CS861: Theoretical Foundations of Machine Learning Lecture 15 - 09/10/2023
University of Wisconsin–Madison, Fall 2023

Lecture 15: ...n, Lower bounds for prediction problems

Lecturer: Kirthe... *Scribed by: Haoyue Bai, Zexuan Sun*

In this lecture, we will first present a lower bound for **nonparametric density estimation**, and then study **kernel density estimation** to upper bound the risk. We will conclude with a framework for proving **lower bounds for prediction problems**.

1 Nonparametric Density Estimation (Cont'd)

1.1 Recap: Nonparametric Density Estimation

Consider the function space $\mathcal{F} : \{f : [0,1] \to [0,B], |f(x_1) - f(x_2)| \leq L|x_1 - x_2|\}$. The constraint ensures the function is Lipschitz continuous with Lipschitz constant L. We are interested in the family of distributions $\mathcal{P}$, which consists of distributions whose densities are L–Lipschitz. That is, $\mathcal{P} : \{p : \text{density } p \text{ of } P \text{ is in } \mathcal{F}\}$.
We observe samples $S = \{X_1, ..., X_n\} \overset{\text{iid}}{\sim} p \in \mathcal{P}$, where the set S represents a random sample of n data points that are independent and identically distributed (iid) according to some unknown density p that belongs to $\mathcal{P}$.
Given the sample S, we wish to estimate the density p of $\mathcal{P}$ in the L_2 loss. That is:

$$\Phi \circ \rho(p_1, p_2) = \int (p_1(t) - p_2(t))^2 \, dt$$

We will show that the minimax risk satisfies:

$$R_n^* = \inf_{\hat{p}} \sup_{p \in \mathcal{F}} \mathbb{E}_S[||\hat{p} - p||_2^2] \in \Theta(n^{-\frac{2}{3}}).$$

1.2 Lower bound

We will first prove a lower bound via Fano's method.
Step 1: Construct alternatives
Consider the function ψ illustrated in Figure 1. The following facts are straightforward to verify.

- ψ is 1-Lipschitz, meaning that for any two inputs λ_1 and λ_2: $|\psi(\lambda_1) - \psi(\lambda_2)| \leq |\lambda_1 - \lambda_2|$;
- $\int \psi = 0$, which indicates that the area under the curve of the function, over its entire domain, sums up to zero;
- $-\frac{1}{4} \leq \psi \leq \frac{1}{4}$, which gives the range of the function;
- $\int \psi^2 = \frac{1}{48}$, which the squared integral of ψ.

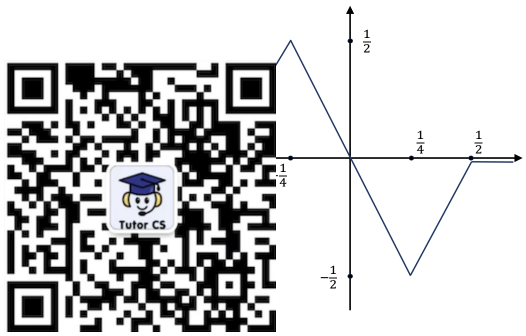

Figure 1: Illustrative figure for the function $\psi(\lambda)$.

To construct the alternatives let h is a positive number ($h > 0$) that we will decide later. Let $m = \frac{1}{n}$. The alternative function space $\mathcal{F}^{'}$ is:

$$\mathcal{F}^{'} = \left\{ P_\omega : P_\omega(\cdot) = 1 + \sum_{i=1}^{m} \omega_j \Phi_j(\cdot) \quad \omega \in \Omega_m \right\}$$

This space defines a set of functions P_ω that are formed by a linear combination of basis function $\Phi_j(t)$. The vector ω resides in some set Ω_m. The basic function is defined as:

$$\Phi_j(t) = Lh\Phi\left(\frac{t - (j - \frac{1}{2}) \cdot h}{h}\right),$$

where L denotes the Lipschitz constant, and h is the bandwidth.

The illustrative example in Figure 2 would provide a visual representation of the approximation.

Figure 2: Illustrative figure for the example p_ω.

