Lecture 4: November 5

Lecturer: Alessandro Rinaldo *Scribes: Alnur Ali*

In this lecture, we continue our discussion of the *Le Cam equation*, which is a method for obtaining minimax lower bounds using Fano's method (see, *e.g.*, `http://projecteuclid.org/download/pdf_1/euclid.aos/1017939142`).

4.1 Brief recap of Fano's method

We begin with a brief recap of Fano's method; in Fano's method, a minimax lower bound (*i.e.*, on the minimax risk) is given as

$$w(\delta)\left(1 - \frac{I(X;V) + \log 2}{\log(m+1)}\right),$$

where $w : \mathbf{R}_+ \to \mathbf{R}_+$ is the loss function assumed to be nondecreasing and satisfying $w(0) = 0$, $\delta > 0$, $m+1$ is the size of our hypothesis class, and $I(X;V)$ is the mutual information of the data X and the random variable V taking values in our hypothesis class.

We have a minimax lower bound if we can show that

$$\frac{I(X;V) + \log 2}{\log(m+1)} \tag{4.1}$$

is less than or equal to (say) 1/2.

Here is the idea behind the Le Cam equation. Let us find an ϵ_n that satisfies

$$n\epsilon_n^2 = \log N(\epsilon_n),$$

where $N(\epsilon_n)$ is the smallest number of balls of radius ϵ needed to cover our hypothesis space in the Hellinger distance sense. Then, if we can show that

$$I(X;V) \le n\epsilon_n^2,$$

by plugging this bound into (4.1), requiring that the resulting quantity be less than or equal to (say) 1/2, and rearranging (we also use the subadditivity of log), we get that we must have

$$\log(m+1) \ge 4n\epsilon_n^2 + 2\log 2 \tag{4.2}$$

in order to have a minimax lower bound; *i.e.*, we choose δ_n, where $m = m(2\delta_n)$, such that (4.2) holds in order to get a minimax lower bound.

4.2 Holder class of functions

Before we see some examples, let us make a few definitions.

Let $\mathcal{X} \subseteq \mathbf{R}^d$ be a closed and convex set.

Let $f : \mathcal{X} \to \mathbf{R}$.

Let D^k be the (higher order partial) differential operator, *i.e.*,

$$D^k = \frac{\partial^k}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}},$$

where $k = \sum_{i=1}^d k_i$.

Let $\beta \in (0, \infty)$ with $\lfloor \beta \rfloor$ denoting the largest integer (strictly) less than β.

Let

$$\|f\|_\beta = \max_{k \le \lfloor \beta \rfloor} \sup_{x,y \in \mathcal{X}} \left| \frac{D^k f(x) - D^k f(y)}{\|x - y\|_2^{\beta - \lfloor \beta \rfloor}} \right|.$$

Finally, let the Holder class of functions $\Sigma(\mathcal{X}, \beta, M)$ be

$$\{f : \mathcal{X} \to \mathbf{R} : \|f\|_\beta \le M\};$$

in words, this is the set of all functions (from $\mathcal{X}$ to $\mathbf{R}$) whose 1st through $\lfloor \beta \rfloor$ (inclusive) derivatives are Lipschitz continuous (when β is integral) with constant M.

4.3 Examples

4.3.1 Density estimation

Suppose $X_1, \ldots, X_n$ are drawn i.i.d. from some distribution $f \in \Sigma(\mathcal{X}, \beta, M)$, and we wish to estimate f. Assume there exist constants $c, C > 0$ s.t. $c \le f(x) \le C$ for all $x \in \mathcal{X}$ and all $f \in \Sigma(\mathcal{X}, \beta, M)$. Then if ϵ_n is such that $n\epsilon_n^2 = \log N(\epsilon_n)$ (*i.e.*, it satisfies the Le Cam equation), it turns out that ϵ_n^2 gives the minimax rate in ℓ_2.

4.3.2 Nonparametric regression

Suppose $y_i = f(x_i) + \epsilon_i$, where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ and $X_1, \ldots, X_n$ are deterministic, and $f \in \Sigma(\mathcal{X}, \beta\, M)$. It turns out that the Le Cam equation gives the minimax rate $n^{-\beta/(2\beta+d)}$, which is the classic rate.

