Learning Theory from First Principles

Solutions to exercises (and extra exercises)

December 16, 2024

Francis Bach

francis.bach@inria.fr

The solutions to exercises have been adapted from students from my Master class: Camille Dubois, Nicolas Thiou.

Chapter 1

Mathematical Preliminaries

Exercise 1.1 *For $\alpha \in \mathbb{R}$ such that $\alpha \neq -1/n$ and $1_n \in \mathbb{R}^n$ the vector of all 1s, show that we have* $(I + \alpha 1_n 1_n^\top)^{-1} = I - \frac{\alpha}{1+n\alpha} 1_n 1_n^\top$.

Exercise 1.2 (♦) *Show that we can diagonalize by blocks the matrices M and M^{-1} as*

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} I & 0 \\ CA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & M/A \end{pmatrix} \begin{pmatrix} I & A^{-1}B \\ 0 & I \end{pmatrix}$$

$$M^{-1} = \begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} I & -A^{-1}B \\ 0 & I \end{pmatrix} \begin{pmatrix} A^{-1} & 0 \\ 0 & (M/A)^{-1} \end{pmatrix} \begin{pmatrix} I & 0 \\ -CA^{-1} & I \end{pmatrix}.$$

Exercise 1.3 *Show that* $\det\left(\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right)\right) = \det(M/A)\det(A) = \det(M/D)\det(D)$.

Exercise 1.4 (♦) *Prove the identities $\mu_{y|x'} = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x' - \mu_x)$ and covariance matrix $\Sigma_{y|x} = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$.*

Exercise 1.5 *Express the eigenvectors of $XX^\top$ and $X^\top X$ using the singular vectors of X.*

Exercise 1.6 *Express the eigenvectors of $\left(\begin{smallmatrix} 0 & X \\ X^\top & 0 \end{smallmatrix}\right)$ using the singular vectors of X.*

Exercise 1.7 *Show that for the logistic regression objective function defined as $F(\theta) = \frac{1}{n}\sum_{i=1}^n \log(1 + \exp(-y_i(X\theta)_i)$ with $X \in \mathbb{R}^{n\times d}$ and $y \in \{-1, 1\}^n$, then $F'(\theta) = \frac{1}{n}X^\top g$, where $g \in \mathbb{R}^n$ is defined as $g_i = -y_i\sigma(-y_i(X\theta)_i)$, with $\sigma(u) = (1 + e^{-u})^{-1}$ the sigmoid function. Show that the Hessian is $\frac{1}{n}X^\top \mathrm{Diag}(h)X$, with $h \in \mathbb{R}^n$ defined as $h_i = \sigma(y_i(X\theta)_i)\sigma(-y_i(X\theta)_i)$.*

Exercise 1.8 (Functions on matrices) *Let A be a symmetric matrix. Show that the* *gradient of the function* $M \mapsto \mathrm{tr}(AM^{-1})$*, defined on invertible symmetric matrices, is equal to* $M \mapsto -M^{-1}AM^{-1}$*. Show that the gradient of* $M \mapsto \log\det(M)$ *is* $M \mapsto M^{-1}$*.*

Exercise 1.9 *Let* Y *be a nonnegative random variable with finite expectation, and* $\varepsilon > 0$*. Show that* $\varepsilon 1_{Y \geqslant \varepsilon} \leqslant Y$ *almost surely and prove Markov's inequality:*

$$\mathbb{P}(Y \geqslant \varepsilon) \leqslant \frac{1}{\varepsilon}\mathbb{E}[Y].$$

Exercise 1.10 (Chernoff bound) *Let* X *be a random variable. Show that for any* $t \in \mathbb{R}$ *and* $s > 0$*, we have* $\mathbb{P}(X \geqslant t) \leqslant e^{-st}\mathbb{E}[e^{sX}]$*.*

Exercise 1.11 *Let* Y *be a nonnegative random variable with finite expectation. Show that* $\mathbb{E}[Y] = \int_0^\infty \mathbb{P}(Y \geqslant t)dt$*.*

Exercise 1.12 (♦) *For* X *a Gaussian random variable with mean* 0 *and variance* 1*, show that for* $t \geqslant 0$*,* $\frac{1}{2}\exp(-t^2) \leqslant \mathbb{P}(X \geqslant t) \leqslant \exp(-t^2/2)$*.*

Exercise 1.13 *Show the one-sided inequality: with probability greater than* $1-\delta$*,* $\frac{1}{n}\sum_{i=1}^n Z_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i] < \frac{|a-b|}{\sqrt{2n}}\sqrt{\log\left(\frac{1}{\delta}\right)}$*.*

Exercise 1.14 (Azuma's inequality (♦)) *Assume that the sequence of random variables* $(Z_i)_{i\geqslant 0}$*, satisfies* $\mathbb{E}(Z_i|\mathcal{F}_{i-1}) = 0$ *for an increasing sequence of increasing "σ-fields"* $(\mathcal{F}_i)_{i\geqslant 0}$*,*[1] *and* $|Z_i| \leqslant c_i$ *almost surely, for* $i \geqslant 1$*. Then*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n Z_i \geqslant t\right) \leqslant \exp\left(\frac{-n^2t^2}{2(c_1^2 + \cdots + c_n^2)}\right).$$

Exercise 1.15 *Show that a Gaussian random variable with variance* σ^2 *is sub-Gaussian with constant* σ^2*.*

Exercise 1.16 *If* $Z_1, \dots, Z_n$ *are independent random variables which are sub-Gaus-sian with constant* τ^2*, show that* $\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i]\right| \geqslant t\right) \leqslant 2\exp(-\frac{nt^2}{2\tau^2})$ *for any* $t \geqslant 0$*.*

Exercise 1.17 (♦) *Let* Z *be a random variable that is sub-Gaussian with constant* τ^2*. Then, by using the tail bound* $\mathbb{P}(|Z - \mathbb{E}[Z]| \geqslant t) \leqslant 2\exp(-\frac{t^2}{2\tau^2})$*:*

$$\forall t \geqslant 0,\ \mathbb{P}(|Z - \mathbb{E}[Z]| \geqslant t) \leqslant 2\exp\left(-\frac{t^2}{2\tau^2}\right).$$

Show that for any positive integer q*,* $\mathbb{E}[(Z - \mathbb{E}[Z])^{2q}] \leqslant (2q)q!(2\tau^2)^q$*.*

[1] See more details in `https://en.wikipedia.org/wiki/Azuma's_inequality`.

Exercise 1.18 (♦♦) *Let Z be a random variable such that for any positive integer q, $\mathbb{E}[(Z-\mathbb{E}[Z])^{2q}] \leqslant (2q)q!(2\tau^2)^q$. Then show that Z is sub-Gaussian with parameter $24\tau^2$.*

Exercise 1.19 *Assume that the random variable Z has almost surely nonnegative values and finite second-order moment. Show that for any $s \geqslant 0$, we have $\log\big(\mathbb{E}[e^{-sZ}]\big) \leqslant -s\mathbb{E}[Z] + \frac{s^2}{2}\mathbb{E}[Z^2]$.*

Exercise 1.20 (♦) *Use McDiarmid's inequality to prove a Hoeffding-type bound for vectors: If $Z_1, \dots, Z_n \in \mathbb{R}^d$ are independent centered vectors such that $\|Z_i\|_2 \leqslant c$ almost surely, then with probability greater than $1-\delta$, we have*

$$\Big\|\frac{1}{n}\sum_{i=1}^n Z_i\Big\|_2 \leqslant \frac{c}{\sqrt{n}}\Big(1+\sqrt{2\log\frac{1}{\delta}}\Big).$$

Exercise 1.21 (♦) *Prove the inequality*

$$\mathbb{P}\bigg(\bigg|\frac{1}{n}\sum_{i=1}^n Z_i - \frac{1}{n}\sum_{i=1}^n \mathbb{E}[Z_i]\bigg| \geqslant t\bigg) \leqslant 2\exp\Big(-\frac{nt^2}{2\sigma^2+2ct/3}\Big).$$

Exercise 1.22 *Assume that $Z_1, \dots, Z_n$ are random variables that are sub-Gaussian with constant τ^2 and have zero means. Show that $\mathbb{E}\big[\max\{|Z_1|, \dots, |Z_n|\}\big] \leqslant \sqrt{2\tau^2\log(2n)}$. Prove the same result up to a universal constant using the tail bounds $\mathbb{P}(|Z_i| \geqslant t) \leqslant 2\exp(-\frac{t^2}{2\tau^2})$ together with the union bound, and the property $\mathbb{E}[|Y|] = \int_0^{+\infty}\mathbb{P}(|Y| \geqslant t)dt$ for any random variable Y such that $\mathbb{E}[|Y|]$ exists.*

Exercise 1.23 (♦♦) *Assume that $Z_1, \dots, Z_n$ are independent Gaussian random variables with mean zero and variance σ^2. Provide a lower bound for $\mathbb{E}[\max\{Z_1, \dots, Z_n\}]$ of the form $c\sqrt{\log n}$ for $c > 0$.*

Exercise 1.24 *Assume that $Z_1, \dots, Z_n$ are sub-Gaussian random variables with common sub-Gaussianity parameter τ, and potentially different means $\mu_1, \dots, \mu_n$. For a fixed set of nonnegative weights $\pi_1, \dots, \pi_n$ that sum to 1, and $\delta \in (0,1)$, show that with probability greater than $1-\delta$, for all $i \in \{1, \dots, n\}$, $|z_i - \mu_i| \leqslant \tau\sqrt{2\log(1/\pi_i)} + \tau\sqrt{2\log(2/\delta)}$. If $\hat{\imath} \in \arg\min_{i\in\{1,\dots,n\}}\big\{z_i + \tau\sqrt{2\log(1/\pi_i)}\big\}$, show that with probability greater than $1-\delta$, $\mu_{\hat{\imath}} \leqslant \min_{i\in\{1,\dots,n\}}\big\{\mu_i + 2\tau\sqrt{2\log(1/\pi_i)}\big\} + 2\tau\sqrt{2\log(2/\delta)}$.*

Exercise 1.25 (♦♦) *Consider a convex function $f: \mathbb{R}^d \to \mathbb{R}$ such that $f(0) = 0$ and f is L-smooth with respect to the norm Ω; that is, f is continuously differentiable and for all $\theta, \eta \in \mathbb{R}^d$, $f(\theta) \leqslant f(\eta) + f'(\eta)^\top(\theta-\eta) + \frac{L}{2}\Omega(\theta-\eta)^2$. Let $Z_i \in \mathbb{R}^d$ be independent zero-mean random vectors with $\mathbb{E}[\Omega(Z_i)^2] \leqslant \sigma^2$, for $i = 1, \dots, n$. Show by induction in n that $\mathbb{E}[f(Z_1 + \dots + Z_n)] \leqslant nL\frac{\sigma^2}{2}$.*

Exercise 1.26 *Consider a function $g : [0,1] \to \mathbb{R}$. Show that the piecewise interpolant based on values at $\{0,1\}$ equals $\tilde{g} : x \mapsto (1-x)g(0) + xg(1)$ and that its integral equals $\frac{1}{2}g(0) + \frac{1}{2}g(1)$. Assuming g is twice differentiable with second-derivative bounded in magnitude by L, show that for all $x \in [0,1]$, $|g(x) - \tilde{g}(x)| \leqslant \frac{L}{2}x(1-x)$.*

Exercise 1.27 *Show that the trapezoidal rule leads to an error in $O(1/n)$ if we assume only one bounded derivative.*

Exercise 1.28 (♦) *Show that for 1-periodic functions, the trapezoidal rule leads to an error in $O(1/n^s)$ if we assume s bounded derivatives.*

Exercise 1.29 *Assume that the matrices $M_i \in \mathbb{R}^{d_1 \times d_2}$ are independent, have zero mean, and $\|M_i\|_{\rm op} \leqslant c$ almost surely for all $i \in \{1, \dots, n\}$. Show that*

$$\mathbb{P}\bigg(\bigg\| \frac{1}{n} \sum_{i=1}^n M_i \bigg\|_{\rm op} \geqslant t \bigg) \leqslant (d_1 + d_2) \cdot \exp\Big(- \frac{nt^2}{8c^2} \Big).$$

Moreover, with $\sigma^2 = \max\big\{ \lambda_{\max}\big(\frac{1}{n}\sum_{i=1}^n M_i^\top M_i\big), \lambda_{\max}\big(\frac{1}{n}\sum_{i=1}^n M_i M_i^\top\big)\big\}$, show that

$$\mathbb{P}\bigg(\bigg\| \frac{1}{n} \sum_{i=1}^n M_i \bigg\|_{\rm op} \geqslant t \bigg) \leqslant (d_1 + d_2) \cdot \exp\Big(- \frac{nt^2/2}{\sigma^2 + ct/3} \Big).$$

