STA3000F: Homework 3

Due: December 5, 2025, 11:59pm on Quercus

Q1: oracle inequality

During the lecture, we mostly focus on nonparametric estimation with the true function belonging to the function class. In practice, all models are wrong, but some are useful. Concretely, let us consider the fixed-design regression model with $Y_i = f^*(x_i) + \varepsilon_i$, where ε_i are i.i.d. sub-Gaussian noise with parameter σ^2. Let $\mathcal{F}$ be a convex class of functions. Define the least-squares estimator over the function class $\mathcal{F}$ as

$$\widehat{f}_n \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(x_i))^2.$$

Show that with high probability, we have

$$\left\| \widehat{f}_n - f^* \right\|_n^2 \leq \inf_{f \in \mathcal{F}} \| f - f^* \|_n^2 + c\delta_n^2,$$

where δ_n is the critical radius defined via the fixed-point equation $\delta_n^2 = \mathcal{G}_n(\delta_n)$; here $\mathcal{G}_n(\delta)$ is the localized Gaussian complexity. (Please note that the pre-factor in the approximation error is exactly 1.)

Q2: Gaussian smoothing

Define the function class

$$\mathcal{F} := \left\{ f : \mathbb{R}^d \to \mathbb{R} \mid f(x) = \mathbb{E}_{Z \sim \mathcal{N}(0, I_d)}[g(x + Z)], \text{ for some function } g \text{ with } \|g\|_\infty \le 1 \right\}.$$

Let $x_1, x_2, \cdots, x_n \in [0,1]^d$ be n fixed design points, and consider the observation model

$$Y_i = f^*(x_i) + \varepsilon_i, \quad i = 1, 2, \cdots, n,$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $f^* \in \mathcal{F}$. Let $\widehat{f}_n$ be the least-squares estimator

$$\widehat{f}_n \in \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(x_i))^2.$$

Show that for any $\varepsilon > 0$, there exists a constant $c_\varepsilon > 0$ that depends on ε and d, such that

$$\left\| \widehat{f}_n - f^* \right\|_n \le c_\varepsilon n^{-\frac{1}{2} + \varepsilon},$$

with high probability.
[Remark: Indeed, you can prove a sharper result with logarithmic factors, but we are fine with this near-optimal rate here.]

Q3: another density estimation method

Let $\mathcal{F}$ be a class of densities on the domain $\mathbb{X}$. For simplicity, let us assume that $\mathcal{F}$ is finite. Given i.i.d. samples $X_1, X_2, \cdots, X_n$ from an unknown density p^*, the goal is to estimate p^*. In doing so, we define the set

$$\mathcal{A} := \{A \subseteq \mathbb{X} \mid A = \{x \in \mathbb{X} \mid p(x) > q(x)\}, \text{ for some } p, q \in \mathcal{F}\}.$$

We can then define the estimator $\widehat{p}_n$ as

$$\widehat{p}_n \in \arg\min_{p \in \mathcal{F}} \sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \in A} - \int_A p(x) \mathrm{d}x \right|.$$

Show that with high probability, we have

$$d_{\mathrm{TV}}(\widehat{p}_n, p^*) \leq 3 \inf_{p \in \mathcal{F}} d_{\mathrm{TV}}(p, p^*) + c\sqrt{\frac{\log |\mathcal{F}|}{n}}.$$

[Hint: an equivalent definition of the total variation distance is $d_{\mathrm{TV}}(p, q) = \sup_{A \subseteq \mathbb{X}} |\int_A p(x) \mathrm{d}x - \int_A q(x) \mathrm{d}x|$, where the supremum is attained at the set $A = \{x \in \mathbb{X} \mid p(x) > q(x)\}$.]

Q4: estimating a probability transition kernel

Let $(X_i, Y_i)_{i=1}^n$ be i.i.d. samples from a joint distribution over $[0,1]^2$, with density $p^*(x,y)$. The goal is to estimate the conditional density $p^*(y \mid x)$. We impose the following assumptions on the joint density p^*.

- (marginal) The marginal density $p^*(x)$ satisfies $0 < p_{\min} \le p^*(x) < p_{\max} < +\infty$ for all $x \in [0,1]$.
- (smoothness) The conditional density $p^*(y \mid x)$ is Lipschitz in x and second-order smooth in y, i.e., there exist constants $L_1, L_2 > 0$ such that for any $x, y \in [0,1]$,

$$|\partial_x p^*(y \mid x)| \le L_1, \quad |\partial_y^2 p^*(y \mid x)| \le L_2.$$

Construct an estimator $\widehat{p}_n(y \mid x)$ and prove its convergence rate under the integrated squared error loss

$$R(\widehat{p}_n; p^*) := \mathbb{E}\Big[d_{\mathrm{TV}}\big(\widehat{p}_n(\cdot \mid X), p^*(\cdot \mid X)\big)^2\Big].$$

Make the convergence rate as sharp as possible, but you do not need to prove minimax lower bounds. You can also see the constants $p_{\min}, p_{\max}, L_1, L_2$ as universal constants that do not depend on n.