$ of customers who rated it, so $X$ stands here for the number of happy customers among the $n$ who rated it. Assuming that the customers opinions are independent, we have $X \sim \text{Binomial}(n, \theta)$. The question is the following: how can we choose a restaurant? Is a restaurant with a good rating but few rates better than a restaurant with a slightly worse rating but more rates?

Using the previous recipe, we know that the density of the posterior distribution is proportional to the density of the joint distribution of the data and prior

$$p(x, \theta) = p(x|\theta)p(\theta) = \binom{n}{x} \theta^x (1 - \theta)^{n-x} 1_{[0,1]}(\theta) 1_{\{0,...,n\}}(x)$$

with respect to the product measure $\nu \otimes \lambda$. This means that the posterior density is proportional to $\theta^x (1 - \theta)^{n-x} 1_{[0,1]}(\theta)$. We recognize the density $\theta \mapsto \frac{1}{\beta(a,b)} \theta^{a-1} (1 - \theta)^{b-1} 1_{[0,1]}(\theta)$ of the Beta$(a, b)$ distribution, that we introduced in Section 3.3.1 of Chapter 3, where we recall that $\beta(a, b) = \int_0^1 t^{a-1} (1 - t)^{b-1} dt = \Gamma(a) \Gamma(b) / \Gamma(a + b)$. Therefore, we know that the posterior distribution is Beta$(x + 1, n - x + 1)$, namely

$$p(\theta|x) = \frac{1}{\beta(x + 1, n - x + 1)} \theta^x (1 - \theta)^{n-x} 1_{[0,1]}(\theta)$$
If $B \sim \text{Beta}(a, b)$, we know that
\[
\mathbb{E}[Z] = \frac{a}{a+b} \quad \text{and} \quad \mathbb{V}[Z] = \frac{ab}{(a+b)^2(a+b+1)}.
\]

If we consider the square loss for the estimation of $\theta$, we know from Equation (4.9) that the Bayesian estimator is given by the expectation of the posterior, namely
\[
\hat{\theta} = \frac{X + 1}{X + 1 + n - X + 1} = \frac{X + 1}{n + 2}.
\]

Note the difference with the frequentist (non-Bayesian) estimator $X/n$ that we considered all along Chapter 2 (it was denoted $S_n/n$ back then). This estimator gives a cute Bayesian answer to the restaurant selection problem. This estimator is also known as the Laplace’s rule of succession (see [rule-of-succession] and Figure 4.2).

Using the bias-variance formula (2.2) from Chapter 2, we can compute the quadratic risk of $\hat{\theta}$ as follows:
\[
\mathbb{E}_\theta[(\hat{\theta} - \theta)^2] = \mathbb{V}_\theta[\hat{\theta}] + (\mathbb{E}_\theta[\hat{\theta}] - \theta)^2
\]
\[
= \frac{n\theta(1-\theta)}{(n+2)^2} + \left(\frac{1-2\theta}{n+2}\right)^2
\]
\[
= \frac{n\theta - n\theta^2 + 1 - 4\theta + 4\theta^2}{(n+2)^2},
\]
so that the Bayes risk with uniform prior $\mu(d\theta) = \rho(\theta)d\theta$ where $\rho(\theta) = 1_{[0,1]}(\theta)$ is given by
\[
R_B(\hat{\theta}, \mu) = \int_0^1 \mathbb{E}_\theta[(\hat{\theta} - \theta)^2]d\theta = \frac{1}{6(n+2)}.
\]

Using the exact same arguments, it is easy to see that if we use the prior distribution $\text{Beta}(a, b)$ instead of Uniform([0, 1]) (which is a particular case, since Uniform([0, 1]) = $\text{Beta}(1, 1)$) the posterior distribution is given by
\[
p(\theta|x) = \text{Beta}(a + x, b + n - x + b),
\]
which leads to a Bayesian estimator (for the square loss) given by
\[
\hat{\theta} = \frac{X + a}{n + a + b}.
\]

In this example, the prior and the posterior both belong to the family of Beta distributions. In such a case, we say that the Binomial and Beta distributions are conjugated, which corresponds to a situation where the posterior distribution can be explicitly computed.