Chapter 2

Introduction to Supervised Learning

Exercise 2.1 *Consider binary classification with* $\mathcal{Y} = \{-1, 1\}$ *with the loss function* $\ell(-1,-1) = \ell(1,1) = 0$ *and* $\ell(-1,1) = c_- > 0$ *(cost of a false positive),* $\ell(1,-1) = c_+ > 0$ *(cost of a false negative). Compute a Bayes predictor at* x *as a function of* $\mathbb{E}[y|x]$*.*

Solution. Given $x \in \mathcal{X}$, we compute

$$\underset{z\in\{-1,1\}}{\operatorname{argmin}} \ \mathbb{E}[\ell(y,z)\ |x = x'].$$

We have

$$\mathbb{E}[y|x] = \mathbb{P}(y = 1|x) - \mathbb{P}(y = -1|x).$$

Therefore, computing $\mathbb{E}[\ell(y,z')\ |x = x']$ for $z' = -1$, we obtain :

$$\mathbb{E}[\ell(y,z')\ |x = x'] = \mathbb{E}[l(y,-1)|x = x'] = c_-\mathbb{P}(y = -1|x = x') = c_-\frac{1 - \mathbb{E}[y|x = x']}{2}.$$

With $z' = 1$, it yields :

$$\mathbb{E}[\ell(y,z')\ |x = x'] = \mathbb{E}[\ell(y,1)|x = x'] = c_+\mathbb{P}(y = 1|x = x') = c_+\frac{1 + \mathbb{E}[y|x = x']}{2}.$$

This gives a choice for a Bayes estimator $f : X \to \mathbb{R}$ such that, for all $x' \in \mathbb{R}$,

$$f(x') = \operatorname{sign}\Big(\mathbb{E}[y|x = x'] - \frac{c_- - c_+}{c_- + c_+}\Big).$$

Exercise 2.2 *We consider a learning problem on* $\mathcal{X} \times \mathcal{Y}$*, with* $\mathcal{Y} = \mathbb{R}$ *and the absolute loss defined as* $\ell(y,z) = |y - z|$*. Compute a Bayes predictor* $f_* : \mathcal{X} \to \mathbb{R}$*.*

Solution. Let $\mathcal{X}, \mathcal{Y}, \boldsymbol{l}$ be as defined in the text. We assume that y given x has a density function $p(y, x)$.

Let $x \in \mathcal{X},\ z \in \mathbb{R}$:

$$\begin{aligned} e(z) = \mathbb{E}(|y - z|\ |x = x) &= \int_{-\infty}^{+\infty} |y - z| p(y, x) dy \\ &= \int_{-\infty}^{z} (z - y) p(y, x) dy + \int_{z}^{+\infty} (y - z) p(y, x) dy. \end{aligned}$$

By the Leibnitz rule, the derivative yields: $e'(z) = \int_{-\infty}^{z} p(y, x) dy - \int_{z}^{+\infty} p(y, x) dy$, which shows that the minimum of e is reached on the median of y given x. Therefore, the Bayes predictor f_* is, in our case, the median of y given x.

Exercise 2.3 *We consider a learning problem on $\mathcal{X} \times \mathcal{Y}$, with $\mathcal{Y} = \mathbb{R}$ and the "pinball" loss $\ell(y, z) = \alpha(y - z)_+ + (1 - \alpha)(z - y)_+$, for $\alpha \in (0, 1)$. Compute a Bayes predictor $f_* : \mathcal{X} \to \mathbb{R}$. Provide an interpretation in terms of quantiles.*

Solution. For all $z, y \in \mathcal{Y}$, the loss function $\ell(z, y)$ is defined as, for $\alpha \in (0, 1)$:

$$\ell(z, y) = \alpha(y - z)_+ + (1 - \alpha)(z - y)_+ = \alpha(y - z) 1_{y>z} + (1 - \alpha)(z - y) 1_{y<z}.$$

The Bayes predictor at x is given by:

$$f^*(x) \in \arg\min_{z \in \mathcal{Y}} \mathbb{E}[\ell(y, z)|x]$$

We have :

$$\begin{aligned} \mathbb{E}[\ell(y, z) \mid x] &= \mathbb{E}[\alpha(y - z) 1_{y>z} + (1 - \alpha)(z - y) 1_{y<z}|x] \\ &= \alpha \int_{\mathbb{R}} (y - z) 1_{y>z} p(y|x) dy + (1 - \alpha) \int_{\mathbb{R}} (z - y) 1_{y<z} p(y|x) dy \\ &= \alpha \int_{z}^{+\infty} (y - z) 1_{y>z} p(y|x) dy + (1 - \alpha) \int_{-\infty}^{z} (z - y) 1_{y<z} p(y|x) dy. \end{aligned}$$

Since we want to find the minimum of this with respect to z (and the loss is convex in z), we compute a subgradient with respect to z and set it to 0. We have:

$$\begin{aligned} &\frac{\partial}{\partial z} \mathbb{E}[\alpha(y - z)_+ + (1 - \alpha)(z - y)_+|x] = 0 \\ \Leftrightarrow\ &(1 - \alpha) \int_{-\infty}^{z} p(y|x) dy - \alpha \int_{z}^{+\infty} p(y|x) dy = 0 \\ \Leftrightarrow\ &\int_{-\infty}^{z} p(y|x) dy = \alpha \end{aligned}$$

We thus have a minimizer by taking z as the quantile of order α of the conditional distribution of y given x.

Exercise 2.4 (♦) *Characterize Bayes predictors for regression with the "ε-insensitive" loss defined as $\ell(y,z) = \max\{0, |y-z| - \varepsilon\}$. If for each x, y is supported in an interval of length less than 2ε, what are the Bayes predictors?*

Solution. Assume $\varepsilon > 0$. Let $x' \in \mathbb{R}$. Let $z \in \mathbb{R}$.

$$\begin{aligned}\mathbb{E}(\ell(y,z)|x = x') &= \int_{|y-z|\geq\varepsilon} (|y-z| - \varepsilon)p(y,x)dy \\ &= \int_{y-z\geq\varepsilon} (y-z-\varepsilon)p(y,x)dy + \int_{z-y\geq\varepsilon} (z-y-\varepsilon)p(y,x)dy.\end{aligned}$$

Derivating the expresion w.r.t z yields:

$$\int_{y-z\geq\varepsilon} p(y,x)dy - \int_{z-y\geq\varepsilon} p(y,x)dy = \mathbb{P}(y-z \geq \varepsilon|x = x') - \mathbb{P}(y-z \leq -\varepsilon|x = x').$$

Therefore, a Bayes estimator can be interpreted as a balance between the number of overestimated and underestimated predictions, above a specific threshold (ε).

Let y be supported in an interval of less than 2ε for all x. For a given x, we assume that (a,b) is the smallest interval supporting y given x ($b - a \leq 2\varepsilon$). As we cannot have both $\mathbb{P}(y-z \leq -\varepsilon|x = x') > 0$ and $\mathbb{P}(y-z \geq \varepsilon|x = x') > 0$, we need

$$\mathbb{P}(y-z \leq -\varepsilon|x = x') = \mathbb{P}(y-z \geq \varepsilon|x = x') = 0.$$

Therefore, $f : \mathcal{X} \to \mathbb{R}$ is a Bayes estimator for this problem if, and only if, for all x, $f(x) \in (b-\varepsilon, a+\varepsilon)$, where a and b are $x-$dependent as defined before.

Exercise 2.5 (Inverting predictions) *Consider the binary classification problem with $\mathcal{Y} = \{-1, 1\}$ and the 0–1 loss. Relate the risk of a prediction f and to that of its opposite $-f$.*

Exercise 2.6 ("Chance" predictions) *Consider binary classification problems with the 0–1 loss. What is the risk of a random prediction rule where we predict the two classes with equal probabilities independent of input x? Address the same question with multiple categories.*

Exercise 2.7 (♦) *Consider a random prediction rule where we predict from the probability distribution of y given x. When is this achieving the Bayes risk?*

Exercise 2.8 *How would the curve move when n increases (assuming the same balance between classes)?*

Chapter 3

Linear Least-Squares Regression

Exercise 3.1 *In the Gaussian model given above, show that $\tilde{\sigma}^2$ the maximum likelihood estimator of σ^2 is equal to $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (y_i - \varphi(x_i)^\top \hat{\theta})^2$.*

Exercise 3.2 *Show that the expected empirical risk is equal to $\mathbb{E}[\widehat{\mathcal{R}}(\hat{\theta})] = \frac{n-d}{n}\sigma^2$. In particular, when $n > d$, deduce that an unbiased estimator of the noise variance σ^2 is given by $\frac{1}{n-d}\|y - \Phi\hat{\theta}\|_2^2$.*

Solution. We want to compute $\hat{\mathcal{R}}(\hat{\theta})$.

$$
\begin{aligned}
n\hat{\mathcal{R}}(\hat{\theta}) &= \mathbb{E}(\|y - \Phi\hat{\theta}\|_2^2) \\
&= \mathbb{E}(\|y - \Phi(\Phi^\top\Phi)^{-1}\Phi^\top y\|_2^2), \text{ as } \hat{\theta} = (\Phi^\top\Phi)^{-1}\Phi^\top y \\
&= \mathbb{E}(\|(I - \Pi)y\|_2^2), \text{ where } \Pi = \Phi(\Phi^\top\Phi)^{-1}\Phi^\top \\
&= \mathbb{E}(\mathrm{tr}(y^\top(I - \Pi)y)), \text{ as } I - \Pi \text{ is symmetric and } (I-\Pi)^2 = I - \Pi \\
&= \mathrm{tr}((I - \Pi)\mathbb{E}(yy^\top)) \\
&= \sigma^2 \,\mathrm{tr}((I - \Pi)), \text{ as } \mathbb{E}(yy^\top) = \Phi\theta_*\theta_*\Phi^\top\sigma^2 I \\
&= \sigma^2(n-d), \text{ as } I \in \mathbb{R}^{n\times n}, \text{ and } \Pi \text{ is a projector on a space of dimension } d.
\end{aligned}
$$

This gives the expected result. Isolating σ^2 in the previous equation, we actually compute $\sigma^2 = \mathbb{E}(\frac{1}{n-d}\|y - \Phi\hat{\theta}\|_2^2)$ which means that $\frac{1}{n-d}\|y - \Phi\hat{\theta}\|_2^2$ is an unbiased estimator of σ^2.

Exercise 3.3 (General noise) *Consider the fixed design regression model $y = \Phi\theta_* + \varepsilon$ with ε with zero mean and covariance matrix equal to $C \in \mathbb{R}^{n\times n}$ (not $\sigma^2 I$ anymore). Show that the expected excess risk of the OLS estimator is equal to $\frac{1}{n}\,\mathrm{tr}\left[\Phi(\Phi^\top\Phi)^{-1}\Phi^\top C\right]$.*

Exercise 3.4 (Multivariate regression (♦)) *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 $S \in \mathbb{R}^{k\times k}$. 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 obtained from i.i.d. $\varepsilon_i \in \mathbb{R}^k$, $i = 1,\dots,n$, derive the OLS estimator minimizing $\frac{1}{n}\|Y - \Phi\theta\|_{\mathrm{F}}^2$ and its excess risk (where $\|M\|_{\mathrm{F}}$ denotes the Frobenius norm defined as the square root of the sum the squared components of M).*

Exercise 3.5 *Using the matrix inversion lemma (discussed in section* **??***), show that the ridge regression estimator given in proposition* **??** *can also be written as $\hat\theta_\lambda = (\Phi^\top\Phi + n\lambda I)^{-1}\Phi^\top y = \Phi^\top(\Phi\Phi^\top + n\lambda I)^{-1}y$. What could be the computational benefits?*