Step 2: Lower bound the distance $\rho(p_\omega, p_{\omega'})$
The objective of this step is to determine a lower bound for the difference between p_ω and $p_{\omega'}$. We can bound the density below:

$$
\begin{aligned}
& \int^1 (p_\omega - p_{\omega'})^2 \\
& \int_{\frac{j-1}{m}}^{\frac{j}{m}} (1 + \omega_j \Phi_j - (1 + \omega_j^{'} \Phi_j))^2 \\
& \mathbb{I}(\omega_j \neq \omega_j^{'}) \int_{\frac{j-1}{m}}^{\frac{j}{m}} \Phi_j^2 \\
&= \frac{H(\omega, \omega^{'}) L^2 h^3}{48}
\end{aligned}
$$

By the Varshamov-Gilbert lemma, we have $H(\omega, \omega^{'}) \geq \frac{m}{8} = \frac{1}{8h}$ as,

$$\min_{\omega, \omega^{'}} ||p_\omega - p_{\omega'}|| = \sqrt{\frac{L^2 h^3}{48}}$$

$$\min_{\omega, \omega^{'}} \sqrt{H(\omega, \omega^{'})} \geq \frac{Lh}{8\sqrt{6}} \triangleq \delta$$

Step 3: Upper bound KL

In this step, the goal is to determine an upper bound for the Kullback-Leibler (KL) divergence between two functions, p_ω and $p_{\omega'}$. Based on the definition of KL divergence and expanding KL for p_ω and $p_{\omega'}$:

$$
\begin{aligned}
KL(p_\omega, p_{\omega'}) &= \int_0^1 p_\omega \log(\frac{p_\omega}{p_{\omega'}}) \\
&= \sum_{j=1}^{m} (\int_{\frac{j-1}{m}}^{\frac{j}{m}} (1 + \omega_j \Phi_j) \log\left(\frac{1 + \omega_j \Phi_j}{1 + \omega_j^{'} \Phi_j}) \mathbb{I}(\omega_j \neq \omega_j^{'}\right)
\end{aligned}
$$

After some algebra (you will do this in the homework), we have the following upperbound:

$$KL(p_\omega, p_{\omega'}) \leq H(\omega, \omega^{'}) \frac{L^2 h^3}{48} \Rightarrow \forall \omega, \omega^{'}, KL(p_\omega, p_{\omega'}) \leq \frac{L^2 h^2}{48}$$

This suggests that regardless of the particular values of ω and $\omega^{'}$, the KL divergence between any two functions p_ω and $p_{\omega'}$ from the considered set is always less than or equal to $\frac{L^2 h^2}{48}$.

A formal proof of this statement is left as an exercise in a homework assignment.

Step 4: Apply local Fano

Applying local Fano's inequality in this step, we derive conditions and constraints for the estimation problem. We want $\max_{\omega, \omega^{'}} KL(p_\omega, p_{\omega'}) \leq \frac{\log(\omega)}{4n}$. This relation is upper bound the maximum KL divergence between any two functions in the set by a term that diminishes with increasing sample size n.

Sufficient if we have the equation $\frac{L^2 h^2}{48} \leq \frac{\log(2^{\frac{m}{8}})}{4n} = \frac{\log(2)}{32nh}$. Choose $h = C, \frac{1}{n^{\frac{1}{3}} L^{\frac{2}{3}}}$, which determines the choice of h as a function of n and L. Then, we have the equation:

$$R_n^* \geq \frac{1}{2} \Phi(\frac{Lh}{2 \times 8\sqrt{6}}) = C \frac{L^{\frac{2}{3}}}{n^{\frac{2}{3}}}.$$

This offers a lower bound on the risk, R_n^*, which quantifies the error in the estimation.

The given requirements ensure the validity and applicability of the above relations:

- $m \geq \frac{1}{h} \geq 8$. This is necessary for the Varshamov-Gilbert lemma to be applicable.

- Cardinality of $\mathcal{F}^{'}$: $|\mathcal{F}^{'}| \geq 16 \Leftarrow 2^{\frac{m}{8}} \geq 16 \Leftarrow h \leq \frac{1}{32} \Rightarrow$ satisfied if $n \geq \frac{c}{L^2}$. This ensures a sufficient number of observations given the Lipschitz constant L.

- KL bounding cond ... nposes the KL divergence between functions remains bounded, and ties the band ... z constant L.

2 Upper Bou ... Density Estimation

Kernel Density Estimat ... metric technique to estimate the probability density function of a continuous random ... Density Estimation (KDE) has the following form:

$$\hat{p}(t) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{t - x_i}{h}\right),$$

where $\hat{p}(t)$ is the estimated density at point t, n is the number of data points, and x_i are the observed data points. h is bandwidth parameters, which plays a critical role in KDE. K is a (smoothing) kernel with the following properties:

(1) Normalization:

$$\int K(u)du = 1,$$

This ensures the result will integrate to 1 over its entire domain, maintaining the fundamental property of a probability density function.

