CS861: Theoretical Foundations of Machine Learning Lecture 6 - 09/16/2024

University of Wisconsin–Madison, Fall 2023

Lecture 06: Proof of Sauer's lemma, Statistical Lower Bound

Lecturer: Kirthevasan Kandasamy *Scribed by: Jiaqi Tang*

In this lecture, we will first prove Sauer's Lemma. We will then start studying statistical lower bounds for point estimation by introducing its general framework.

1 Proof of Sauer's Lemma

Recall the growth function $g(n, \mathcal{H}) = \max_{|A|=n} |\mathcal{H}(A)|$, and the Sauer's lemma

Lemma 1 (Sauer's Lemma)**.** *Define* $\Phi_d(n) \coloneqq \sum_{i=0}^{d} \binom{n}{i}$*. If the VC-dimension of a hypothesis class* $\mathcal{H}$ *is* d*, then*

$$g(n, \mathcal{H}) \leq \Phi_d(n)$$

We will now provide a proof for Sauer's lemma.

The setup for our proof of Sauer's lemma will be via induction on $k = n + d$; where n is the number of i.i.d samples and d is the VC-dimension of our hypothesis class.

1. Base case: Show Sauer's lemma to be true for
   (a) $n = 0$ and $\forall d$
   (b) $d = 0$ and $\forall n$
2. Inductive case: Assume Sauer's lemma holds for all $n + d \leq k - 1$. Show Sauer's lemma to be true for $\forall n, d$ such that $n + d = k$. See Figure 1 for a visual demo of the induction strategy.

Step 1: Base Case

(a) let $n = 0$, and consider some $\mathcal{H}$ with VC dimension d. Note that $g(0, \mathcal{H}) = 1$ as we can label 0 points only one way. (Specifically speaking, if $A = \{\}, \mathcal{H}(A) = \{[\,]; h \in \mathcal{H}\} = \{[\,]\}$ and thus, $|\mathcal{H}(\{\})| = 1$.)

$$\Phi_d(n) = \sum_{i=0}^{d} \binom{n}{i} = \sum_{i=0}^{d} \binom{0}{i} = \binom{0}{0} + \sum_{i=1}^{d} \binom{0}{i} = 1 + 0 = 1 = g(0, \mathcal{H})$$

Therefore, Sauer's lemma is satisfied for $n = 0$.

(b) Let $\mathcal{H}$ be a hypothesis class with VC dimension $d = 0$. The hypothesis class cannot shatter a set of size 1 and all classifiers in $\mathcal{H}$ must generate the same label.

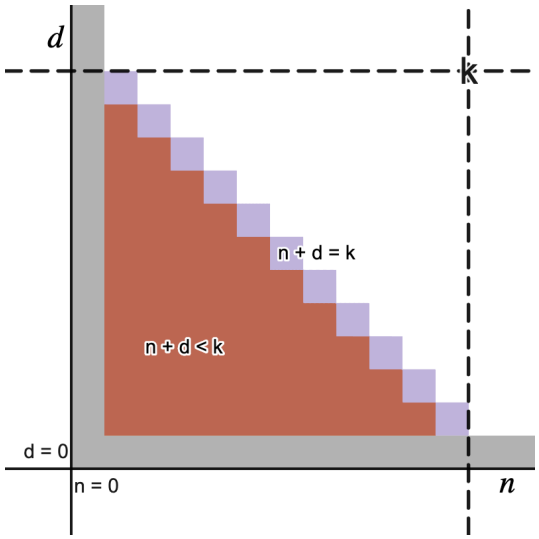

Figure 1: Visual demo of the proof by induction. The axes are n and d. The gray region represents the base case for n and d (where $n = 0$ and $d = 0$). The brownish region represents the induction hypothesis (where $n + d \leq k$). The purple region represents the inductive step (where $n + d = k$).

Then, for any dataset $A \subseteq \mathcal{X}$, you can label them exactly one way, i.e. $g(n, \mathcal{H}) = 1$.

$$\Phi_d(n) = \sum_{i=0}^{0} \binom{n}{i} = \binom{n}{0} = 1 = g(n, \mathcal{H})$$

The lemma is satisfied for $d = 0$.

Step 2: Inductive Case

Assume that Sauer's lemma holds true for all d, n such that $d + n \leq k - 1$. Let d, n be such that $d + n = k$.

