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

Lecture 4: September 14

Lecturer: Csaba Szepesvári *Scribes: Zixin Zhong*

4.1 Outline

1. Concentration inequalities:
   Chernoff's inequality, multiplicative Chernoff's inequality; Bernett's inequality, Bernstein inequality
2. PAC-learning:
   PAC learnability based on 'fitness'/union bounds

4.2 Concentration inequalities

Theorem 4.1 (Chernoff's Inequality)**.** *Let* $X_1, \ldots, X_n \in [0,1]$ *be i.i.d. random variables,* $\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$*,* $\mu = \mathbb{E}X_1$*. We have*

(a) $\forall \delta \in (0,1)$*, with probability* $1-\delta$*,*

$$\bar{X}_n \leq \mu + \sqrt{\frac{\log(1/\delta)}{2n}};$$

(b) $\forall \delta \in (0,1)$*, with probability* $1-\delta$*,*

$$\bar{X}_n \geq \mu - \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. Since $X_1 \in [a,b]$ implies that X_1 is $\sigma(X_1)$-SG with $\sigma(X_1) = \frac{b-a}{n}$, $X_1 \in [0,1]$ indicates that

$$\sigma(\bar{X}_n) = \frac{\sigma(X_1)}{\sqrt{n}} = \frac{1}{2\sqrt{n}}.$$

Applying this fact with Hoeffding inequality, the Chernoff's inequality is proven. □

Theorem 4.2 (Multiplicative Chernoff's Inequality)**.** *Let* $X_1, \ldots, X_n \in [0,1]$ *be i.i.d. random variables,* $\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$*,* $\mu = \mathbb{E}X_1$*. We have*

(a) $\forall \delta \in (0,1)$*, with probability* $1-\delta$*,*

$$\bar{X}_n \leq \mu + \sqrt{\frac{2\mu\log(1/\delta)}{n}} + \frac{1}{3n};$$

(b) $\forall \delta \in (0,1)$*, with probability* $1-\delta$*,*

$$\bar{X}_n \geq \mu - \sqrt{\frac{2\mu\log(1/\delta)}{n}}. \qquad (*)$$

Remark 4.3.

(a) How big can μ be?
By (*): $\mu \leq \bar{X}_n + \sqrt{\frac{2\mu \log(1/\delta)}{n}}$.

(b) Let

$$f(a,c) = \max\{u : u \leq a + \sqrt{u \cdot c}\}, \text{ where } a = \bar{X}_n,\ c = \frac{2\log(1/\delta)}{n}.$$

Then

$$\begin{aligned}
&\mu + \frac{\log(1/\delta)}{2n} \geq \sqrt{\frac{2\mu\log(1/\delta)}{n}} \text{ and equality holds when } \mu = \frac{\log(1/\delta)}{2n},\\
&\Rightarrow \inf_{0<\gamma<1} \gamma\mu + \frac{\log(1/\delta)}{2\gamma n} = \sqrt{\frac{2\mu\log(1/\delta)}{n}},\\
&\Rightarrow \mu - \sqrt{\frac{2\mu\log(1/\delta)}{n}} = \sup_{0<\gamma<1} (1-\gamma)\mu - \frac{\log(1/\delta)}{2\gamma n}.
\end{aligned}$$

Let $\gamma = 1/2$, then with (*) we have

$$\bar{X}_n \geq \frac{\mu}{2} - \frac{\log(1/\delta)}{n}.$$

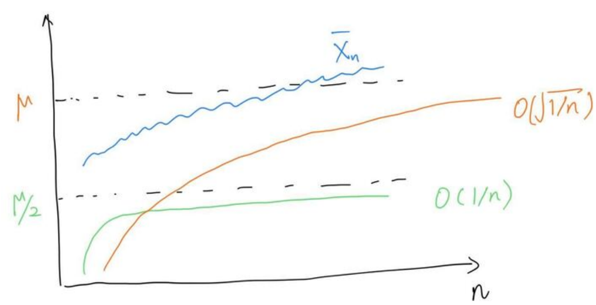

Figure 4.1: Example: set $\gamma = 1/2$.

