Machine Learning Theory Lecture 2: Concentration Inequality

Richard Xu

August 30, 2021

1 Motivation for this lecture

let's look at this recent NTK paper: `https://arxiv.org/abs/2012.11654`. It uses the following inequality/bound/definitions:

1. Hoeffding inequality
2. Chernoff bound
3. sub-Gaussian

To motivate the audience, today's lecture is centered around these terms

1.1 A revision exercise for last week

QUESTION if we do know the upper bound of $\mathbb{E}[\|X\|_1] \leq C$, then, how would you proceed to bound $\|X\|_2$?

2 Simple question: how to tightly bound Gaussian

if $X \sim \mathcal{N}(0, \sigma^2)$, then:

$$\Pr(X > t) = \frac{1}{\sqrt{(2\pi)}\sigma} \int_{x=t}^{\infty} \exp^{\frac{-x^2}{\sigma^2}} \,\mathrm{d}x \tag{3}$$

The integral is a problem. But we can apply some trick to it: as t is the smallest integral limit, then $\frac{x}{t} > 1 \quad \forall x > t$:

$$\begin{aligned}
\Pr(X > t) &< \frac{1}{\sqrt{(2\pi)}\sigma} \int_{x=t}^{\infty} \frac{x}{t} \exp^{\frac{-x^2}{\sigma^2}} \,\mathrm{d}x \\
&= \frac{1}{\sqrt{(2\pi)}\sigma\, t} \int_{x=t}^{\infty} x \exp^{\frac{-x^2}{\sigma^2}} \,\mathrm{d}x \\
&= \frac{1}{\sqrt{(2\pi)}\sigma\, t} \int_{x=t}^{\infty} \left(-\frac{\mathrm{d}}{\mathrm{d}x} \exp^{\frac{-x^2}{\sigma^2}} \right) \mathrm{d}x \quad \text{easy to check it's the same} \\
&= \frac{1}{\sqrt{(2\pi)}\sigma\, t} \left[-\exp^{\frac{-x^2}{\sigma^2}} \right]_{x=t}^{\infty} \\
&= \frac{1}{\sqrt{(2\pi)}\sigma\, t} \exp^{\frac{-t^2}{\sigma^2}}
\end{aligned} \tag{4}$$

we will compare this result with bound derived from generic subG(σ^2) case.

3 Use MGF to bound: Chernoff bounds

Theorem 1

$$\begin{aligned}
\Pr(X - \mathbb{E}(X) \geq \epsilon) &\leq \min_{\lambda \geq 0} \left[\mathbb{E}\big[\exp^{\lambda(X - \mathbb{E}[X])} \big] \exp^{-\lambda\epsilon} \right] \\
&= \min_{\lambda \geq 0} \frac{\mathbb{E}\big[\exp^{\lambda(X - \mathbb{E}[X])} \big]}{\exp^{\lambda\epsilon}}
\end{aligned} \tag{6}$$

1. note that Chernoff bound does **not** assume $X - \mathbb{E}(X) \geq 0$
2. however, it's important to realize that in Chernoff bound, $\lambda \geq 0$

3.1 Proof for Chernoff bounds

proof for **theorem 1** is really simple, it's just apply Markov Inequality to $\exp^{(\cdot)}$:

$$\begin{aligned}
\Pr(X - \mathbb{E}(X) \geq \epsilon) &= \Pr\Big(\exp^{\lambda(X - \mathbb{E}(X))} \geq \exp^{(\lambda\epsilon)} \Big) \quad \exp^{\lambda x} \text{ is monotonically increasing, when } \lambda \geq 0 \\
&\leq \frac{\mathbb{E}[\exp^{\lambda(X - \mathbb{E}(X))}]}{\exp^{(\lambda\epsilon)}} \quad \text{Markov Inequality} \\
&= \mathbb{E}[\exp^{\lambda(X - \mathbb{E}(X))}] \exp^{-\lambda\epsilon}
\end{aligned} \tag{7}$$

QUESTION What if we do **not** restrict $\lambda \geq 0$?

QUESTION Does it still work if: $X - \mathbb{E}(X) < 0$?

QUESTION If it can be bounded by every $\lambda \geq 0$, then which one would you choose?

QUESTION What is $\mathbb{E}[\exp^{\lambda(X-\mathbb{E}(X))}]$?