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.*

Solution. Replacing the regularization term $\lambda\|\theta\|_2^2$ by $\|\theta\|_\Lambda^2$, with Λ positive definite, we obtain that $\hat\theta_\Lambda = \frac{1}{n}(\Sigma + \Lambda)^{-1}\Phi^\top y$. Therefore, $\mathbb{E}(\hat\theta_\Lambda) = \theta_* - (I + \Lambda)^{-1}\Lambda\theta_*$. As in the book, we decompose the excess risk in bias B and variance V. The computations yield :

$$
\begin{aligned}
B &= \|\mathbb{E}(\hat\theta_\Lambda) - \theta_*\|_{\hat\Sigma}^2 = \|(\hat\Sigma + \Lambda)^{-1}\Lambda\theta_*\|_{\hat\Sigma}^2 \\
B &= \theta_*^\top\Lambda(\hat\Sigma + \Lambda)^{-1}\hat\Sigma(\hat\Sigma + \Lambda)^{-1}\Lambda\theta_* \\
V &= \mathbb{E}(\|\frac{1}{n}(\hat\Sigma + \Lambda)^{-1}\Phi^\top\varepsilon\|_{\hat\Sigma}^2) \\
&= \mathbb{E}(\frac{1}{n^2}\operatorname{tr}(\varepsilon^\top\Phi(\hat\Sigma + \Lambda)^{-1}\hat\Sigma(\hat\Sigma + \Lambda)^{-1}\Phi^\top\varepsilon)) \\
&= \frac{1}{n^2}\operatorname{tr}(\sigma^2(\hat\Sigma + \Lambda)^{-1}\hat\Sigma(\hat\Sigma + \Lambda)^{-1}\hat\Sigma) \\
V &= \frac{\sigma^2}{n}\operatorname{tr}(((\hat\Sigma + \Lambda)^{-1}\hat\Sigma)^2).
\end{aligned}
$$

By summing the preceding terms, we have :

$$
\mathbb{E}(\mathcal{R}(\hat\theta)) = \mathcal{R}^* + B + V = \sigma^2 + \theta_*^\top\Lambda(\hat\Sigma + \Lambda)^{-1}\hat\Sigma(\hat\Sigma + \Lambda)^{-1}\Lambda\theta_* + \frac{\sigma^2}{n}\operatorname{tr}(((\hat\Sigma + \Lambda)^{-1}\hat\Sigma)^2).
$$

Exercise 3.7 (♦) *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*

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

where $\operatorname{Diag}(h)$ *denotes the diagonal matrix with h as its diagonal. Hint: use Woodbury matrix identities from section* **??***.*

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

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

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 PCA. Using the singular value decomposition of Φ, 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 of D; compute the corresponding updates.*

Exercise 3.10 (K-means clustering) *Given $\Phi \in \mathbb{R}^{n \times d}$, we consider minimizing the objective $\|\Phi - AD\|_\mathrm{F}^2$ 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 1. Compute the updates of an alternating optimization algorithm that minimizes $\|\Phi - AD\|_\mathrm{F}^2$.*

Chapter 4

Empirical Risk Minimization

Exercise 4.1 (♦) *On top of the assumptions made in this section, assume that* $a(0) = 0$*. Show that if* a^* *is the Fenchel conjugate of* a*, then for any function* $g : \mathcal{X} \to \mathbb{R}$*, we have* $a^*\big(\mathcal{R}(g) - \mathcal{R}^*\big) \leqslant \mathcal{R}_\Phi(g) - \mathcal{R}^*_\Phi$*.*

Exercise 4.2 (♦♦) *Consider a convex function* $\Phi : \mathbb{R} \to \mathbb{R}$*, which is differentiable at zero with* $\Phi'(0) < 0$*. Define* $G(z) = \Phi(0) - \inf_{u \in \mathbb{R}} \big\{ \frac{1+z}{2}\Phi(u) + \frac{1-z}{2}\Phi(-u) \big\}$*. Show that* G *is convex,* $G(0) = 0$*, and* $G\big[\mathcal{R}(g) - \mathcal{R}^*\big] \leqslant \mathcal{R}_\Phi(g) - \mathcal{R}^*_\Phi$ *for any function* $g : \mathcal{X} \to \mathbb{R}$*. Compute* G *for the exponential loss.*

Exercise 4.3 (♦) *Assume that* $|2\eta(x) - 1| > \varepsilon$ *almost surely for some* $\varepsilon \in (0, 1]$*. Show that for any smooth convex classification-calibrated function* $\Phi : \mathbb{R} \to \mathbb{R}$ *of the form* $\Phi(v) = a(v) - v$ *as in this section, then we have* $\mathcal{R}(g) - \mathcal{R}(g_*) \leqslant \frac{\varepsilon}{a^*(\varepsilon)}\big[\mathcal{R}_\Phi(g) - \mathcal{R}^*_\Phi\big]$ *for any function* $g : \mathcal{X} \to \mathbb{R}$*.*

Exercise 4.4 *For the logistic loss, show that for data generated with class-conditional densities of* $x|y = 1$ *and* $x|y = -1$*, which are Gaussians with the same covariance matrix, the function* $g(x)$ *minimizing the expected logistic loss is affine in* x*. This model is often referred to as "linear discriminant analysis (LDA)." Provide an extension to the multicategory setting.*

Exercise 4.5 *Show that for* $\Theta = \{\theta \in \mathbb{R}^d,\ \|\theta\|_1 \leqslant D\}$ *(*ℓ_1*-norm instead of the* ℓ_2*-norm), we have*

$$\inf_{\theta \in \Theta} \mathcal{R}(f_\theta) - \inf_{\theta \in \mathbb{R}^d} \mathcal{R}(f_\theta) \leqslant G\, \mathbb{E}\big[\|\varphi(x)\|_\infty\big](\|\theta_*\|_1 - D)_+.$$

Generalize to all norms.

Exercise 4.6 (♦) *Provide an explicit bound on* $\sup_{\|\theta\|_2 \leqslant D} |\mathcal{R}(f) - \widehat{\mathcal{R}}(f)|$*, and compare it to using Rademacher complexities in section* **??***. The concentration of averages of matrices from section* **??** *can be used.*

Exercise 4.7 (♦) *In terms of expectation, show the following (using the proof of the max of random variables from section* **??***, which applies because bounded random variables are sub-Gaussian):*

$$E\Big[\sup_{f\in\mathcal{F}}\big|\widehat{\mathcal{R}}(f)-\mathcal{R}(f)\big|\Big] \leqslant \ell_\infty\sqrt{\frac{\log(2|\mathcal{F}|)}{2n}}.$$

Exercise 4.8 *Let $m(\varepsilon)$ be the covering number of a unit ball of $\mathbb{R}^d$ by balls of radius ε for an arbitrary norm. Using comparisons of volumes, show that $\left(\frac{1}{\varepsilon}\right)^d \leqslant m(\varepsilon) \leqslant \left(1+\frac{2}{\varepsilon}\right)^d$.*

Exercise 4.9 *Show the following properties of Rademacher complexities (see* **?***, for more details):*

- *If $\mathcal{H}\subset\mathcal{H}'$, then $\mathrm{R}_n(\mathcal{H})\leqslant \mathrm{R}_n(\mathcal{H}')$.*
- $\mathrm{R}_n(\mathcal{H}+\mathcal{H}') = \mathrm{R}_n(\mathcal{H})+\mathrm{R}_n(\mathcal{H}')$.
- *If $\alpha\in\mathbb{R}$, $\mathrm{R}_n(\alpha\mathcal{H}) = |\alpha|\cdot\mathrm{R}_n(\mathcal{H})$.*
- *If $h_0:\mathcal{Z}\to\mathbb{R}$, $\mathrm{R}_n(\mathcal{H}+\{h_0\}) = \mathrm{R}_n(\mathcal{H})$.*
- $\mathrm{R}_n(\mathcal{H}) = \mathrm{R}_n(\text{convex hull}(\mathcal{H}))$.

Solution.

- We define $\mathcal{H},\mathcal{H}'$ s.t $\mathcal{H}\subset\mathcal{H}'$. Let $\mathcal{Y}_\mathcal{H} = \sup_{h\in\mathcal{H}}\varepsilon^\top(h(z_1),\dots,h(z_n))$ and $\mathcal{Y}_{\mathcal{H}'}$ defined similarly. Since $\mathcal{Y}_{\mathcal{H}'}$ is obtained by maximizing over a larger set, it is larger, hence the result.
- We define $\mathcal{H},\mathcal{H}'$ and want to compute $\mathcal{R}_n(\mathcal{H}+\mathcal{H}')$. We have $\mathcal{H}+\mathcal{H}' = \{h+h', h\in\mathcal{H}, h'\in\mathcal{H}'\}$. Therefore, by linearity of the expectation, and additivity of the evaluated expression w.r.t h (meaning that $\sup_{h\in\mathcal{H},h'\in\mathcal{H}'}\cdots = \sup_{h\in\mathcal{H}}\cdots + \sup_{h'\in\mathcal{H}'}\dots$ here), we get $\mathcal{R}_n(\mathcal{H}+\mathcal{H}') = \mathcal{R}_n(\mathcal{H})+\mathcal{R}_n(\mathcal{H}')$.
- Let $\alpha\in\mathbb{R}$. If $\alpha\geqslant 0$, the result is obvious. If $\alpha\leqslant 0$, let's consider the expectation w.r.t the Rademacher variables $(\varepsilon'_1,\dots,\varepsilon'_n)\sim-(\varepsilon_1,\dots,\varepsilon_n)$ (by symmetry). We therefore have to compute $\mathbb{E}_{\varepsilon',\mathcal{D}}(\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^n\varepsilon'_i(-\alpha)h(z_i))$. As $-\alpha=|\alpha|$ is positive, we clearly have $\mathcal{R}_n(\alpha\mathcal{H}) = |\alpha|\mathcal{R}_n(\mathcal{H})$. This concludes the proof.
- We have

  $$\mathcal{R}_n(\{h_0\}) = \mathbb{E}_{\varepsilon,\mathcal{D}}(\frac{1}{n}\sum_{i=1}^n\varepsilon_i h_0(z_i)) = \mathbb{E}_\mathcal{D}(\frac{1}{n}\sum_{i=1}^n\mathbb{E}_\varepsilon(\varepsilon_i)h_0(z_i)) = 0,$$

  using that we evaluate the sup on a singleton. The result follows from the second property shown in this exercise.
- We clearly have $\mathcal{R}_n(\mathcal{H})\leqslant\mathcal{R}_n(\text{convex hull}(\mathcal{H}))$ by using $\mathcal{H}\subset$ convex hull$(\mathcal{H})$ and the first result of this exercise. We therefore want to show $\mathcal{R}_n(\mathcal{H})\geqslant\mathcal{R}_n(\text{convex hull}(\mathcal{H}))$.

Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ be a draw of Rademacher variables ; let $\tilde{h} \in$ convex hull $\mathcal{H}$. There exists $(\alpha_i)_{i\in\{1,\ldots,m\}} \in \mathbb{R}^m$, which sum to 1, and $(h_i)_{i\in\{1,\ldots,m\}} \in \mathcal{H}^m$ s.t. $\tilde{h} = \sum_{k=1}^m \alpha_k h_k$. We have :

$$\begin{aligned}\sum_{i=1}^n \varepsilon_i \tilde{h}(z_i) &= \sum_{i=1}^n \varepsilon_i \sum_{k=1}^m \alpha_k h_k(z_i) \\ &= \sum_{k=1}^m \alpha_k \sum_{i=1}^n \varepsilon_i h_k(z_i) \\ &\leqslant \sum_{k=1}^m \alpha_k \sup_{h\in\mathcal{H}} \sum_{i=1}^n (\varepsilon_i h(z_i)) \\ &\leqslant \sup_{h\in\mathcal{H}} \sum_{i=1}^n \varepsilon_i h(z_i).\end{aligned}$$

Therefore, $\sup_{\tilde{h}\in\text{convex hull}(\mathcal{H})} \sum_{i=1}^n \varepsilon_i \tilde{h}(z_i) \leqslant \sup_{h\in\mathcal{H}} \sum_{i=1}^n \varepsilon_i h(z_i)$. Taking the expectancy concludes the proof.