**Definition 4.5** (Conjugated distributions) Given a prior $\mu(d\theta) = \rho(\theta)\lambda(d\theta)$ and a model $P_\theta(dx) = p(x|\theta)\nu(dx)$, we say that the
distributions of the prior and of the model are \textit{conjugated} whenever the prior and the posterior distribution belong to the same family of distributions.

4.5.2 Gaussian sample with a Gaussian prior

Another classical example is with the Gaussian distribution. Consider data $X_1, \ldots, X_n$ iid with Normal($\theta, \sigma^2$) distribution, namely

$$p(x|\theta) = \text{const}(\sigma) \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 \right),$$

where $x = [x_1 \cdots x_n]^T$ and \text{const}(\sigma) is a constant which depends only on $\sigma$ and a prior Normal(0, $\tau^2$) distribution on $\theta$, namely

$$p(\theta) = \text{const}(\tau) \exp \left( -\frac{\theta^2}{2\tau^2} \right).$$

We proceed as previously and write the joint distribution

$$p(x|\theta)p(\theta) = \text{const}(\sigma, \tau) \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 - \frac{\theta^2}{2\tau^2} \right)$$

$$= \text{const}(\sigma, \tau, x) \exp \left( -\frac{1}{2} \left( \frac{n}{\sigma^2} + \frac{1}{\tau^2} \right) \theta \sum_{i=1}^{n} x_i - \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i^2 \right)$$

$$= \text{const}(\sigma, \tau, x) \exp \left( -\frac{1}{2\gamma} \left( \theta - \gamma \sum_{i=1}^{n} x_i \right)^2 \right),$$

where we put $\gamma = \sigma^2 / (n + \sigma^2 / \tau^2)$. This proves that

$$p(\theta|x) = \text{Normal} \left( \frac{1}{n + \sigma^2 / \tau^2} \sum_{i=1}^{n} x_i, \frac{\sigma^2}{n + \sigma^2 / \tau^2} \right),$$

and that the Bayes estimator for the square loss is given by

$$\hat{\theta} = \frac{1}{n + \sigma^2 / \tau^2} \sum_{i=1}^{n} X_i.$$

This proves, in particular, that the Gaussian family is \textit{conjugated} with itself.

4.5.3 Bayesian linear regression with a Gaussian prior

Another very interesting example is the Gaussian linear regression model that we considered in Section 3.3 of Chapter 3, where

$$Y_i = X_i^\top \theta + \varepsilon_i$$
with deterministic $X_i \in \mathbb{R}^d$ and iid $\epsilon_i \sim \text{Normal}(0, \sigma^2)$. This means that $y \sim \text{Normal}(X \theta, \sigma^2 I_n)$, where we recall that $y = [Y_1 \ldots Y_n]^\top$ and that $X$ is the $n \times d$ matrix with rows given by $X_1, \ldots, X_n$. We consider this model in a Bayesian setting, by using a $\text{Normal}(0, \lambda^{-1} I_d)$ prior on $\theta$, where $\lambda > 0$. The joint distribution of $(\theta, y)$ is given by

$$p(\theta, y) = \text{const}(\sigma, \lambda) \exp \left(- \frac{1}{2\sigma^2} \|y - X \theta\|^2 - \frac{\lambda}{2} \|\theta\|^2\right).$$

What is, in this setting, the posterior distribution $p(\theta | y)$? This is slightly more complicated than what we did in both previous examples, and deserves the next theorem.

**Theorem 4.1** Consider the Gaussian linear model, namely the data distribution

$$p(y | \theta) = \text{Normal}(X \theta, \sigma^2 I_n)$$

with prior

$$p(\theta) = \text{Normal}(0, \lambda^{-1} I_d)$$

for some $\lambda > 0$. Then, we have

$$p(\theta | y) \propto \text{Normal} \left( (X^\top X + \lambda \sigma^2 I_d)^{-1} X^\top y, \sigma^2 (X^\top X + \lambda \sigma^2 I_d)^{-1} \right).$$

The proof of Theorem 4.1 is given in Section 4.6 below. If we consider the square loss $\ell(\theta', \theta) = \|\theta' - \theta\|^2$ where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^d$, we have that the Bayesian estimator is the expectation$^7$ of the posterior distribution, namely

$$\hat{\theta} = (X^\top X + \lambda \sigma^2 I_d)^{-1} X^\top y.$$  

In this example, the Bayes estimator coincides$^8$ with the so-called MAP estimator (Maximum A Posteriori), which is given, when it exists, by the mode of the posterior distribution.