3.1.1 To bound $\Pr(X - \mathbb{E}(X) \leq -\epsilon)$

notice that $X - \mathbb{E}(X) \leq -\epsilon \quad \Leftrightarrow \quad \mathbb{E}(X) - X \geq \epsilon$, therefore: $\forall \lambda \geq 0$:

$$\begin{aligned}
\Pr(X - \mathbb{E}(X) \leq -\epsilon) &= \Pr(\mathbb{E}(X) - X \geq \epsilon) \\
&= \Pr\Big(\exp^{\lambda(\mathbb{E}(X)-X)} \geq \exp^{\lambda\epsilon}\Big) \\
&\leq \frac{\mathbb{E}[\exp^{\lambda(\mathbb{E}(X)-X)}]}{\exp^{\lambda\epsilon}} \quad \text{Markov Inequality} \\
&= \mathbb{E}[\exp^{\lambda(\mathbb{E}(X)-X)}]\exp^{-\lambda\epsilon}
\end{aligned} \tag{8}$$

3.2 summary

in both cases, since any λ works, to make the bound tighter, we may choose:

$$\begin{cases} \Pr(X - \mathbb{E}(X) \geq \epsilon) & \leq \min_{\lambda \geq 0} \frac{\mathbb{E}[\exp^{\lambda(X-\mathbb{E}(X))}]}{\exp^{\lambda\epsilon}} \\ \Pr(X - \mathbb{E}(X) \leq -\epsilon) & \leq \min_{\lambda \geq 0} \frac{\mathbb{E}[\exp^{\lambda(\mathbb{E}(X)-X)}]}{\exp^{\lambda\epsilon}} \end{cases} \tag{9}$$

Note $\Pr(X - \mathbb{E}(X) \geq \epsilon)$ and $\Pr(\mathbb{E}(X) - X \geq \epsilon)$ do **not** have the same bound! So nothing can be said about $\Pr(|X - \mathbb{E}(X)| \leq \epsilon)$

QUESTION : does it work with $\lambda = 0$?

3.3 Chernoff bounds to sum of variables

since we know,

$$\begin{aligned}
\text{MGF}_{X_1+\cdots+X_n}(\lambda) &= \prod_{i=1}^{n} \text{MGF}_{X_i}(\lambda) \\
&= \big(\text{MGF}_{X_i}(\lambda)\big)^n \quad \text{for i.i.d samples}
\end{aligned} \tag{11}$$

therefore, for $X_i \overset{\text{i.i.d}}{\sim} p_X(\cdot)$:

$$\Pr\Big(\sum_{i=1}^{n} X_i - n\mathbb{E}(X) \geq \epsilon\Big) \leq \min_{\lambda \geq 0}\Big[\Big(\mathbb{E}_{X\sim P_X(\cdot)}[\exp^{\lambda(X-\mathbb{E}(X))}]\Big)^n \exp^{-\lambda\epsilon}\Big] \tag{12}$$

3.4 Example: sum of Rademacher R.Vs

It's out of order, but let's assume we know how to **bound** MGF for Rademacher distribution in Eq.(34), we can bound:

$$X = \sum_{i=1}^{n} \sigma_i \tag{13}$$

using **Chernoff bound**, we have:

$$\begin{aligned}
\Pr(X - \mathbb{E}(X) \geq \epsilon) &\leq \min_{\lambda \geq 0} \Big[\mathbb{E}\big[\exp^{\lambda(X - \mathbb{E}[X])} \big] \exp^{-\lambda\epsilon} \Big] \\
\implies \Pr(\sum_{i=1}^{n} \sigma_i - n\mathbb{E}(\sigma_1) \geq \epsilon) &\leq \min_{\lambda \geq 0} \Big[\big(\mathbb{E}\big[\exp^{\lambda(\sigma_1 - \mathbb{E}[\sigma_1])} \big] \big)^n \exp^{-\lambda\epsilon} \Big] \qquad \mathbb{E}(\sigma_1) = 0 \\
&\leq \min_{\lambda \geq 0} \Big[\Big(\exp\Big(\frac{\lambda^2}{2}\Big) \Big)^n \exp^{-\lambda\epsilon} \Big] \quad \text{apply} \quad \text{Eq.(34). Just trust it for now!} \\
&= \min_{\lambda \geq 0} \Big[\exp\Big(\frac{n\lambda^2}{2} - \lambda\epsilon \Big) \Big]
\end{aligned} \tag{14}$$

to minimize, we just need to minimize $\frac{n\lambda^2}{2} - \lambda\epsilon$: **QUESTION** why this is true in here?

