36-789: Topics in High Dimensional Statistics II	Fall 2015

Lecture 5: November 10

Lecturer: Alessandro Rinaldo *Scribe: Shashank Singh*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

5.1 Recap and Outline

In the previous lecture, we introduced the Le Cam equation which states the minimax rate for certain problems can be computed by solving

$$N(\varepsilon, \Theta, d) = n\varepsilon^2 \tag{5.1}$$

for ε, where Θ is the hypothesis class on which d is a metric (used as a loss function) and for which $N(\varepsilon, \Theta, d)$ is the ε-covering number. Today, we discuss applications of the Le Cam equation to nonparametric density estimation under L_2 loss, over hypothesis classes consisting of Hölder continuous functions. We also provide some intuition for the Le Cam equation by analyzing the risk of a regression estimator based on empirical risk minimization.

Remark: While all the examples we discuss here are nonparametric problems, the Le Cam equation also holds in many parametric problems. Here, we typically have $\log N(\varepsilon) = d \log \varepsilon^{-1}$. [1] This leads to a minimax lower bound of $\asymp \sqrt{d/n}$, which is hence commonly referred to as the *parametric rate*. [2]

[1] $N(\varepsilon)$ denotes the covering number, suppressing the dependence on the hypothesis class and loss (since these are fixed).

[2] Of course, many nonparametric problems have the parametric rate (e.g. estimating statistical functionals when densities are sufficiently smooth relative to the dimension [K15]).

5.2 Hölder Classes of Smooth Functions

We first provide some notation necessary for defining the Hölder function class. $\mathbb{N}^d$ denotes the set of d-tuples of non-negative integers, which we denote with a vector symbol $\vec{\cdot}$, and, for $\vec{i} \in \mathbb{N}^d$, we define the operators

$$D^{\vec{i}} := \frac{\partial^{|\vec{i}|}}{\partial^{i_1} x_1 \cdots \partial^{i_d} x_d} \quad \text{and} \quad |\vec{i}| = \sum_{k=1}^{d} i_k.$$

Definition 5.1 (Hölder Ball): *Suppose $\mathcal{X} \subseteq \mathbb{R}^n$ is open. For $\beta, L > 0$, a Hölder ball is a set of the form:*

$$\Sigma(\mathcal{X}, \beta, L) := \left\{ f : \mathcal{X} \to \mathbb{R} \middle| \sup_{\substack{x \neq y \in \mathcal{X} \\ |\vec{i}| = \lfloor \beta \rfloor}} \frac{|D^{\vec{i}} f(x) - D^{\vec{i}} f(y)|}{\|x - y\|^{\beta - \lfloor \beta \rfloor}} \leq L \right\}, \tag{5.2}$$

where $\lfloor \beta \rfloor$ is the greatest integer strictly *less than β.*

We will use without proof the following fact about Hölder Balls:

Lemma 5.2 *The* log-*covering number of a Hölder ball over an open set $\mathcal{X} \subseteq \mathbb{R}^d$ is*

$$\log N(\varepsilon, \Sigma(\mathcal{X}, \beta, L), \| \cdot \|_2) \asymp \varepsilon^{-d/\beta}. \tag{5.3}$$

5.3 Application to Nonparametric Density Estimation

Suppose $f \in \Sigma(\mathcal{X}, \beta, L)$ is an unknown probability density function, from which we observe samples $X_1, \ldots, X_n$. We are interested in estimating f. If we make the further assumption that f is lower and upper bounded away from 0 and ∞, respectively (i.e., $c := \inf_{x \in \mathcal{X}} f(x) > 0$ and $C := \sup_{x \in \mathcal{X}} f(x) < \infty$), then, it can be shown that the Le Cam equation (5.1) holds. Then, (5.3) gives $n\varepsilon^2 = \varepsilon^{-d/\beta}$, and solving for ε^2 in terms of n gives the following:

Proposition 5.3 *The L_2 minimax rate for the above nonparametric density estimation problem satisfies*

$$\inf_{\hat{f}} \sup_{\substack{f \in \Sigma(\mathcal{X}, \beta, L) \\ 0 < c \leq f \leq C < \infty}} \mathbb{E} \left[\|f - \hat{f}\|_2 \right] \in \Omega \left(n^{-\frac{\beta}{d + 2\beta}} \right).$$

Remark: With the right choice of bandwidth and additional assumptions on the behavior of f near the boundary of $\mathcal{X}$, the risk of the kernel density estimator is of this order, providing a matching upper bound.