We could have computed the MAP estimator without computing the posterior distribution. Indeed, we know that the posterior distribution $p(y | \theta)$ is proportional to the joint distribution $p(\theta, y)$, hence proportional to

$$\exp \left(- \frac{1}{2\sigma^2} \|y - X \theta\|^2 - \frac{\lambda}{2} \|\theta\|^2\right),$$

so that maximizing this function with respect to $\theta$ corresponds to minimizing

$$F(\theta) = \|y - X \theta\|^2 + \sigma^2 \lambda \|\theta\|^2.$$  

The function $F$ is strongly convex on $\mathbb{R}^d$, since its Hessian satisfies $\nabla^2 F(\theta) = 2X^\top X + 2\sigma^2 \lambda I_d > \sigma^2 \lambda I_d > 0$ for any $\theta \in \mathbb{R}^d$. So, its unique global minimizer cancels out the gradient

$$\nabla F(\theta) = 2X^\top (X \theta - y) + 2\sigma^2 \lambda \theta,$$
and is therefore equal to

\[ \hat{\theta} = (X^\top X + 2\sigma^2 I_d)^{-1} X^\top y, \]

as announced above. This estimator corresponds to a regularized or penalized version of the least-squares estimator. Any minimizer of

\[ \hat{\theta}_{\text{pen}} = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \| y - X \theta \|^2 + \text{pen} (\theta) \right\}, \]

where \( \text{pen} : \mathbb{R}^d \to \mathbb{R}^+ \) is a so-called penalization, is called a penalized least-squares estimator.\(^9\) A penalization function \( \text{pen} \) typically satisfies \( \text{pen}(0) = 0 \) and that \( \text{pen}(\theta) \) is a non-increasing function with respect to the absolute value of each coordinate of \( \theta \), so that \( \text{pen} \) penalizes the fact that \( \theta \) has large coordinates.

**Ridge penalization.** Whenever \( \text{pen}(\theta) = \lambda \| \theta \|^2 \), we call \( \text{pen} \) the ridge penalization and the problem is called ridge regression. This penalization “forces” the coordinates of \( \theta \in \mathbb{R}^d \) to remain “small”. It is the most widely used form of penalization in statistics and machine learning and it is used way beyond the setting of least-squares regression. For instance, this penalization is used in deep learning under the name Weight decay.

**A prior is a form of regularization.** Interestingly, we observe in this example that for the model of Gaussian linear regression, an isotropic Gaussian prior \( p(\theta) = \text{Normal}(0, \lambda^{-1} I_d) \) acts exactly as a ridge penalization, which forbids the coordinates of \( \theta \) to be free.\(^10\) Given \( \lambda > 0 \), we define the minimizer of the ridge regression problem as

\[ \hat{\theta}_\lambda = \arg\min_{\theta \in \mathbb{R}^d} \left\{ \| y - X \theta \|^2 + \lambda \| \theta \|^2 \right\}. \tag{4.10} \]

Whenever \( \lambda \) is very small, then the prior is almost “flat” which is equivalent to the fact that the ridge penalization term in (4.10) is negligible. In this case, we expect \( \hat{\theta}_\lambda \) to be close to the least-squares estimator \( \hat{\theta}_0 \).\(^11\) On the other hand, if \( \lambda \) is large, the prior is highly concentrated around 0, which is equivalent to a very strong ridge penalization in the computation of \( \hat{\theta}_\lambda \).

The parameter \( \lambda > 0 \) used in the ridge penalization correspond to a regularization strength. It is also called in machine learning a hyper-parameter\(^12\), which is tuned in practice using cross-validation.