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\lambda} \Big(\frac{n\lambda^2}{2} - \lambda\epsilon \Big) \\
\implies & n\lambda - \epsilon = 0 \\
\implies & \lambda = \frac{\epsilon}{n}
\end{aligned} \tag{15}$$

after substitution, we have:

$$\begin{aligned}
\Pr(X - \mathbb{E}(X) \geq \epsilon) &\leq \exp\Big(\frac{\epsilon^2}{2n} - \frac{\epsilon^2}{n} \Big) \\
&= \exp\Big(-\frac{\epsilon^2}{2n} \Big)
\end{aligned} \tag{16}$$

3.4.1 alternative expression to make R.H.S simple

making R.H.S simple, i.e., δ, we have:

$$\begin{aligned}
\delta &= \exp\Big(-\frac{\epsilon^2}{2n} \Big) \\
\log(\delta) &= -\frac{\epsilon^2}{2n} \\
\epsilon &= \sqrt{-2n\log(\delta)}
\end{aligned} \tag{17}$$

QUESTION can you see $-2n\log(\delta) \geq 0$?
substitute it back, we have:

$$\Pr\Big((X - \mathbb{E}[X]) \geq \sqrt{-2n\log(\delta)} \Big) \leq \delta \tag{18}$$

or, with probability of at least $1 - \delta$: $X - \mathbb{E}[X]$ is bounded by $\sqrt{-2n\log(\delta)}$

3.4.2 Exercise to use Chernoff Bound

QUESTION : use Chernoff Bound for $\|\mathbf{X}\|_2^2$ when $X_i \sim \mathcal{N}(0, 1)$

3.5 Sub-Gaussian

Definition A mean-zero random variable X is σ^2-sub-Gaussian, or written as $X \sim \text{subG}(\sigma^2)$, if:

$$\mathbb{E}\big[\exp^{\lambda X}\big] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \tag{21}$$

i.e., if the MGF of a zero-meaned X can be bounded by a Gaussian MGF if it was to also have σ^2 variance

the simplest example would be Gaussian itself

3.5.1 Properties 1: bound sum of subGaussian variables

Lemma 2 *let X_i be zero-mean-ed independent random variables (no need to be identical), and $X_i \sim$ subG(σ_i^2). then:*

$$\sum_{i=1}^{n} X_i \sim \textit{subG}\big(\sum_{i=1}^{n}\sigma_i^2\big) \tag{22}$$

3.5.2 combine Chernoff Bound with subGaussian

Lemma 3 *Let $X \sim$ subG(σ^2), then for any $t > 0$, we have:*

$$\Pr(X > t) \leq \exp^{-\frac{t^2}{2\sigma^2}} \tag{23}$$

proof for Lemma 3

$$\begin{aligned}
\Pr(X \geq t) &\leq \min_{\lambda \geq 0}\left[\mathbb{E}[\exp^{\lambda(X)}]\exp^{-\lambda t}\right] \quad \text{by Chernoff bound} \\
&\leq \min_{\lambda \geq 0}\left[\exp^{\frac{\lambda^2\sigma^2}{2}}\ \exp^{-\lambda t}\right] \quad \text{by subGaussian definition} \\
&= \min_{\lambda \geq 0}\left[\exp^{\frac{\lambda^2\sigma^2}{2} - \lambda t}\right]
\end{aligned} \tag{24}$$

by minimizing $\frac{\lambda^2\sigma^2}{2} - \lambda t$:

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\lambda^2\sigma^2}{2} - \lambda t\right) \\
&= \lambda\sigma^2 - t = 0 \\
\implies \lambda &= \frac{t}{\sigma^2}
\end{aligned} \tag{25}$$

$$\begin{aligned}
\Pr(X \geq t) &\leq \exp^{\frac{t^2\sigma^2}{2\sigma^4} - \frac{t^2}{\sigma^2}} \\
&= \exp^{\frac{t^2}{2\sigma^2} - \frac{t^2}{\sigma^2}} \\
&= \exp^{-\frac{t^2}{2\sigma^2}}
\end{aligned} \tag{26}$$

