Exercises Solutions for "Early stopping and non-parametric regression: An optimal data-dependent stopping rule

By Fofana Ladji Idrissa & Zacharia Echchair (Group 13)

Generalisation properties of ML algorithms

IP-PARIS

March 2023

Question 1

- The objective is given by :

$$\mathcal{L}(\omega) = \frac{1}{2n} \left\| y_1^n - \sqrt{n} K \omega \right\|_2^2$$

giving $\theta = \sqrt{K}\omega$ It follows : $\mathcal{L}(\omega) = \frac{1}{2n} \left\| y_1^n - \sqrt{n}\sqrt{K}\theta \right\|_2^2 = \frac{1}{2n} \left\| y_1^n \right\|_2^2 - \frac{1}{\sqrt{n}} \left\langle y_1^n, \sqrt{K}\theta \right\rangle + \frac{1}{2}(\theta)^T K \theta$

By considering the sequence $\{\alpha^t\}_{t=0}^{\infty}$, we can write the following gradient descent algorithm :

$$\theta^{t+1} = \theta^t - \alpha^t \nabla \tilde{\mathcal{L}} \left(\theta^t \right)$$

Question 1

- The objective is given by :

$$\mathcal{L}(\omega) = \frac{1}{2n} \left\| y_1^n - \sqrt{n} K \omega \right\|_2^2$$

giving $\theta = \sqrt{K}\omega$ It follows : $\mathcal{L}(\omega) = \frac{1}{2n} \left\| y_1^n - \sqrt{n}\sqrt{K}\theta \right\|_2^2 = \frac{1}{2n} \left\| y_1^n \right\|_2^2 - \frac{1}{\sqrt{n}} \left\langle y_1^n, \sqrt{K}\theta \right\rangle + \frac{1}{2}(\theta)^T K \theta$

By considering the sequence $\{\alpha^t\}_{t=0}^{\infty}$, we can write the following gradient descent algorithm :

$$\theta^{t+1} = \theta^t - \alpha^t \nabla \tilde{\mathcal{L}} \left(\theta^t \right)$$

- Now, by calculating the gradient of $\widetilde{\mathcal{L}}$, we get :

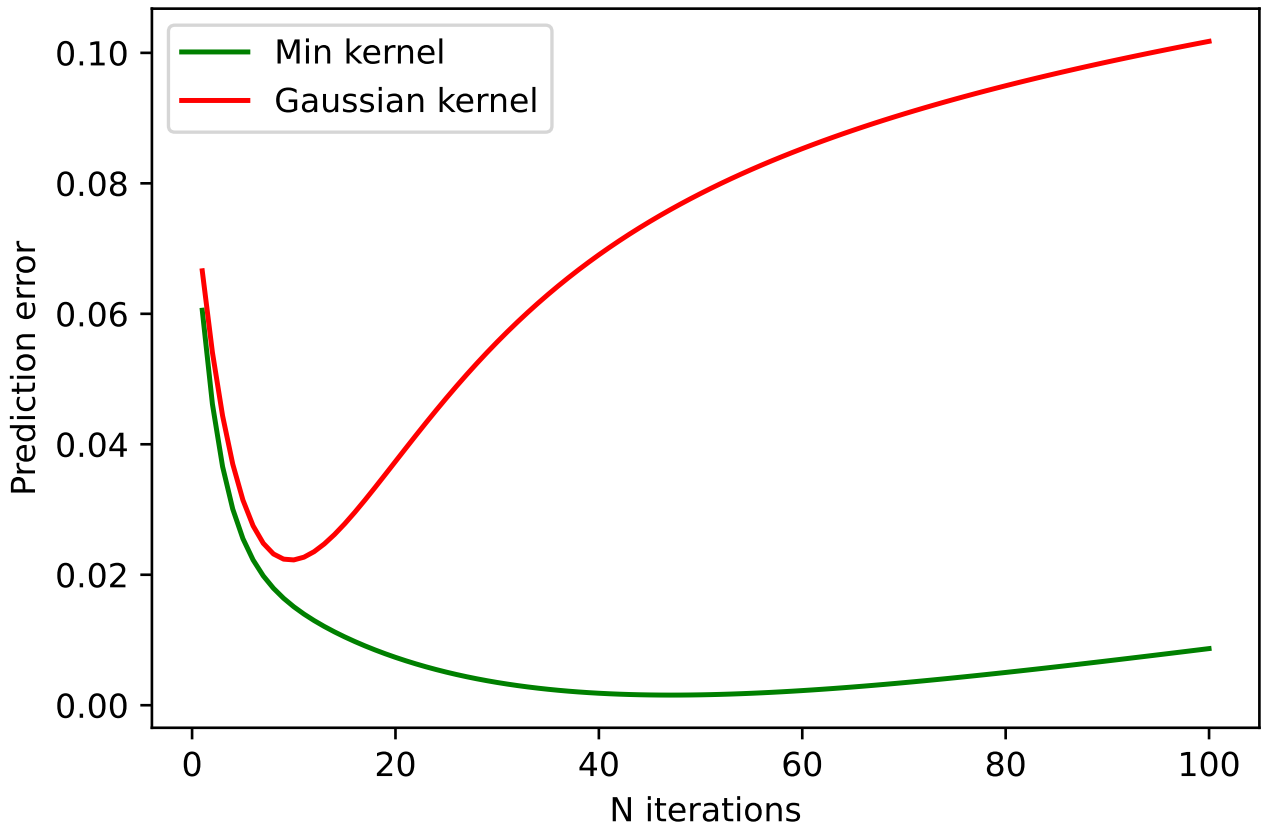
$$\nabla \widetilde{\mathcal{L}}(\theta) = K\theta - \frac{1}{\sqrt{n}} \sqrt{K} y_1^n$$

