CSC 588: Machine learning theory **Spring 2022**

Lecture 5: Hoeffding's Inequality, Bernstein's Inequality, ERM's Guarantee

Lecturer: Chicheng Zhang *Scribe: Yanan Wang*

1 Proof of Lemma 2 for Hoeffding's Inequality

Lemma 2. *If $X_1, \cdots, X_n$ are independent, for each i, X_i is σ_i^2-SG, then $\sum_{i=1}^n a_i X_i$ is $\sum_{i=1}^n a_i^2 \sigma_i^2$-SG $\forall a_1, \cdots, a_n$.*

To prove Lemma 2, We first prove two special cases:

$$aX_i \text{ is } a^2\sigma_i^2\text{-SG}, \tag{2.1}$$

$$X_1 + X_2 \text{ is } (\sigma_1^2 + \sigma_2^2)\text{-SG}. \tag{2.2}$$

Case 2.1 is proved in Lecture 3.

To show the proof of 2.2, let $\mu_1 = \mathbb{E}[X_1]$, $\mu_2 = \mathbb{E}[X_2]$, $Y = X_1 + X_2, \mathbb{E}[Y] = \mu_1 + \mu_2$.

$$\begin{aligned}
&\mathbb{E}[\mathrm{e}^{\lambda(Y-(\mu_1+\mu_2))}] \\
&= \mathbb{E}[\mathrm{e}^{\lambda(X_1-\mu_1)}\mathrm{e}^{\lambda(X_2-\mu_2)}] \\
&= \mathbb{E}[\mathrm{e}^{\lambda(X_1-\mu_1)}]\mathbb{E}[\mathrm{e}^{\lambda(X_2-\mu_2)}] && \text{(independence of } X_1 \text{ and } X_2) \\
&\le \mathrm{e}^{\frac{\lambda^2\sigma_1^2}{2}}\mathrm{e}^{\frac{\lambda^2\sigma_2^2}{2}} && (X_i \text{ is } \sigma_i^2\text{-SG}) \\
&= \mathrm{e}^{\frac{\lambda^2(\sigma_1^2+\sigma_2^2)}{2}}.
\end{aligned}$$

Therefore, $X_1 + X_2$ is $(\sigma_1^2 + \sigma_2^2)$-SG.

It is now straightforward to prove Lemma 2 using 2.1 and 2.2 inductively.

2 Proof of Lemma 3 for Hoeffding's Inequality

Lemma 3. *$\forall$ random variable (r.v.) X taking value in interval $[a, b]$, X is $\frac{(b-a)^2}{4}$-SG.*

Proof. Want to show $\forall\lambda$,

$$\mathbb{E}[\mathrm{e}^{\lambda(\mathrm{X}-\mu)}] \le \mathrm{e}^{\frac{(b-a)^2\lambda^2}{8}}.$$

Let $\psi(\lambda) = \ln \mathbb{E}[\mathrm{e}^{\lambda(\mathrm{X}-\mu)}]$, $\psi(\lambda)$ is called the cumulant generating function (cgf) of $Y = X - \mu$. It suffices to show $\forall\lambda$, $\psi(\lambda) \le \frac{(b-a)^2\lambda^2}{8}$.

Using second-order Taylor expansion (with Lagrange Remainder) at 0, $\exists\xi$ between 0 and λ,

$$\psi(\lambda) = \psi(0) + \psi'(0)\lambda + \frac{\psi''(\xi)}{2}\lambda^2$$

$\psi(0) = 0$.

$$\begin{aligned}
&\psi'(\lambda) \\
&= \frac{1}{\mathbb{E}[e^{\lambda Y}]}\frac{\partial \mathbb{E}[e^{\lambda Y}]}{\partial\lambda} && (Y = X - \mu) \\
&= \frac{\mathbb{E}[Ye^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]},
\end{aligned}$$

where, $\psi'(0) = \mathbb{E}[Y] = 0$.

$$\psi''(\lambda) = \underbrace{\frac{\mathbb{E}[e^{\lambda Y} Y^2]}{\mathbb{E}[e^{\lambda Y}]}}_{*1} - (\underbrace{\frac{\mathbb{E}[e^{\lambda Y} Y]}{\mathbb{E}[e^{\lambda Y}]}}_{*2})^2.$$

Let Z be r.v. with probability density function:

$$P_Z(y) = \frac{P_Y(y) e^{\lambda y}}{\int_{\mathbb{R}} P_Y(y) e^{\lambda y} dy}.$$

Exercise: Show $\mathbb{E}[Z] = *2, \mathbb{E}[Z^2] = *1$.
Then,

$$\begin{aligned}
\psi''(\lambda) &= \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \\
&= \mathrm{var}(Z) \\
&= \mathbb{E}[(Z - \mathbb{E}[Z])^2] \\
&\le \mathbb{E}[(Z - w)^2] && (\mathbb{E}[Z] = \mathrm{argmin}_w \mathbb{E}[(Z - w)^2]) \\
&= \mathbb{E}[(Z - (\frac{a+b}{2} - \mu))^2] && (\text{set } w = (\tfrac{a+b}{2} - \mu), \text{ and notice } Z \in [a - \mu, b - \mu]) \\
&\le \frac{(b-a)^2}{4}.
\end{aligned}$$

□

3 Proof of Hoeffding's Inequality

Hoeffding's Inequality: Suppose $Z_1, \cdots, Z_n$ are iid, $\forall i, Z_i \in [a, b], \bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i, \mu = \mathbb{E}[Z_i]$, then, for all $\epsilon > 0$:

$$\mathbb{P}(|\bar{Z} - \mu| \ge \epsilon) \le 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

Proof. As $Z_i \in [a, b]$, Z_i is $\frac{(b-a)^2}{4}$-SG according to Lemma 3.