Exercise 4.10 (Massart's lemma) *Assume that $\mathcal{H} = \{h_1, \ldots, h_m\}$, and almost surely we have the bound $\frac{1}{n}\sum_{i=1}^n h_j(x_i)^2 \leqslant R^2$ for all $j \in \{1, \ldots, m\}$. Show that the Rademacher complexity of the class of functions $\mathcal{H}$ satisfies $\mathrm{R}_n(\mathcal{H}) \leqslant \sqrt{\frac{2\log m}{n}}R$.*

Exercise 4.11 (♦) *The Gaussian complexity of a class of functions $\mathcal{H}$ from $\mathcal{Z}$ to $\mathbb{R}$ is defined as $\mathrm{G}_n(\mathcal{H}) = \mathbb{E}_{\varepsilon,\mathcal{D}}\big[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \varepsilon_i h(z_i)\big]$, where $\varepsilon \in \mathbb{R}^n$ is a vector of independent Gaussian variables with mean zero and variance 1. Show that (1) $\mathrm{R}_n(\mathcal{H}) \leqslant \sqrt{\frac{\pi}{2}} \cdot \mathrm{G}_n(\mathcal{H})$ and (2) $\mathrm{G}_n(\mathcal{H}) \leqslant \sqrt{2\log(2n)} \cdot \mathrm{R}_n(\mathcal{H})$.*

Exercise 4.12 (ℓ_1-norm) *Assume that almost surely, $\|\varphi(x)\|_\infty \leqslant R$. Show that the Rademacher complexity $\mathrm{R}_n(\mathcal{F})$ for $\mathcal{F} = \{f_\theta(x) = \theta^\top\varphi(x),\ \Omega(\theta) \leqslant D\}$, with $\Omega = \|\cdot\|_1$, is upper-bounded by $RD\big(\frac{2\log(2d)}{n}\big)^{1/2}$.*

Solution. We use $\mathcal{R}_n(\mathcal{F}) = \frac{D}{n}\mathbb{E}(||\Phi^\top\varepsilon||_\infty)$ ($\|\cdot\|_\infty$ is the dual norm of $\|\cdot\|_1$). We want to upper-bound

$$\mathbb{E}[||\Phi^\top\varepsilon||_\infty] = \mathbb{E}\Big[\max_{1\leqslant i\leqslant d} \max\Big\{\sum_{j=1}^n \varphi_j(x_i)\varepsilon_i, -\sum_{j=1}^n \varphi_j(x_i)\varepsilon_i\Big\}\Big],$$

which is the maximum of $2d$ random variables.

Since $\|\varphi(x)\|_\infty \leqslant R$ almost surely, we have $|\varphi_j(x)| \leqslant R$ for all j. The random variables $\varepsilon_i\varphi_j(x_i)$ are therefore bounded by R and $-R$ and are sub-Gaussian with a sub-Gaussian parameter $\sigma^2 = R^2$. The sum $\sum_{j=1}^n \varphi_j(x_i)\varepsilon_i$ is therefore also sub-Gaussian (as the summed random variables are independent) with a parameter $\tau^2 = nR^2$. So is $-\sum_{j=1}^n \varphi_j(x_i)\varepsilon_i$.

Using the result from section 1.2.4, we can bound the expectation of the maximum of these $2d$ variables by $\sqrt{2\tau^2 \log d} = R\sqrt{2n \log 2d}$.

Combining it to the first result, we obtain :

$$\mathcal{R}_n(\mathcal{F}) = RD\sqrt{\frac{2\log 2d}{n}}.$$

Exercise 4.13 (♦) *Let $p > 1$, and q such that $1/p + 1/q = 1$. Assume that almost surely, $\|\varphi(x)\|_q \leqslant R$. Show that the Rademacher complexity $\mathrm{R}_n(\mathcal{F})$ for $\mathcal{F} = \{f_\theta(x) = \theta^\top \varphi(x),\ \Omega(\theta) \leqslant D\}$, with $\Omega = \|\cdot\|_p$, is upper-bounded by $\frac{RD}{\sqrt{n}}\frac{1}{\sqrt{p-1}}$ (hint: use exercise 1.25). Recover the result of exercise 4.12 by taking $p = 1 + \frac{1}{\log(2d)}$.*

Exercise 4.14 *Consider a learning problem with 1-Lipschitz-continuous loss (with respect to the second variable), a function class $f_\theta(x) = \theta^\top\varphi(x)$, $\|\theta\|_1 \leqslant D$, and $\varphi : \mathcal{X} \to \mathbb{R}^d$, with $\|\varphi(x)\|_\infty$ almost surely less than R. Given the expected risk $\mathcal{R}(f_\theta)$ and the empirical risk $\widehat{\mathcal{R}}(f_\theta)$. Show that $\mathbb{E}\big[\mathcal{R}(f_{\hat\theta})\big] \leqslant \inf_{\|\theta\|_1 \leqslant D} \mathcal{R}(f_\theta) + 4RD\sqrt{2\log(2d)/n}$, for the constrained empirical risk minimizer $f_{\hat\theta}$.*

Exercise 4.15 (♦♦) *Extend the result in proposition **??** to features that are almost surely bounded in the ℓ_p-norm by R, and a regularizer ψ that is strongly convex with respect to the ℓ_p-norm; that is, such that for all $\theta, \eta \in \mathbb{R}^d$, $\psi(\theta) \geqslant \psi(\eta) + \psi'(\eta)^\top(\theta - \eta) + \frac{\mu}{2}\|\theta - \eta\|_p^2$, for some $\mu > 0$, where $\psi'(\eta)$ is a subgradient of ψ at η. Hint: use exercise 4.13.*

Exercise 4.16 (♦) *Consider a learning algorithm and a distribution p on (x, y) such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and two outputs $f, g : \mathcal{X} \to \mathcal{Y}$ of the learning algorithm on datasets of n observations that differ by a single observation, $|\ell(y, f(x)) - \ell(y, g(x))| \leqslant \beta_n$, an assumption referred to as "uniform stability." Show that the expected deviation between the expected risk and the empirical risk of the algorithm's output is bounded by β_n. With the same assumptions as in proposition **??**, show that we have $\beta_n = \frac{2G^2R^2}{\lambda n}$ (see **?**, for more details).*

Chapter 5

Optimization for Machine Learning

Exercise 5.1 *Let μ_+ be the smallest nonzero eigenvalue of H. Show that GD is linearly convergent with a convergence rate proportional to $(1-\mu_+/L)^t$ after t iterations.*

Solution. We have, for any $\lambda \in \Lambda(H)$ the eigenvalues of H :

$$\left|\lambda\left(1-\frac{\lambda}{L}\right)^{2t}\right| \leqslant \max_{\substack{\lambda' \in \Lambda(H) \\ \lambda > 0}} \left|\lambda'\left(1-\frac{\lambda'}{L}\right)^{2t}\right| \leqslant L \max_{\substack{\lambda' \in \Lambda(H) \\ \lambda > 0}} \left(1-\frac{\lambda'}{L}\right)^{2t},$$

where we use between the first and second terms that $\lambda = 0$ (if it exists) can not be a maximizer, and between the second and third terms that for a, b positive, $\max(ab) \leqslant \max(a)\max(b)$.

As $\Lambda(H) \cap \mathbb{R}^* \subset [\mu_+, L]$, this gives the expected result directly, having

$$|F(\theta_t) - F(\eta_*)| \leqslant \frac{L}{2}\left(1-\frac{\mu_+}{L}\right)^{2t} \|\theta_0 - \eta_*\|_2^2.$$

Exercise 5.2 (Exact line search (♦)) *For the quadratic objective in equation (**??**), show that the optimal step size γ_t in equation (**??**) is equal to $\gamma_t = \frac{\|F'(\theta_{t-1})\|_2^2}{F'(\theta_{t-1})^\top H F'(\theta_{t-1})}$. Show that when F is strongly convex, we have $F(\theta_t) - F(\eta_*) \leqslant \left(\frac{\kappa-1}{\kappa+1}\right)^2 \left[F(\theta_{t-1}) - F(\eta_*)\right]$, and compare the rate with constant step size GD. Hint: prove and use the Kantorovich inequality $\sup_{\|z\|_2=1} z^\top H z z^\top H^{-1} z = \frac{(L+\mu)^2}{4\mu L}$.*

Exercise 5.3 *Assume that function $F : \mathbb{R}^d \to \mathbb{R}$ is strictly convex; that is, $\forall \theta, \eta \in \mathbb{R}^d$ such that $\theta \neq \eta$ and $\alpha \in (0,1)$, $F(\alpha\eta + (1-\alpha)\theta) < \alpha F(\eta) + (1-\alpha)F(\theta)$. Show that there* *is equality in Jensen's inequality in equation (**??**) if and only if the random variable θ is almost surely constant.*

Exercise 5.4 *Identify all stationary points in the function in $\mathbb{R}^2$ depicted here:*

Exercise 5.5 *Show that function $F : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly-convex if and only if function $\theta \mapsto F(\theta) - \frac{\mu}{2}\|\theta\|_2^2$ is convex.*

Exercise 5.6 *Show that if function $F : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly-convex, then it has a unique minimizer.*

Exercise 5.7 (♦) *Show that the differentiable function $F : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly convex if and only if for all $\theta, \eta \in \mathbb{R}^d$, $\|F'(\theta) - F'(\eta)\|_2 \geqslant \mu\|\theta - \eta\|_2$.*

Exercise 5.8 (♦) *Consider angle α between the descent direction $-F'(\theta)$ and the deviation to optimum $\theta - \eta_*$, defined through $\cos\alpha = \frac{F'(\theta)^\top(\theta-\eta_*)}{\|F'(\theta)\|\cdot\|\theta-\eta_*\|_2}$. Show that for a μ-strongly-convex, L-smooth quadratic function, $\cos\alpha \geqslant \frac{2\sqrt{\mu L}}{L+\mu}$. (Hint: prove and use the Kantorovich inequality $\sup_{\|z\|_2=1} z^\top H z z^\top H^{-1} z = \frac{(L+\mu)^2}{4\mu L}$.) (♦♦) Show that the same result holds without the assumption that F is quadratic. (Hint: use the co-coercivity of the function $\theta \mapsto F(\theta) - \frac{\mu}{2}\|\theta\|_2^2$ from proposition **??**.)*

Exercise 5.9 *Compute all constants for ℓ_2-regularized logistic regression and for ridge regression.*

Solution. For ridge regression with data matrix $X \in \mathbb{R}^{n\times d}$, the Hessian of the cost function is $X^\top X/n + \lambda I$, thus the objective function has smoothness constant $\lambda_{\max}(X^\top X/n) + \lambda$, and strong convexity constant $\lambda_{\min}(X^\top X/n) + \lambda$.

For logistic regression, we define the regularized empirical risk as

$$\widehat{\mathcal{R}}(\theta) = \frac{1}{n}\sum_{i=1}^n \log(1+\exp(-y_i\theta^\top x_i)) + \frac{\lambda}{2}\|\theta\|_2^2,$$

with gradient

$$\widehat{\mathcal{R}}''(\theta) = -\frac{1}{n}\sum_{i=1}^n \frac{\exp(-y_i\theta^\top x_i)}{1+\exp(-y_i\theta^\top x_i)} y_i x_i + \lambda\theta = -\frac{1}{n}\sum_{i=1}^n \frac{1}{1+\exp(y_i\theta^\top x_i)} y_i x_i + \lambda\theta,$$

and Hessian

$$\widehat{\mathcal{R}}''(\theta) = \frac{1}{n}\sum_{i=1}^n \frac{\exp(y_i\theta^\top x_i)}{(1+\exp(-y_i\theta^\top x_i))^2} x_i x_i^\top + \lambda I.$$

The scalar $\frac{\alpha}{(1+\alpha)^2} = \frac{\alpha}{1+\alpha} \times \left(1 - \frac{\alpha}{\alpha+1}\right)$ is always between 0 and 1/4, thus, we have, using the Löwner ordering between symmetric matrices:

$$\lambda I \preccurlyeq \widehat{\mathcal{R}}''(\theta) \preccurlyeq \frac{1}{4n} X^\top X + \lambda I$$

with $X \in \mathbb{R}^{n\times d}$ the data matrix. Thus, $\widehat{\mathcal{R}}$ has a smoothness constant less than $\frac{1}{4}\lambda_{\max}(X^\top X/n)+\lambda$ and is λ-strongly-convex.