Question 2

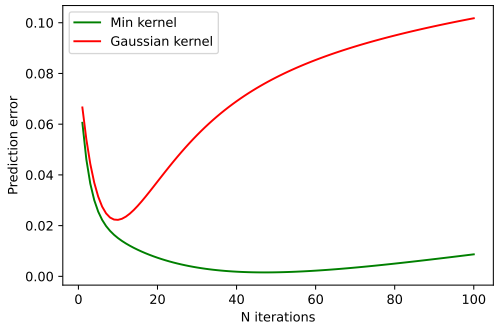
- We will let you see the python code and also this exercises solution sheet in our **github repository**.

Question 2

- We will let you see the python code and also this exercises solution sheet in our **github repository**.
- The resulting plots for the kernels is given by :

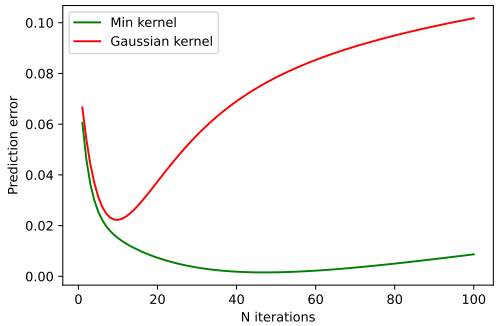


Question 2 # Comments



- The first plot shows the min_kernel model starts with high error of 0.8 and decreases rapidly within the first 10 iterations. After that, the error decreases at a slower rate and reaches 0.06 after 100 iterations. The second plot shows the gaussian_kernel model starts with a lower error of 0.5 and decreases slowly but steadily throughout the 100 iterations. The final error achieved after 100 iterations is approximately 0.03, which is lower than that of the min_kernel.

Question 2 # Comments



- The first plot shows the min_kernel model starts with high error of 0.8 and decreases rapidly within the first 10 iterations. After that, the error decreases at a slower rate and reaches 0.06 after 100 iterations. The second plot shows the gaussian_kernel model starts with a lower error of 0.5 and decreases slowly but steadily throughout the 100 iterations. The final error achieved after 100 iterations is approximately 0.03, which is lower than that of the min_kernel.
- Overall, the gaussian kernel appears to perform better than the min kernel in this particular experiment. However, the performance of each kernel may vary depending on the specific dataset and task at hand.