9: it might be unique or not, depending on \( \text{pen} \) and \( X \)

10: and eventually take arbitrary large values, whenever the conditioning of \( X \) is bad for instance

11: which is \( \hat{\theta}_\lambda \) with \( \lambda = 0 \)

12: since it “parametrizes” the parameters...
4.6 Proofs

4.6.1 Proof of Theorem 4.1

Let us first recall that the prior is given by

\[ p(\theta) = \text{const}(\lambda) \exp \left( -\frac{\lambda}{2} \|\theta\|^2 \right) \]

and that the model is given by

\[ p(y | \theta) = \text{const}(\sigma) \exp \left( -\frac{1}{2\sigma^2} \|y - X \theta\|^2 \right), \]

so that the logarithm of the joint density of \((\theta, y)\) writes

\[
\log p(\theta, y) = \log p(\theta) + \log p(y | \theta) = \text{const}(\sigma^2, \lambda) - \frac{1}{2\sigma^2} (y - X \theta)^\top (y - X \theta) - \frac{\lambda}{2} \theta^\top \theta.
\]

Let us develop and rewrite this expression as a quadratic form with respect to \((\theta, y)\) (forgetting about the constant terms):

\[
\frac{1}{\sigma^2} y^\top y - \frac{2}{\sigma^2} y^\top X \theta + \frac{1}{\sigma^2} \theta^\top X^\top X \theta + \lambda \theta^\top \theta
= \begin{bmatrix} \theta \\ y \end{bmatrix}^\top \begin{bmatrix} \frac{1}{\sigma^2} X^\top X + \lambda I_d & -\frac{1}{\sigma^2} X^\top \\ -\frac{1}{\sigma^2} X & \frac{1}{\sigma^2} I_n \end{bmatrix} \begin{bmatrix} \theta \\ y \end{bmatrix}.
\]

This computation proves that the joint distribution of \((\theta, y)\) is Gaussian with precision matrix \(K\), namely

\[ p(\theta, y) = \text{Normal}(0, K^{-1}). \]

Now, in order to obtain the posterior distribution \(p(\theta | y)\), we need to compute the conditional density of \(\theta\) knowing \(y\) from the joint distribution of \((\theta, y)\). Since the joint distribution is Gaussian, it turns out to be particularly easy, as explained in the following proposition.

**Proposition 4.2** Let \(Z\) be a Gaussian vector \(Z \sim \text{Normal}(\mu, \Sigma)\) on \(\mathbb{R}^m\) with \(\Sigma > 0\). We consider the decomposition of \(Z\), and of its expectation and covariance matrix, into two blocks \(X_a\) and \(X_b\) as follows

\[
X = \begin{bmatrix} X_a \\ X_b \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} \Sigma_{a,a} & \Sigma_{a,b} \\ \Sigma_{b,a} & \Sigma_{b,b} \end{bmatrix}.
\]

The precision matrix is the inverse of the covariance matrix.
We decompose in the same way the precision matrix, namely

\[
K = \Sigma^{-1} = \begin{bmatrix}
K_{a,a} & K_{a,b} \\
K_{b,a} & K_{b,b}
\end{bmatrix}.
\] (4.11)

Then, the conditional density of \(X_a\) knowing \(X_b\) is given by

\[
p_{X_a|X_b}(x_a|x_b) = \text{Normal}\left(\mu_a - K_{a,b}^{-1}K_{a,b}(x_b - \mu_b), \ K_{a,a}^{-1}\right),
\]

where we used the precision matrix \(K\). We can also compute the conditional density as

\[
p_{X_a|X_b}(x_a|x_b) = \text{Normal}\left(\mu_a + \Sigma_{a,b} \Sigma_{b,b}^{-1}(x_b - \mu_b), \ \Sigma_{a,a} - \Sigma_{a,b} \Sigma_{b,b}^{-1} \Sigma_{a,b}^T\right),
\]

where we used this time the covariance \(\Sigma\).

The proof of Proposition 4.2 is given below. In order to compute the posterior \(p(\theta|y)\), we use Proposition 4.2 with \(X_a = \theta, X_b = y\), \(\mu_a = 0, \mu_b = 0\) and the precision matrix

\[
K = \begin{bmatrix}
K_{a,a} & K_{a,b} \\
K_{b,a} & K_{b,b}
\end{bmatrix} = \begin{bmatrix}
\frac{1}{\sigma^2}X^\top X + \lambda I_d & -\frac{1}{\sigma^2}X \\
-\frac{1}{\sigma^2}X & \frac{1}{\sigma^2}I_n
\end{bmatrix}.
\]

Namely, we obtain that \(p(\theta|y) = \text{Normal}(\mu_{\theta|y}, \Sigma_{\theta|y})\) with

\[
\mu_{\theta|y} = \mu_a - K_{a,a}^{-1}K_{a,b}(x_b - \mu_b) = (X^\top X + \lambda \sigma^2 I_d)^{-1}X^\top y,
\]

and

\[
\Sigma_{\theta|y} = K_{a,a}^{-1} = \sigma^2\left(X^\top X + \lambda \sigma^2 I_d\right)^{-1},
\]

which concludes the proof of Theorem 4.1.