Given the VC dimension d, consider any $A = \{x_1, \ldots, x_n\}$, i.e. $A \subseteq \mathcal{X}$ such that $|A| = n$. Let $\mathcal{G} \subseteq \{g : A \to y\}$ be a class of functions, defined only over A as follows: for each element in $\mathcal{H}(A)$, add one element from $\mathcal{H}$ to $\mathcal{G}$. In more detail, for each label $\{y_1, ..., y_n\} \in \mathcal{H}(A)$, there exists $h \in \mathcal{H}$ such that $\{y_1, ..., y_n\} = \{h(x_1), ..., h(x_n)\}$ and we add one such h to $\mathcal{G}$. Therefore, $\mathcal{G}$ will have exactly one function that generates each labeling in $\mathcal{H}(S^X)$. It follows that

$$|\mathcal{G}| = |\mathcal{G}(A)| = |\mathcal{H}(A)|$$

Next, we will partition $\mathcal{G}$ into the sets $\mathcal{G}_1$ and $\mathcal{G}_2$ as follows:

1. $\underline{\mathcal{G}_1}$: For every possible labeling of $\{x_1, ..., x_{n-1}\}$, add one function from $\mathcal{G}$ to $\mathcal{G}_1$.
2. $\underline{\mathcal{G}_2}$: Let $\mathcal{G}_2 = \mathcal{G} \setminus \mathcal{G}_1$.

Remarks: There could be 2 hypothesis in $\mathcal{G}$ which achieve the same labeling of $\{x_1, \ldots, _{n-1}\}$. Furthermore, $\forall g_2 \in \mathcal{G}_2$, there exists $g_1 \in \mathcal{G}_1$ such that $g_1(x_i) = g_2(x_i)$ for $i \in \{1, \ldots, n-1\}$ and $g_1(x_n) \neq g_2(x_n)$.

By our construction, $\mathcal{H}(A) = \mathcal{G}(A) = \mathcal{G}_1(A) \cup \mathcal{G}_2(A)$. Hence,

$$|\mathcal{H}(A)| = |\mathcal{G}(A)| = |\mathcal{G}_1(A)| + |\mathcal{G}_2(A)| \tag{1}$$

We also have

$$d_{\mathcal{G}_1} \underbrace{\leq}_{\mathcal{G}_1 \subseteq \mathcal{G}} d_{\mathcal{G}} \underbrace{\leq}_{\mathcal{G} \subseteq \mathcal{H}} d_{\mathcal{H}} = d \tag{2}$$

Claim 1. $|\mathcal{G}_1(A)| = |\mathcal{G}_1(\{x_1, ..., x_{n-1}\})|$

Proof For every labeling $\{g(x_1), ..., g(x_{n-1})\}$ of $\{x_1, \ldots, x_{n-1}\}$ in $\mathcal{G}_1(\{x_1, ..., x_{n-1}\})$, we have exactly one of either $\{g(x_1), ..., g(x_{n-1}), 0\}$ or $\{g(x_1), ..., g(x_{n-1}), 1\}$ in $\mathcal{G}_1(A)$. □

Claim 2. $|\mathcal{G}_2(A)| = |\mathcal{G}_2(\{x_1, ..., x_{n-1}\})|$

Proof Via identical argument. □

Now, we will apply the inductive hypothesis.

First, note that (2) indicates that $d_{\mathcal{G}_1} + n - 1 \leq d + n - 1 - 1$, so we could apply the assumption on $(d_{\mathcal{G}_1}, n-1)$. By Claim (1), we have

$$\begin{aligned}
|\mathcal{G}_1(A)| &= |\mathcal{G}_1(\{x_1, ..., x_{n-1}\})| && \text{Definition of growth function} \\
&\leq g(n-1, \mathcal{G}_1) && \text{Inductive Hypothesis} \\
&\leq \Phi_{d_{\mathcal{G}_1}}(n-1) && \Phi_d \text{ increases with } d \\
&\leq \Phi_d(n-1)
\end{aligned}$$

Next, note that for every $g_2 \in \mathcal{G}_2, \exists g_1 \in \mathcal{G}_1 \subseteq G$ which disagrees g_2 only on x_n. Therefore, if $T \subseteq \{x_1, ..., x_{n-1}\}$ is shattered by $\mathcal{G}_2$, $T \cup \{x_n\}$ must be shattered by $\mathcal{G}$. Because $|T \cup \{x_n\}| \leq d_{\mathcal{G}}$, $|T| \leq d_{\mathcal{G}} - 1 \leq d_{\mathcal{H}} - 1 = d - 1$. Hence, $d_{\mathcal{G}_2} \leq d - 1$. Therefore, by Claim (2),

$$\begin{aligned}
|\mathcal{G}_2(A)| &= |\mathcal{G}_2(\{x_1, ..., x_{n-1}\})| && \text{Definition of growth function} \\
&\leq g(n-1, \mathcal{G}_2) && \text{Inductive Hypothesis} \\
&\leq \Phi_{d_{\mathcal{G}_2}}(n-1) && \Phi_d \text{ increases with } d \\
&\leq \Phi_{d-1}(n-1)
\end{aligned}$$