Exercise 5.10 *Let F be an L-smooth convex function on $\mathbb{R}^d$. Show that its Fenchel conjugate is $(1/L)$-strongly convex.*

Exercise 5.11 (Fenchel-Young inequality) *Let F be an L-smooth convex function on $\mathbb{R}^d$ and F^* be its Fenchel conjugate. Show that for any $\theta, z \in \mathbb{R}^d$, we have $F(\theta) + F^*(z) - z^\top\theta \geqslant 0$, if and only if $z = F'(\theta)$. (♦) Show in addition that we have the lower bound $F(\theta) + F^*(z) - z^\top\theta \geqslant \frac{1}{2L}\|z - F'(\theta)\|_2^2$.*

Exercise 5.12 (Alternative convergence proof - I) *Consider an L-smooth convex function with a global minimizer η_*, and GD with step size $\gamma_t = 1/L$:*

- *Using proposition **??**, show that $\|\theta_t - \eta_*\|_2^2 \leqslant \|\theta_{t-1} - \eta_*\|_2^2 - \frac{1}{L}F'(\theta_{t-1})^\top(\theta_{t-1} - \eta_*)$.*
- *Show that $F(\theta_t) \leqslant F(\theta_{t-1})$.*
- *Using a telescoping sum, show that $F(\theta_t) - F(\eta_*) \leqslant \frac{L}{t+1}\|\theta_0 - \eta_*\|_2^2$.*

Exercise 5.13 (Alternative convergence proof - II (♦)) *Consider an L-smooth convex function with a global minimizer η_*, and GD with step size $\gamma_t = 1/L$:*

- *Show that $\|\theta_t - \eta_*\|_2^2 \leqslant \|\theta_{t-1} - \eta_*\|_2^2$ for all $t \geqslant 1$.*
- *Show that $F(\theta_t) \leqslant F(\theta_{t-1}) - \frac{1}{2L}\|F'(\theta_{t-1})\|_2^2$ for all $t \geqslant 1$.*
- *Denoting $\Delta_t = F(\theta_t) - F(\eta_*)$, show that $\Delta_t \leqslant \Delta_{t-1} - \frac{1}{2L\|\theta_0-\eta_*\|_2^2}\Delta_{t-1}^2$ for all $t \geqslant 1$. Conclude that $\Delta_t \leqslant \frac{2L}{t+4}\|\theta_0 - \eta_*\|_2^2$.*

Exercise 5.14 (♦♦) *For the updates in equations (**??**) and (**??**), show that for the Lyapunov function $V(\theta,\eta) = f(\theta) - f(\eta_*) + \frac{\mu}{2}\big\|\theta - \eta_* + (1+\sqrt{L/\mu})(\eta - \theta)\big\|_2^2$, then we have $V(\theta_t,\eta_t) \leqslant (1 - \sqrt{\mu/L})V(\theta_{t-1},\eta_{t-1})$. Show that this implies a convergence rate proportional to $(1-\sqrt{\mu/L})^t$.*

Exercise 5.15 (♦♦) *For the updates in equations (**??**) and (**??**), show that for the Lyapunov function $V_t(\theta,\eta) = \big(\frac{t+1}{2}\big)^2\big[f(\theta) - f(\eta_*)\big] + \frac{L}{2}\big\|\eta - \eta_* + \frac{t}{2}(\eta - \theta)\big\|_2^2$, then we have $V_t(\theta_t,\eta_t) \leqslant V_{t-1}(\theta_{t-1},\eta_{t-1})$. Show that this implies a convergence rate proportional to $1/t^2$.*

Exercise 5.16 (♦) *Assume that function F is μ-strongly convex, twice-differentiable, and such that the Hessian is Lipschitz-continuous: $\|f''(\theta) - f''(\eta)\|_{\rm op} \leqslant M\|\theta - \eta\|_2$. Using Taylor's formula with an integral remainder, show that for the iterates of Newton's method, $\|\nabla F(\theta_t)\|_2 \leqslant \frac{M}{2\mu^2}\|\nabla F(\theta_{t-1})\|_2^2$. Show that this implies local quadratic convergence.*

Exercise 5.17 (Convergence of proximal gradient method) *Consider a convex L-smooth function G and a convex function H defined on $\mathbb{R}^d$. We consider the update in* *equation (**??**) and a minimizer η_* of $G + H$.*

- *Show that $G(\theta_t) \leqslant G(\theta_{t-1}) + G'(\theta_{t-1})^\top(\theta_t - \theta_{t-1}) + \frac{L}{2}\|\theta_t - \theta_{t-1}\|_2^2$.*
- *Show that $G(\theta_{t-1}) \leqslant G(\eta_*) + G'(\theta_{t-1})^\top(\theta_{t-1} - \eta_*)$.*
- *Show that $H(\theta_t) \leqslant H(\eta_*) + (L\theta_{t-1} - L\theta_t - G'(\theta_{t-1}))^\top(\theta_t - \eta_*)$.*
- *Deduce that $G(\theta_t) + H(\theta_t) \leqslant G(\eta_*) + H(\eta_*) + \frac{L}{2}\|\theta_{t-1} - \eta_*\|_2^2 - \frac{L}{2}\|\theta_t - \eta_*\|_2^2$.*
- *Conclude that for $t \geqslant 1$, $G(\theta_t) + H(\theta_t) - \big[G(\eta_*) + H(\eta_*)\big] \leqslant \frac{L}{2t}\|\theta_0 - \eta_*\|_2^2$.*

Exercise 5.18 *Show that if F is differentiable, B-Lipschitz-continuity is equivalent to the assumption $\|F'(\theta)\|_2 \leqslant B$, $\forall \theta \in \mathbb{R}^d$.*

Exercise 5.19 *Compute the subdifferential of $\theta \mapsto |\theta|$ and $\theta \mapsto (1 - y\theta^\top x)_+$ for the label $y \in \{-1, 1\}$ and the input $x \in \mathbb{R}^d$.*

Exercise 5.20 *Consider the iteration $\theta_t = \theta_{t-1} - \frac{\gamma_t'}{\|F'(\theta_{t-1})\|_2} F'(\theta_{t-1})$. Show that with the step size $\gamma_t' = D/\sqrt{t}$ (independent of B), we get the following guarantee: $\min_{0 \leqslant s \leqslant t-1} F(\theta_s) - F(\eta_*) \leqslant DB\frac{2+\log(t)}{2\sqrt{t}}$.*

Exercise 5.21 *Let $K \subset \mathbb{R}^d$ be a convex closed set, and denote as $\Pi_K(\theta)$ the orthogonal projection of θ onto K, defined as $\Pi_K(\theta) = \arg\min_{\eta \in K} \|\eta - \theta\|_2^2$. Show that function Π_K is contractive; that is, for all $\theta, \eta \in \mathbb{R}^d$, $\|\Pi_K(\theta) - \Pi_K(\eta)\|_2 \leqslant \|\theta - \eta\|_2$. For the algorithm $\theta_t = \Pi_K(\theta_{t-1} - \gamma_t F'(\theta_{t-1}))$, and with η_* being a minimizer of F on K, show that the guarantee of proposition **??** still holds.*

Exercise 5.22 (♦) *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a differentiable function, and $\psi : \mathbb{R}^d \to \mathbb{R}$ a strictly convex function.*

- *Show that the minimizer of $F(\theta) + F'(\theta)^\top(\eta - \theta) + \frac{1}{2\gamma}\|\eta - \theta\|_2^2$ is equal to $\eta = \theta - \gamma F'(\theta)$.*
- *Show that the Bregman divergence $D_\psi(\eta, \theta)$, defined as $D_\psi(\eta, \theta) = \psi(\eta) - \psi(\theta) - \psi'(\theta)^\top(\eta - \theta)$, is nonnegative and equal to zero if and only if $\eta = \theta$.*
- *Show that the minimizer of $F(\theta) + F'(\theta)^\top(\eta - \theta) + \frac{1}{\gamma}D_\psi(\eta, \theta)$ satisfies $\psi'(\eta) = \psi'(\theta) - \gamma F'(\theta)$. Show that the same conclusion holds if ψ is only defined on an open convex set $K \subset \mathbb{R}^d$, and the gradient ψ' is a bijection from K to $\mathbb{R}^d$.*
- *Provide an explicit form of the resulting algorithm when $\psi(\theta) = \sum_{i=1}^d \theta_i \log \theta_i$.*

Exercise 5.23 (♦) *Consider the same assumptions as exercise 5.21 and the same algorithm with orthogonal projections. With D being the diameter of K, show that for the average iterate $\bar\theta_t = \frac{1}{t}\sum_{s=0}^{t-1}\theta_s$, we have $F(\bar\theta_t) - F(\theta_*) \leqslant \frac{3BD}{2\sqrt{t}}$.*

Exercise 5.24 (Doubling trick for subgradient method) *Consider an algorithm that successively applies the SGD iteration with step size $\gamma = D/(B\sqrt{2^k})$ during 2^k itera-*

tions, for $k = 0, 1, \dots$. Show that after t subgradient iterations, the observed best expected value of F is less than a constant times $DB/\sqrt{t}$.

Exercise 5.25 *Compute all constants for ℓ_2-regularized logistic regression and the support vector machine (SVM) with linear predictors (section **??**).*

Exercise 5.26 (High-probability bound for SGD (♦)) *Using the same assumptions and notations as in proposition **??**, we consider the projected SGD iteration: $\theta_t = \Pi_D(\theta_{t-1} - \gamma_t g_t)$, where Π_D is the orthogonal projection on the ℓ_2-ball with center 0 and radius D. Denoting $z_t = -\gamma_t(\theta_{t-1} - \theta_*)^\top[g_t - F'(\theta_{t-1})]$, show that $\mathbb{E}[z_t|\mathcal{F}_{t-1}] = 0$ and $|z_t| \leqslant 4\gamma_t BD$ almost surely, and*

$$\gamma_t[F(\theta_{t-1}) - F(\theta_*)] \leqslant \frac{1}{2}\Big(\mathbb{E}\big[\|\theta_{t-1} - \theta_*\|_2^2\big] - \mathbb{E}\big[\|\theta_t - \theta_*\|_2^2\big]\Big) + \frac{1}{2}\gamma_t^2 B^2 + z_t.$$

*Using Azuma's inequality (see exercise 1.14), show that with probability at least $1 - \delta$, then, for the weighted average $\bar{\theta}_t$ defined in proposition **??**, for any step sizes γ_t:*

$$F(\bar{\theta}_t) - F(\theta_*) \leqslant \frac{2D^2}{\sum_{s=1}^t \gamma_s} + B^2 \frac{\sum_{s=1}^t \gamma_s^2}{2\sum_{s=1}^t \gamma_s} + 4BD \frac{\big(\sum_{s=1}^t \gamma_s^2\big)^{1/2}}{\sum_{s=1}^t \gamma_s}\sqrt{2\log\frac{1}{\delta}},$$

and for a constant step size, $\gamma_t = \gamma$, $F(\bar{\theta}_t) - F(\theta_*) \leqslant \frac{2D^2}{\gamma T} + \frac{\gamma B^2}{2} + \frac{4DB}{\sqrt{t}}\sqrt{2\log\frac{1}{\delta}}$ (for the uniformly averaged iterate).

Exercise 5.27 (Minibatch SGD) *Consider the mini-batch version of SGD, where at every iteration, we replace $g_t(\theta_{t-1})$ by the average of m independent samples of stochastic gradients at θ_{t-1}. Show that the convergence result of proposition **??** still holds.*
(♦) Which assumption on gradients would improve the convergence rate?

Exercise 5.28 (SGD for smooth functions (♦)) *Consider independent and identically distributed (i.i.d.) convex L-smooth random functions $f_t : \mathbb{R}^d \to \mathbb{R}$, $t \geqslant 1$, with expectation $F : \mathbb{R}^d \to \mathbb{R}$, which has a minimizer $\theta_* \in \mathbb{R}^d$. Consider the SGD recursion $\theta_t = \theta_{t-1} - \gamma_t f_t'(\theta_{t-1})$, with γ_t being a deterministic step-size sequence. Using co-coercivity (proposition **??**), show that*

$$\mathbb{E}\big[\|\theta_t - \theta_*\|_2^2\big] \leqslant \mathbb{E}\big[\|\theta_{t-1} - \theta_*\|_2^2\big] - 2\gamma_t(1 - \gamma_t L)\mathbb{E}\big[F'(\theta_{t-1})^\top(\theta_{t-1} - \theta_*)\big] + 2\gamma_t^2\mathbb{E}\big[\|f_t'(\theta_*)\|_2^2\big].$$

*Extend the proof of proposition **??** to obtain an explicit rate in $O(1/\sqrt{t})$. (♦) Show that the minibatch version leads to an improvement in the rate (as opposed to the nonsmooth case in exercise 5.27).*

Exercise 5.29 (Nonuniform sampling (♦)) *Consider the function $F : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$, which is convex with respect to the first variable, with a subgradient $F'(\theta, z)$ with respect to the first variable that is bounded in the ℓ_2-norm by a constant $B(z)$ that depends on z. Consider a distribution p on $\mathcal{Z}$. We aim to minimize $\mathbb{E}_{z\sim p}[F(\theta, z)]$, but we sample from a distribution q, with density $dq/dp(z)$ with respect to p to get i.i.d. random z_t, $t \geqslant 1$.*

Consider the recursion $\theta_t = \theta_{t-1} - \frac{\gamma}{dq/dp(z_t)} F'(\theta_{t-1}, z_t)$*. Provide a convergence rate for this algorithm and show how a good choice of* q *leads to significant improvements over the choice* $q = p$ *when* $B(z)$ *is far from uniform in* z*. Apply this result to the SVM when applying SGD to the empirical risk.*

Exercise 5.30 (SGD for nonconvex functions) *Consider an* L*-smooth potentially nonconvex function* F*, and the SGD recursion with constant step size* γ*, with unbiased and bounded gradient estimates (e.g., assumptions (H-1) and (H-2)).*

- *Show that* $\mathbb{E}\big[F(\theta_t)\big] \leqslant \mathbb{E}\big[F(\theta_{t-1})\big] - \gamma \mathbb{E}\big[\|F'(\theta_{t-1}\|_2^2\big] + \frac{LB^2\gamma^2}{2}$.
- *Show that* $\frac{1}{t}\sum_{s=1}^t \mathbb{E}\big[F(\theta_{s-1})\big] \leqslant \frac{1}{\gamma t}\big[F(\theta_0) - \inf_{\eta\in\mathbb{R}^d} F(\eta)\big] + \frac{LB^2\gamma}{2}$.