**Proof of Proposition 4.2.** We have that

\[
\log p(X_a,X_b)(x_a, x_b) = \text{const}(K) - \frac{1}{2}(x - \mu)^\top K(x - \mu)
\]

and that

\[
(x - \mu)^\top K(x - \mu) = (x_a - \mu_a)^\top K_{a,a}(x_a - \mu_a) + (x_a - \mu_a)^\top K_{a,b}(x_b - \mu_b) \\
+ (x_b - \mu_b)^\top K_{b,a}(x_a - \mu_a) + (x_b - \mu_b)^\top K_{b,b}(x_b - \mu_b)
\]

\[
= (x_a - \mu_a + K_{a,a}^{-1}K_{a,b}(x_b - \mu_b))^\top K_{a,a} \\
\times (x_a - \mu_a + K_{a,a}^{-1}K_{a,b}(x_b - \mu_b)) + \text{const}(\mu, K, x_b).
\]
We know that \( p_{X_a | X_b}(x_a | x_b) \) is proportional to \( p_{(X_a, X_b)}(x_a, x_b) \), so we already know from the previous computation that

\[
p_{X_a | X_b}(x_a | x_b) = \text{Normal} \left( \mu_a - K_{a,a}^{-1} K_{a,b} (x_b - \mu_b), K_{a,a}^{-1} \right).
\]

Now, in order to express \( p_{X_a | X_b} \) through the covariance \( \Sigma \), we need to compute the inverse of \( \Sigma \). Let us recall the following classical block inversion formula

\[
\begin{bmatrix}
A & B \\
C & D
\end{bmatrix}^{-1} = \begin{bmatrix}
S & -SBD^{-1} \\
-D^{-1}CS & D^{-1}CSBD^{-1}
\end{bmatrix},
\]

where \( S = (A - BD^{-1}C)^{-1} \) is called the Schur complement with respect to the block \( D \). We use this formula to compute

\[
K = \begin{bmatrix}
K_{a,a} & K_{a,b} \\
K_{a,b} & K_{b,b}
\end{bmatrix} = \begin{bmatrix}
\Sigma_{a,a} & \Sigma_{a,b} \\
\Sigma_{a,b}^T & \Sigma_{b,b}
\end{bmatrix}^{-1}.
\]

This gives us

\[
K_{a,a} = S = (A - BD^{-1}C)^{-1} = (\Sigma_{a,a} - \Sigma_{a,b} \Sigma_{b,b}^{-1} \Sigma_{a,b})^{-1}
\]

and

\[
K_{a,a}^{-1} K_{a,b} = -S^{-1} SB D^{-1} = -BD^{-1} = -\Sigma_{a,b} \Sigma_{b,b}^{-1},
\]

so that

\[
\mu_a - K_{a,a}^{-1} K_{a,b} (x_b - \mu_b) = \mu_a + \Sigma_{a,b} \Sigma_{b,b}^{-1}(x_b - \mu_b),
\]

which concludes the proof of Proposition 4.2.

4.6.2 Proof of the lower bound from Theorem 3.4

We have now all the tools required to prove the lower bound involved in Theorem 3.4 from Chapter 3, namely that

\[
\inf_{\tilde{\theta}} \sup_{P \in \mathcal{G}(P_X, \sigma^2)} \mathbb{E}[||\tilde{\theta} - \theta^*||_2^2] \geq \frac{\sigma^2}{n} \mathbb{E}[\text{tr}(\Sigma^{-1})],
\]

(4.12)

where we recall that \( \Sigma = \mathbb{E}[XX^T] \succeq 0 \) and that \( \mathcal{G}(P_X, \sigma^2) \) is the set of joint distributions \( P \) for \((X, Y)\) satisfying \( X \sim P_X, Y = X^T \theta^* + \varepsilon \) almost surely and \( \varepsilon \) independent of \( X \) and such that \( \varepsilon \sim \text{Normal}(0, \sigma^2) \). Let us recall also that \( \tilde{\theta} \) is any estimator, namely any measurable function of \((X_1, Y_1), \ldots, (X_n, Y_n)\) iid with the same distribution \( P \in \mathcal{G}(P_X, \sigma^2) \).

