Machine Learning Theory Lecture 2: Concentration Inequality

Richard Xu

March 26, 2022

1 Motivation for this lecture

Initially, I wrote this class by reading 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

So I thought in order to motivate the audience, today's lecture is centered around these terms. However, then throughout the Saturday Night Live (SNL) class, I decided to add all common inequalities to the class.

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 $\|X\|_2$?

$$\|\mathbf{x}\|_2 = \underbrace{\sqrt{\sum_i |x_i|^2} \leq \sum_i |x_i|}_{\text{Eq.(2)}} = \|\mathbf{x}\|_1 \tag{1}$$

you can see $\sqrt{\sum_i |x_i|^2} \leq \sum_i |x_i|$ from the two dimensions case, recursively derive the rest:

$$\begin{aligned}
|x_1|^2 + |x_2|^2 &\leq |x_1|^2 + 2|x_1||x_2| + |x_2|^2 \\
&= (|x_1| + |x_2|)^2 \\
\implies \sqrt{|x_1|^2 + |x_2|^2} &\leq |x_1| + |x_2| \quad \text{recursively adding dimension also works} \\
\implies \sqrt{\sum_i |x_i|^2} &\leq \sum_i |x_i|
\end{aligned} \tag{2}$$

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 bound by MGF: Chernoff bounds

Theorem 1

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

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 $\exp^{(\cdot)}$:

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

Some questions to consider:

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

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

3. **QUESTION** If it can be bounded by every $\lambda \geq 0$, then which one would you choose?
4. **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

Chernoff bounds (and all its derivatives) are very useful to bound sum of independent (not necessarily identical) random variables. Since we know:

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

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 do know how to **bound** MGF for Rademacher distribution in Eq.(34), we can bound X where:

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

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 $X_i \sim \mathcal{N}(0,1)$

$$\|\mathbf{X}\|_2^2 = \sum_i^k X_i^2 \tag{19}$$

apply Eq.(20), let $Y_i = X_i^2$:

$$\begin{aligned}
\Pr\big(\sum_{i=1}^n Y_i - n\mathbb{E}(Y) \geq \epsilon\big) &\leq \min_{\lambda \geq 0}\Big[\Big(\mathbb{E}_{Y\sim P_Y(\cdot)}[\exp^{\lambda(Y-\mathbb{E}(Y))}]\Big)^n \exp^{-\lambda\epsilon}\Big] \\
&= \min_{\lambda \geq 0}\Big[\big(\mathrm{MGF}_{\chi^2(Y)}(\lambda)\big)^n \exp^{-\lambda\epsilon}\Big] \\
&= \min_{\lambda \geq 0}\Big[(1-2\lambda)^{-\frac{n}{2}} \exp^{-\lambda\epsilon}\Big] \\
&= \min_{\lambda \geq 0}\Big[\frac{\exp^{-\lambda\epsilon}}{(1-2\lambda)^{\frac{n}{2}}}\Big] \quad \text{for } \lambda \leq \frac{1}{2}
\end{aligned} \tag{20}$$

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:

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

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

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

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

1. when apply Chernoff bound, RHS contains MGF. Then hoeffding lemma and Bernstein lemma can further upper bound the MGF (given certain conditions)

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)

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

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 (1)

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: $\sigma \sim$ 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:

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

both achieves the above derivations

4.2.3 bound it in a hard-way (2)

first to note that $\exp(\lambda x)$ is a convex function. We then pick tangent of two points $\big(-1, \exp(-\lambda)\big)$ and $\big(1, \exp(\lambda)\big)$:

$$
\begin{aligned}
\exp\big(\lambda(-1\times(1-\theta) + 1\times\theta)\big) &\le (1-\theta)\exp(-\lambda) + \theta\exp(\lambda) \qquad 0\le\theta\le 1 \\
\exp\big((2\theta-1)\lambda\big) &\le (1-\theta)\exp(-\lambda) + \theta\exp(\lambda) \qquad 0\le\theta\le 1
\end{aligned} \tag{38}
$$

realizing $0 \le \theta \le 1 \implies \theta = \frac{X+1}{2} \quad -1 \le X \le 1$:

$$
\begin{aligned}
\exp\left(x\lambda\right) &\leq \left(\frac{1-x}{2}\right)\exp(-\lambda) + \left(\frac{x+1}{2}\right)\exp(\lambda) \qquad -1 \leq x \leq 1 \\
\mathbb{E}_X\left[\exp\left(x\lambda\right)\right] &\leq \mathbb{E}_X\left[\left(\frac{1-x}{2}\right)\exp(-\lambda) + \left(\frac{x+1}{2}\right)\exp(\lambda)\right] \qquad -1 \leq x \leq 1 \\
&= \mathbb{E}_X\left[\frac{1-x}{2}\right]\exp(-\lambda) + \mathbb{E}_X\left[\frac{x+1}{2}\right]\exp(\lambda) \qquad -1 \leq x \leq 1 \\
&= \frac{1}{2}\Big(\exp(-\lambda) + \exp(\lambda)\Big) \qquad \mathbb{E}[X] = 0 \\
&= \frac{1}{2}\Big(\big(1 - \frac{\lambda}{1!} + \frac{\lambda^2}{2!} - \frac{\lambda^3}{3!} + \frac{\lambda^4}{4!} + \dots\big) + \big(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \frac{\lambda^4}{4!} + \dots\big)\Big) \\
&= 1 + \frac{\lambda^2}{2!} + \frac{\lambda^4}{4!} + \frac{\lambda^6}{6!} + \dots \\
&= \sum_{k=0,1,2,\dots}^{\infty} \frac{\lambda^{2k}}{(2k)!}
\end{aligned}
\tag{39}
$$

the rest should just follow Eq. 37, so when $|X| \leq 1$:

$$
\mathbb{E}_X[\exp^{\lambda X}] \leq \exp\left(\frac{\lambda^2}{2}\right) \tag{40}
$$

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

$$
\begin{aligned}
&(g \circ f)(x) \quad \textit{is convex} \\
\textit{i.e.,} \quad &(g \circ f)\big(\theta x + (1-\theta)y\big) \leq \theta(g \circ f)(x) + (1-\theta)(g \circ f)(y)
\end{aligned}
\tag{42}
$$

proof of Lemma (6)

$$
\begin{aligned}
(g \circ f)\big(\theta x + (1-\theta)y\big) &= g\big(f\left(\theta x + (1-\theta)y\right)\big) \\
&\leq g\Big(\theta \underbrace{f(x)}_{x'} + (1-\theta)\underbrace{f(y)}_{y'}\Big) \quad f \text{ is convex and } g \text{ non-decreasing} \\
&\leq \theta g\big(f(x)\big) + (1-\theta) g\big(f(y)\big) \quad g \text{ is convex} \\
&= \theta(g \circ f)(x) + (1-\theta)(g \circ f)(y)
\end{aligned}
\tag{43}
$$

the **example** here:

$$\begin{cases} f = \lambda(C - Z) & \text{convex} \\ g = \exp(\cdot) & \text{convex and non-decreasing} \end{cases} \tag{44}$$

4.3.2 the Z' trick

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

$$\begin{aligned}
&\mathbb{E}_Z\big[\exp^{\lambda(Z-\mathbb{E}[Z])}\big] \quad \text{MGF of } Z \\
&= \mathbb{E}_Z\big[\exp^{\lambda(Z-\mathbb{E}[Z'])}\big] \quad Z' \textbf{ trick: } \text{ since } Z, Z' \text{ from same distribution} \\
&\leq \mathbb{E}_Z\big[\mathbb{E}_{Z'}[\exp^{\lambda(Z-Z')}]\big] \\
&\qquad\qquad \because \exp^{\lambda(Z-\mathbb{E}[Z'])} \text{ is convex (by lemma 6), we apply Jensen's inequality}
\end{aligned} \tag{45}$$

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

$$\begin{aligned}
\mathbb{E}_Z\big[\mathbb{E}_{Z'}[\exp^{(\lambda(Z-Z'))}]\big] &\leq \mathbb{E}_Z\big[\mathbb{E}_{Z'}[\exp^{(\lambda(b-a))}]\big] \\
&= \exp^{(\lambda(b-a))} \quad \text{assume } \lambda(Z-Z') \leq \lambda(b-a) \quad \forall Z, Z', \lambda > 0
\end{aligned} \tag{46}$$

