Math of ML : Exercises 5 *

October 13, 2025

Exercise 1 (Kernel ridge regression)**.** *Let $x_1, \dots, x_n \in \mathcal{X}$ be the observed design vectors and let $y_1, \dots, y_n \in \mathbb{R}$ be the observed response variables. Let $\phi : \mathcal{X} \to \mathbb{R}^p$ be a feature mapping, where p is possibly equal to infinity, and let $k(x, y) = \langle \phi(x), \phi(y) \rangle$ be the associated kernel. In this exercise, we are interested in obtaining the predictor $f_{\widehat{\theta}}(x) = \langle \widehat{\theta}, \phi(x) \rangle$, where $\widehat{\theta}$ is the solution to the optimization problem*

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (\theta^\top \phi(x_i) - y_i))^2 + \lambda \|\theta\|_2^2, \tag{1}$$

where $\lambda > 0$ is a regularization parameter. We will consider two separate approaches for solving the above optimization problem.

1. *(Explicit feature vector manipulation approach)*
   (a) *Write an explicit formula for the unique $\widehat{\theta}$ that minimizes the objective (1). What is the computational complexity of computing $\widehat{\theta}$?*
   (b) *Given a new point $x \in \mathcal{X}$, what is the computational complexity of computing $f_{\widehat{\theta}}(x)$?*
2. *(Kernel approach)*
   (a) *Suggest a way to solve the problem (1) without ever explicitly manipulating the feature vectors $\phi(x_i)$ (think of how to represent the output function $f_{\widehat{\theta}}$). Assuming that a computation of $k(x, y)$ takes $O(\kappa)$ number of arithmetic operations, what is the computational complexity of computing $f_{\widehat{\theta}}$ using your method? How does the computational complexity compare with the "explicit feature vector manipulation method" as $p \to \infty$?*
   Hint: Let $\Phi \in \mathbb{R}^{n \times p}$ be an arbitrary matrix. Check that for any $\lambda > 0$ it holds that
   $$\left(\Phi^\top \Phi + \lambda I_p\right)^{-1} \Phi^\top = \Phi^\top \left(\Phi \Phi^\top + \lambda I_n\right)^{-1}.$$
   (b) *Given a new point $x \in \mathcal{X}$, what is the computational complexity of computing $f_{\widehat{\theta}}(x)$?*

Solution 1.

1. (a) *Let $\Phi \in \mathbb{R}^{n \times d}$ be a matrix whose i-th row is equal to $\phi(x_i)$. We have*
   $$\widehat{\theta} = \left(\Phi^\top \Phi + n\lambda I_p\right)^{-1} \Phi^\top y. \tag{2}$$

*Lénaïc Chizat EPFL lenaic.chizat@epfl.ch

Inverting the $p \times p$ matrix $\left(\Phi^\top \Phi + n\lambda I_p\right)^{-1}$ can be done in $O(p^3)$ operations. Computing $\Phi^\top \Phi$ takes $O(p^2 n)$ operations Computing $\Phi^\top y$ takes $O(np)$. Finally, multiplying the matrix $\left(\Phi^\top \Phi + n\lambda I_p\right)^{-1}$ with a vector $\Phi^\top y$ takes $O(p^2)$ operations. Hence, the total computational cost scales as $O(p^3 + p^2 n)$.

(b) Computing $f_{\widehat{\theta}}(x)$ amounts to computing the inner product $\langle \widehat{\theta}, \phi(x) \rangle$ which takes $O(p)$ operations.

2. *(a) To check the hint, note that*

$$\Phi^\top \left(\Phi\Phi^\top + \lambda I_n\right) = \left(\Phi^\top \Phi + \lambda I_p\right) \Phi^\top,$$

then left-multiply both sides by $\left(\Phi^\top \Phi + \lambda I_p\right)^{-1}$, and then right-multiply both sides by $\left(\Phi\Phi^\top + \lambda I_n\right)^{-1}$.
Let $K \in \mathbb{R}^{n \times n}$ be the kernel matrix given by $K_{ij} = k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. Observe that $K = \Phi\Phi^\top$. Using the hint, the exact expression for the $\widehat{\theta}$ vector (2) may be written as

$$\widehat{\theta} = \left(\Phi^\top \Phi + n\lambda I_p\right)^{-1} \Phi^\top y = \Phi^\top \left(\Phi\Phi^\top + n\lambda I_n\right)^{-1} y = \Phi^\top \left(K + n\lambda I_n\right)^{-1} y.$$