First, let us remark that \( \sup_{P \in \mathcal{G}(P_X, \sigma^2)} \) corresponds to \( \sup_{\theta^* \in \mathbb{R}^d} \), so that denoting \( P_{\theta^*} = P_{X,Y} \) and the corresponding expectation \( \mathbb{E}_{\theta^*} \), we
need to lower bound
\[ \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \| \hat{\theta} - \theta \|^2 \Sigma. \]

The first, and certainly most important trick, is to lower bound the minimax risk by the Bayes risk. Let us choose some prior distribution \( \Pi(d\theta) = p(\theta) d\theta \) for \( \theta \) and write
\[
\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \| \hat{\theta} - \theta \|^2 \Sigma \geq \inf_{\hat{\theta}} \int_{\mathbb{R}^d} \mathbb{E}_\theta \| \hat{\theta} - \theta \|^2 \Sigma p(\theta) d\theta
\]
\[
= \inf_{\theta} \mathbb{E}_{\theta \sim \Pi} \| \hat{\theta} - \theta \|^2 \Sigma. \quad (4.13)
\]

Let us reason conditionally on \( X_1, \ldots, X_n \) in what follows to simplify notations and use Bayesian reasoning where the data has density
\[ p(y | \theta) = \text{Normal}(X \theta, \sigma^2 I_n) \]
and where the prior on \( \theta \) is given by \( \Pi_\lambda(d\theta) = p_\lambda(\theta) d\theta \) with
\[ p_\lambda(\theta) = \text{Normal}\left(0, \frac{\sigma^2}{\lambda n} I_d\right) \]
for some \( \lambda > 0 \). Note that this example is exactly the one considered in Section 4.5.3 with \( \lambda' = n \lambda / \sigma^2 \) instead of \( \lambda \). So, using Theorem 4.1, we have that
\[ p(\theta | y) = \text{Normal}\left(\hat{\theta}_\lambda, \frac{\sigma^2}{n} (X^\top X + \lambda I_d)^{-1}\right) \]
where
\[
\hat{\theta}_\lambda = (X^\top X + n\lambda I_d)^{-1} X^\top y
\]
\[
= \arg\min_{\theta \in \mathbb{R}^d} \left( \frac{1}{n} \| y - X \theta \|^2 + \lambda \| \theta \|^2 \right) \quad (4.14)
\]
is the ridge-penalized least squares estimator. The second trick is that we know how to minimize the Bayes risk (4.13): it can be minimized by looking for
\[ \hat{\theta} \in \arg\min_{\theta' \in \mathbb{R}^d} \int_{\mathbb{R}^d} \| \theta' - \theta \|^2 \Sigma p(\theta | y) d\theta, \]
as explained in Section 4.4. But, let us remark that if \( Z \) is a random vector such that \( \mathbb{E}\| Z \|^2 < \infty \), then the function \( F : \mathbb{R}^d \rightarrow \mathbb{R}^+ \) given by \( F(t) = \mathbb{E}\| Z - t \|^2 \Sigma \) is minimized at \( t^* = \mathbb{E}[Z] \) whenever \( \Sigma > 0 \). This entails that the Bayes estimator for the loss \( \ell(\theta', \theta) = \| \theta' - \theta \|^2 \Sigma \) is indeed \( \hat{\theta}_\lambda \). So, we end up with the lower bound
\[
\inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \| \hat{\theta} - \theta \|^2 \Sigma \geq \int_{\mathbb{R}^d} \mathbb{E}_\theta \| \hat{\theta}_\lambda - \theta \|^2 \Sigma \Pi_\lambda(d\theta)
\]
\[
= \mathbb{E}_{\theta \sim \Pi_\lambda} \mathbb{E}_\theta [\mathcal{E}(\hat{\theta}_\lambda)].
\]
for any $\lambda > 0$, that we are able to compute exactly thanks to the next Lemma. Let us recall that $\bar{\Sigma} = n^{-1} X^T X = n^{-1} \sum_{i=1}^n X_i X_i^T$ and introduce $\check{\Sigma}_\lambda = \bar{\Sigma} + \lambda I_d$.