Compare this with bound using Eq.(4) where we have: $\Pr(X > t) < \frac{1}{\sqrt{2\pi)}\sigma\ t}\exp^{\frac{-t^2}{\sigma^2}}$

3.5.3 Bound sum of i.i.d. subG variables using Chernoff Bound

1. expectation version:

$$
\begin{aligned}
&\Pr\big(X \geq t\big) \leq \exp^{-\frac{t^2}{2\sigma^2}} \quad \textbf{Lemma (3)} \\
\implies &\Pr\Big(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\Big) = \Pr\Big(\sum_{i=1}^{n} X_i \geq nt\Big) \\
&\leq \exp^{-\frac{n^2t^2}{2\sum_{i=1}^{n}\sigma_i^2}} \quad \text{apply } \textbf{Lemma (2)} \quad \text{replace } \sigma^2 \to \sum_{i=1}^{n}\sigma_i^2 \\
&= \exp^{-\frac{nt^2}{2\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2}} \quad \text{rewrite denominator as average } \sigma^2 \\
&= \exp^{-\frac{nt^2}{2\bar{\sigma}^2}}
\end{aligned} \tag{27}
$$

2. sum version: if we are just interested in bounding $\Pr\big(\sum_{i=1}^{n} X_i \geq t\big)$:

$$
\implies \Pr\Big(\sum_{i=1}^{n} X_i \geq t\Big) \leq \exp^{-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}} \quad \text{apply } \textbf{Lemma (2)} \quad \text{replace } \sigma^2 \to \sum_{i=1}^{n}\sigma_i^2 \tag{28}
$$

4 bound MGF when $X \in [a, b]$: hoeffding lemma

1. when apply Chernoff bound, RHS contains MGF. Then hoeffding lemma can further upper bound the MGF
2. Markov Inequality assumes R.Vs to have support over $0 \ldots \infty^+$. Let's see what if we place a more restrictive range over its support $[a, b]$ (ideal for hypothesis values)
3. higher the moment one can bound, the tighter the bound, so let's look at bounding movement generation function:

we have two versions of **hoeffding lemma**, for $\lambda \in \mathbb{R}$:

Theorem 4 *loose version: for $\lambda \in \mathbb{R}$:*

$$\mathbb{E}\Big[\exp^{\lambda(X-\mathbb{E}[X])}\Big] \leq \exp\Big(\frac{\lambda^2(b-a)^2}{2}\Big) \tag{29}$$

Theorem 5 *tight version: for $\lambda \in \mathbb{R}$:*

$$\mathbb{E}\Big[\exp^{\lambda(X-\mathbb{E}[X])}\Big] \leq \exp\Big(\frac{\lambda^2(b-a)^2}{8}\Big) \tag{30}$$

a few things to note:

QUESTION what does it tell you about the sub-gaussiantiy of $X - \mathbb{E}[X]$, when it's bounded by (a, b)?

4.1 $\mathbb{E}\Big[\exp^{\lambda(X-\mathbb{E}[X])}\Big]$ and $\mathbb{E}\Big[\exp^{\lambda(\mathbb{E}[X]-X)}\Big]$ has the same bound!

it should be realized that in hoeffding lemma $\lambda \in \mathbb{R}$ instead, this is different to Chernoff bound where $\lambda > 0$. One of the consequnce is that:

$$\begin{aligned}
\mathbb{E}\Big[\exp^{\lambda(\mathbb{E}[X]-X)}\Big] &= \mathbb{E}\Big[\exp^{(-\lambda)(X-\mathbb{E}[X])}\Big] \\
&\leq \exp\Big(\frac{(-\lambda)^2(b-a)^2}{8}\Big) \quad \because \text{Theorem (5)} \\
&= \exp\Big(\frac{\lambda^2(b-a)^2}{8}\Big)
\end{aligned} \tag{33}$$

Eq.(33) is the key why Hoeffding inequality has the same bound for $\Pr(X-\mathbb{E}[X] \geq \epsilon)$ and $\Pr(\mathbb{E}[X]-X \leq \epsilon)$