Let

$$\alpha = \left(K + n\lambda I_n\right)^{-1} y. \tag{3}$$

Thus, we have

$$\widehat{\theta} = \Phi^\top \alpha = \sum_{i=1}^{n} \alpha_i \phi(x_i)$$

Hence,

$$f_{\widehat{\theta}}(\cdot) = \sum_{i=1}^{n} \alpha_i k(x_i, \cdot). \tag{4}$$

Computing the kernel matrix K takes $O(n^2 \kappa)$ operations. Inverting the matrix $\left(K + n\lambda I_n\right)^{-1}$ takes $O(n^3)$ operations. Then, computing the vector α (3) takes $O(n^2)$ operations. Hence, the total computational complexity for computing the function $f_{\widehat{\theta}}$ represented by (4) is $O(n^2(n + \kappa))$.

(b) By (4), we have for any $x \in \mathcal{X}$

$$f_{\widehat{\theta}}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x).$$

The computation of the above expression can be done in $O(n\kappa)$ number of operations.

Exercise 2 (Covering number bounds for the supremum of a stochastic process)**.** *Let $(X_t)_{t\in T}$ be a collection of zero-mean σ^2-sub-Gaussian random variables, indexed by some set T equipped with a metric d. Suppose we wish to upper-bound $\sup_{t\in T} X_t$.*

Remark: *In week 3 of the course, we showed how to upper bound the excess risk of the Empirical Risk Minimization estimator over a hypothesis class $\mathcal{F}$, by controlling the supremum of an empirical process: $\sup_{f\in\mathcal{F}} \hat{\mathcal{R}}(f) - \mathcal{R}(f) = \sup_{f\in\mathcal{F}} \left\{ \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i) - \mathbf{E}_Z \ell(f, Z) \right\}$. This setting corresponds to taking $T = \mathcal{F}$ and $\forall f \in \mathcal{F}, X_f = \frac{1}{n}\sum_{i=1}^n \ell(f, Z_i) - \mathbf{E}_Z \ell(f, Z)$.*

We already know how to handle the case where the index set T is finite (recall exercise 1.5 in exercise sheet 3). We have also seen how to deal with some well-behaved infinite index sets T and stochastic processes $(X_t)_{t\in T}$ via the use of Rademacher complexities (e.g., when bounding the Rademacher complexity of norm-constrained linear predictors, as in section 4.2 of lecture 3).

The purpose of this exercise is to introduce a general technique for bounding $\mathbf{E}[\sup_{t\in T} X_t]$ for infinite index sets T. The key idea is to approximate the infinite index set with a finite set, use exercise 1.5 of exercise sheet 3 to bound the maximum of the finite set, and pay an additional approximation error term. Observe that there is a trade-off between the size of the approximating finite set and the incurred approximation error. Before proceeding with the exercise, we need one additional definition.

Definition 1 (ε-net and ε-covering number)**.** *A set $\{t_1, \ldots, t_N\} \subseteq T$ is said to be an ε-net for the metric space (T, d) if for any $t \in T$ there exists $j \in \{1, \ldots, N\}$ such that $d(t, t_j) \leq \varepsilon$. The cardinality of the smallest ε-net, denoted $N(\varepsilon, T, d)$, is an ε-covering number of (T, d).*

1. *Suppose that there exists a random variable L such that for any $t, s \in T$ we have $|X_t - X_s| \leq L d(t, s)$. Prove that*

$$\mathbf{E}\left[\sup_{t\in T} X_t\right] \leq \inf_{\varepsilon>0} \left\{ \mathbf{E}[L]\varepsilon + \sqrt{2\sigma^2 \log N(\varepsilon, T, d)} \right\}. \tag{5}$$

2. *We will now consider a simple application of the bound (5). Let $M \in \mathbb{R}^{n\times m}$ be a rectangular matrix such that each entry M_{ij} is an independent zero-mean τ^2-sub-Gaussian random variable. The operator norm of M is defined by $\|M\| = \sup_{x\in B_n, y\in B_m} \langle x, My\rangle$, where $B_d = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$ is the unit ball in the d-dimensional Euclidean space. Prove that $\mathbf{E}\|M\| \leq c\tau\sqrt{n+m}$, where $c > 0$ is a universal constant.*

   Hint: *Let $x, y \in B_d$ and consider the metric induced by the Euclidean norm $d_2(x, y) = \|x - y\|_2$. Then, for any $\varepsilon \in (0, 1)$ we have*

$$\left(\frac{1}{\varepsilon}\right)^d \leq N(\varepsilon, B_d, d_2) \leq \left(1 + \frac{2}{\varepsilon}\right)^d. \tag{6}$$

   The above inequality is classical; see, e.g., [Wainwright, 2019, Example 5.8] for a proof.