**Lemma 4.3** The excess risk of the ridge estimator $\widehat{\theta}_\lambda$ given by (4.14) is given by

\[
\mathbb{E}_\theta [\mathcal{E}(\widehat{\theta}_\lambda)] = \lambda^2 \mathbb{E} \|\theta\|_2^2 (\check{\Sigma}_\lambda)^{-1} \check{\Sigma}(\check{\Sigma}_\lambda)^{-1} + \frac{\sigma^2}{n} \mathbb{E} \text{tr} \left( (\check{\Sigma}_\lambda)^{-1} \check{\Sigma}(\check{\Sigma}_\lambda)^{-1} \bar{\Sigma} \right)
\]

under the assumption that $Y_i = X_i^T \theta + \varepsilon_i$ for $\varepsilon_i \sim \text{Normal}(0, \sigma^2)$.

We inject the formula given by Lemma 4.3 to end up with the lower bound

\[
\mathbb{E}_{\theta \sim \Pi_\lambda} \left[ \lambda^2 \mathbb{E} \|\theta\|_2^2 (\check{\Sigma}_\lambda)^{-1} \check{\Sigma}(\check{\Sigma}_\lambda)^{-1} + \frac{\sigma^2}{n} \mathbb{E} \text{tr} \left( (\check{\Sigma}_\lambda)^{-1} \check{\Sigma}(\check{\Sigma}_\lambda)^{-1} \bar{\Sigma} \right) \right].
\]

So, using Fubini, and since $\mathbb{E}_{\theta \sim \Pi_\lambda} [\theta \theta^T] = \frac{\sigma^2}{n} I_d$ by definition of $\Pi_\lambda$, we end up with

\[
\mathbb{E}_{\theta \sim \Pi_\lambda} \left[ \lambda^2 \mathbb{E} \|\theta\|_2^2 (\check{\Sigma}_\lambda)^{-1} \check{\Sigma}(\check{\Sigma}_\lambda)^{-1} \right] = \lambda^2 \mathbb{E} \mathbb{E}_{\theta \sim \Pi_\lambda} \text{tr} \left( \theta^T (\check{\Sigma}_\lambda)^{-1} \check{\Sigma}(\check{\Sigma}_\lambda)^{-1} \theta \right)
\]

using $\text{tr}[x] = x$ for $x \in \mathbb{R}$

\[
= \lambda^2 \mathbb{E} \mathbb{E}_{\theta \sim \Pi_\lambda} \text{tr} \left( (\check{\Sigma}_\lambda)^{-1} \check{\Sigma}(\check{\Sigma}_\lambda)^{-1} \theta \theta^T \right)
\]

\[
= \lambda^2 \mathbb{E} \text{tr} \left( (\check{\Sigma}_\lambda)^{-1} \check{\Sigma}(\check{\Sigma}_\lambda)^{-1} \lambda I_d \right),
\]

so that the lower bound becomes now

\[
\frac{\sigma^2}{n} \mathbb{E} \text{tr} \left[ (\check{\Sigma}_\lambda)^{-1} \check{\Sigma}(\check{\Sigma}_\lambda)^{-1} (\check{\Sigma} + \lambda I_d) \right] = \frac{\sigma^2}{n} \mathbb{E} \text{tr} \left[ (\check{\Sigma}_\lambda)^{-1} \check{\Sigma} \right].
\]

We proved that the lower bound

\[
\inf_{\widehat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta \|\widehat{\theta} - \theta\|^2_\Sigma \geq \frac{\sigma^2}{n} \mathbb{E} \text{tr} \left[ (\check{\Sigma}_\lambda)^{-1} \check{\Sigma} \right]
\]

holds for any $\lambda > 0$. Since $P_X$ is non-degenerate, we know from Theorem 3.1 that $\check{\Sigma} \succ 0$ almost surely and we have that the function

\[
\lambda \mapsto \text{tr} \left[ (\check{\Sigma} + \lambda I_d)^{-1} \check{\Sigma} \right] = \text{tr} \left[ (\Sigma^{-1/2} \check{\Sigma} \Sigma^{-1/2} + \lambda \Sigma^{-1})^{-1} \right]
\]