4.2 Example: MGF for Rademacher R.V.

4.2.1 apply hoeffding lemma (strong version)

$$\begin{aligned}
&\mathbb{E}\Big[\exp^{\lambda X}\Big] \leq \exp^{\lambda\mathbb{E}[X]+\frac{\lambda^2(b-a)^2}{8}} \\
\implies &\mathbb{E}_{\sigma\sim\text{Rad}}[\exp(\lambda\sigma)] \leq \exp^{\lambda\times 0+\frac{\lambda^2(1-(-1))^2}{8}} \\
&= \exp^{\frac{\lambda^2}{2}}
\end{aligned} \tag{34}$$

as a note: $\text{MGF}_{\sigma\sim Rad}(\lambda) = \cosh(\lambda) = \frac{\exp^{\lambda}+\exp^{-\lambda}}{2}$

4.2.2 bound it in a hard-way

Moment Generation Function in general:

$$\mathbb{E}_X[\exp^{\lambda X}] = \sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E}[X^k]}{k!} \tag{35}$$

in the case: $\sigma \sim \text{Rad}$, we have:

$$\mathbb{E}[\sigma^k] = \begin{cases} p(\sigma=-1)s^k + p(\sigma=1)s^k = \frac{1}{2}\times 1 + \frac{1}{2}\times 1 = 1 & \text{if } k \text{ is even} \\ p(\sigma=-1)s^k + p(\sigma=1)s^k = \frac{1}{2}\times(-1) + \frac{1}{2}\times 1 = 0 & \text{if } k \text{ is odd} \end{cases} \tag{36}$$

since odd terms of $\lambda^k \mathbb{E}[\sigma^k]$ in the sum is gone, then Rademacher MGF only has even terms:

$$\begin{aligned} \mathbb{E}_{\sigma\sim\text{Rad}}[\exp^{\lambda\sigma}] &= \sum_{k=0,2,4,\ldots}^{\infty} \frac{\lambda^k}{k!} \\ &= \sum_{k=0,1,2,\ldots}^{\infty} \frac{\lambda^{2k}}{(2k)!} \quad \text{put back to increment by 1} \\ &\text{the following is try to put the form back, to be bounded by } \exp(\cdot) \\ &\leq \sum_{k=0,1,2,\ldots}^{\infty} \frac{\lambda^{2k}}{2^k \times k!} \qquad \because \frac{1}{(2k)!} \leq \frac{1}{2^k \times k!} \\ &= \sum_{k=0,1,2,\ldots}^{\infty} \left(\frac{\lambda^2}{2}\right)^k \frac{1}{k!} \quad \text{this is in form of exp} \\ &= \exp\left(\frac{\lambda^2}{2}\right) \end{aligned} \tag{37}$$

both achieves the above derivations

4.3 Proof for hoeffding lemma: the loose version

4.3.1 fact: composite "non-decreasing convex function" of convex function, is also convex

To do so, recognizing $\exp^{\lambda(C-Z)}$ is convex function. Also, in general the following lemma holds:

Lemma 6 *f and g are both convex, and g is non-decreasing, then:*

$$\begin{aligned} &(g\circ f)(x) \quad \textit{is convex} \\ \textit{i.e.,}\quad &(g\circ f)\big(\theta x + (1-\theta)y\big) \leq \theta(g\circ f)(x) + (1-\theta)(g\circ f)(y) \end{aligned} \tag{39}$$

proof of Lemma (6)

$$
\begin{aligned}
(g\circ f)\big(\theta x+(1-\theta)y\big) &= g\big(f\,(\theta x+(1-\theta)y)\,\big)\\
&\le g\Big(\theta\underbrace{f(x)}_{x'}+(1-\theta)\underbrace{f(y)}_{y'}\Big) \quad f \text{ is convex and } g \text{ non-decreasing}\\
&\le \theta g\big(f(x)\big)+(1-\theta)g\big(f(y)\big) \quad g \text{ is convex}\\
&= \theta(g\circ f)(x)+(1-\theta)(g\circ f)(y)
\end{aligned}
\tag{40}
$$

the **example** here:

$$
\begin{cases} f=\lambda(C-Z) & \text{convex}\\ g=\exp(\cdot) & \text{convex and non-decreasing}\end{cases}
\tag{41}
$$