Solution 2.

1. *For any $\varepsilon > 0$, let $T_\varepsilon = \{t_1, \ldots, t_{N_\varepsilon}\}$ be the minimum cardinality ε-net for (T, d). Then, for any $t \in T$ there exists some $f_\varepsilon(t) \in T_\varepsilon$ such that $d(t, f_\varepsilon(t)) \leq \varepsilon$. It follows that*

$$\begin{aligned}
\mathbf{E}\left[\sup_{t\in T} X_t\right] &= \mathbf{E}\left[\sup_{t\in T} X_{f_\varepsilon(t)} + (X_t - X_{f_\varepsilon(t)})\right] \\
&\leq \mathbf{E}\left[\sup_{t\in T} X_{f_\varepsilon(t)} + \sup_{t\in T}(X_t - X_{f_\varepsilon(t)})\right]
\end{aligned}$$

$$\leq \varepsilon \mathbf{E}[L] + \mathbf{E}\left[\sup_{t\in T} X_{f_\varepsilon(t)}\right]$$
$$\leq \varepsilon \mathbf{E}[L] + \mathbf{E}\left[\sup_{t\in T_\varepsilon} X_t\right]$$
$$\leq \varepsilon \mathbf{E}[L] + \sqrt{2\sigma^2 \log |T_\varepsilon|},$$

where the last line follows via exercise 1.5 in exercise sheet 3. Observing that $|T_\varepsilon| = N(\varepsilon, T, d)$ *and minimizing the above bound over* $\varepsilon > 0$ *yields the desired result.*

2. *Let* $T = B_n \times B_m$ *and for* $(x, y), (x', y') \in T$ *define* $X_{(x,y)} = \langle x, My\rangle$ *and* $d((x, y), (x', y')) = \|x - x'\|_2 + \|y - y'\|_2$. *Observe that by exercise 1.4 in exercise sheet 3, for any* $(x, y) \in T$ *the random variable* $X_{(x,y)}$ *is* τ^2*-sub-Gaussian.*

   We will now attempt to find the random variable L *such that* $X_{(x,y)} - X_{(x',y')} \leq Ld((x, y), (x', y'))$. *Observe that for any* $(x, y), (x', y') \in T$ *we have*

$$\begin{aligned}
|X_{(x,y)} - X_{(x',y')}| &= \left|\langle x, My\rangle - \langle x', My'\rangle\right| \\
&= \left|\langle x - x', My\rangle - \langle x', M(y' - y)\rangle\right| \\
&\leq \left|\langle x - x', My\rangle\right| + \left|\langle x', M(y' - y)\rangle\right| \\
&\leq \|x - x'\|_2 \|M\| \|y\|_2 + \|x'\|_2 \|M\| \|y' - y\|_2 \\
&\leq \|x - x'\|_2 \|M\| + \|M\| \|y' - y\|_2 \\
&= \|M\| d((x, y), (x', y')).
\end{aligned}$$

   Thus, we have $L = \|M\|$. *Applying the first part of this exercise, it follows that for any* $\varepsilon > 0$ *we have*

$$\mathbf{E}\|M\| \leq \varepsilon \mathbf{E}\|M\| + \sqrt{2\tau^2 \log N(\varepsilon, T, d)}.$$

   In particular, we have

$$\mathbf{E}\|M\| \leq \inf_{\varepsilon>0} \frac{\sqrt{2\tau^2 \log N(\varepsilon, T, d)}}{1 - \varepsilon}. \tag{7}$$

   It remains to bound the ε*-covering numbers* $N(\varepsilon, T, d)$. *Observe that if* $T^n_{\varepsilon/2}$ *is a* $\varepsilon/2$*-net for* B_n *and* $T^m_{\varepsilon/2}$ *is a* $\varepsilon/2$*-net for* B_m, *then* $T^n_{\varepsilon/2} \times T^n_{\varepsilon/2}$ *is a* ε*-net for* (T, d). *Using the bound* (6), *it follows that*

$$N(\varepsilon, T, d) \leq N(\varepsilon/2, B_n, d_2) N(\varepsilon/2, B_m, d_2) \leq \left(1 + \frac{4}{\varepsilon}\right)^{n+m}.$$