(2) Symmetry:

$$K(u) = K(-u),$$

This property ensures that the kernel is symmetric around zero. As a result, the estimated density will not be biased towards any direction from the point of estimation.

For the problem at hand, the kernel selected is $K(t) = \mathbb{I}(|t| \leq \frac{1}{2})$, which is sufficient for Lipschitz functions. This is a simple uniform kernel.

We can bound the risk as follows:

$$\begin{aligned}
\mathbb{E}[||p - \hat{p}||_2^2] &= \mathbb{E}[\int (p - \hat{p})^2 \, dt] \\
&= \mathbb{E}[\int (p - \mathbb{E}(\hat{p}))^2 \, dt + \int (\mathbb{E}(\hat{p}) - \hat{p})^2 \, dt + 2\int (p - \mathbb{E}(\hat{p}))(\mathbb{E}(\hat{p}) - \hat{p}) \, dt] \\
&= \int_0^1 \underbrace{(p(t) - \mathbb{E}(\hat{p}(t)))^2}_{\text{bias}(t)} dt + \int_0^1 \underbrace{\mathbb{E}[(\hat{p}(t) - \mathbb{E}(\hat{p}(t)))^2]}_{\text{Var}(t)} dt + 2\int_0^1 (p - \mathbb{E}(\hat{p})) \underbrace{\mathbb{E}[\mathbb{E}(\hat{p}) - \hat{p}]}_{=0} dt
\end{aligned}$$

The bias and variance terms can be written and bounded below. The bias of an estimator indicates how far on average the estimate is from the true value. In our context, the bias term is derived from:

$$
\begin{aligned}
\text{bia}\ldots &= \mathbb{E}[\hat{p}(t) \ldots p(t)] \\
&\ldots K \ldots)] - p(t) \quad (\text{Apply } \mathbb{E}[\frac{1}{n}\sum_i Z_i] = \mathbb{E}[Z_i]) \\
&\ldots x)\, dx - p(t) \\
&\ldots)\, du - p(t) \int K(u)\, du \\
&= \int K(u)\,(p(t+uh) - p(t))\, du
\end{aligned}
$$

The bias depends on the choice of kernel and bandwidth h. Then we can bound the bias term:

$$
\begin{aligned}
|\text{bias}(t)| &\leq \int K(u)|p(t+uh) - p(t)|\, du \\
&\leq \int K(u) L |uh|\, du \\
&= Lh \int K(u)|u|\, du \\
&= C_1 L h
\end{aligned}
$$

where $C_1 = \int K(u)|u|\, du$ is a constant that essentially depends on the kernel, and L is the Lipschitz constant, signifying the bound on the rate of change of our function.

Variance quantifies the dispersion of an estimator around its expected value. In the context of KDE, variance arises from the randomness in the sample. For our setup, the variance term is captured by:

$$
\begin{aligned}
\text{Var}(t) &= \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K(\frac{t - X_i}{h})\right) \\
&= \frac{1}{n}\text{Var}_{X\sim P}\left(\frac{1}{h}K(\frac{t-X}{h})\right) \quad (\text{Apply Var}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right) = \frac{1}{n}\text{Var}(Z_1)) \\
&\leq \frac{1}{n}\mathbb{E}_{X\sim P}[\frac{1}{h^2}K^2(\frac{t-X}{h})] \quad (\text{Var}(Z) = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \leq \mathbb{E}[Z^2]) \\
&= \frac{1}{nh^2}\int K^2(\frac{x-t}{h})p(x)dx \\
&= \frac{1}{nh}\int K^2(u)p(t+uh)du \\
&\leq \frac{B}{nh}\int K^2(u)du = \frac{B}{nh}
\end{aligned}
$$

Where B is a constant that bounds the product of the squared kernel and the true density.
Combine these bounds for bias and variance terms, we have

$$
\begin{aligned}
\mathbb{E}_S[\|\rho - \hat{\rho}\|_2^2] &\leq \int_0^1 \text{bias}^2(t)dt + \int_0^1 \text{Var}(t)dt \\
&\leq C_1^2 L^2 h^2 + \frac{B}{nh} \\
&= (B + C_1^2)\frac{L^{2/3}}{n^{2/3}} \quad (\text{if } h = \frac{1}{n^{1/3}L^{2/3}})
\end{aligned}
$$