Question 3

- In this proof, we aim to show that the critical radius $\widehat{\varepsilon}_n$ exists, is unique, and lies in the interval $(0, \infty)$. The stopping point $\widehat{T}$ is defined as the minimum value of ϵ such that $\widehat{\mathcal{R}}_K(\epsilon) \leq \epsilon^2/(2e\sigma)$. We can rearrange and substitute to obtain an equivalent expression for $\widehat{\varepsilon}_n$:

$$\widehat{\varepsilon}_n := \arg\min \epsilon > 0 \mid \sum_{i=1}^{n} \min \epsilon^{-2}\widehat{\lambda}_i, 1 > n\epsilon^2 / \left(4e^2\sigma^2\right)$$

Question 3

- In this proof, we aim to show that the critical radius $\widehat{\varepsilon}_n$ exists, is unique, and lies in the interval $(0, \infty)$. The stopping point $\widehat{T}$ is defined as the minimum value of ϵ such that $\widehat{\mathcal{R}}_K(\epsilon) \leq \epsilon^2/(2e\sigma)$. We can rearrange and substitute to obtain an equivalent expression for $\widehat{\varepsilon}_n$:

$$\widehat{\varepsilon}_n := \arg\min \epsilon > 0 \mid \sum_{i=1}^{n} \min \epsilon^{-2}\widehat{\lambda}_i, 1 > n\epsilon^2 / \left(4e^2\sigma^2\right)$$

- We note that $\sum_{i=1}^{n} \min \epsilon^{-2}\widehat{\lambda}_{i,1}$ is non-increasing in ϵ, while $n\epsilon^2$ is increasing in ϵ. Also, $\widehat{\varepsilon}n$ exists since for $\epsilon = 0$, we have $0 = n\epsilon^2 < \sum_{i=1}^{n} \min \epsilon^{-2}\widehat{\lambda}i, 1 > 0$, while for $\epsilon = \infty$, we have $\sum_{i=1}^{n} \min \eta_t\widehat{\lambda}_{i,1} < n\epsilon^2$.

Question 3 following

- Moreover, since $\widehat{\mathcal{R}}_K(\epsilon)$ is the sum of n continuous functions, it is also continuous. Therefore, the critical radius $\widehat{\varepsilon}_n$ exists, is unique, and satisfies the fixed-point equation $\widehat{\mathcal{R}}_K\left(\widehat{\varepsilon}_n\right) = \widehat{\varepsilon}_n^2/(2e\sigma)$.

Question 3 following

- Moreover, since $\widehat{\mathcal{R}}_K(\epsilon)$ is the sum of n continuous functions, it is also continuous. Therefore, the critical radius $\widehat{\varepsilon}_n$ exists, is unique, and satisfies the fixed-point equation $\widehat{\mathcal{R}}_K\left(\widehat{\varepsilon}_n\right) = \widehat{\varepsilon}_n^2/(2e\sigma)$.
- To show that the integer $\widehat{T}$ belongs to the interval $[0, \infty)$ and is unique for any valid sequence of step-sizes, we use the definition of $\widehat{T}$ given by stopping rule (1) and $\widehat{\varepsilon}n$, which implies that $\frac{1}{\eta \widehat{T}+1} \leq \hat{\varepsilon} n^2 \leq \frac{1}{\eta \widehat{T}}$. Since $\eta_0 = 0$ and $\eta_t \to \infty$ as $t \to \infty$ and $\widehat{\varepsilon}_n \in (0, \infty)$, there exists a unique stopping point $\widehat{T}$ in the interval $[0, \infty)$.