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 try to prove $\lambda^2(Z - Z')^2 \leq \lambda^2(a - b)^2$ instead, then it will work (just like Theorem 4 which we try to prove)

4.3.3 the $\times\sigma$ trick

continue from Eq.(45), here comes the $\times\sigma$ trick. Let's look at only the inner-most term, where Z and Z' are treated as constants:

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

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.(45) 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

$$\begin{aligned}
\mathbb{E}_{\sigma\sim\text{Rad}}[\exp^{\lambda\sigma(Z-Z')}] &= \text{MGF}_\sigma(\lambda(Z-Z')) \\
&\equiv \text{MGF}_{\sigma(Z-Z')}(\lambda)
\end{aligned} \tag{48}$$

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:

$$
\begin{aligned}
&\mathbb{E}_{\sigma\sim\text{Rad}}[\exp^{\lambda\sigma(Z-Z')}] \qquad \lambda \to \lambda(Z-Z') \\
&= \text{MGF}_{\sigma}(\lambda(Z-Z')) \\
&\leq \exp\Big(\frac{\lambda^2(Z-Z')^2}{2}\Big)
\end{aligned}
\tag{49}
$$

added an alternative derivation

4.3.5 back to the proof

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

$$
\begin{aligned}
\mathbb{E}_Z\big[\exp(\lambda(Z-\mathbb{E}[Z]))\big] &\leq \mathbb{E}_Z\big[\mathbb{E}_{Z'}\big[\mathbb{E}_{\sigma\sim\text{Rad}}\big[\exp^{(\lambda\sigma(Z-Z'))}\big]\big]\big] \\
&\leq \mathbb{E}_Z\big[\mathbb{E}_{Z'}\big[\exp^{\frac{\lambda^2(Z-Z')^2}{2}}\big]\big] \\
&\leq \mathbb{E}_Z\Big[\mathbb{E}_{Z'}\Big[\exp\Big(\frac{\lambda^2(a-b)^2}{2}\Big)\Big]\Big] \quad \text{safely insert the range} \\
&= \exp\Big(\frac{\lambda^2(a-b)^2}{2}\Big)
\end{aligned}
\tag{51}
$$

compare with Eq.(46), we achieve the above since we transformed:

$$
\lambda(Z-Z') \leq \lambda(a-b) \quad \longrightarrow \quad \lambda^2(Z-Z')^2 \leq \lambda^2(a-b)^2 \tag{52}
$$

alternative expression:

$$
\begin{aligned}
\mathbb{E}_Z\big[\exp(\lambda(Z-\mathbb{E}[Z]))\big] = \frac{\mathbb{E}_Z\big[\exp(\lambda Z)\big]}{\exp(\lambda\mathbb{E}[Z])} &\leq \exp\Big(\frac{\lambda^2(a-b)^2}{2}\Big) \\
\implies \mathbb{E}_Z\big[\exp(\lambda Z)\big] &\leq \exp\Big(\lambda\mathbb{E}[Z] + \frac{\lambda^2(a-b)^2}{2}\Big)
\end{aligned}
\tag{53}
$$

4.4 tight version

look at bounding movement generation function using Taylor expansion:

$$
\begin{aligned}
\mathbb{E}\big[\exp^{\lambda(X-\mathbb{E}[X])}\big] &\leq \exp\Big(\frac{\lambda^2(b-a)^2}{8}\Big) \\
\implies \mathbb{E}\big[\exp^{\lambda X}\big] &\leq \exp\Big(\lambda\mathbb{E}[X] + \frac{\lambda^2(b-a)^2}{8}\Big)
\end{aligned}
\tag{54}
$$

We now can extend Eq.(59) further to be of the lemma:

Lemma 7 *let X be a random varaible over the sample space $[a,b]$ s.t. $\mathbb{E}[X]=0$. For any $\lambda>0$, we have:*

$$\mathbb{E}\big[\exp^{\lambda X}\big] \leq \frac{b}{b-a}\exp(\lambda a) - \frac{a}{b-a}\exp(\lambda b) \tag{55}$$

proof of Lemma 7 it is just a generalization of Eq.(59):

$$\exp\big(\lambda(-1\times(1-\theta)+1\times\theta)\big) \leq (1-\theta)\exp(-\lambda)+\theta\exp(\lambda) \qquad 0\leq\theta\leq 1 \tag{56}$$

realizing $0\leq\theta\leq 1 \implies \begin{cases}\theta & = \frac{X-a}{b-a} \\ 1-\theta & = 1-\frac{X-a}{b-a} = \frac{b-a-(X-a)}{b-a} = \frac{b-X}{b-a}\end{cases} \quad b\leq X\leq a$:

$$\begin{aligned}
\exp\big(X\lambda\big) &= \exp\lambda\big((1-\theta)a+\theta b\big) \qquad \because 0\leq\theta\leq 1 \text{ and } a\leq X\leq b \\
&\leq (1-\theta)\exp(\lambda a)+\theta\exp(\lambda b) \\
&= \frac{b-X}{b-a}\exp(\lambda a)+\frac{X-a}{b-a}\exp(\lambda b) \qquad \text{substitute } \theta \\
\implies \mathbb{E}_X\big[\exp\big(X\lambda\big)\big] &\leq \mathbb{E}_X\Big[\frac{b-X}{b-a}\exp(\lambda a)+\frac{X-a}{b-a}\exp(\lambda b)\Big] \qquad a\leq X\leq b \\
&= \frac{b-\mathbb{E}[X]}{b-a}\exp(\lambda a)+\frac{\mathbb{E}[X]-a}{b-a}\exp(\lambda b) \\
&= \frac{b}{b-a}\exp(\lambda a)-\frac{a}{b-a}\exp(\lambda b) \qquad \because \mathbb{E}[X]=0
\end{aligned} \tag{57}$$

it obviously works for when $a=-1, b=1$:

$$\begin{aligned}
\exp\big(X\lambda\big) &\leq \frac{1}{2}\exp(\lambda(-1))+\frac{1}{2}\exp(\lambda\times 1) \\
&= \frac{\exp(-\lambda)+\exp(\lambda)}{2}
\end{aligned} \tag{58}$$

Lemma 8 *for $a<0<b$, we have:*

$$\frac{b}{b-a}\exp(\lambda a)-\frac{a}{b-a}\exp(\lambda b) \leq \exp\Big(\frac{\lambda^2(b-a)^2}{8}\Big) \tag{59}$$

Obviously, combining both Lemma 7 and Lemma 8, we derive Hoeffding Lemma.
to make it simpler, we introduce variable t:

$$\begin{aligned}
t &= \frac{-a}{b-a} \\
\implies 1-t &= \Big(\frac{-a}{b-a}\Big) = \frac{(b-a)-(-a)}{b-a} = \frac{b}{b-a}
\end{aligned} \tag{60}$$

note since $a<0<b \implies 0\leq t\leq 1$

$$
\begin{aligned}
&\frac{b}{b-a}\exp(\lambda a)-\frac{a}{b-a}\exp(\lambda b)\\
&=\frac{b}{b-a}\exp(\lambda a)-\frac{a}{b-a}\exp(\lambda b)\frac{\exp(\lambda a)}{\exp(\lambda a)}\\
&=\exp(\lambda a)\Big(\frac{b}{b-a}-\frac{a}{b-a}\exp\big(\lambda(b-a)\big)\Big)\\
&=\exp(\lambda a)\Big(1+\frac{a}{b-a}-\frac{a}{b-a}\exp\big(\lambda(b-a)\big)\Big)\quad 1+\frac{a}{b-a}=\frac{b-a+a}{b-a}=\frac{b}{b-a}\\
&=\exp\big(\lambda(-t(b-a))\big)\Big(1-t+t\exp\big(\lambda(b-a)\big)\Big)\qquad \text{let } t=-\frac{a}{b-a}\\
\implies\ & f=\lambda(-t(b-a))+\log\Big(1-t+t\exp\big(\lambda(b-a)\big)\Big)\quad \text{taking } \log(\cdot)
\end{aligned}
\tag{61}
$$