3 Lower Bounds for Prediction Problems

So far, we have talked a[...]e metrics, for instance

$$\ldots - \theta_2|, \theta(P) \in \mathbb{R}$$
$$\ldots - \theta_2\|_{L_2}, \theta(P) \in \{f : \mathcal{X} \to \mathbb{R}\}$$

Next, we will develop a[...] lower bounds for prediction problems (among others). The framework is comprised[...]nents:

1. Data space $\mathcal{Z}$. Th[...]ich data samples arise.
2. A family of distribution $\mathcal{P}$, where $\forall P \in \mathcal{P}, \text{supp}(P) \subseteq \mathcal{Z}$. This represents a collection of probability distributions from which the data can be drawn.
3. A hypothesis/parameter space $\mathcal{H}$. This contains all the potential hypotheses or models that we might use to make predictions.
4. An "instance loss", $f : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$, where $f(h, Z)$ is the loss of hypothesis h on instance Z. This measures how well a particular hypothesis h from $\mathcal{H}$ performs on an instance Z from $\mathcal{Z}$.
5. The *population loss*, $F(h, P) = \mathbb{E}_{Z\sim P}[f(h, Z)]$; *Excess population loss*, $L(h, P) = F(h, P) - \inf_{h'\in\mathcal{H}} F(h', P)$. These terms indicate how well our hypothesis does on average (under distribution P) and how it compares to the best possible hypothesis in H, respectively.
6. A dataset S is drawn from some $P \in \mathcal{P}$.
7. An *estimator* $\hat{h}$, which maps the dataset S to a hypothesis in $\mathcal{H}$. Note that we overload notation here for $\hat{h}$

$$\hat{h} : \text{a map form data to } \mathcal{H} \quad \text{(estimator)}$$
$$\hat{h} : \text{as the estimate} \quad (\hat{h} \in \mathcal{H})$$

8. Risk of estimator $\hat{h}$

$$\begin{aligned} R(\hat{h}, P) &= \mathbb{E}[L(\hat{h}(S), P)] \\ &= \mathbb{E}[F(\hat{h}(S), P)] - \inf_{h\in\mathcal{H}} F(h, P) \end{aligned}$$

9. Minimax risk:

$$R^\star = \inf_{\hat{h}} \sup_{P\in\mathcal{P}} R(\hat{h}, P)$$

Next, let us see an example.

Example 1. Excess risk in classification/regression

- Data space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $\mathcal{H} : \{h : \mathcal{X} \to \mathcal{Y}\}$. This is the set of all possible prediction functions mapping from features $\mathcal{X}$ to outcomes Y.
- Instance loss (in classification), $f(h, (X, Y)) = \mathbb{1}(h(X) \neq Y)$. This measures the discrepancy between a predicted and actual class label.
- $F(h, P) = \mathbb{E}_{X,Y\sim P}[\mathbb{1}(h(X) \neq Y)]$ is the "risk". This is the expected value of the instance loss over the joint distribution of $\mathcal{X}$ and $\mathcal{Y}$, and can be understood as the overall error rate of the classifier.

- $L(h, P) = F(h, P) - F(h^\star, P)$, where $h^\star$ is the Bayes optimal classifier, i.e. $h^\star = \arg\max_{y \in \mathcal{Y}} \mathbb{P}(Y = y|X = x)$, and $L(h, P)$ is the "excess risk" in classification. The excess risk quantifies how much worse our classifier h pe[illegible] Bayes optimal classifier $h^\star$.
- In regression we d[illegible] $(X) - Y)^2$. The typical loss function used is the squared loss.
- $L(h, P) = \mathbb{E}_{X \sim P}[$[illegible]$(h^\star(X) - Y)^2]$, where $h^\star(x) = \mathbb{E}[Y|X = x]$, $L(h, P)$ is the "excess risk" in re[illegible] measures how much worse our regression function performs compared to the [illegible]ion $h^*(x)$ which is the conditional expectation.
- Note that this is [illegible] $= \|h_1 - h_2\|_2^2$, which captures the squared difference between two hypotheses h_1 and h_2, and $\mathbb{E}_S[\Phi \circ \rho(\hat{h}, h^\star)]$ measures how much estimator $\hat{h}$ deviates from the optimal hypothesis h^*.