CS861: Theoretical Foundations of Machine Learning Lecture 13 - 10/04/2023
University of Wisconsin–Madison, Fall 2023

Lecture 13: ert lemma, Nonparametric regression

Lecturer: Kirthe *Scribed by: Elliot Pickens, Yuya Shimizu*

Disclaimer: *These n jected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

In this lecture, we provide further methods to derive a lower bound for the minimax risk. First, we will continue our previous discussion on constructing alternatives via `tight packings`. Then, we will introduce the `Varshamov-Gilbert lemma`, which is another method to construct well-separated alternatives. We also briefly mention `other methods for lower bounds`. Finally, we will discuss `nonprarametric regression`.

1 Method 1: Constructing alternatives via tight packings (continued)

In the previous lecture, we have learned ε-packing numbers. We will see ε-covering numbers that behave in the equivalent order to packing numbers as $\varepsilon \downarrow 0$.

Definition 1. *(Covering number, metric entropy)*

- *An ε-**covering** of set $\mathcal{X}$ with respect to a metric ρ is a set $\{x_1, \cdots, x_N\}$ such that for all $x \in \mathcal{X}$, there exists some $x_i \in \{x_1, \cdots, x_N\}$ s.t. $\rho(x, x_i) \leq \varepsilon$.*
- *The ε-**covering number** $N(\varepsilon, \mathcal{X}, \rho)$ is the size of the smallest covering.*
- *The **metric entropy** is $\log(N(\varepsilon, \mathcal{X}, \rho))$.*

We have the following lemma that relates covering numbers and packing numbers.

Lemma 1. *A covering number $N(\cdot, \mathcal{X}, \rho)$ and a packing number $M(\cdot, \mathcal{X}, \rho)$ satisfy*

$$M(2\varepsilon, \mathcal{X}, \rho) \leq N(\varepsilon, \mathcal{X}, \rho) \leq M(\varepsilon, \mathcal{X}, \rho).$$

Remark This lemma is useful since we can apply prior work on bounding the metric entropy $\log(N(\varepsilon, \mathcal{X}, \rho))$.

2 Method 2: Varshamov–Gilbert Lemma

To get a lower bound for the minimax risk, it is often convenient to consider alternatives indexed with a hypercube:

$$\left\{P_\omega; \omega = (\omega_1, \ldots, \omega_d) \in \{0,1\}^d\right\}.$$

Example 1. (Normal mean estimation in $\mathbb{R}^d$) Consider a hypercube:

$$\left\{N\left(\delta\omega, \sigma^2 I\right); \omega \in \{0,1\}^d\right\}.$$

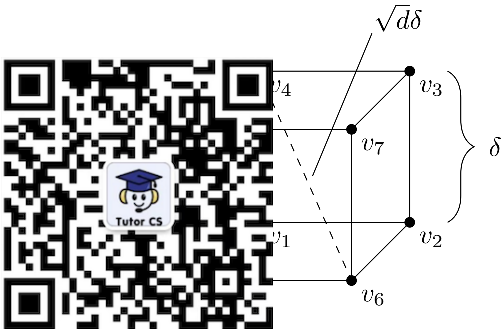

Figure 1: A hypercube that we will use to generate our alternatives by removing a few vertices from the cube.

For these alternatives, we can calculate the following values:

$$\begin{aligned}
\min_{\omega \neq \omega'} \rho\left(\theta\left(P_\omega\right), \theta\left(P_{\omega'}\right)\right) &= \min_{\omega \neq \omega'} \left\|\delta\omega - \delta\omega'\right\|_2 \\
&= \delta, \qquad (\omega \text{ and } \omega' \text{ differ on only one coordinate}) \\
\max_{\omega, \omega'} \mathrm{KL}\left(P_\omega, P_{\omega'}\right) &= \frac{\max_{\omega,\omega'} \left\|\delta\omega - \delta\omega'\right\|_2^2}{2\sigma^2} \\
&= \frac{d\delta^2}{2\sigma^2} \qquad (\omega \text{ and } \omega' \text{ differ on all coordinates})
\end{aligned}$$

The problem here is the Kullback-Leibler divergence could be large relative to the minimum distance, thus, we cannot simply apply the local Fano's method.

This example motivates us to introduce Varshamov-Gilbert Lemma. The Varshamov-Gilbert lemma states that we can find a *large* subset of $\{0,1\}^d$ such that the minimum distance between any two points in the subset is also *large*. Before stating the lemma, we define the Hamming distance.

