Learning Theory from First Principles
Exercises correction

December 5, 2024

Francis BACH (book), Camille DUBOIS (correction)

This pdf intends to give a correction to some of the exercises of Francis Bach's book : Learning Theory from First Principles, which *presents old and recent results in learning theory for the most widely used learning* and proves *many results starting from first principles*.

Although most of the proposals in this pdf are used in the official correction, alternative formulations as well as solutions to exercises I have not done can be found on the official website. Please refer to the official correction for a proof-read by Pr Francis Bach version.

Chapter 2

1. E2.1

Ans: We take l, c_+ and c_- as defined above. Given $x \in X$, we compute

$$\operatorname*{argmin}_{z \in \{-1,1\}} \mathbb{E}[l(y,z) \mid x = x'].$$

We have

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

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

$$\begin{aligned}\mathbb{E}[l(y,z') \mid x = x'] &= \mathbb{E}[l(y,-1)|x = x'] \\ &= c_- \mathbb{P}(y=-1|x=x') \\ &= c_- \frac{1 - \mathbb{E}[y|x=x']}{2}\end{aligned}$$

And with $z' = -1$, it yields :

$$\begin{aligned}\mathbb{E}[l(y,z') \mid x = x'] &= \mathbb{E}[l(y,1)|x = x'] \\ &= c_+ \mathbb{P}(y=1|x=x') \\ &= c_+ \frac{1 + \mathbb{E}[y|x=x']}{2}\end{aligned}.$$

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

$$f(x') = 2 \, \mathbb{1}_{\frac{c_-}{c_+} \leq \frac{1+\mathbb{E}[y|x=x']}{1-\mathbb{E}[y|x=x']}} - 1.$$

2. E2.2

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

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

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

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.

3. E2.4 (*)

Ans: Having the same notations as in the problem statement. Let $\epsilon > 0$.

Let $x' \in \mathbb{R}$. Let $z \in \mathbb{R}$.

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

Derivating the expresion w.r.t z yields :

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

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 ($b - a \leq 2\epsilon$). As we cannot have both $\mathbb{P}(y - z \leq -\epsilon|x = x') > 0$ and $\mathbb{P}(y - z \geq \epsilon|x = x') > 0$, we need

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

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

Chapter 3

1. E3.2

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

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

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

2. E3.6

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

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

By summing the preceding terms, we have :

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

Chapter 4

1. E4-9

Ans:

(a) 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 with the same logic. We clearly have $\mathcal{Y}_{\mathcal{H}} \leq \mathcal{Y}_{\mathcal{H}'}$ (for all $\varepsilon, \mathcal{D}$), which yields the expected result.

(b) We define $\mathcal{H}, \mathcal{H}'$ and want to compute $\mathcal{R}_n(\mathcal{H} + \mathcal{H}')$. As we have $\mathcal{H} + \mathcal{H}' = \{\tilde{h} | \tilde{h} = h + h', h \in \mathcal{H}, h' \in \mathcal{H}'\}$. Therefore, by linearity of the expectation, and additivity of the evaluated expression w.r.t h (meaning $\sup_{h \in \mathcal{H}, h' \in \mathcal{H}'} \dots = \sup_{h \in \mathcal{H}} \dots + \sup_{h' \in \mathcal{H}'} \dots$ here). This gives $\mathcal{R}_n(\mathcal{H} + \mathcal{H}') = \mathcal{R}_n(\mathcal{H}) + \mathcal{R}_n(\mathcal{H}')$.

(c) Let $\alpha \in \mathbb{R}$. If $\alpha \geq 0$, the result is obvious. If $\alpha \leq 0$, 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 $\mathbf{E}_{\varepsilon', \mathcal{D}}(\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \varepsilon'_i (-\alpha) h(z_i))$. As $-\alpha = |\alpha|$ is positive, we clearly have $\mathcal{R}_n(\alpha \mathcal{H}) = |\alpha| \mathcal{R}_n(\mathcal{H})$. This concludes the proof.

(d) Using result b), we just have to show that for h_0 as defined in the book, $\mathcal{R}_n(h_0) = 0$. The expression simplifies as

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

using that we evaluate the sup on a singleton.