4.3.3 Least squares

Here, we want to find

$$\hat{f} = \operatorname*{argmin}_{g \in \mathcal{F}} (1/n) \sum_{i=1}^n (y_i - g(x_i))^2 .$$

Let us consider the task of upper bounding

$$\mathbf{E}_{f^*}\left[(1/n)\|\hat{f}-f^*\|_2^2\right] = \mathbf{E}_{f^*}\left[(1/n)\sum_i \left(\hat{f}(x_i)-f^*(x_i)\right)^2\right].$$

Now, since $\hat{f}$ is optimal and feasible and f^* is (clearly) feasible, we have

$$(1/n)\|y-\hat{f}(X)\|_2^2 = (1/n)\sum_i \left(y_i-\hat{f}(x_i)\right)^2 \leq (1/n)\|y-f^*(X)\|_2^2, \tag{4.3}$$

where we written $y=(y_1,\ldots,y_n)$, $X=[x_1,\ldots,x_n]$. Similarly, letting $\epsilon=(\epsilon_1,\ldots,\epsilon_n)$, we write $y=f^*(X)+\epsilon$, plug this quantity into (4.3), and eventually get that

$$(1/n)\|\hat{f}-f^*\|_2^2 \leq (2/n)\sum_i \epsilon_i\left(\hat{f}(x_i)-f^*(x_i)\right) \rightsquigarrow \left((2\sigma)/\sqrt{n}\right)\sum_i w_i\left(\hat{f}(x_i)-f^*(x_i)\right)/\sqrt{n},$$

where $\rightsquigarrow$ denotes convergence in distribution and $w_1,\ldots,w_n$ are i.i.d. $\mathcal{N}(0,1)$.

Now, letting $G=\{(f-f')/\sqrt{n} : f, f' \in \mathcal{F}\}$, we have that this quantity is upper bounded by

$$\left((2\sigma)/\sqrt{n}\right)\sup_{g\in G}\sum_i w_i g(x_i),$$

which implies that

$$\mathbf{E}_{f^*}\left[(1/n)\|\hat{f}-f^*\|_2^2\right] \leq \left((2\sigma)/\sqrt{n}\right)\mathbf{E}_{f^*}\sup_{g\in G}\sum_i w_i g_i(x_i).$$

Now, let $g^{(1)},\ldots,g^{(N(\epsilon))}$ be an ϵ-cover of G. Then for all $g\in G$

$$\sum_i w_i g(x_i) = \sum_i w_i g^{(j)}(x_i) + \sum_i w_i\left(g(x_i)-g^{(j)}(x_i)\right),$$

where $g^{(j)}$ is the closest point to g amongst $g^{(1)},\ldots,g^{(N(\epsilon))}$.

This implies that

$$\sup_{g\in G}\sum_i w_i g(x_i) \leq \max_{j=1,\ldots N(\epsilon)}\sum_i w_i g^{(j)}(x_i).$$

By Jensen's inequality, we have that

$$\begin{aligned}
\mathbf{E}_{f^*}\sup_{g\in G}\sum_i w_i g(x_i) &\leq \mathbf{E}_{f^*}\max_j\sum_i w_i g^{(j)}(x_i) + \epsilon\sqrt{\mathbf{E}_{f^*}\sum_i w_i^2} \\
&\leq \mathbf{E}_{f^*}\left[\text{max of } N(\epsilon) \text{ normal r.v.'s}\right] + \epsilon\sqrt{n} \\
&\leq \sqrt{2\left(\max_j\sum_i g^{(j)}(x_i)\right)^2\log N(\epsilon)} + \epsilon\sqrt{n},
\end{aligned}$$

which has the form of a Le Cam equation.

Combining this result with the metric entropy for the Holder class of functions, *i.e.*,

$$\log N\left(\epsilon, \Sigma(\mathcal{X},\beta,M), \|\cdot\|_\infty\right) \asymp (1/\epsilon)^{d/\beta},$$

we finally get that

$$\mathbf{E}_{f^*}\left[(1/n)\|\|\hat{f}-f\|_2^2\right] = \left((2\sigma)/\sqrt{n}\right)\left(\sqrt{2\sigma^2 n^{1/3}} + n^{-1/3}\sqrt{n}\right) \asymp n^{-1/3}.$$