4.3.2 the Z' trick

first to apply Z' trick: let Z and Z' from identical distributions, we have:

$$
\begin{aligned}
&\mathbb{E}_Z\big[\exp^{\lambda(Z-\mathbb{E}[Z])}\big] \quad \text{MGF of } Z\\
&=\mathbb{E}_Z\big[\exp^{\lambda(Z-\mathbb{E}[Z'])}\big] \quad Z' \textbf{ trick: } \text{since } Z, Z' \text{ from same distribution}\\
&\le \mathbb{E}_Z\big[\mathbb{E}_{Z'}[\exp^{\lambda(Z-Z')}]\big] \qquad \exp^{\lambda(Z-\mathbb{E}[Z'])} \text{ is convex, so Jensen's inequality}
\end{aligned}
\tag{42}
$$

we have introduced the $\le$ sign, but there is no easy way to bound the above. If we attempt the following:

$$
\begin{aligned}
\mathbb{E}_Z\big[\mathbb{E}_{Z'}[\exp^{(\lambda(Z-Z'))}]\big] &\le \mathbb{E}_Z\big[\mathbb{E}_{Z'}[\exp^{(\lambda(b-a))}]\big]\\
&=\exp^{(\lambda(b-a))} \quad \text{assume } \lambda(Z-Z')\le\lambda(b-a) \quad \forall Z, Z', \lambda>0
\end{aligned}
\tag{43}
$$

however, the above does **not** work for $\lambda < 0$, as $\lambda(Z-Z')$ is **not** universally less than $\lambda(b-a)$, when $\lambda < 0$.

the intuition is that if we can bring $\lambda \to \lambda^2$, then it will work

4.3.3 the $\times\sigma$ trick

continue from Eq.(42), here comes the $\times\sigma$ trick. Let's look at only the inner-most term, where Z and Z' are treated as constants:

$$
\begin{aligned}
\mathbb{E}_Z\big[\exp^{\lambda(Z-\mathbb{E}[Z])}\big] &\le \mathbb{E}_Z\big[\mathbb{E}_{Z'}[\exp^{\lambda(Z-Z')}]\big]\\
&=\mathbb{E}_Z\big[\mathbb{E}_{Z'}\big[\mathbb{E}_{\sigma\sim\text{Rad}}[\exp^{\lambda\sigma(Z-Z')}]\big]\big]
\end{aligned}
\tag{44}
$$

the reason to bring Z' to the equation has been two folds:

1. we can apply Jensen's inequality. we already show this in Eq.(42) i.e., Z' **trick part**
2. it also allowed us to construct a new random variable $Z-Z'$, that is symmetric around 0, for all $p(Z)$. Of course, if $Z-\mathbb{E}[z]$ is already a symmetric, then we can times σ directly
3. now that we have $(Z-Z')$ is symmetric around 0, here comes the $\times\sigma$ **trick**: multiply by Rademacher R.V. $\sigma\sim$ Rad doesn't change the distribution of $Z-Z'$.
4. note that the same $\times\sigma$ trick will be used again in Rademacher Complexity section $\sum_{i=1}^n\big(h(Z_i')-h(Z_i)\big)=\sum_{i=1}^n\sigma_i\big(h(Z_i')-h(Z_i)\big)$

4.3.4 inner most expectation if MGF of Radmarcher distribution

$\mathbb{E}_{\sigma\sim\text{Rad}}[\exp^{\lambda\sigma(Z-Z')}]$ is $\text{MGF}_\sigma(\lambda(Z-Z'))$ which is bounded by either Eq.(34), or Eq.(37).

However, since we are proving looser version of Hoeffding Lemma here, we can't claim it is bounded by a derivation using (stronger version) Heoffding Lemma, i.e., Eq.(34), otherwise, it is "nested" prove!. Therefore, we claim we used Eq.(37) instead:

$$\begin{aligned}&\mathbb{E}_{\sigma\sim\text{Rad}}[\exp^{\lambda\sigma(Z-Z')}] \qquad \lambda\to\lambda(Z-Z')\\&=\text{MGF}_\sigma(\lambda(Z-Z'))\\&\le\exp\Big(\frac{\lambda^2(Z-Z')^2}{2}\Big)\end{aligned}\tag{45}$$