   Plugging the above bound into (7) *yields*

$$\mathbf{E}\|M\| \leq \sqrt{2}\tau \inf_{\varepsilon>0} \frac{\sqrt{(n + m) \log(1 + \frac{4}{\varepsilon})}}{1 - \varepsilon} \leq c\tau\sqrt{n + m},$$

   where the constant c *(not optimized) can be obtained, for instance, by taking* $\varepsilon = 1/2$.

Remark 2. *Exercise* 2 *is based on Section 5.2 in the excellent lecture notes on high-dimensional probability by Van Handel [2014, Section 5.2].*

Exercise 3 (Random Fourier features)**.** *Let* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *be a continuous real-valued translation-invariant kernel scaled so that for any* $x \in \mathbb{R}^d$ *we have* $k(x, x) = 1$. *Then, by Bochner's theorem there exists a probability measure* P *such that*

$$k(x, x') = q(x - x') = \int_{\mathbb{R}^d} \exp\left(i\omega^\top (x - x')\right) P(d\omega) = \mathbf{E}_{\omega\sim P}\left[\exp\left(i\omega^\top (x - x')\right)\right]. \tag{8}$$

The aim of this exercise is to exploit the representation (8) *to build an approximate (and random) feature mapping* $\widehat{\phi} : \mathbb{R}^d \to \mathbb{R}^{2p}$ *such that* $k(x, x') \approx \widehat{\phi}(x)^\top \widehat{\phi}(x')$.

1. *Using the fact that the kernel k is real-valued, show using (8) that*

$$k(x,x') = \mathbf{E}_{\omega\sim P}\left[\cos(\omega^\top x)\cos(\omega^\top x') + \sin(\omega^\top x)\sin(\omega^\top x')\right]. \tag{9}$$

2. *The identity (9) suggests drawing p i.i.d. samples $\omega_1,\dots,\omega_p$ from the distribution P to build the approximate (random) feature map*

$$\widehat{\phi}(x) = \frac{1}{\sqrt{p}}\left(\cos(\omega_1^\top x), \sin(\omega_1^\top x), \cos(\omega_2^\top x), \sin(\omega_2^\top x), \dots, \cos(\omega_p^\top x), \sin(\omega_p^\top x)\right)^\top \in \mathbb{R}^{2p}.$$

Let $\widehat{k}(x,x') = \widehat{\phi}(x)^\top\widehat{\phi}(x')$ and let $r>0$ be a positive constant and let $B_r = \{x\in\mathbb{R}^d : \|x\|_2 \le r\}$. Let $\sigma_k = \mathbf{E}_{\omega\sim P}[\|\omega\|_2]$. Prove that

$$\mathbf{E}_{\omega_1,\dots,\omega_p}\left[\sup_{x,x'\in B_r}\left\{k(x,x') - \widehat{k}(x,x')\right\}\right] \le c\sqrt{\frac{d}{p}\log\left(1+\frac{r\sigma_k\sqrt{p}}{\sqrt{d}}\right)},$$

where $c>0$ is some universal constant. In particular, the above result shows that to get an ε-approximation of the kernel k, it suffices to take $p = \tilde{O}(d/\varepsilon^2)$, where the notation $\tilde{O}(\cdot)$ hides logarithmic factors.

Hint: *use the bounds (5) and (6) stated in Exercise 2.*

Solution 3.

1. *Since the kernel k is real-valued, we have*

$$\begin{aligned} k(x,x') &= \operatorname{Re} k(x,x') \\ &= \mathbf{E}_{\omega\sim P}\left[\operatorname{Re}\exp\left(i\omega^\top(x-x')\right)\right] \\ &= \mathbf{E}_{\omega\sim P}\left[\cos(\omega^\top x - \omega^\top x')\right] \\ &= \mathbf{E}_{\omega\sim P}\left[\cos(\omega^\top x)\cos(\omega^\top x') + \sin(\omega^\top x)\sin(\omega^\top x')\right]. \end{aligned}$$