4.4.1 approach one: $u = (b-a)$

$$
\begin{aligned}
f&=\lambda(-t(b-a))+\log\Big(1-t+t\exp\big(\lambda(b-a)\big)\Big)\\
\implies f(\lambda)&=\lambda(-tu)+\log\Big(1-t+t\exp\big(\lambda u\big)\Big)\\
f'(\lambda)&=-tu+\frac{tu\exp\big(\lambda u\big)}{1-t+t\exp\big(\lambda u\big)}\\
f''(\lambda)&=\frac{tu^2\exp\big(\lambda u\big)}{1-t+t\exp\big(\lambda u\big)}+\frac{t^2u^2\exp\big(2\lambda u\big)}{\Big(1-t+t\exp\big(\lambda u\big)\Big)^2}
\end{aligned}
\tag{62}
$$

does not give us the desired form

4.4.2 approach one: $u = \lambda(b-a)$

$$
\begin{aligned}
f&=\lambda(-t(b-a))+\log\Big(1-t+t\exp\big(\lambda(b-a)\big)\Big)\\
\implies f(u)&=-tu+\log\Big(1-t+t\exp\big(u\big)\Big)\qquad\implies f(0)=0\\
f'(u)&=-tu+\frac{t\exp\big(u\big)}{1-t+t\exp\big(u\big)}\qquad\implies f'(0)=0\\
f''(u)&=\frac{t\exp\big(u\big)}{1-t+t\exp\big(u\big)}-\frac{t^2\exp\big(2u\big)}{\Big(1-t+t\exp\big(u\big)\Big)^2}\\
&=\frac{t\exp\big(u\big)}{1-t+t\exp\big(u\big)}\Big(1-\frac{t\exp\big(u\big)}{1-t+t\exp\big(u\big)}\Big)\\
&=\alpha(1-\alpha)\quad \text{where } \alpha=\frac{t\exp\big(u\big)}{1-t+t\exp\big(u\big)}\\
&\le\frac{1}{4}
\end{aligned}
\tag{63}
$$

last line because $0 \le t \le 1, \exp(u) \ge 0$:

$$
\max(\alpha(1-\alpha))=\frac{1}{4}
\tag{64}
$$

therefore, we have:

$$
\begin{aligned}
f(u) &= -tu + \log\Big(1 - t + t\exp\big(u\big)\Big) \\
&\approx f(0) + f'(0)u + f''(0)\frac{u^2}{2} \\
&\le \frac{u^2}{2} \times \frac{1}{4} = \frac{u^2}{8} \\
\implies \frac{b}{b-a}\exp(\lambda a) - \frac{a}{b-a}\exp(\lambda b) &\le \exp\Big(\frac{\lambda^2(b-a)^2}{8}\Big) \quad \text{take } \exp(\cdot), \text{ substitute } u = \lambda(b-a)
\end{aligned}
\tag{65}
$$

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 9 *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:*

$$
\begin{aligned}
\Pr\Big(\overline{X} - \mu \ge \epsilon\Big) &\le \exp\Big(- \frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big) \\
\Pr\Big(\big|\overline{X} - \mu\big| \ge \epsilon\Big) &\le 2\exp\Big(- \frac{2n^2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big) \qquad \textit{by Eq.(33)} \\
&= 2\exp\Big(- 2nC\epsilon^2\Big) \qquad \textit{where } C = \frac{n}{\sum_{i=1}^n (b_i - a_i)^2}
\end{aligned}
\tag{66}
$$

5.1.2 sum version

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

Theorem 10 *X_i are strictly bounded by intervals $[a_i, b_i]$, and $S_n = \sum_i X_i$ of the random variables:*

$$
\begin{aligned}
\Pr(S_n - \mathbb{E}[S_n] \ge \epsilon) &\le \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \\
\Pr\big(\big|S_n - \mathbb{E}[S_n]\big| \ge \epsilon\big) &\le 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)
\end{aligned}
\tag{67}
$$