According to Lemma 2, $\bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i$ is $\frac{(b-a)^2}{4n}$-SG.

Finally, as $\mathbb{E}[\bar{Z}] = \mu$, according to Lemma 1,

$$\mathbb{P}(|\bar{Z} - \mu| \ge \epsilon) \le 2\exp\left(-\frac{\epsilon^2}{2 \cdot \frac{(b-a)^2}{4n}}\right) = 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

□

Important Corollary For a classifier h, training set S: m iid training samples, $\forall \epsilon$:

$$\mathbb{P}(|\,\mathrm{err}(h, S) - \mathrm{err}(h, D)| \ge \epsilon) \le 2\exp\left(-2m\epsilon^2\right)$$

This is obtained by setting $n = m, a = 0, b = 1$, and each $Z_i = I(h(x_i) \ne y_i)$ is the mistake indicator of h on example (x_i, y_i).

Equivalently, by setting $\delta = 2\exp\left(-2m\epsilon^2\right)$:

$$\forall \delta, \mathbb{P}\left(|\,\mathrm{err}(h, S) - \mathrm{err}(h, D)| \ge \sqrt{\frac{\ln \frac{2}{\delta}}{2m}}\right) \le \delta$$

4 Bernstein's Inequality (taking r.v.'s refined information into account)

Let $X_1, \cdots, X_n$ be iid random variables, $\forall i, |X_i - \mathbb{E}[X_i]| \leq R, \mu = \mathbb{E}[X_i], \sigma^2 = \text{var}[X_i]$, then, $\forall \epsilon > 0$:

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu|) \geq \epsilon) \leq \underbrace{2\text{exp}\left(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2}{3}R\epsilon}\right)}_{(*)}.$$

If $\sigma^2 \ll (b-a)^2$, then $(*) \ll \exp\left(-\frac{n\epsilon^2}{(b-a)^2}\right)$, which indicates a more tighter bound than Hoeffding's inequality.

Set a small ϵ s.t. $(*) \leq \delta$:

$$\begin{aligned} & 2\text{exp}\left(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2}{3}R\epsilon}\right) \leq \delta \\ & \Leftarrow n\epsilon^2 \geq (2\sigma^2 + \frac{2}{3}R\epsilon)\text{ln}\frac{2}{\delta} \\ & \Leftarrow n\epsilon^2 \geq 4\sigma^2\text{ln}\frac{2}{\delta} \text{ and } n\epsilon^2 \geq \frac{4}{3}R\epsilon\text{ln}\frac{2}{\delta} \\ & \Leftarrow \epsilon \geq \sqrt{\frac{4\sigma^2\text{ln}\frac{2}{\delta}}{n}} \text{ and } \epsilon \geq \frac{4R\text{ln}\frac{2}{\delta}}{3n}, \end{aligned}$$

Chicheng notes after lecture: the constants presented in the lecture were off by a factor of 2, which is corrected here. This is because in the second $\Leftarrow$, we use $A \geq B + C \Leftarrow A \geq 2B$ and $A \geq 2C$, which introduces extra constants 2.

So we can select $\epsilon \geq \sqrt{\frac{4\sigma^2\text{ln}\frac{2}{\delta}}{n}} + \frac{4R\text{ln}\frac{2}{\delta}}{3n}$:

$$\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu|) \geq \sqrt{\frac{4\sigma^2\text{ln}\frac{2}{\delta}}{n}} + \frac{4R\text{ln}\frac{2}{\delta}}{3n}\right) \leq \delta.$$

As $\frac{4R\text{ln}\frac{2}{\delta}}{3n}$ is a lower order term, compared with Hoeffding's inequality's result:

$$\mathbb{P}\left(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu|) \geq \sqrt{\frac{(b-a)^2\text{ln}\frac{2}{\delta}}{2n}}\right) \leq \delta,$$

it is more tight when $\sigma^2 \ll (b-a)^2$.

5 ERM's Guarantee

Theorem (ERM's Guarantee). *ERM with $\mathcal{H}$ has an agnostic PAC sample complexity of $f(\epsilon, \delta) = \frac{2}{\epsilon^2}(\text{ln}|\mathcal{H}| + \text{ln}\frac{2}{\delta})$; in other words, given $m \geq f(\epsilon, \delta)$ iid training examples, w.p. $1 - \delta$:*

$\hat{h}$(ERM output) satisfies:

$$\text{err}(\hat{h}, D) \leq \min_{h \in \mathcal{H}} \text{err}(h, D) + \epsilon.$$

***Proof sketch**. define*

$$E = \cap_{h \in \mathcal{H}} \left\{|\text{ err}(h, S) - \text{err}(h, D)| \leq \frac{\epsilon}{2}\right\}.$$

If we show $P(E) \geq 1 - \delta$, then we are done: indeed, using the key observation last time with $\mu = \frac{\epsilon}{2}$, then when E happens,

$$\text{err}(\hat{h}, D) \leq \min_{h \in \mathcal{H}} \text{err}(h, D) + \epsilon.$$