CMPUT 654 Fa 23: Theoretical Foundations of Machine Learning **Fall 2023**

Lecture 7: September 26

Lecturer: Csaba Szepesvári *Scribes: Kushagra Chandak*

7.1 Outline

- Recap of lower bracketing cover and using it in Chernoff bounds.
- Linear threshold class example.
- Bounded variance class.

7.2 Recap

In the last lecture, we started talking about infinite function classes and the fact that we just need to account for a finite *cover* of the infinite function class in the union bound to get uniform deviation bounds. The covering happens at some accuracy or *scale* ε, and there is a tradeoff between the approximation error introduced by ε and the number of elements in the cover.

We also talked about the *lower bracketing cover* which is useful in obtaining one-sided uniform deviation bounds. Recall the setting for defining a lower bracketing cover: Let $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ and $P \in \mathcal{M}_1(\mathcal{Z})$.

Definition 7.1 (Lower bracketing cover)**.** Fix $\varepsilon > 0$. Then $g_1, \dots, g_m : \mathcal{Z} \to \mathbb{R}$ is a lower bracketing cover of $\mathcal{G}$ with distribution P and scale ε (shorthand: $\mathcal{G}@P@\varepsilon$) such that for any $g \in \mathcal{G}$ there exists $j \in [m]$ such that: (1) $g_j \le g$; and (2) $Pg \le Pg_j + \varepsilon$.

The minimum number of functions in the lower bracketing cover is called the lower bracketing number denoted by $N_\varepsilon = N_{\mathrm{LB}}(\varepsilon, \mathcal{G}, P)$. Note that the cover $g_1, \dots, g_m$ may or may not be in $\mathcal{G}$. A lower bracketing cover can be used for one-sided uniform deviations, which we used to analyze ERM.

Further, recall that the empirical distribution is given by $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{z_i}$ where $z_1, \dots, z_n \sim P$ iid. Then the ERM is

$$\hat{g}_n = \arg\min_{g \in \mathcal{G}} P_n g \,,$$

which implicitly selects a predictor that underlies the loss. Using Chernoff's inequality, we get the following bound for ERM.

Proposition 7.2. *Let $\mathcal{G} \subseteq [0,1]^{\mathcal{Z}}$. For every $\delta \in (0,1)$:*

1. *w.p. $1-\delta$, $P\hat{g}_n \le \inf_{g\in\mathcal{G}} Pg + \inf_{\varepsilon>0}\left(\varepsilon + 2\sqrt{\frac{\ln(N_\varepsilon+1)/\delta}{2n}}\right)$.*
2. *for every $\varepsilon > 0$, w.p. $1-\delta$, $P\hat{g}_n \le \inf_{g\in\mathcal{G}}\left(Pg + \sqrt{\frac{2Pg\ln(N_\varepsilon+1)/\delta}{n}}\right) + \varepsilon + \sqrt{\frac{2P\hat{g}_n\ln(N_\varepsilon+1)/\delta}{n}} + \frac{\ln(N_\varepsilon+1)/\delta}{3n}$.*

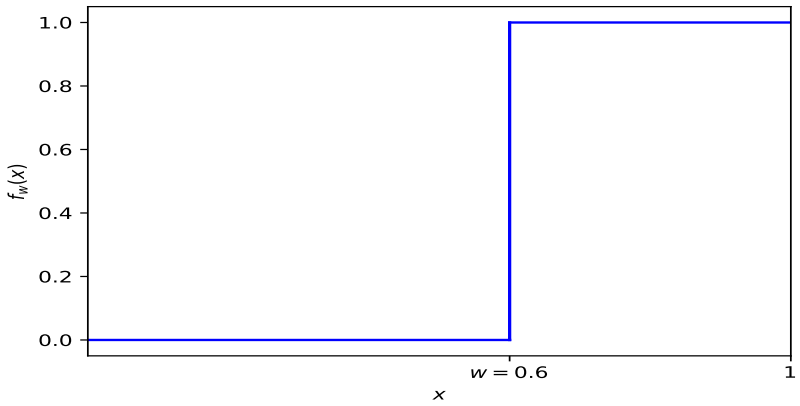

Figure 7.1: An example of a linear threshold function $f_w(x) = \mathbb{I}(x \geq w)$ with $w = 0.6$.