(c) When we apply γ that does not maximize the term $(1-\gamma)\mu - \frac{\log(1/\delta)}{2\gamma n}$, we cannot claim that we get a better 'convergence' rate, because when $n \to \infty$, $(1-\gamma)\mu - \frac{\log(1/\delta)}{2\gamma n}$ and $\bar{X}_n$ converges to different values. In detail, $\bar{X}_n$ converges to μ regardless of the value of γ, and $(1-\gamma)\mu - \frac{\log(1/\delta)}{2\gamma n}$ converges to $(1-\gamma)\mu \neq \mu$ when $0 < \gamma < 1$.
To say something about the convergence of $\bar{X}_n$, we need to have the coefficient of μ be 1.

Theorem 4.4 (Bernett's Inequality)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables. Set $\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$ and $\mu = \mathbb{E}X_1$. If $X_1 - \mu \leq b$, with probability $1-\delta$, we have*

$$\bar{X}_n \leq \mu + \sqrt{\frac{2\,\mathrm{Var}(X_1)\log(1/\delta)}{n}} + \frac{b}{3n}.$$

4.3 PAC-learning (L. Valiant)

Let function $f_* : \{0,1\}^d \to \{0,1\}$, $X_1, X_2, \ldots, X_n \in \{0,1\}^d := \underline{2}^d$ be i.i.d. random variables drawn from distribution P_X, data set $D_n = \{(X_1, f_*(X_1)), \ldots, (X_n, f_*(X_n))\}$.

Let $f_* \in \mathcal{F} \subset \underline{2}^{\underline{2}^d}$ and $f \in \underline{2}^{\underline{2}^d}$, $P_X^{f_*} := P(X_1, f_* X_1)$, and

$$\begin{aligned}
&L(f) = \mathbb{P}\left(f(X) \neq f_*(X)\right) = L(P_X^{f_*}, f),\\
&l : \underline{2} \times \underline{2} \to \underline{2}, \quad l(y, y') = \mathbf{1}(y \neq y'),\\
&L(P_X^{f_*}, f) = \int P(\mathrm{d}x, \mathrm{d}y)\, l(f(x), y).
\end{aligned}$$

Definition 4.5 (PAC-Learning)**.** Fix $\mathcal{C} = (\mathcal{C}_d)_{d\geq 1}$, where $\mathcal{C}_d \subset \underline{2}^{\underline{2}^d}$. $\mathcal{C}$ is **PAC-learnable (Proabably Approximately Correctly)** if $\exists$ polynomial $p \in \mathbb{R}[x, y, z]$ and $\mathcal{A} = (\mathcal{A}_{n,d})_{n\geq 1, d\geq 1}$ where $\mathcal{A}_{n,d} : \left(\underline{2}^d \times \underline{2}\right)^n \to \underline{2}^{\underline{2}^d}$

$$\begin{aligned}
\text{s.t.}\quad &\forall \varepsilon \in (0,1),\ \delta \in (0,1),\ d \geq 1,\ P \in \mathcal{M}_1(\underline{2}^d),\ f_* \in \mathcal{C}_d,\\
&n \geq \lceil p(1/\varepsilon, 1/\delta, d) \rceil,\\
&X_1, X_2, \ldots, X_n \sim P_X,\\
&f_n = \mathcal{A}_{n,d}\left(\underbrace{\left(X_1, f_*(X_1)\right), \ldots, \left(X_n, f_*(X_n)\right)}_{D_n}\right),
\end{aligned}$$

we have

$$\mathbb{P}\left(L\left(P_X^{f_*}, f_n\right) \geq \varepsilon\right) \leq \delta.$$

In other words, with probability $1 - \delta$, $\mathbb{P}\left(f_n(X) \neq f_*(x) | D_n\right) \leq \varepsilon$.

Remark 4.6. (a) Example:

$$\mathcal{C}^d_{\text{AND}} = \left\{ f : \underline{2}^d \to \underline{2} \mid \exists u \subset [d],\ \forall x \in \underline{2}^d : f(x) = \min_{j \in u} X_j \right\}.$$

(b) (i) $L(f_*) = 0$. (ii) When $Y_i = f_*(X_i)$, there is **NO** noise and this will make learning **faster**.