Definition 2. *(Hamming distance) The hamming distance between two binary vectors* ω, ω' *is* $H\left(\omega, \omega'\right) = \sum_{j=1}^d \mathbf{1}\left\{\omega_j \neq \omega_j'\right\}$ *for* $\omega, \omega' \in \mathbb{R}^d$*. It counts the number of coordinate where* ω_j *and* ω_j' *differ.*

Lemma 2. *(Varshamov-Gilbert) Let* $m \geq 8$*. Then there exists* $\Omega_m \subseteq \{0,1\}^m$ *such that the followings are true: (i)* $\left|\Omega_m\right| \geq 2^{m/8}$*. (ii)* $\forall \omega, \omega' \in \Omega_m, H\left(\omega, \omega'\right) \geq m/8$*. We will call* Ω_m *the Varshamov-Gilbert pruned hypercube of* $\{0,1\}^m$*.*

We will revisit the normal mean estimation example to illustrate an application of the Varshamov-Gilbert lemma.

Example 2. (Normal mean estimation in $\mathbb{R}^d$) Let an i.i.d. data $S = \{x_1, \cdots, x_n\} \sim P^n$ where $P \in \mathcal{P}$ and $\mathcal{P} = \left\{N(\mu, \Sigma); \mu \in \mathbb{R}^d, \Sigma \preceq \sigma^2 I\right\}$. Consider $\Phi \circ \rho\left(\theta_1, \theta_2\right) = \left\|\theta_1 - \theta_2\right\|_2^2$. Let Ω_d be the Varshamov-Gilbert pruned hypercube of $\{0,1\}^d$. Define

$$\mathcal{P}' = \left\{N\left(\sqrt{\frac{8}{d}}\delta\omega, \sigma^2 I\right); \omega \in \Omega_d\right\}.$$

For these alternatives, we have the following bound:

$$
\begin{aligned}
{}_{P_\omega\neq P} \text{[illegible]} (P_{\omega'})) &= \min_{\omega\neq\omega'} \sqrt{\sum_{j=1}^{d}\left(\sqrt{\frac{8}{d}}\delta\omega_j - \sqrt{\frac{8}{d}}\delta\omega'_j\right)^2} \\
&= \sqrt{\frac{8}{d}}\delta \min_{\omega\neq\omega'} \sqrt{H\left(\omega,\omega'\right)} \\
&\geq \sqrt{\frac{8}{d}}\delta\sqrt{\frac{d}{8}} = \delta,
\end{aligned}
$$

where the inequality follows from the property (ii) of the Varshamov-Gilbert pruned hypercube. Since the maximum ℓ_2-distance over a hypercube is the length of a diagonal, we also have

$$
\max_{P_\omega, P_{\omega'}\in\mathcal{P}'} \mathrm{KL}\left(P_\omega, P_{\omega'}\right) = \frac{\left(\sqrt{d}\times\left(\sqrt{\frac{8}{d}}\delta\right)\right)^2}{2\sigma^2} = \frac{4\delta^2}{\sigma^2}.
$$

Choose $\delta = \sigma\sqrt{\frac{d\log(2)}{128n}}$. Then,

$$
\begin{aligned}
\max_{P_\omega, P_{\omega'}\in\mathcal{P}'} \mathrm{KL}\left(P_\omega, P_{\omega'}\right) &= \frac{4\delta^2}{\sigma^2} = \frac{d\log(2)}{32n} \\
&= \frac{\log\left(2^{d/8}\right)}{4n} \\
&\leq \frac{\log\left(|\mathcal{P}'|\right)}{4n},
\end{aligned}
$$

where the inequality follows from the property (i) of the Varshamov-Gilbert pruned hypercube: $|\mathcal{P}'| = |\Omega_d| \geq 2^{d/8}$. Therefore, by the local Fano's method (here we also require $d \geq 32$),

$$
R_n^* \geq \frac{1}{2}\Phi\left(\frac{\delta}{2}\right) = \frac{\log(2)}{1024}\cdot\frac{d\sigma^2}{n}.
$$

3 Other methods for lower bounds

Theorem 3. *Informal theorem (Ch 2, Tsyhakov)*

Let $\mathrm{S} = \{(\mathrm{x}_1,\mathrm{y}_1),\ldots,(\mathrm{x}_n,\mathrm{y}_n)\} \sim P \in \mathcal{P}$, $\{P_0,\ldots,P_N\} \subseteq P$, *and* $\delta = \max_{j\neq k}\Phi\circ\rho(\theta(P_j),\theta(P_k))$. *Then if*

$$
\frac{1}{N}\sum_{j=1}^{N} KL(P_j,P_0) \leq C_1\log(N)
$$

we can say that

$$
R_n^* \geq C_2\Phi\left(\frac{\delta}{2}\right).
$$

Roughly, this informal theorem says that if the average KL distance between $P_j \forall j$ and some "central" distribution P_0 is small enough we get a the lower bound of on the minimax risk seen above.

A related method, *Assouad's method*, can be found in chapter 9 of John Duchi's "Lecture Notes on Information Theory." This method applies when there is additional structure in the problem.