Exercise 5.31 (♦) *Consider the minimization of* $F(\theta) = \frac{1}{2}\theta^\top H\theta - c^\top\theta$*, where* $H \in \mathbb{R}^{d\times}$ *is positive-definite (and thus invertible), and the recursion* $\theta_t = \theta_{t-1} - \gamma[F'(\theta_{t-1}) + \varepsilon_t]$*, where all* ε_t*'s are independent, with zero mean and covariance matrix equal to* C*. Compute explicitly* $\mathbb{E}\big[F(\theta_t) - F(\theta_*)\big]$*, and provide an upper bound of* $\mathbb{E}\big[F(\bar\theta_t) - F(\theta_*)\big]$*, where* $\bar\theta_t = \frac{1}{t}\sum_{s=0}^{t-1}\theta_s$*.*

Exercise 5.32 *With the same assumptions as proposition* **??***, show that with the step size* $\gamma_t = \frac{2}{\mu(t+1)}$*, and with* $\bar\theta_t = \frac{2}{t(t+1)}\sum_{s=1}^t s\theta_{s-1}$*, we have* $\mathbb{E}\big[G(\bar\theta_t) - G(\theta_*)\big] \leqslant \frac{8B^2}{\mu(t+1)}$*.*

Exercise 5.33 *Consider the minimization of* $F(\theta) = \mathbb{E}\big[\|\theta - z\|_2^2/2\big]$ *from i.i.d. observations* $z_1, \dots, z_t$*. Show that the* t*-th iterate of SGD equals* $\frac{1}{t}(z_1 + \cdots + z_t)$*.*

Exercise 5.34 (♦♦) *With the same assumptions as in proposition* **??***, with step size* $\gamma_t = 1/(B^2\sqrt{t} + \mu t)$*, provide a convergence rate for the averaged iterate.*

Exercise 5.35 (Weaker assumptions) *Consider a joint distribution on* $(x, y) \in \mathcal{X}\times\mathbb{R}$*, and a feature map* $\varphi : \mathcal{X} \to \mathbb{R}^d$ *bounded by* R *in the* ℓ_2*-norm. Denoting* θ_* *a minimizer of* $\mathbb{E}\big[(y - \varphi(x)^\top\theta)^2\big]$ *with respect to* θ*, show that the bound in equation (***??***) applies with* $\sigma^2 = \mathbb{E}\big[(y - \varphi(x)^\top\theta_*)^2\big]$*.*

Exercise 5.36 *Check the homogeneity of all quantities of this section (step size and convergence rates).*

Chapter 6

Local Averaging Methods

Exercise 6.1 *For k-nearest-neighbors and partitioning estimates, what is the pattern of nonzeros in the smoothing matrix $H \in \mathbb{R}^{n\times n}$?*

Solution. Common to both cases, we have a sparse pattern of nonzeros in the smoothing matrix. This comes from the fact that in usual settings, we have either k small compared to the number of points (k-NN) or J, the number of sets, big enough to capture meaningful patterns in the data (partitions).

For k-NN, following the book's notations, we have :

$$w_i(x) = \begin{cases} \frac{1}{k} & \text{if } i \in \{i_1(x), \dots, i_k(x)\}, \\ 0 & \text{otherwise} \end{cases},$$

where $\{i_1(x), \dots, i_k(x)\}$ are the indices of the k-closest elements of $(x_j)_{1\le j\le n}$ to x.

Therefore, unless we specify (by convention) that $w_i(x_i) = 0$, we have $\mathrm{diag}(H)_i = 1/k$. This means that on each column, $k-1$ other cases are equal to $1/k$, and the rest equal to 0, but no specific pattern can be found.

Moreover, the smoothing matrix is not symmetric (the point x_i being among the closest points to a certain x_j does not necessarily mean that the opposite stands).

For the partitioning case, unlike k-NN, in the case of partitions, the space segmentation is the same for all points (whereas it is local for KNN, as explained before). Therefore, the smoothing matrix H is symmetric.

Moreover, by rearranging the points' indices s.t, if φ is a permutation of $\{1, \dots, n\}$, we have $(x_{\varphi(1)}, \dots, x_{\varphi(n_{A_1})} \in A_1$, $(x_{\varphi(n_{A_1}+1)}, \dots, x_{\varphi(n_{A_1}+n_{A_2})}) \in A_2$, etc..., and we thus obtain a block-diagonal matrix.

Exercise 6.2 *For the binary classification problem, with $\mathcal{Y} = \{-1, 1\}$, assume that $f_*(x) = \mathbb{E}[y|x]$ is B-Lipschitz-continuous. Using section **??**, show that the excess risk* *of the majority vote is upper-bounded by*

$$\Big(B^2\int_{\mathcal{X}}\mathbb{E}\Big[\sum_{i=1}^n \hat{w}_i(x)\Delta(x_i,x)^2\Big]dp(x)+\sigma^2\sum_{i=1}^n\int_{\mathcal{X}}\mathbb{E}[\hat{w}_i(x)^2]dp(x)\Big)^{1/2}.$$

Exercise 6.3 *Show that if the Bayes rate is* 0 *(i.e.,* $\sigma = 0$*), then the* 1*-nearest-neighbor predictor is consistent.*

Solution. We note $\hat{f}_n$ the 1-NN estimator computed on n samples, and want to show that $(\hat{f}_n)_n$ converges in probability to f_*. Using proposition 6.2, we show that having $\sigma = 0$, the expected risk tends to 0 when n tends to infinity. Therefore, as the convergence in L_p norm ($p > 1$, here $p = 2$) implies the convergence in probability, we directly obtain the expected result.

Exercise 6.4 *Assume that the support* $\mathcal{X}$ *of the density* p *of inputs is bounded and that* p *is strictly positive and continuously differentiable on* $\mathcal{X}$*. Show that for* h *small enough (with an explicit upper bound), then* $C_h = \int_{\mathcal{X}} \frac{p(x)}{[q_h * p](x)}dx \leqslant \frac{1}{2}\mathrm{vol}(\mathcal{X})$*.*

Exercise 6.5 *If* $Z_1, \ldots, Z_m$ *are i.i.d. Bernoulli random variables with parameter* $\rho \in (0, 1]$*. Show that* $\mathbb{E}\big[\frac{1}{1+Z_1+\cdots+Z_m}\big] \leqslant \frac{1}{(m+1)\rho}$*.*

Exercise 6.6 (♦) *For the Nadaraya Watson estimator, show that when the target function and the kernel are twice continuously differentiable, then the bias term is bounded by a constant times* h^4*. Show that the optimal bandwidth selection leads to a rate proportional to* $n^{-4/(4+d)}$*.*

Chapter 7

Kernel Methods

Exercise 7.1 (♦♦) *Let $\mathcal{H}$ be a Hilbert space of real-valued functions on $\mathcal{X}$ endowed with a dot product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, such that for any $x \in \mathcal{X}$, the linear form $f \mapsto f(x)$ is bounded (i.e., $\sup_{f \in \mathcal{H},\ \|f\|_{\mathcal{H}} \leqslant 1} |f(x)|$ is finite). Using the Riesz representation theorem, show that this is an RKHS.*

Exercise 7.2 *Show that if $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive-definite kernel, so is the function $(x, x') \mapsto e^{k(x,x')}$.*

Solution. We have $e^{k(x,x')} = \sum_{i=0}^{+\infty} \frac{k(x,x')^i}{i!}$. Each $(x, x') \mapsto k(x, x')^i$ is a positive-definite kernel as the product of positive-definite kernels. So is their sum, hence the result. Note that this is different from the matrix exponential.

Exercise 7.3 *Show that kernel $k(x, x') = (1 + x^\top x')^s$ corresponds to the set of all monomials $x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ such that $\alpha_1 + \cdots + \alpha_d \leqslant s$. Also, show that the dimension of the feature space is $\binom{d+s}{s}$.*

Exercise 7.4 *Show that for $s = 2$, we have for all $x, x' \in [0, 1]$, $k(x, x') = q(x - x')$, with $q(t) = 1 - \frac{(2\pi)^4}{24}\big(\{t\}^4 - 2\{t\}^3 + \{t\}^2 - \frac{1}{30}\big)$.*

Exercise 7.5 (♦♦♦) *Show that we have $k(x, x') = \sum_{m \in \mathbb{Z}} \frac{e^{2im\pi(x-x')}}{1+\alpha^2|m|^2} = q(x - x')$ for $q(t) = \frac{\pi}{\alpha} \frac{\cosh \frac{\pi}{\alpha}(1-2|\{t+1/2\}-1/2|)}{\sinh \frac{\pi}{\alpha}}$. Hint: use the Cauchy residue formula.*[1]

Exercise 7.6 (Mercer kernels) *Consider a probability distribution p on a set $\mathcal{X}$, an orthonormal basis $(\varphi_i)_{i \in I}$ of the Hilbert space $L_2(p)$ of square-integrable functions (with I countable), and a summable positive sequence $(\lambda_i)_{i \in I}$. Show that the function defined as*

[1] See https://francisbach.com/cauchy-residue-formula/.

$k(x,x') = \sum_{i\in I} \lambda_i \varphi_i(x)\varphi_i(x')$ is a positive-definite kernel and describe an associated feature space.

Exercise 7.7 (Mercer decomposition (♦♦)) *Consider a probability distribution p on a set $\mathcal{X}$, a positive-definite kernel $k : \mathcal{X}\times\mathcal{X}\to\mathbb{R}$, and the operator T defined on $L_2(p)$ as $Tf(y) = \int_{\mathcal{X}} k(x,y) f(x) dp(x)$.*

- *Show that if $\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)^2 dp(x)dp(y)$ is finite, then the operator T is bounded (it is an instance of Hilbert-Schmidt integral operator[2]).*
- *Given an orthonormal basis $(e_i)_{i\in I}$ of $L_2(p)$ composed of eigenvectors for T (which is assumed to exist), show that the corresponding eigenvalues $(\lambda_i)_{i\in I}$ are nonnegative and $k(x,x') = \sum_{i\in I} \lambda_i \varphi_i(x)\varphi_i(x')$ (convergence meant in the norm $L_2(p)$).*

Exercise 7.8 (♦) *Show that column sampling corresponds to approximating optimally each $\varphi(x_j)$, $j \notin I$, by a linear combination of $\varphi(x_i)$, $i \in I$.*

Exercise 7.9 *Show that the matrix $K - K(V,I)K(I,I)^{-1}K(I,V)$ is positive-definite. If $\|M\|_*$ denotes the nuclear norm (sum of absolute values of eigenvalues of symmetric matrix M), show that the approximation error $\|K - K(V,I)K(I,I)^{-1}K(I,V)\|_*$ can be computed without the need to compute the entire matrix K.*

Exercise 7.10 *In the setup of exercise 7.6, provide a random feature expansion of Mercer kernels.*

Exercise 7.11 *(a) For ridge regression, compute the dual problem and compare the condition number of the primal problem and the condition number of the dual problem; (b) compare the two formulations to the use of normal equations as in chapter 3, and relate the two using the matrix inversion lemma $(\Phi\Phi^\top + n\lambda I)^{-1}\Phi = \Phi(\Phi^\top\Phi + n\lambda I)^{-1}$.*

Solution. Using the same notations as those of chapter 3, the primal problem of ridge regression can be expressed as

$$\min_{\substack{\theta\in\mathbb{R}^d \\ y-\Phi\theta=u}} \frac{1}{2n}\|u\|^2 + \frac{\lambda}{2}\|\theta\|^2,$$

where $\Phi \in \mathbb{R}^{n\times d}$ is the design matrix.