(e) We clearly have $\mathcal{R}_n(\mathcal{H}) \leq \mathcal{R}_n(\text{convex hull } \mathcal{H})$ by using $\mathcal{H} \subset$ convex hull $\mathcal{H}$ and a). We therefore want to show $\mathcal{R}_n(\mathcal{H}) \geq \mathcal{R}_n(\text{convex hull } \mathcal{H})$. Let $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ be a draw of Rademacher variables ; let $\tilde{h} \in$ convex hull $\mathcal{H}$. There exists $(\alpha_i)_{i \in [|1,m|]} \in \mathbb{R}^m$, which sum to 1, and $(h_i)_{i \in [|1,m|]} \in \mathcal{H}^m$ s.t. $\tilde{h} = \sum_{k=1}^m \alpha_k h_k$. We have :

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

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

2. Q4-12

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

$$\mathbb{E}(||\Phi^\top \varepsilon||_\infty) = \mathbb{E}(\max_{1 \leq i \leq d} |\sum_{j=1}^n \varphi_j(x_i)\varepsilon_i|).$$

As $||\varphi(x)|| \leq R$ almost surely, the same applies tho its components (i.e for $1 \leq j \leq n,\ |\varphi_j(x)| \leq R$). 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 $\sigma_S^2 = nR^2$.

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

Combining it to the first result, we obtain :

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

Chapter 5

1. E5-1

Ans: Given the same notations as in the book, we want to show that :

$$\exists \alpha \in \mathbb{R},\ |F(\theta_t) - F(\eta_*)| \leq \alpha \left(1 - \frac{\mu_+}{L}\right)^{2t} \|\theta_0 - \eta_*\|_2.$$

According to already proven results, we need to show that

$$|\lambda(1 - \frac{\lambda}{L})^{2t}| \leq \alpha' \left(1 - \frac{\mu_+}{L}\right)^{2t}$$

for λ any eigenvalue of $H = \frac{1}{n}\Phi^\top \Phi$.

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

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

where we use between terms 1 and 2 that $\lambda = 0$ (if it exists) can not be a maximizer, and between terms 2 and 3 that for a, b positive, $\max(ab) \leq \max(a)\max(b)$.

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

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

2. E5-9

Ans:

(a) Let's show that an $\ell 2$-regularized logistic regression is strongly convex and smooth.

Gradient and Hessian computations : We define the loss function as

$$\mathcal{L}(y, \theta) = -\frac{1}{n} \sum_{i=1}^{n} \left(y_i \log(\sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i))\right) + \frac{\lambda}{2} \|\theta\|_2^2,$$

where σ is the sigmoid function.

To facilitate the computation, we consider, for $i \in [|1, n|]$,

$$-l_i(\theta) = y_i \log(\sigma(\theta^\top x_i) + (1 - y_i) \log(1 - \sigma(\theta^\top x_i)).$$

The function l_i is twice differentiable. Letting $z_i = \theta^\top x_i$, and using $\sigma'(z) = \sigma(z)(1 - \sigma(z))$, we have

$$\frac{\partial l_i}{\partial z}(\theta) = -y_i(1 - \sigma(z_i)) + (1 - y_i)\sigma(z_i) = \sigma(z_i) - y_i.$$

By the chain rule, we obtain

$$\nabla l_i(\theta) = (\sigma(z_i) - y_i)\frac{\partial z_i}{\partial \theta} = (\sigma(\theta^\top x_i) - y_i)x_i.$$

The Hessian of this function is $\nabla^2 l_i(\theta) = \sigma(z_i)(1 - \sigma(z_i)x_i x_i^\top \geq 0$.

Strong convexity: This expression ensures that l_i is convex, therefore, $\sum_{i=1}^n l_i$ is convex and adding the $\ell 2-$regularization term, we obtain that $\mathcal{L}$ is λ-strongly-convex.

