Group number : Amine ELKARI - Marouane NAJID

MAP670L - Validation

Chosen article : "Early stopping and non-parametric regression: An optimal data-dependent stopping rule"

---

1 Exercise

Our goal is to estimate f^*. Equivalently, we observe samples of the form :

$$y_i = f^*(x_i) + \omega_i, \text{ for } i = 1, 2, ..., n$$

Where $\omega_i = y_i - f^*(x_i)$ is a zero-mean noise random variable.

We are going to focus on non parametric regression in a a reproducing kernel Hilbert space (RKHS). Considering a Hilbert space $\mathcal{H}$, we suppose that :

$$\exists\ \mathbb{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+, \text{ symetric}$$

Such that :
a) $\forall x \in \mathcal{X},\ \mathbb{K}(., x) \in \mathcal{H}$
b) $\forall f \in \mathcal{H},\ f(x) = \langle f, \mathbb{K}(., x)\rangle$

Under this conditions, we say $\mathcal{H}$ is a RKHS.
We can write $\mathbb{K}$ in this form :

$$\mathbb{K}\left(x, x'\right) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k\left(x'\right)$$

Where $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq 0$ are non negative sequence of eigenvalues and $\{\phi_k\}_{k=1}^{\infty}$ are the associated eigen functions.
Since the eigen functions $\{\phi_k\}_{k=1}^{\infty}$ form an orthonormal basis.

$$f(x) = \sum_{k=1}^{\infty} \sqrt{\lambda_k} a_k \phi_k(x), \text{we suppose that} \sum_{k=1}^{\infty} a_k \leq 1$$

Under this considerations and over some subset of the hilbert space $\mathcal{H}$, it suffices to restrict attention to functions f belonging to the span of the kernel functions $\{\mathbb{K}(., x_i)\ ;\ i = 1, ..., n\}$
.

We adopt that parametrization :

$$f(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i \mathbb{K}(., x_i)$$

Considering the loss function :

$$\mathcal{L}(f) := \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - f\left(x_i\right)\right)^2$$

Also the empirical Kernel matrix $K \in \mathbb{R}^{n\times n}$ with entries:

$$[K]_{ij} = \frac{1}{n} K\left(x_i, x_j\right) \quad \text{ for } i, j = 1, 2, \ldots, n.$$

For any positive semidefinite kernel function, this matrix must be positive semidefinite, and so has a unique symmetric square root denoted by $\sqrt{K}$.
Introducing the convenient shorthand $y_1^n := (y_1 y_2 \cdots y_n) \in \mathbb{R}^n$, we can then write the least-squares loss in the form :

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

Question 1 :

Proove that the gradient descent algorithm could be written as :

$$\theta^{t+1} = \theta^t - \alpha^t \left(K\theta^t - \frac{1}{\sqrt{n}} \sqrt{K} y_1^n \right),$$

Where $\{\alpha_t\}_0^\infty$ is a sequence of positive step size and θ **to be determined** .

Question 2 :

Implement an example of gradient descent and visualize the error as a function of the iterations. Conclude.
To over come this problem, lets find a data dependent stopping rule.
First lets define two quantities :

$$\eta_t := \sum_{\tau=0}^{t-1} \alpha^\tau,$$

$$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_n \geq 0 \;\text{ eigen values of } K.$$

$$\widehat{\mathcal{R}}_K(\varepsilon) := \left[\frac{1}{n} \sum_{i=1}^{n} \min\left\{ \hat{\lambda}_i, \varepsilon^2 \right\} \right]^{1/2} .$$

For a given noise variance $\sigma > 0$, a closely related quantity-one of central importance to our analysis is critical empirical radius $\widehat{\varepsilon}_n > 0$, defined to be the smallest positive solution to the inequality :

$$\widehat{\mathcal{R}}_K(\varepsilon) \leq \varepsilon^2/(2e\sigma).$$

Our stopping rule is defined in terms of an analogous inequality that involves the running sum $\eta_t = \sum_{\tau=0}^{t-1} \alpha^\tau$ of the step sizes.

We assume that the step sizes are chosen to satisfy the following properties:

- Boundedness: $0 \leq \alpha^\tau \leq \min\left\{1, 1/\hat{\lambda}_1\right\}$ for all $\tau = 0, 1, 2, \ldots$
- Non-increasing: $\alpha^{\tau+1} \leq \alpha^\tau$ for all $\tau = 0, 1, 2, \ldots$.
- Infinite travel: the running sum $\eta_t = \sum_{\tau=0}^{t-1} \alpha^\tau$ diverges as $t \to +\infty$.

We refer to any sequence $\{\alpha^\tau\}_{\tau=0}^{\infty}$ that satisfies these conditions as a valid stepsize sequence.
We then define the stopping time :

$$\widehat{T} := \arg\min \left\{ t \in \mathbb{N} \mid \widehat{\mathcal{R}}_K \left(1/\sqrt{\eta_t}\right) > \left(2e\sigma\eta_t\right)^{-1} \right\} - 1. \tag{1}$$

Question 3 :

Prove the existence and uniqueness of $\hat{\epsilon}_n$ and $\widehat{T}$. $\left(\text{Assuming that : } \frac{1}{\eta_{\hat{T}+1}} \leq \hat{\epsilon_n}^2 \leq \frac{1}{\eta_{\hat{T}}}\right)$

Question 4 :

Given the stopping time $\widehat{T}$ defined by the rule (1), there are universal positive constants (c_1, c_2) such that the following events both hold with probability at least $1 - c_1 \exp\left(-c_2 n \bar{\varepsilon}_n^2\right)$, Proove the following inequalities :

(a) For all iterations $t = 1, 2, \ldots, \widehat{T}$:

$$\|f_t - f^*\|_n^2 \leq \frac{4}{e\eta_t}.$$

(Assuming that $\mathbb{E}\left[\|f_t - f^*\|_n^2\right] \geq \mathbb{E}\left[V_t\right]$, where V_t is the variance)

(b) At the iteration $\widehat{T}$ chosen according to the stopping rule (1), we have

$$\left\|f_{\widehat{T}} - f^*\right\|_n^2 \leq 12\widehat{\varepsilon}_n^2.$$

Question 5 :

Knowing that also for all $t > \widehat{T}$, we have

$$\mathbb{E}\left[\|f_t - f^*\|_n^2\right] \geq \frac{\sigma^2}{4}\eta_t \widehat{\mathcal{R}}_K\left(\eta_t^{-1/2}\right).$$

Interpret this result.