4 Nonparametric regression

The model: Let $\mathcal{F}$ be ipschitz functions in $[0, 1]$. That is

$$, B] \; ; \; |f(x_1) - f(x_2)| \leq L|x_1, x_2|\}$$

where L is the Lipsc

We observe some S } drawn i.i.d. from $P_{xy} = \mathcal{P}$ where

$$\begin{aligned} \mathcal{P} = \{P_{xy} \; ; \; 0 < \alpha_0 \leq p(x) \leq \alpha_1 < \infty \\ \text{the regression function } f(x) \triangleq \mathbb{E}[Y|X = x] \in \mathcal{F} \\ Var(Y|X = x) \leq \sigma^2\} \end{aligned}$$

We wish to estimate the regression function via the following loss

$$\ell(P_{xy}, y) = \int (f(x) - g(x))^2 p(x) dx$$

where f is the regression function and $g : [0, 1] \to \mathbb{R}$. The risk of $\hat{f}$ is

$$\begin{aligned} R(P_{xy}, \hat{f}) &= \mathbb{E}_S\left[\ell(P_{xy}, \hat{f})\right] \\ &= \mathbb{E}_S\left[\int (f(x) - \hat{f}(x))^2 p(x) dx\right] \end{aligned}$$

and the minimax risk of $\hat{f}$ is

$$R_n^* = \inf_{\hat{f}} \sup_{P_{xy} \in \mathcal{P}} R(P_{xy}, \hat{f}).$$

We want to show that the minimax risk is $\Theta\left(n^{-\frac{2}{3}}\right)$. We will do this in two steps:

1. Establish a lower bound with Fano's method
2. Get and upper bound using Nadaraya-Watson Estimation

4.1 Lower Bound

To begin it should be noted that we have a problem where the loss does cannot be written as $\ell = \Phi \circ \rho$. To circumvent this, denote $\mathcal{P}'' = \{P_{xy} \in \mathcal{P} \; ; \; p(x) = 1\}$.[1] Then

$$R_n^* = \inf_{\hat{f}} \sup_{P_{xy} \in \mathcal{P}''} \mathbb{E}_S\left[\underbrace{\int (f(x) - \hat{f}(x))^2 p(x) dx}_{\Phi \circ \rho}\right]$$

where $\Phi \circ \rho(f_1, f_2) = ||f_1 - f_2||_2^2$.

Now we proceed with proving the lower bound via the following three steps:

---

[1] We choose $p(x) = 1$ here for simplicity, but any uniform pdf will work.

1. Constructing the alternatives
2. Lower bounding ρ
3. Upper bounding I

4.1.1 Constructing

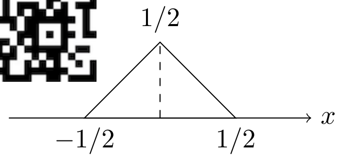

Figure 2: $\Psi(x)$

Let

$$\Psi(x) = \begin{cases} x + \frac{1}{2} & \text{if } x \in [-1/2, 0] \\ -x + \frac{1}{2} & \text{if } x \in [0, 1/2] \\ 0 & \text{o.w.} \end{cases}$$

where Ψ is 1−Lipschitz and $\int \Psi^2(x)dx = \frac{1}{12} < \infty$.

Now let $h > 0$ (we will choose it more precisely later), and let $m = \frac{1}{h}$. Define $\mathcal{F}'$

$$\mathcal{F}' = \left\{ f_w \; ; \; f_w(\cdot) = \sum_{j=1}^{m} w_j \phi_j(\cdot), \; w \in \Omega_m \right\}$$

where Ω_m is the Vashamov-Gilbert pruned subset of $\{0,1\}^d$. And ϕ_j as

$$\phi_j(x) = Lh\Psi\left(\frac{(x-(j-1/2)h)}{h}\right).$$

Figure 3: Depiction of ϕ_j and w when $w = \{0,0,1,0,1\}$.

Now we need to check that $\mathcal{F}' \subseteq \mathcal{F}$ (to show that f_w is L−Lipschitz). To do this, it is sufficient to check within one of the "bumps." By the chain rule,

$$|\phi_j'| = |Lh\Psi'\left(\frac{(x-(j-1/2)h)}{h}\right)\frac{1}{h}| = L.$$

Finally we define our set of alternatives $\mathcal{P}'$ as follows.

$$\mathcal{P}' = \Big\{\mathcal{P} \,;\,$$ $$\triangleq \mathbb{E}[Y|X=x] \in \mathcal{F}', \quad Y|X=x \sim \mathcal{N}(f(x),\sigma^2)\Big\}.$$

We see that $\mathcal{P}' \subseteq \mathcal{P}'' \subseteq$