5.4 The Le Cam Equation for Nonparametric Regression

Here we provide some intuition for the Le Cam equation by showing that it arises naturally when tuning a nonparametric regression estimator based on empirical risk minimization.

Setup: Fix a real-valued function class $\mathcal{F}$ on a domain $\mathcal{X}$, and fix $f^* \in \mathcal{F}$ and $X_1, \dots, X_n \in \mathcal{X}$. Let $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$, and suppose we observe $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathcal{X} \times \mathbb{R}$, where each $Y_i = f^*(X_i) + \varepsilon_i$. Define the empirical L_2 risk

$$R_n(f, f') := \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \left(\hat{f}(X_i) - f^*(X_i)\right)\right], \quad \forall f, f' \in \mathcal{F}$$

(noting that R_n is a (pseudo)metric) and the empirical risk minimization estimator

$$\hat{f} := \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n \left(f(X_i) - Y_i\right)^2 .^3$$

Analysis: By construction, of $\hat{f}$,

$$\frac{1}{n}\sum_{i=1}^n \left(\hat{f}(X_i) - f^*(X_i)\right) \leq \frac{2}{n}\sum_{i=1}^n \varepsilon_i \left(\hat{f}(X_i) - f^*(X_i)\right) \stackrel{D}{=} \frac{2\sigma}{n}\sum_{i=1}^n w_i \left(\hat{f}(X_i) - f^*(X_i)\right) = \frac{2\sigma}{\sqrt{n}} \sup_{g \in \mathcal{G}} \sum_{i=1}^n w_i g(X_i),$$

where $w \sim \mathcal{N}(0_n, I_n)$, $\mathcal{G} := \left\{\frac{f - f^*}{\sqrt{n}} : f \in \mathcal{F}\right\}$, and $\stackrel{D}{=}$ denotes equality in distribution.

Let $\delta > 0$ (to be chosen later), and suppose that $(\mathcal{G}, R_n)$ has a finite covering number $N_\delta := N(\delta, \mathcal{G}, R_n)$. Suppose $\{g^{(1)}, \dots, g^{(N_\delta)}\} \subseteq \mathcal{G}$ is a minimal δ-covering of $(\mathcal{G}, R_n)$. Then, for any $g \in \mathcal{G}$, letting

$$j := \operatorname*{argmin}_{i \in \{1, \dots, N_\delta\}} R_n\left(g^{(i)}, g\right),$$

by definition of a δ-covering,

$$\sum_{i=1}^n w_i g(X_i) = \sum_{i=1}^n w_i g^{(j)}(X_i) + w_i\left(g(X_i) - g^{(j)}(X_i)\right) \leq \max_{\ell \in \{1, \dots, N_\delta\}} \sum_{i=1}^n w_i g^{(\ell)}(X_i) + \delta \|w\|_2.$$

Chaining together the above inequalities and taking expectations on both sides gives

$$R_n(f, f') \leq \frac{2\sigma}{\sqrt{n}}\left(\mathbb{E}\left[\max_{\ell \in \{1, \dots, N_\delta\}} \sum_{i=1}^n w_i g^{(\ell)}(X_i)\right] + \delta \mathbb{E}\left[\|w\|_2\right]\right) \leq \frac{2\sigma}{\sqrt{n}}\left(\sqrt{2\nu \log\left(N_\delta\right)} + \delta\sqrt{n}\right),$$

where $\nu := \max_{\ell \in \{1, \dots, N_\delta\}} \sum_{i=1}^n g^{(\ell)}(X_i)$, using a standard bound on the expectation of a maximum of a sum of independent Gaussian random variables. Note that the first term is non-increasing in δ (since N_δ is non-increasing in δ), while the second term is non-decreasing in δ. Hence, to minimize the rate of this expression in δ, we equate the two terms:

$$\sqrt{\log N_\delta} = \delta\sqrt{n}.$$

Squaring both sides gives the Le Cam equation.

Next time, we will formally prove this minimax lower bound for nonparametric regression with L_2 loss, and begin discussing Assouad's method.

---

[3] $\hat{f}$ may very well not be computable in practice; here, we are interested only in analyzing its statistical performance.

References

[T08] Tsybakov, Alexandre B. *Introduction to nonparametric estimation.* Springer Science & Business Media, 2008.

[YB99] Yang, Yuhong, and Andrew Barron. "Information-theoretic determination of minimax rates of convergence." Annals of Statistics (1999): 1564-1599.

[K15] Kandasamy, Kirthevasan, et al. "Influence Functions for Machine Learning: Nonparametric Estimators for Entropies, Divergences and Mutual Informations." arXiv preprint arXiv:1411.4342 (2014).