L-smoothness: As $u(1 - u) \leq \frac{1}{4}$ for $u \in \mathbb{R}$, we have, for $\theta, \theta' \in \mathbb{R}$:

$$\begin{aligned}
\|\mathcal{L}'(\theta) - \mathcal{L}'(\theta')\|_2 &= \|\frac{1}{n}\sum_{i=1}^n ((\sigma(z_i) - \sigma(z_i'))x) + \lambda(\theta - \theta')\|_2 \\
&\leq \frac{1}{n}\sum_{i=1}^n |(\sigma(z_i) - \sigma(z_i'))|\|x\|_2 + \lambda\|\theta - \theta'\|_2, \\
&\text{by triangle inequality} \\
&\leq \sum_{i=1}^n \frac{\|x_i\|_2}{4n}(\theta - \theta')^\top x_i + \lambda\|\theta - \theta'\|_2, \\
&\text{as sigmoid is } 1/4\text{-Lipschitz} \\
&\leq \left(\sum_{i=1}^n \frac{\|x_i\|_2^2}{4n} + \lambda\right)\|\theta - \theta'\|_2.
\end{aligned}$$

$\mathcal{L}$ is therefore $\left(\sum_{i=1}^n \frac{\|x_i\|_2^2}{4n} + \lambda\right)$-smooth.

(b) We now tackle the ridge-regression case, for which we will shorten the intermediate computations.

Gradient and Hessian computations: Using the same notation as in previous chapters of the book, we define

$$\begin{aligned}\mathcal{L}(y,\theta) &= \frac{1}{2}\|y-\Phi\theta\|_2^2 + \frac{\lambda}{2}\|\theta\|_2^2 \\ \nabla_\theta \mathcal{L}(y,\theta) &= \frac{1}{n}(\phi^\top\phi\theta - \phi^\top y) + \lambda\theta \\ \nabla_\theta^2 \mathcal{L}(y,\theta) &= H + \lambda I\end{aligned}$$

Strong convexity: The Hessian form os the objective function shows that, denoting $\lambda_{\min}(H)$ the minimum eigenvalue of H, the objective function is $(\lambda_{\min}(H)+\lambda)$-strongly-convex.

L-smoothness: We have

$$\begin{aligned}\|\mathcal{L}'(\theta) - \mathcal{L}'(\theta')\|_2 &= \|(H+\lambda I)(\theta-\theta')\|_2 \\ &\leq (\lambda_{\max}(H)+\lambda)\|(\theta-\theta')\|_2\end{aligned}$$

And therefore, $\mathcal{L}$ is $(\lambda_{\max}(H)+\lambda)$-smooth.

Chapter 6

1. E6-1

Ans: 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).

(a) **KNN case**
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 $(x_j)_{1 \leq j \leq n}$ to x.

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

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

(b) **Partition case**
Unlike KNN, 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 symmetrical.

Moreover, by rearranging the points' indices s.t, if φ is a permutation of $[|1, n|]$, 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. . . , we obtain a block matrix.

With the normalization rule we defined on the weights, the nonzeros of the smoothing matrix are not all equal but both rows and columns sum to 1.

2. E6-3

Ans: 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.

Chapter 7

1. E7-1 **

Ans: The Riesz representation theorem states that having :

- a Hilbert space $\mathcal{H}$ with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$,
- ψ a continuous linear form on $\mathcal{H}$,

there exists a unique $u \in \mathcal{H}$ s.t. for all $h \in \mathcal{H}$, $\psi(h) = \langle u, h \rangle_{\mathcal{H}}$.

Assuming we were able to apply this theorem to our problem - meaning we need to show the continuity of $f \mapsto f(x)$ - we would obtain the following statement, for a fixed $x \in \mathcal{X}$:

$$\exists! u_x \in \mathcal{H},\ \forall f \in \mathcal{H},\ f(x) = \langle u_x, f \rangle_{\mathcal{H}}$$

Notice that the application $x \mapsto u_x$ is well defined. The theorem's results would directly yield the reproducing properties (noticing that $u_y(x)$ would be identified to $k(x, y)$, for $x, y \in \mathcal{H}$):

- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H},\ f(x) = \langle u_x, f \rangle_{\mathcal{H}}$
- $\forall x \in \mathcal{X}, \forall y \in \mathcal{X},\ u_y(x) = \langle u_x, u_y \rangle_{\mathcal{H}}$, as for any $y \in \mathcal{X}$, u_y belongs to $\mathcal{H}$.