The associated Lagrangian is therefore

$$\mathcal{L}(\theta,u,\alpha) = \frac{1}{2n}\|u\|^2 + \frac{\lambda}{2}\|\theta\|^2 + \lambda\alpha^\top(y - \Phi\theta - u).$$

As our primal optimization problem is convex, using the saddle point theorem, we can express our minimization problem as the maximization of the dual function

$$g : \alpha \mapsto \min_{\theta\in\mathbb{R}^d,\ u\in\mathbb{R}^n} \mathcal{L}(\theta,r,\alpha).$$

[2]See https://en.wikipedia.org/wiki/Hilbert-Schmidt_integral_operator.

We compute

$$\frac{\partial \mathcal{L}}{\partial \theta} = \lambda\theta - \lambda\Phi^\top\alpha = 0 \Longleftrightarrow \theta = \Phi^\top\alpha,$$
$$\frac{\partial \mathcal{L}}{\partial r} = u/n - \alpha \Longleftrightarrow u = \alpha n.$$

Which yields $g(\alpha) = -\frac{n\lambda^2\|\alpha\|^2}{2} - \frac{\lambda\|\Phi^\top\alpha\|^2}{2} + \lambda\alpha^\top y = \lambda\alpha^\top y - \frac{\lambda}{2}\alpha^\top(\Phi\Phi^\top + n\lambda I)\alpha$.
Finally, computing $g'(\alpha) = 0$ gives

$$\hat{\alpha} = (\Phi\Phi^\top + n\lambda I)^{-1}y,$$

and therefore

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

For the condition numbers, we are interested in comparing the eigenvalues of the kernel matrix $\Phi\Phi^\top \in \mathbb{R}^{n\times n}$ and of the rescaled empirical covariance matrix $\Phi^\top\Phi \in \mathbb{R}^{d\times d}$, which share the same non-zero eigenvalue. Let nL bet the largest eigenvalue.

If $n > d$ (with $\Phi\Phi^\top$ rank-deficient), denoting $n\mu$ the smallest eigenvalue of $\Phi^\top\Phi$, the condition number of the primal problem in θ is $\frac{L+\lambda}{\mu+\lambda}$, while the one of the dual problem in α is $\frac{L+\lambda}{\lambda}$, and the one of the primal problem after having used the representer theorem to obtain a minimization problem in α is infinite (this is the minimization of $\frac{1}{n}\|y - K\alpha\|_2^2 + \frac{\lambda}{2}\alpha^\top K\alpha$, where $K = \Phi\Phi^\top$). Thus using the representer theorem is not advantageous.

If $n < d$ (with $\Phi^\top\Phi$ rank-deficient), denoting $n\mu$ the smallest eigenvalue of the kernel matrix $\Phi\Phi^\top$, the condition number of the primal problem in θ is $\frac{L+\lambda}{\lambda}$, while the one of the dual problem in α is $\frac{L+\lambda}{\mu+\lambda}$, and the one of the primal problem after having used the representer theorem to obtain a minimization problem in α is $\frac{L^2+\lambda L}{\lambda\mu}$. Again, using the representer theorem is not advantageous.

Exercise 7.12 *Write down the dual problem in equation (**??**) for the logistic loss and the for the hinge loss (compare the results to section **??**).*

Exercise 7.13 (Unregularized constant term) *Consider the minimization problem* $\min_{\theta\in\mathcal{H},c\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle\varphi(x_i),\theta\rangle + c) + \frac{\lambda}{2}\|\theta\|^2$. *If the loss function is convex with respect to the second variable, show that the dual problem is the one in equation (**??**) with the additional constraint that* $\sum_{i=1}^n \alpha_i = 0$. *Without any assumption on the loss function, show that we can restrict the search space for* θ *to all combinations* $\sum_{i=1}^n \alpha_i\varphi(x_i)$ *with the same constraint that* $\sum_{i=1}^n \alpha_i = 0$.

Exercise 7.14 (Limit of Gaussian kernel for infinite bandwidth) *Consider the minimization problem* $\min_{\theta\in\mathcal{H},c\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \ell(y_i, \langle\varphi(x_i),\theta\rangle + c) + \frac{\lambda}{2}\|\theta\|^2$ *from exercise 7.13. For the Gaussian kernel* $k(x,x') = \exp(-\|x-x'\|_2^2/r^2)$, *show that when* r *tends to infinity, the resulting prediction function is the same as the one obtained by the linear kernel* $k(x,x') = x^\top x'$ *with the regularization parameter* $\lambda r^2/2$.

Exercise 7.15 (Optimization of the kernel) *Show that for convex loss functions, the maximal value in equation (??) is a convex function of the kernel matrix K. For the square loss, show that it is equal to $\frac{\lambda}{2} y^\top (K + n\lambda I)^{-1} y$.*

Exercise 7.16 (♦) *Consider the minimization of $F(\theta) = \mathbb{E}\big[\ell(y, \langle \varphi(x), \theta \rangle)\big]$ using constant step-size SGD for a convex G-Lipschitz-continuous loss and features almost surely bounded by R. Show that after t steps (initialized at $\theta_0 = 0$ and with step size γ), the averaged iterate $\bar{\theta}_t$ satisfies $\mathbb{E}\big[F(\bar{\theta}_t)\big] \leqslant \inf_{\theta \in \mathcal{H}} \big\{F(\theta) + \frac{\|\theta\|_{\mathcal{H}}^2}{2\gamma t}\big\} + \frac{\gamma G^2 R^2}{2}$.*

Exercise 7.17 (Kernel PCA) *We consider n observations $x_1, \dots, x_n$ in a set $\mathcal{X}$ equipped with a positive-definite kernel and feature map φ from $\mathcal{X}$ to $\mathcal{H}$. Show that the largest eigenvector of the empirical noncentered covariance operator $\frac{1}{n} \sum_{i=1}^n \varphi(x_i) \otimes \varphi(x_i)$ is proportional to $\sum_{i=1}^n \alpha_i \varphi(x_i)$, where $\alpha \in \mathbb{R}^n$ is an eigenvector of the $n \times n$ kernel matrix associated with the largest eigenvalue. Given the RKHS $\mathcal{H}$ associated with kernel k, relate this eigenvalue problem to the maximizer of $\frac{1}{n} \sum_{i=1}^n f(x_i)^2$ subject to $\|f\|_{\mathcal{H}} = 1$.*

Exercise 7.18 (Kernel K-means) *Show that the K-means clustering algorithm[3] can be expressed only using dot products.*

Exercise 7.19 (Kernel quadrature) *We consider a probability distribution p on a set $\mathcal{X}$ equipped with a positive-definite kernel k with feature map $\varphi : \mathcal{X} \to \mathcal{H}$. For a function f that is linear in φ, we want to approximate $\int_{\mathcal{X}} f(x) dp(x)$ from a linear combination $\sum_{i=1}^n \alpha_i f(x_i)$ with $\alpha \in \mathbb{R}^n$.*
(a) Show that

$$\Big| \int_{\mathcal{X}} f(x) dp(x) - \sum_{i=1}^n \alpha_i f(x_i) \Big| \leqslant \|f\| \cdot \Big\| \int_{\mathcal{X}} \varphi(x) dp(x) - \sum_{i=1}^n \alpha_i \varphi(x_i) \Big\|.$$

(b) Express the square of the right side with the kernel function and show how to minimize it with respect to $\alpha \in \mathbb{R}^n$.
(c) Show that if the points $x_1, \dots, x_n$ are sampled i.i.d. from p and $\alpha_i = 1/n$ for all i, then $\mathbb{E}\Big[\big\| \int_{\mathcal{X}} \varphi(x) dp(x) - \sum_{i=1}^n \alpha_i \varphi(x_i) \big\|^2\Big] \leqslant \frac{1}{n} \mathbb{E}[k(x,x)]$.

Exercise 7.20 *Consider a binary classification problems with data $(x_1, y_1), \dots, (x_n, y_n)$ in $\mathcal{X} \times \{-1, 1\}$, with a positive kernel k defined on $\mathcal{X}$ with feature map $\varphi : \mathcal{X} \to \mathcal{H}$. Let μ_+ (μ_-) be the mean of all feature vectors for positive (negative) labels. We consider the classification rule that predicts 1 if $\|\varphi(x) - \mu_+\|_{\mathcal{H}}^2 < \|\varphi(x) - \mu_-\|_{\mathcal{H}}^2$ and -1 otherwise. Compute the classification rule only using kernel functions and compare it to local averaging methods from chapter 6.*

Exercise 7.21 (♦) *Find an upper bound of $\widetilde{A}(\mu, f_*)$ for the same assumption on f_*, but with the Gaussian kernel.*

[3] See https://en.wikipedia.org/wiki/K-means_clustering.

Exercise 7.22 *Consider the optimization problem* $\min_{\theta,\eta} \frac{1}{2n}\|y - \Phi\theta - \eta 1_n\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2$ *in the variables* $\theta \in \mathbb{R}^d$ *and* $\eta \in \mathbb{R}$*, where* $\Phi \in \mathbb{R}^{n\times d}$ *is the design matrix obtained from feature map* φ *and data points* $x_1, \dots, x_n$*,* $y \in \mathbb{R}^n$*, and* $1_n \in \mathbb{R}^n$ *is the vector of all* 1*s. Show that the optimal values of* θ *and* η *are* $\theta = \Phi^\top \alpha$ *and* $\eta = \frac{1}{n} 1_n^\top (y - \Phi\theta)$*, with* $\alpha = \Pi_n(\Pi_n K \Pi_n + n\lambda I)^{-1}\Pi_n y$*, and* $\Pi_n = I - \frac{1}{n} 1_n 1_n^\top$*. Show that the prediction function* $f(x) = \varphi(x)^\top\theta + \eta$ *takes the form* $\sum_{i=1}^n \hat{w}_i(x) y_i$ *with weights that sum to* 1*.*

Exercise 7.23 (♦) *For* $x_1, \dots, x_n$ *equally spaced in* $[0, 1]$ *and for a translation-invariant kernel from section* **??**, *compute the eigenvalues of the kernel matrix and the smoothing matrix.*

Chapter 8

Sparse Methods

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

Exercise 8.2 *Assume that* $\hat{\theta} \in \Theta$ *is such that* $\frac{1}{n}\|y - \Phi\hat{\theta}\|_2^2 \leqslant \inf_{\theta\in\Theta} \frac{1}{n}\|y - \Phi\theta\|_2^2 + \rho$. *Show that* $\|\Phi(\hat{\theta} - \theta_*)\|_2^2 \leqslant 4 \sup_{\theta\in\Theta} \left[\varepsilon^\top \left(\frac{\Phi(\theta-\theta_*)}{\|\Phi(\theta-\theta_*)\|_2}\right)\right]^2 + 2n\rho$ *(with notations from section* **??***).*

Solution. Let $\tilde{\theta} \in \operatorname{argmin}_{\theta\in\Theta} \|y - \Phi\theta\|_2^2$. If $\theta_* \in \Theta$, we have

$$\|y - \Phi\hat{\theta}\|_2^2 - n\rho \leq \|y - \Phi\tilde{\theta}\|_2^2 \leq \|y - \Phi\theta_*\|_2^2,$$

using the approximation error on $\hat{\theta}$.

We develop the expressions as in section (8.1.1) and obtain, before taking the square of the expression,

$$\|\Phi(\hat{\theta} - \theta_*)\|_2^2 - n\rho \leq 2\|\Phi(\hat{\theta} - \theta_*)\|_2 \sup_{\theta\in\Theta} \left[\varepsilon^\top \frac{\Phi(\theta - \theta_*)}{\|\Phi(\theta - \theta_*)\|_2}\right]^2.$$

We divide by $\|\Phi(\theta - \theta_*)\|_2$ and take the square of the expression. The left term is :

$$\begin{aligned}\left(\|\Phi(\hat{\theta} - \theta_*)\|_2 - \frac{n\rho}{\|\Phi(\theta - \theta_*)\|_2}\right)^2 &= \|\Phi(\hat{\theta} - \theta_*)\|_2^2 + \left(\frac{n\rho}{\|\Phi(\hat{\theta} - \theta_*)\|_2}\right)^2 - 2n\rho \\ &\geq \|\Phi(\hat{\theta} - \theta_*)\|_2^2 - 2n\rho.\end{aligned}$$

This concludes the proof, as one just needs to rearange the terms.