Remark 7.3.
1. For the second bound above, we did not take $\inf$ over $\varepsilon > 0$ in the RHS, since $P\hat{g}_n$ is random on the RHS. But we can first solve for $\inf$ over $\mathcal{G}$, solve a quadratic to get a bound for $P\hat{g}_n$, and then take the $\inf$ over ε.
2. If $g_1, \ldots, g_m$ were chosen from the class $\mathcal{G}$, then the size of the cover might increase slightly and we might lose some constant factors in the bounds.

7.3 Covering the Linear Threshold Class

Setting. Let $\mathcal{X} = [0, 1]$ and $\mathcal{Y} \in \{0, 1\}$. Let the function class be $\mathcal{F} = \{f_w : \mathcal{X} \to \mathcal{Y} : w \in [0, 1]\}$ where $f_w(x) = \mathbb{I}(x \geq w)$ is the *linear threshold function* (Fig. 7.1). We consider the zero-one loss where the loss class is defined as $\mathcal{G} = \{\ell_{01} \circ f_w : w \in [0, 1]\}$. The loss function is a map $\ell_{01} \circ f_w : \mathcal{X} \times \{0, 1\} \to \{0, 1\}$ defined as $(x, y) \mapsto \mathbb{I}(f(x) \neq y)$. We also write the loss $\ell_{01} \circ f_w$ as g_w.

Lower bracketing cover. We first discretize the one-dimensional space of w's at the scale of ε. Let $w_j = \{\varepsilon, 2\varepsilon, \ldots, N_\varepsilon \varepsilon\}$, where $N_\varepsilon = \lceil \frac{1}{\varepsilon} \rceil$.

The elements of the cover are given by $g_j = g_{w_j} \mathbb{I}(x \notin [w_j - \varepsilon, w_j])$. To check the first condition for a lower bracketing cover, notice that for any function g_w, outside $[w_j - \varepsilon, w_j]$ we have $g_w = g_{w_j}$ as $f_w = f_{w_j}$. Inside $[w_j - \varepsilon, w_j]$, we have $g_i = 0$ by definition. Therefore $g_i \leq g_w$. To check the second condition for a lower bracketing cover, we follow the same argument as above and notice that the length of the interval $[w_j - \varepsilon, w_j]$ is ε and the difference between g_w and g_j is bounded by 1, which gives $Pg_w \leq Pg_i + \varepsilon$. We illustrate these conditions in Fig. 7.2.

We can also define two-sided bracketing.

Definition 7.4 (Bracketing)**.** Fix $\varepsilon > 0$. The functions $g_1^U, \ldots, g_m^U$ and $g_1^L, \ldots, g_m^L$ from $\mathcal{Z} \to \mathbb{R}$ is a cover of $\mathcal{G}@P@\varepsilon$ if $\forall g \in \mathcal{G}$ there exists $j \in [m]$ such that (1) $g_j^L \leq g \leq g_j^U$; and (2) $Pg_j^U - \varepsilon \leq Pg \leq Pg_j^L + \varepsilon$ (CHECK!).

We remark that we lose some constant factors in the bounds while using bracketing.

7.4 Bounded Variance Condition

Multiplicative Chernoff gave fast convergence rates $(1/n)$, however, it was restricted to the case when the predictors/losses are bounded and the loss is small. For some special cases when the variance of the loss is