Question 3 following

- Moreover, since $\widehat{\mathcal{R}}_K(\epsilon)$ is the sum of n continuous functions, it is also continuous. Therefore, the critical radius $\widehat{\varepsilon}_n$ exists, is unique, and satisfies the fixed-point equation $\widehat{\mathcal{R}}_K\left(\widehat{\varepsilon}_n\right) = \widehat{\varepsilon}_n^2/(2e\sigma)$.
- To show that the integer $\widehat{T}$ belongs to the interval $[0, \infty)$ and is unique for any valid sequence of step-sizes, we use the definition of $\widehat{T}$ given by stopping rule (1) and $\widehat{\varepsilon}n$, which implies that $\frac{1}{\eta\widehat{T}+1} \leq \hat{\varepsilon}n^2 \leq \frac{1}{\eta\widehat{T}}$. Since $\eta_0 = 0$ and $\eta_t \to \infty$ as $t \to \infty$ and $\widehat{\varepsilon}_n \in (0, \infty)$, there exists a unique stopping point $\widehat{T}$ in the interval $[0, \infty)$.
- Conclusion : This establishes at first the existence, uniqueness, location of critical radius and stopping point in optimization problem using key observations. And seconly it shows rigorous demonstration of convergence properties of algorithm.

Question 4

- Firstly, let us rewrite the gradient update in a new form. For each iteration $t = 0, 1, 2, \ldots$, we define the shorthand $f^t\left(x_1^n\right)$ as the n-vector obtained by evaluating the function f^t at all design points $x_1, x_2, \ldots, x_n$. Additionally, we define w as the vector of zero mean sub-Gaussian noise random variables $[w_1, w_2, \ldots, w_n]$.

Question 4

- Firstly, let us rewrite the gradient update in a new form. For each iteration $t = 0, 1, 2, \ldots$, we define the shorthand $f^t\left(x_1^n\right)$ as the n-vector obtained by evaluating the function f^t at all design points $x_1, x_2, \ldots, x_n$. Additionally, we define w as the vector of zero mean sub-Gaussian noise random variables $[w_1, w_2, \ldots, w_n]$.
- Using the equation $f(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i \mathbb{K}\left(., x_i\right)$, we can obtain the relation $f^t\left(x_1^n\right) = \frac{1}{\sqrt{n}} K\omega^t = \frac{1}{\sqrt{n}} \sqrt{K}\theta^t$. By multiplying both sides of the gradient update $\theta^{t+1} = \theta^t - \alpha^t \left(K\theta^t - \frac{1}{\sqrt{n}}\sqrt{K}y_1^n\right)$ by $\sqrt{K}$, we find that the sequence $f^t\left(x_1^n\right)\ t = 0^\infty$ evolves according to the recursion $f^{t+1}\left(x_1^n\right) = f^t\left(x_1^n\right) - \alpha^t K\left(f^t\left(x_1^n\right) - y_1^n\right) = \left(In \times n - \alpha^t K\right) f^t\left(x_1^n\right) - \alpha^t K y_1^n$.

Question 4 following 1

- We begin by initializing the sequence with $f^0\left(x_1^n\right) = 0$, since $\theta^0 = 0$. Using the rank of the empirical kernel matrix K, which has an eigen-decomposition of $K = U\Lambda U^T$, where $U \in \mathbb{R}^{n\times n}$ is an orthonormal matrix $\left(UU^T = U^TU = I_{n\times n}\right)$ and $\Lambda := \text{diag}\left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_r, 0, 0, \cdots, 0\right)$ is the diagonal matrix of eigenvalues, augmented with $n - r$ zero eigenvalues as needed, we create a sequence of diagonal shrinkage matrices S^t.

Question 4 following 1

- We begin by initializing the sequence with $f^0\left(x_1^n\right) = 0$, since $\theta^0 = 0$. Using the rank of the empirical kernel matrix K, which has an eigen-decomposition of $K = U\Lambda U^T$, where $U \in \mathbb{R}^{n\times n}$ is an orthonormal matrix $(UU^T = U^TU = I_{n\times n})$ and $\Lambda := \text{diag}\left(\hat{\lambda}_1, \hat{\lambda}_2, \ldots, \hat{\lambda}_r, 0, 0, \cdots, 0\right)$ is the diagonal matrix of eigenvalues, augmented with $n - r$ zero eigenvalues as needed, we create a sequence of diagonal shrinkage matrices S^t.
- S^t is defined as follows:

$$S^t := \prod_{\tau=0}^{t-1}\left(I_{n\times n} - \alpha^\tau\Lambda\right) \in \mathbb{R}^{n\times n}$$