Exercise 8.3 (♦) *Consider a linear model $f(x) = \theta^\top \varphi(x)$ with a G-Lipschitz-continuous loss function and features almost surely bounded in ℓ_∞-norm by R. Using section* **??**, *show that the minimizer of the empirical risk over all $\theta \in \mathbb{R}^d$, such that $\|\theta\|_0 \leqslant k$ and $\|\theta\|_2 \leqslant D$, has an expected risk less than the minimum expected risk over this same set with an additive term proportional to $GRD\sqrt{k\log(d)/n}$.*

Exercise 8.4 (♦♦) *With a penalty proportional to $\|\theta\|_0 \log \frac{d}{\|\theta\|_0}$, show the same bound as for k known.*

Exercise 8.5 *Provide a closed-form expression for the iteration of the coordinate descent algorithm described just above.*

Exercise 8.6 *Assume that $\lambda \geqslant \left\|\frac{1}{n}\Phi^\top y\right\|_\infty$. Show that $\theta = 0$ is a minimizer of the Lasso objective function in equation (**??**).*

Solution. Using notations from the book, let $H(\theta) = \frac{1}{2n}\|y - \Phi\theta\|_2^2 + \lambda\|\theta\|_1$. This function is convex, therefore, one just has to show that all its directional derivatives in 0 are nonnegative to show that 0 is a minimizer.

Let $\varepsilon > 0$ and $\Delta \in \mathbb{R}^d$. We have

$$\frac{1}{\varepsilon}\left(H(0) - H(\varepsilon\Delta)\right) = \frac{1}{2n}\left(\varepsilon\|\Phi\Delta\|_2^2 + 2y^\top\Phi\Delta\right) + \lambda\|\Delta\|_1.$$

Therefore,

$$\begin{aligned}
\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(H(0) - H(\varepsilon\Delta)\right) &= \lambda\|\Delta\|_1 - \frac{1}{n}y^\top\Phi\Delta \\
&\geq (\lambda - \|\frac{1}{n}y^\top\Phi\|_\infty)\|\Delta\|_1, \quad \text{as } \frac{1}{n}y^\top\Phi\Delta \leq \|\frac{1}{n}y^\top\Phi\|_\infty\|\Delta\|_1 \\
\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(H(0) - H(\varepsilon\Delta)\right) &\geq 0, \quad \text{as } \lambda \geq \|\frac{1}{n}y^\top\Phi\|_\infty.
\end{aligned}$$

Exercise 8.7 *For $p \in [1, \infty]$, show that the dual of the ℓ_p-norm is the ℓ_q-norm for $\frac{1}{p} + \frac{1}{q} = 1$.*

Exercise 8.8 (♦) *With the same assumptions as proposition* **??**, *and with the choice of the regularization parameter $\lambda = 4\sigma\sqrt{\frac{\log(dn)}{n}}\sqrt{\|\widehat{\Sigma}\|_\infty}$, use lemma* **??** *to provide an upper bound of $\mathbb{E}\left[\frac{1}{n}\|\Phi(\hat{\theta} - \theta_*)\|_2^2\right]$.*

Exercise 8.9 (♦♦) *With the same assumptions as proposition* **??**, *with the choice of the regularization parameter $\lambda = 4\sigma\sqrt{\frac{\log(dn)}{n}}\sqrt{\|\widehat{\Sigma}\|_\infty}$, provide an upper bound on the expectation of the excess risk $\mathbb{E}\left[\frac{1}{n}\|\Phi(\hat{\theta} - \theta_*)\|_2^2\right]$.*

Exercise 8.10 (♦♦♦) *If sampling $\varphi(x)$ from a Gaussian with mean zero and covariance matrix identity, then with large probability, for n greater than a constant times $k^2 \frac{\log d}{n}$, the mutual incoherence property in equation (??) is satisfied.*

Exercise 8.11 *With the notations of section ??, show that if $\mu = 0$, from equation (??), we can recover the slow rate $\mathcal{R}(\hat{\theta}_\lambda) - \mathcal{R}(\theta^*) \leqslant \frac{4R\|\theta_*\|_1}{\sqrt{n}}(3\sigma + 2R\|\theta_*\|_1)\sqrt{2\log\frac{4d^2}{\delta}}$.*

Exercise 8.12 *Assuming that the design matrix Φ is orthogonal, compute the minimizer of $\frac{1}{2n}\|y - \Phi\theta\|_2^2 + \lambda\sum_{i=1}^m \|\theta_{A_i}\|_2$.*

Exercise 8.13 *Consider the d (overlapping) sets $A_i = \{1, \dots, i\}$ and the norm $\sum_{i=1}^d \|\theta_{A_i}\|_2$. Show that penalization with this norm will tend to select patterns of nonzeros of the form $\{i+1, \dots, d\}$.*

Exercise 8.14 *Compute the minimizer of $\frac{1}{2n}\|Y - \Theta\|_{\mathrm{F}}^2 + \lambda\|\Theta\|_*$, where $\|M\|_{\mathrm{F}}$ is the Frobenius norm and $\|M\|_*$ is the nuclear norm.*

Exercise 8.15 *Show that $\|M\|_*$ is the minimum of $\frac{1}{2}\|U\|_{\mathrm{F}}^2 + \frac{1}{2}\|V\|_{\mathrm{F}}^2$ over all decompositions of $M = UV^\top$.*

Exercise 8.16 (♦) *Consider m feature vectors $\varphi_j : \mathcal{X} \to \mathcal{H}_j$, associated with kernels $k_j : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ for $j \in \{1, \dots, m\}$. Show that*

$$\inf_{\theta_1,\dots,\theta_m} \frac{1}{n}\sum_{i=1}^n \ell\big(y_i, \langle\theta_1, \varphi_1(x_i)\rangle + \cdots + \langle\theta_m, \varphi_m(x_i)\rangle\big) + \frac{\lambda}{2}\big(\|\theta_1\| + \cdots + \|\theta_m\|\big)^2$$

is equivalent to $\inf_{\eta\in\Delta_m} \inf_{\alpha\in\mathbb{R}^n} \frac{1}{n}\sum_{i=1}^n \ell(y_i, (K(\eta)\alpha)_i) + \frac{\lambda}{2}\alpha^\top K(\eta)\alpha$, *where $K(\eta) \in \mathbb{R}^{n\times n}$ is the kernel matrix associated with the kernel $\eta_1 k_1 + \cdots + \eta_m k_m$ and Δ_m is the simplex in dimension m.*

Exercise 8.17 *Show that for $\alpha \in (0,1)$, $\frac{1}{\alpha}u^\alpha = \inf_{\eta>0} \frac{u}{\eta} + \big(\frac{1}{\alpha} - 1\big)\eta^{\alpha/(1-\alpha)}$, and derive both a reweighted ℓ_1-minimization and a reweighted ℓ_2-minimization algorithm for the penalty $\sum_{i=1}^d |\theta_i|^\alpha$.*

Chapter 9

Neural Networks

Exercise 9.1 (♦) *Provide a bound similar to proposition* **??** *for the alternative constraint* $\|w_j\|_1 + |b_j|/R = 1$, *where* R *denotes the supremum of* $\|x\|_\infty$ *over all* x *in the support of its distribution.*

Solution. Using the same computations as in the book, we obtain

$$R_n(\mathcal{G}) \leq 2GD\mathbb{E}\left[\sup_{\|w\|_1+|c|=1}\left|w^\top\left(\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i\right) + c\left(\frac{R}{n}\sum_{i=1}^n \varepsilon_i\right)\right|\right],$$

after using η's bounds, the G-Lipschitz property of the loss function, and the result on Rademacher complexities defined by a absolute value.

Let's upper-bound the expression we have to maximize :

$$\begin{aligned}|w^\top z + ct| &\leq |w^\top z| + |c||t| \\ &\leq \|z\|_\infty\|w\|_1 + |c||t|, \text{ using Hölder's inequality,} \\ &\leq \|z\|_\infty + |c|(|t| - \|z\|_\infty), \text{ as } \|w\|_1 + |c| = 1.\end{aligned}$$

Therefore,

$$\sup_{\|w\|_1+|c|=1}\left|w^\top z + ct\right| = \max\left(\|z\|_\infty, |t|\right).$$

Let's compute each :

$$\mathbb{E}(\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_i\|_\infty) = \mathbb{E}(\max_{1\leq j\leq d}|\frac{1}{n}\sum_{i=1}^n \varepsilon_i x_{ij}|) = \sqrt{\frac{2R^2\log 2d}{n}}$$

using the results from Chap. 1 on the expectation of maximum; see Exercise 4.12 for a more detailed explanation ;

$$\frac{R}{n}\mathbb{E}(|\sum_{i=1}^n \varepsilon_i|) \leq \frac{R}{\sqrt{n}},$$

using Jensen's inequality.

This leads to an upper bound of the form :

$$R_n(\mathcal{G}) \leq 2GDR\frac{\sqrt{2\log 2d}}{\sqrt{n}} \leq \frac{4GDR\sqrt{\log 2d}}{\sqrt{n}}.$$

This leads to a bound whose expression close to the book's one, but with a dependance in the log of the number of parameters.

Exercise 9.2 (♦) *We consider a 1-Lipschitz-continuous activation function σ such that $\sigma(0) = 0$, and the classes of functions defined recursively as $\mathcal{F}_0 = \{x \mapsto \theta^\top x, \|\theta\|_2 \leqslant D_0\}$, and, for $i = 1, \dots, M$, $\mathcal{F}_i = \{x \mapsto \sum_{j=1}^{m_i} \theta_j \sigma(f_j(x)), f_j \in \mathcal{F}_{i-1}, \|\theta\|_1 \leqslant D_i\}$, corresponding to a neural network with M layers. Assuming that $\|x\|_2 \leqslant R$ almost surely, show by recursion that the Rademacher complexity satisfies $\mathrm{R}_n(\mathcal{F}_M) \leqslant 2^M \frac{R}{\sqrt{n}} \prod_{i=0}^M D_i$.*

Exercise 9.3 (♦♦) *Assume $-R = x_1 < \dots < x_n = R$, $y_1, \dots, y_n \in \mathbb{R}$, show that the piecewise-affine interpolant on $[-R, R]$ is a minimum norm interpolant.*

Exercise 9.4 (Step activation function (♦)) *Consider the step activation function defined as $\sigma(u) = 1_{u>0}$. Show that the corresponding variation norm can be upper-bounded by a constant times $\int_{\mathbb{R}^d} |\hat{f}(\omega)|(1 + R\|\omega\|_2) d\omega$.*

Exercise 9.5 *Show that if we replace equation (??) with $f_t = \frac{t-1}{t} f_{t-1} + \frac{1}{t}\bar{f}_t$, f_t is the uniform convex combination of $\bar{f}_1, \dots, \bar{f}_t$, and we have the convergence rate $J(f_t) - \inf_{f \in \mathcal{K}} J(f) \leqslant \frac{L}{t}(1 + \log t)\mathrm{diam}_{\mathcal{H}}(\mathcal{K})^2$.*

Exercise 9.6 (Frank-Wolfe with line search) *The update in equation (??) is often replaced by $f_t = (1 - \rho_t) f_{t-1} + \rho_t \bar{f}_t$ with $\rho_t = \arg\min_{\rho \in [0,1]} \rho \langle J'(f_{t-1}), \bar{f}_t - f_{t-1} \rangle_{\mathcal{H}} + \frac{L}{2}\rho^2 \|\bar{f}_t - f_{t-1}\|_{\mathcal{H}}^2$. Show that we have $J(f_t) - \inf_{f \in \mathcal{K}} J(f) \leqslant \frac{4L}{t+1}\mathrm{diam}_{\mathcal{H}}(\mathcal{K})^2$.*

Exercise 9.7 *Extend the bound in equation (??) to all activation functions.*

Exercise 9.8 *Consider target functions of the form $f_*(x) = \sum_{j=1}^k f_j(w_j^\top x)$ for one-dimensional Lipschitz-continuous functions $f_1, \dots, f_k$. Provide an upper bound on excess risk proportional to $k/n^{1/6}$.*

Exercise 9.9 (Link with kernel learning (♦)) *With the setup presented in this section, show that the infimum of $\int_K \left|\frac{d\nu(w,b)}{d\tau(w,b)}\right|^2 d\tau(w,b)$ over probability distributions τ on K is equal to $\left(\int_K |d\nu(w,b)|\right)^2$. Using exercise 8.16, show how the penalty γ_1 can be interpreted as kernel learning.*

Exercise 9.10 (Step activation function) *Consider, instead of equation (??), the kernel $k(x, x') = \int_K 1_{w^\top x + b \geqslant 0} 1_{w^\top x' + b \geqslant 0} d\tau(w, b)$. Show that it can be expressed in closed form as $k(x, x') = \frac{1}{2} - \frac{1}{4R}\frac{\Gamma(1)\Gamma(\frac{d}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{d}{2}+\frac{1}{2})}\|x - y\|_2$.*