4.3.5 back to the proof

as $a\le Z,Z'\le b\Leftrightarrow|Z-Z'|\le|b-a|$:

$$\begin{aligned}\mathbb{E}_Z\big[\exp(\lambda(Z-\mathbb{E}[Z]))\big]&\le\mathbb{E}_Z\big[\mathbb{E}_{Z'}\big[\mathbb{E}_{\sigma\sim\text{Rad}}\big[\exp^{(\lambda\sigma(Z-Z'))}\big]\big]\big]\\&\le\mathbb{E}_Z\big[\mathbb{E}_{Z'}\big[\exp^{\frac{\lambda^2(Z-Z')^2}{2}}\big]\big]\\&\le\mathbb{E}_Z\Big[\mathbb{E}_{Z'}\Big[\exp\Big(\frac{\lambda^2(a-b)^2}{2}\Big)\Big]\Big]\\&=\exp\Big(\frac{\lambda^2(a-b)^2}{2}\Big)\end{aligned}\tag{47}$$

compare with Eq.(43), we achieve the above since we transformed:

$$\lambda(a-b)\to\lambda^2(a-b)^2\tag{48}$$

alternative expression:

$$\begin{aligned}\mathbb{E}_Z\big[\exp(\lambda(Z-\mathbb{E}[Z]))\big]&=\frac{\mathbb{E}_Z\big[\exp(\lambda Z)\big]}{\exp(\lambda\mathbb{E}[Z])}\le\exp\Big(\frac{\lambda^2(a-b)^2}{2}\Big)\\&\implies\mathbb{E}_Z\big[\exp(\lambda Z)\big]\le\exp\Big(\lambda\mathbb{E}[Z]+\frac{\lambda^2(a-b)^2}{2}\Big)\end{aligned}\tag{49}$$

4.4 tight version

look at bounding movement generation function using Taylor expansion:

$$\begin{aligned}&\mathbb{E}\big[\exp^{\lambda(X-\mathbb{E}[X])}\big]\le\exp\Big(\frac{\lambda^2(b-a)^2}{8}\Big)\\&\implies\mathbb{E}\big[\exp^{\lambda X}\big]\le\exp\Big(\lambda\mathbb{E}[X]+\frac{\lambda^2(b-a)^2}{8}\Big)\end{aligned}\tag{50}$$

proof is left as an exercise.

5 hoeffding inequality

5.1 definition

bounding the tail distribution when condition exist for $X_i \in [a_i, b_i]$. In the context of bounding $\hat{R}_S$, the condition is set for value of R. This is different to McDiarmid, where condition is set on relationship between input and output.

5.1.1 mean version

Theorem 7 *When it is known that X_i are strictly bounded by intervals $[a_i, b_i]$, we let $\mu = \mathbb{E}[\overline{X}]$, it is used to bound sample means of random variables:*

$$\begin{aligned}
\Pr\Big(\overline{X} - \mu \geq \epsilon\Big) &\leq \exp\Big(- \frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big) \\
\Pr\Big(\big|\overline{X} - \mu\big| \geq \epsilon\Big) &\leq 2\exp\Big(- \frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big) \qquad \textit{by Eq.(33)} \\
&= 2\exp\Big(- 2nC\epsilon^2\Big) \qquad \textit{where } C = \frac{n}{\sum_{i=1}^n (b_i - a_i)^2}
\end{aligned} \tag{51}$$

5.1.2 sum version

hoeffding inequality can also be used to bound the sum instead of the sample mean:

Theorem 8 *X_i are strictly bounded by intervals $[a_i, b_i]$, and $S_n = \sum_i X_i$ of the random variables:*

$$\begin{aligned}
\Pr(S_n - \mathbb{E}[S_n] \geq \epsilon) &\leq \exp\left(- \frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \\
\Pr\big(\big|S_n - \mathbb{E}[S_n]\big| \geq \epsilon\big) &\leq 2\exp\left(- \frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)
\end{aligned} \tag{52}$$