2. *Observe that for any $x,x'\in B_r$, we have $x-x'\in B_{2r}$. We have $k(x,x') = q(x-x')$ and $\widehat{k}(x,x') = \widehat{q}(x-x') = \frac{1}{p}\sum_{i=1}^p \cos\left(\omega_i^\top(x-x')\right)$.*

We will apply the first part of Exercise 2 with $T = B_{2r}$, $d(z,z') = \|z-z'\|_2$ and $X_z = q(z) - \widehat{q}(z)$ for $z\in B_{2r}$. In particular, we have

$$\sup_{x,x'\in B_r}\left\{k(x,x') - \widehat{k}(x,x')\right\} = \sup_{z\in B_{2r}}\{X_z\}.$$

Furthermore, observe that for any $z\in B_{2r}$, the random variable X_z is zero-mean and $\frac{4}{p}$-sub-Gaussian (by Hoeffding's lemma and exercise 1.4 in exercise sheet 3). Finally, let

$$L = \sup_{z\neq z'\in B_{2r}}\frac{X_z - X_{z'}}{\|z-z'\|_2}.$$

By the first part of Exercise 2, and by the covering number bound of Euclidean balls (6) we have

$$\mathbf{E}\left[\sup_{x,x'\in B_r}\left\{k(x,x') - \widehat{k}(x,x')\right\}\right] = \mathbf{E}\left[\sup_{z\in B_{2r}}\{X_z\}\right]$$

$$
\begin{aligned}
&\leq \inf_{\varepsilon>0}\left\{\varepsilon\mathbf{E}[L]+\sqrt{\frac{8\log N(\varepsilon,B_{2r},\|\cdot\|_2)}{p}}\right\}\\
&\leq \inf_{\varepsilon>0}\left\{\varepsilon\mathbf{E}[L]+\sqrt{\frac{8\log N(\varepsilon/(2r),B_1,\|\cdot\|_2)}{p}}\right\}\\
&\leq \inf_{\varepsilon>0}\left\{\varepsilon\mathbf{E}[L]+\sqrt{\frac{8d\log(1+4r/\varepsilon)}{p}}\right\}. \qquad (10)
\end{aligned}
$$

It remains to obtain an upper bound on the $\mathbf{E}[L]$*. Because* q *and* $\widehat{q}$ *are continuously differentiable, we have*

$$
L=\sup_{z\neq z'\in B_{2r}}\frac{X_z-X_{z'}}{\|z-z'\|_2}=\sup_{z\in B_{2r}}\|\nabla q(z)-\nabla\widehat{q}(z)\|_2.
$$

We will now upper bound the expectation of the above quantity.

$$
\begin{aligned}
\mathbf{E}[L]&=\mathbf{E}\left[\sup_{z\in B_{2r}}\|\nabla q(z)-\nabla\widehat{q}(z)\|_2.\right]\\
&=\frac{1}{p}\mathbf{E}\left[\sup_{z\in B_{2r}}\left\|\sum_{i=1}^{p}\mathbf{E}_\omega\left[\omega\sin(\omega^\top z)\right]-\omega_i\sin(\omega_i^\top z)\right\|_2\right]\\
&\leq\frac{1}{p}\mathbf{E}\left[\sup_{z\in B_{2r}}\sum_{i=1}^{p}\mathbf{E}_\omega\left[\left\|\omega\sin(\omega^\top z)\right\|_2\right]+\left\|\omega_i\sin(\omega_i^\top z)\right\|_2\right]\\
&\leq\frac{1}{p}\mathbf{E}\left[\sup_{z\in B_{2r}}\sum_{i=1}^{p}\mathbf{E}_\omega\left[\|\omega\|_2\right]+\|\omega_i\|_2\right]\\
&=\frac{1}{p}\mathbf{E}\left[\sum_{i=1}^{p}\mathbf{E}_\omega\left[\|\omega\|_2\right]+\|\omega_i\|_2\right]\\
&=2\mathbf{E}_\omega\left[\|\omega\|_2\right]\\
&=2\sigma_k.
\end{aligned}
$$

Taking $\varepsilon=\sqrt{d/p}/\sigma_k$ *in the bound* (10) *completes the proof.*

Remark 3. *For a high-probability version of the above result and for pointers to related literature we refer to [Mohri, Rostamizadeh, and Talwalkar, 2018, Sections 6.6 and 6.7].*

References

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Ramon Van Handel. Probability in high dimension. Technical report, Princeton University, 2014. URL https://web.math.princeton.edu/~rvan/APC550.pdf.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.