This matrix S^t represents the degree of shrinkage towards the origin. Since $0 \leq \alpha^t \leq \min 1, 1/\hat{\lambda}1$ for all iterations t, we have the sandwich relation in the positive semi-definite ordering:

$$0 \preceq S^{t+1} \preceq S^t \preceq In \times n$$

Question 4 following 2

- The algorithm produces a sequence of functions $(f_t)_{t\geq 0}$, and we can conclude that this sequence satisfies a bound with a high probability. The bound is given by the formula $\left|\widehat{f}_T - f^*\right|_n^2 \leq \frac{4\widehat{\varepsilon}_n}{n^2} \ln\left(\frac{4}{\delta}\right)$, where $\widehat{T}$ and $\widehat{\varepsilon}_n$ are determined by the algorithm and δ controls the probability of error.

Question 4 following 2

- The algorithm produces a sequence of functions $(f_t)_{t\geq 0}$, and we can conclude that this sequence satisfies a bound with a high probability. The bound is given by the formula $\left|\widehat{f}_T - f^*\right|_n^2 \leq \frac{4\widehat{\varepsilon}_n}{n^2} \ln\left(\frac{4}{\delta}\right)$, where $\widehat{T}$ and $\widehat{\varepsilon}_n$ are determined by the algorithm and δ controls the probability of error.
- We can write finally that $\left|\widehat{f}_T - f^*\right|_n^2 \leq \frac{4}{e\eta\widehat{T}} \leq 12\widehat{\varepsilon}_n^2$

Question 4 following 2

- The algorithm produces a sequence of functions $(f_t)_{t\geq 0}$, and we can conclude that this sequence satisfies a bound with a high probability. The bound is given by the formula $\left|\widehat{f}_T - f^*\right|_n^2 \leq \frac{4\widehat{\varepsilon}_n}{n^2} \ln\left(\frac{4}{\delta}\right)$, where $\widehat{T}$ and $\widehat{\varepsilon}_n$ are determined by the algorithm and δ controls the probability of error.
- We can write finally that $\left|\widehat{f}_T - f^*\right|_n^2 \leq \frac{4}{e\eta\widehat{T}} \leq 12\widehat{\varepsilon}_n^2$
- The lower bound on the expected error of our method requires some additional assumptions and analysis, but it appears to be based on bounding the variance of the estimator and showing that this provides a lower bound on the error.

Question 5

- Applying the gradient update reduces the prediction error if the total of the step-sizes η_t is less than the limit defined by the definition of $\widehat{T}$. It's worth noting that for Hilbert spaces with a more complex kernel, the stopping time $\widehat{T}$ is smaller. This is because fitting functions in a larger class carries a greater risk of overfitting.

Question 5

- Applying the gradient update reduces the prediction error if the total of the step-sizes η_t is less than the limit defined by the definition of $\widehat{T}$. It's worth noting that for Hilbert spaces with a more complex kernel, the stopping time $\widehat{T}$ is smaller. This is because fitting functions in a larger class carries a greater risk of overfitting.
- The inequality $\left|f_{\widehat{T}} - f^*\right| n^2 \leq \frac{4}{e\eta\widehat{T}} \leq 12\widehat{\varepsilon}_n^2$ holds.

Question 5

- Applying the gradient update reduces the prediction error if the total of the step-sizes η_t is less than the limit defined by the definition of $\widehat{T}$. It's worth noting that for Hilbert spaces with a more complex kernel, the stopping time $\widehat{T}$ is smaller. This is because fitting functions in a larger class carries a greater risk of overfitting.
- The inequality $\left|f_{\widehat{T}} - f^*\right| n^2 \leq \frac{4}{e\eta\widehat{T}} \leq 12\widehat{\varepsilon}_n^2$ holds.
- Comments : This inequality highlights the trade-off between the kernel complexity and the stopping time: a larger kernel complexity requires a smaller stopping time to avoid overfitting. Overall, this result provides useful insights into the factors that affect the accuracy of gradient-based prediction methods in Hilbert spaces.