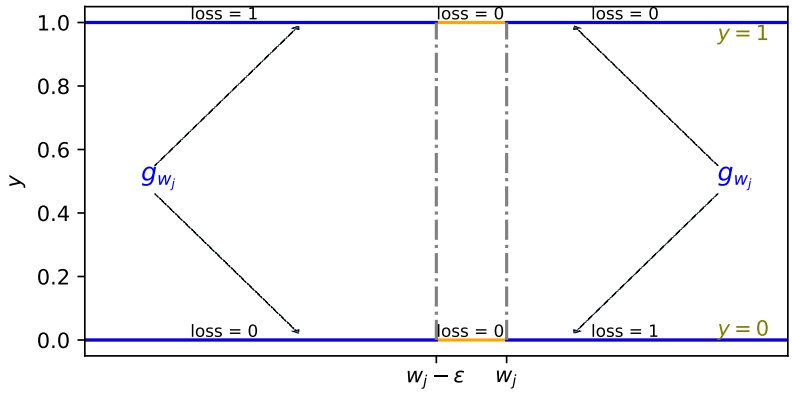

Figure 7.2: Lower bracketing cover for linear threshold functions. The goal is to cover a function g_w so that $w \in [w_j - \varepsilon, w_j]$. The cover element is $g_j = w_j \mathbb{I}(x \not\in [w_j - \varepsilon, w_j])$. The indicated losses are for g_j for the cases $y = 0$ and $y = 1$. The blue horizontal lines show the function. The blue horizontal lines indicate the intervals on which $g_j = g_{w_j}$ and the orange lines show the interval when $g_j = 0$.

bounded, we can still get fast rates by exploiting the bounded variance property. This is given by Bernstein's inequality. But before delving into Bernstein's, let us define the bounded variance class and discuss some examples of loss classes that have bounded variance.

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ for some sets $\mathcal{X}, \mathcal{Y}$ and $g : \mathcal{Z} \to \mathcal{R}$ be a measurable function. Further, let $P \in \mathcal{M}_1(\mathcal{Z})$. The variance of g measured against P is given by $\mathrm{Var}_P(g) = P(g - Pg)^2$.

Definition 7.5 (Bounded variance class)**.** Fix $c_0, c_1 > 0$. Then the bounded variance class is defined as

$$Var_{\mathcal{Z}}(c_0, c_1, P) = \left\{ g : \mathcal{Z} \to \mathbb{R} : \mathrm{Var}_P(g) \leq c_0^2 + c_1 Pg \right\} .$$

Examples.

1. **Bounded functions.** If $0 \leq g \leq 1$, then $g \in \mathrm{Var}_{\mathcal{Z}}(0, 1, P)$.

2. **Convex function class.** For some set $\mathcal{X}$, let $\mathcal{F} \subset \mathcal{M}(\mathcal{X}, \mathbb{R})$ and assume that $\mathcal{F}$ is convex (i.e., for any $\alpha \in [0, 1]$, $f, g \in \mathcal{F}$, $\alpha f + (1 - \alpha) g \in \mathcal{F}$ also holds). Let $\mathcal{Z} = \mathcal{X} \times \mathbb{R}$ and

   $$\mathcal{G} = \{ \ell_f \, : \, \ell_f : \mathcal{Z} \to \mathbb{R}, \ell_f(x, y) = (f(x) - y)^2, f \in \mathcal{F} \} .$$

   By abusing notation, we also write for this set $\mathcal{G} = \ell_{\mathrm{sq}} \circ \mathcal{F}$. Let $P \in \mathcal{M}_1(\mathcal{Z})$ be such that for some $M > 0$ constant, for any $g \in \mathcal{G}$, $g(Z) \leq M^2$ with probability one, where $Z \sim P$. Define $g_* = \arg\min_{g \in \mathcal{G}} Pg$ (which is assumed to exist) and

   $$\tilde{\mathcal{G}} = \{ g - g_* \, : \, g \in \mathcal{G} \} \quad (= \mathcal{G} - \{ g_* \}) ,$$

   so that $\inf_{\tilde{g} \in \tilde{\mathcal{G}}} P\tilde{g} = 0$. Then,

   $$\tilde{\mathcal{G}} \subset \mathrm{Var}_{\mathcal{Z}}(0, 4M^2, P).$$

3. **Best predictor not in the function class.** Fix $M > 0$. Let $\mathcal{F} \subset \mathcal{M}(\mathcal{X}, [0, M])$ be set set of functions bounded in the interval $[0, M]$, $\mathcal{Z} = \mathcal{X} \times [0, M]$, $P \in \mathcal{M}_1(\mathcal{Z})$ be a probability distribution over $\mathcal{Z}$.

Now let $\mathcal{G} = \ell_{\text{sq}} \circ \mathcal{F}$ and $f_*(x) = \mathbb{E}[Y|X = x]$ for $x \in \mathcal{X}$ be the best predictor which may not be in the $\mathcal{F}$. Define

$$\tilde{\mathcal{G}} = \mathcal{G} - \{\ell_{f_*}\}\,.$$

Then,

$$\tilde{\mathcal{G}} \subset \text{Var}_{\mathcal{Z}}(0, 2M^2, P)\,.$$