5.2 proof of hoeffding inequality

for all $\lambda > 0$:

$$\begin{aligned}
\Pr\big(S_n - \mathbb{E}[S_n] \geq \epsilon\big) &= \Pr\big(\exp^{\lambda(S_n - \mathbb{E}[S_n])} \geq \exp^{\lambda\epsilon}\big) \\
&\leq \exp^{-\lambda\epsilon}\,\mathbb{E}\big[\exp^{\lambda(S_n - \mathbb{E}[S_n])}\big] \qquad \text{Markov or Chernoff require } \lambda \geq 0 \\
&= \exp^{-\lambda\epsilon} \prod_{i=1}^n \mathbb{E}\big[\exp^{\lambda(X_i - \mathbb{E}[X_i])}\big] \\
&\leq \exp^{-\lambda\epsilon} \prod_{i=1}^n \exp^{\frac{\lambda^2(b_i - a_i)^2}{8}} \qquad \text{strong version of hoeffding lemma} \\
&= \exp\big(- \lambda\epsilon + \frac{1}{8}\lambda^2 \sum_{i=1}^n (b_i - a_i)^2\big) \\
&\equiv \exp\big(- \lambda\epsilon + C\lambda^2\big) \qquad \text{let } C = \frac{1}{8}\sum_{i=1}^n (b_i - a_i)^2
\end{aligned} \tag{53}$$

then we optimize λ:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}\big(C\lambda^2 - \lambda\epsilon\big) &= 2C\lambda - \epsilon = 0 \\
\implies \lambda &= \frac{\epsilon}{2C}
\end{aligned} \tag{54}
$$

after substitution:

$$
\begin{aligned}
\Pr\big(S_n - \mathbb{E}[S_n] \geq \epsilon\big) &\leq \exp\Big(-\frac{\epsilon}{2C}\epsilon + \Big(\frac{\epsilon}{2C}\Big)^2 C\Big) \\
&= \exp\Big(-\frac{\epsilon^2}{2C} + \frac{\epsilon^2}{4C}\Big) \\
&= \exp\Big(-\frac{\epsilon^2}{4C}\Big) \\
&= \exp\left(-\frac{8 \times \epsilon^2}{4\sum_{i=1}^{n}(b_i - a_i)^2}\right) \\
&= \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)
\end{aligned} \tag{55}
$$

5.2.1 to bound $S_n - \mathbb{E}[S_n] \leq -\epsilon$:

$$
\begin{aligned}
\Pr\big(S_n - \mathbb{E}[S_n] \leq -\epsilon\big) &= \Pr\big(\mathbb{E}[S_n] - S_n \geq \epsilon\big) \\
&= \Pr\left(\exp^{\lambda(\mathbb{E}[S_n] - S_n)} \geq \exp^{\lambda\epsilon}\right) \\
&\leq \exp^{-\lambda\epsilon}\,\mathbb{E}\left[\exp^{\lambda(\mathbb{E}[S_n] - S_n)}\right] \quad \text{Markov or Chernoff} \\
&= \exp^{-\lambda\epsilon}\prod_{i=1}^{n}\mathbb{E}\left[\exp^{\lambda(\mathbb{E}[X_i] - X_i)}\right] \\
&\leq \exp^{-\lambda\epsilon}\prod_{i=1}^{n}\exp\Big(\frac{\lambda^2(b_i - a_i)^2}{8}\Big) \quad \text{same bound for: } \mathbb{E}[X_i] - X_i \quad \text{Eq.(33)} \\
&= \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \quad \text{rest of the proof is same as Eq.(55)}
\end{aligned} \tag{56}
$$

5.3 obvious application of hoeffding inequality

looking at empirical risk:

$$
\hat{R}_S(h) = \frac{1}{n}\sum_{i}^{n}\mathbf{1}(y_i \neq h(x_i)) \tag{57}
$$

we also know $\mathbb{E}[\hat{R}(h)] = R(h)$, substituting this into Hoeffding Inequality: and $a_i = 0, b_i = 1 \quad \forall i$:

$$
\begin{aligned}
&\Pr\left(\left|\hat{R}_n(h) - R(h)\right| \geq \epsilon\right) \\
&\leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \\
&= 2\exp^{-\frac{2n^2\epsilon^2}{n}} \\
&= 2\exp^{-2n\epsilon^2}
\end{aligned} \tag{58}
$$

6 homework

Read up the following:

1. general concept of Rademacher Complexity

7 references

in this tutorial, I have paraphrased a number of existing courses and notes, I encourage people to see the original notes too.

1. `http://cs229.stanford.edu/extra-notes/hoeffding.pdf`
2. various Wikipedia pages