Combining the above results, we have

$$\begin{aligned}
|\mathcal{H}(A)| &= |\mathcal{G}_1(A)| + |\mathcal{G}_2(A)| \\
&\leq \Phi_d(n-1) + \Phi_{d-1}(n-1) \\
&= \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\
&= \binom{n-1}{0} + \sum_{i=1}^{d} \binom{n-1}{i} + \sum_{i=1}^{d} \binom{n-1}{i-1} \\
&= \binom{n}{0} + \sum_{i=1}^{d} \left(\binom{n-1}{i} + \binom{n-1}{i-1} \right) \quad \cdots \text{using } \binom{n}{i} = \binom{n-1}{i} + \binom{n-1}{i-1} \\
&= \binom{n}{0} + \sum_{i=1}^{d} \binom{n}{i} \\
&= \sum_{i=0}^{d} \binom{n}{i} \\
&= \Phi_d(n)
\end{aligned}$$

$A \subseteq \mathcal{X}$ is arbitrary, so $g(n, \mathcal{H}) = \max_{|A|=n} |\mathcal{H}(A)| \leq \Phi_d(n)$. Therefore, Sauer's lemma holds true in the inductive case.

2 Statistical Lower Bound

Statistical lower bound is motivated by the following questions: How hard is a learning/estimation problem? Is our learning optimal? While we will utilize lower bound for regression, classification, density estimation, online leaning, bandit, etc, we will start with point estimation.

2.1 Point Estimation

We are interested in estimating a single parameter of a distribution (e.g mean of the distribution) using data drawn from that distribution.

A point estimation problem consists of **the following components**:

1. A family of distributions $\mathcal{P}$.
2. A data set $S = \{x_1, \ldots, x_n\}$ drawn i.i.d for a distribution $P \in \mathcal{P}$.
3. A parameter of interest $\theta = \theta(P) \in \mathbb{R}$.
4. An estimator for the parameter based on the drawn data set S: $\hat{\theta} = \hat{\theta}(S) \in \mathbb{R}$.
5. A loss function $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$.
6. Risk $R(P, \hat{\theta}) = \mathbb{E}_{S \sim P}\left[\ell(\theta(P), \hat{\theta}(S))\right]$.

Example 1 (Normal Mean Estimation)**.**

1. Distribution family $\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}\}$, where σ^2 is known.
2. Data $S = \{x_1, x_2, \ldots, x_n\}$ i.i.d from P, $P \in \mathcal{P}$.
3. Parameter $\theta = \theta(P) = \mathbb{E}_{X \sim P}[X]$, $P \in \mathcal{P}$
4. Loss function: $\ell(\theta_1, \theta_2) = (\theta_1 - \theta_2)^2$
5. Risk: $R(P, \hat{\theta}) = \mathbb{E}_{S \sim P}\left[(\theta(P) - \hat{\theta}(S))^2\right]$

In HW0, we saw two possible estimators:

$$\hat{\theta}_1(S) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$R(\theta, \hat{\theta}_1) = \mathbb{E}_{S \sim P}\left[(\theta - \hat{\theta}_1(S))^2\right] = \frac{\sigma^2}{n}$$

and

$$\hat{\theta}_2(S) = \frac{\alpha}{n}\sum_{i=1}^{n} x_i, \alpha \in [0, 1)$$

$$R(\theta, \hat{\theta}_2) = \mathbb{E}_{S \sim P}\left[(\theta - \hat{\theta}_2(S))^2\right] = \theta^2(1-\alpha)^2 + \frac{\alpha^2\sigma^2}{n}$$

(In both cases, when we take the expectation, θ is fixed, and $\hat{\theta}$ is the random variable since $\hat{\theta} = \hat{\theta}(S)$, and the expectation is w.r.t data S)

Notice that $R(\hat{\theta}_1, \theta) > R(\hat{\theta}_2, \theta)$ for small θ, while $R(\hat{\theta}_2, \theta) > R(\hat{\theta}_1, \theta)$ for large θ, which brings difficulty to find a optimal estimator $\hat{\theta}$ that estimate well for an arbitrary parameter θ.

Extending this example, we see that the estimator $\hat{\theta} = \mu$ for $\mu \in \mathbb{R}$ will achieve 0 risk when $\theta = \mu$ but will perform poorly elsewhere. This illustrates that we cannot find a uniformly good estimator $\widehat{\theta}$ which minimizes $R(\theta, \theta)$ for all θ (i.e. $\forall \theta, R(\theta, \hat{\theta}) \leq R(\theta, \hat{\theta}')$ for any estimator $\hat{\theta}'$).

Hence, it is customary to resort to other versions of optimality. The following to notions are common in the literature:

1. **Minimax Optimality.** $\hat{\theta}$ minimizes the maximum risk over the class of distribution $\mathcal{P}$, (i.e $\hat{\theta}$ minimizes $\sup_{P \in \mathcal{P}} R(\theta(P), \hat{\theta})$).
2. **Average Risk Optimality.** Average (Bayesian) Risk optimality.