Therefore, we just have to show - or know - that as the linear form $\Psi_x : f \mapsto f(x)$ is bounded, it is continuous. Let's give a quick proof.

We fix $x \in \mathcal{X}$. As the linear form Ψ_x is bounded, there exists M that upper-bounds the set $\{f \in \mathcal{H},\ ||f||_{\mathcal{H}} \leq 1\}$.

Now, for any $f \in \mathcal{H}$, we have $f = ||f||_{\mathcal{H}} \frac{f}{||f||_{\mathcal{H}}}$ which means, by linearity of the norm and of Ψ_x, that we have $|\Psi_x(f)| \leq ||f||_{\mathcal{H}} M$ (applying the upper-bound to $\frac{f}{||f||_{\mathcal{H}}}$). This result stands for any couple $(f, f') \in \mathcal{H}^2$, which shows that Ψ_x is M-Lipschitz, and therefore continuous.

2. E7-2

Ans: We note K the kernel matrix. As k is a positive-definite kernel, K is positive-definite. We need to show that e^K is also positive definite. This comes

from the following definition of the exponential of matrix :

$$e^K = \sum_{i=0}^{+\infty} \frac{K^i}{i!}.$$

We can diagonalize K and e^K in the same basis (one can be convinced by injecting the diagonalized matrix PKP^{-1} in the sum), and the eigenvalues of e^K will be the exponential of the eigenvalues of K. As K is positive-definite, its eigenvalues are real. Therefore, those of e^K are positive and e^K is also positive-definite.

This shows that $(x, x') \mapsto e^{k(x,x'}$ is a positive-definite kernel.

3. E7-11

Ans:

(a) Using the same notations as, e.g. those of chapter 3, the primal problem of Ridge regression can be expressed as

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

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

The associated Lagrangian is therefore

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

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

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

We compute

$$\frac{\partial \mathcal{L}}{\partial \theta} = \lambda\theta - \Phi^\top \alpha = 0 \iff \theta = \frac{1}{\lambda}\Phi^\top \alpha,$$
$$\frac{\partial \mathcal{L}}{\partial r} = r - \alpha \iff r = \alpha.$$

Which yields $g(\alpha) = -\frac{\|\alpha\|^2}{2} - \frac{\|\Phi^\top \alpha\|^2}{2\lambda} + \alpha^\top y$.

Finally, computing $g'(\alpha) = 0$ gives

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

and therefore

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

Not so sure of this one... Regarding condition numbers, we are interested in comparing the eigenvalues of $\Phi\Phi^\top \in \mathbb{R}^{n\times n}$ and $\Phi^\top\Phi \in \mathbb{R}^{d\times d}$. We assume that we have $n < d$.

Both matrix share the same non-zero eigenvalues (with the same algebraic multiplicity, slightly more complicated to show, I had to look online).

Therefore, it is more likely that $\Phi^\top\Phi \in \mathbb{R}^{d\times d}$ has a null eigenvalue, meaning that we would obtain the following condition numbers (ratio of the extrema of the Hessian's eigenvalues set) :

$$\text{(primal) } \frac{L+\lambda}{\lambda} \geq \frac{L+\lambda}{\mu+\lambda} \text{ (dual).}$$

(b) Using the transpose of the matrix inversion lemma, meaning $\Phi^\top(\Phi\Phi^\top + \lambda I)^{-1} = (\Phi^\top\Phi + \lambda I)^{-1}\Phi^\top$, in the expression we found for $\hat{\theta}_{dual} = \Phi^\top(\lambda I + \Phi\Phi^\top)^{-1} y$, we directly obtain the same minimizer as in the equations of chapter 3.

However, on top of the above-mentioned benefit for the condition number, optimizing over α ($\in \mathbb{R}^n$) instead of θ ($\in \mathbb{R}^d$) can speed up computations !

Chapter 8

1. E8-2

Ans: 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.

2. E8-6

Ans: 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 $\epsilon > 0$ 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,

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

Chapter 9

1. E9-1

Ans: 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.