is decreasing on $(0, +\infty)$ since $\lambda_2 \Sigma^{-1} \succ \lambda_1 \Sigma^{-1}$ whenever $\lambda_2 > \lambda_1$. So, by monotone convergence, we have indeed that

\[
\mathbb{E} \text{tr} \left[ (\check{\Sigma}_\lambda)^{-1} \check{\Sigma} \right] \to \mathbb{E} \text{tr} \left[ (\check{\Sigma})^{-1} \check{\Sigma} \right] = \mathbb{E} \text{tr} \left[ (\Sigma)^{-1} \right]
\]

as $\lambda \to 0^+$. This proves the desired lower bound of the minimax risk, up to the proof of Lemma 4.3.

$\Box$
Proof of Lemma 4.3. Let us recall that \( Y_i = X_i^T \theta + \varepsilon_i \) with \( \varepsilon_i \sim \text{Normal}(0, \sigma^2) \) and that each \( \varepsilon_i \) is independent of \( X_1, \ldots, X_n \). We have

\[
\frac{1}{n} \sum_{i=1}^{n} Y_i X_i = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \theta + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i = \bar{\Sigma} \theta + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i,
\]

so that

\[
E_\theta [\varepsilon] = E_\theta \| \bar{\theta} - \theta \|_2^2 = E \left\| (\bar{\Sigma}_\lambda)^{-1} \left( \bar{\Sigma} \theta + \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right) - \theta \right\|_2^2 \Sigma,
\]

but using \((\bar{\Sigma}_\lambda)^{-1}(\bar{\Sigma} + \lambda I_d - \lambda I_d) = I_d - \lambda(\bar{\Sigma}_\lambda)^{-1}\) we obtain

\[
E_\theta [\varepsilon] = E \left\| (\bar{\Sigma}_\lambda)^{-1} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i - \lambda(\bar{\Sigma}_\lambda)^{-1} \theta \right\|_2^2 \Sigma
\]

\[
= E \left[ E \left( \left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i - \lambda \theta \right\|_2^2 (\bar{\Sigma}_\lambda)^{-1} \Sigma (\bar{\Sigma}_\lambda)^{-1} \left| X_1, \ldots, X_n \right. \right) \right]
\]

\[
= E \left[ E \left( \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i X_i \right)^2 (\bar{\Sigma}_\lambda)^{-1} \Sigma (\bar{\Sigma}_\lambda)^{-1} \left| X_1, \ldots, X_n \right. \right] + \lambda^2 E \left\| \theta \right\|_2^2 (\bar{\Sigma}_\lambda)^{-1} \Sigma (\bar{\Sigma}_\lambda)^{-1}
\]

\[
= \frac{\sigma^2}{n^2} E \left[ \sum_{i=1}^{n} \| X_i \|_2^2 (\bar{\Sigma}_\lambda)^{-1} \Sigma (\bar{\Sigma}_\lambda)^{-1} \right] + \lambda^2 E \left\| \theta \right\|_2^2 (\bar{\Sigma}_\lambda)^{-1} \Sigma (\bar{\Sigma}_\lambda)^{-1},
\]

where we used repeatedly that \( E[\varepsilon_i | X_1, \ldots, X_n] = 0 \), \( E[\varepsilon_i \varepsilon_j | X_1, \ldots, X_n] = 0 \) for any \( i \neq j \) and \( E[\varepsilon_i^2 | X_1, \ldots, X_n] = \sigma^2 \). But

\[
\frac{1}{n} \sum_{i=1}^{n} \| X_i \|_2^2 (\bar{\Sigma}_\lambda)^{-1} \Sigma (\bar{\Sigma}_\lambda)^{-1} = \frac{1}{n} \sum_{i=1}^{n} \text{tr} \left[ (\bar{\Sigma}_\lambda)^{-1} \Sigma (\bar{\Sigma}_\lambda)^{-1} X_i X_i^T \right]
\]

\[
= \text{tr} \left[ (\bar{\Sigma}_\lambda)^{-1} \Sigma (\bar{\Sigma}_\lambda)^{-1} \bar{\Sigma} \right],
\]

which concludes the proof of Lemma 4.3. \( \square \)