5.2 proof of hoeffding inequality

for all $\lambda > 0$:

$$
\begin{aligned}
\Pr\big(S_n - \mathbb{E}[S_n] \geq \epsilon\big) &= \Pr\big(\exp^{\lambda(S_n - \mathbb{E}[S_n])} \geq \exp^{\lambda\epsilon}\big) \\
&\leq \exp^{-\lambda\epsilon}\, \mathbb{E}\big[\exp^{\lambda(S_n - \mathbb{E}[S_n])}\big] \qquad \text{Chernoff require } \lambda \geq 0 \\
&= \exp^{-\lambda\epsilon} \prod_{i=1}^{n} \mathbb{E}\big[\exp^{\lambda(X_i - \mathbb{E}[X_i])}\big] \\
&\leq \exp^{-\lambda\epsilon} \prod_{i=1}^{n} \exp^{\frac{\lambda^2 (b_i - a_i)^2}{8}} \qquad \text{strong version of hoeffding lemma} \\
&= \exp\big(-\lambda\epsilon + \frac{1}{8}\lambda^2 \sum_{i=1}^{n} (b_i - a_i)^2\big) \\
&\equiv \exp\big(-\lambda\epsilon + C\lambda^2\big) \qquad \text{let } C = \frac{1}{8}\sum_{i=1}^{n}(b_i - a_i)^2
\end{aligned} \tag{68}
$$

then we optimize λ:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}\big(C\lambda^2 - \lambda\epsilon\big) &= 2C\lambda - \epsilon = 0 \\
\implies \lambda &= \frac{\epsilon}{2C}
\end{aligned} \tag{69}
$$

after substitution:

$$
\begin{aligned}
\Pr\big(S_n - \mathbb{E}[S_n] \geq \epsilon\big) &\leq \exp\Big(-\frac{\epsilon}{2C}\epsilon + \Big(\frac{\epsilon}{2C}\Big)^2 C\Big) \\
&= \exp\Big(-\frac{\epsilon^2}{2C} + \frac{\epsilon^2}{4C}\Big) \\
&= \exp\Big(-\frac{\epsilon^2}{4C}\Big) \\
&= \exp\left(-\frac{8 \times \epsilon^2}{4\sum_{i=1}^{n}(b_i - a_i)^2}\right) \\
&= \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)
\end{aligned} \tag{70}
$$

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

$$
\begin{aligned}
\Pr\big(S_n - \mathbb{E}[S_n] \leq -\epsilon\big) &= \Pr\big(\mathbb{E}[S_n] - S_n \geq \epsilon\big) \\
&= \Pr\left(\exp^{\lambda(\mathbb{E}[S_n] - S_n)} \geq \exp^{\lambda\epsilon}\right) \\
&\leq \exp^{-\lambda\epsilon}\, \mathbb{E}\left[\exp^{\lambda(\mathbb{E}[S_n] - S_n)}\right] \qquad \text{Chernoff} \\
&= \exp^{-\lambda\epsilon} \prod_{i=1}^{n} \mathbb{E}\left[\exp^{\lambda(\mathbb{E}[X_i] - X_i)}\right] \\
&\leq \exp^{-\lambda\epsilon} \prod_{i=1}^{n} \exp\Big(\frac{\lambda^2 (b_i - a_i)^2}{8}\Big) \qquad \text{same bound for: } \mathbb{E}[X_i] - X_i \quad \text{Eq.(33)} \\
&= \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \qquad \text{rest of the proof is same as Eq.(70)}
\end{aligned} \tag{71}
$$

5.3 obvious application of hoeffding inequality

looking at empirical risk:

$$\hat{R}_S(h) = \frac{1}{n}\sum_{i}^{n} \mathbf{1}(y_i \neq h(x_i)) \tag{72}$$

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

$$\begin{aligned}
&\Pr\left(\left|\hat{R}_n(h) - R(h)\right| \geq \epsilon\right) \\
&\leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \\
&= 2\exp^{-\frac{2n^2\epsilon^2}{n}} \\
&= 2\exp^{-2n\epsilon^2}
\end{aligned} \tag{73}$$

6 Azuma inequality

6.1 Martingale and Martingale difference

Definition 1 *Let $(X_i)_{i=1}^n$ be sequence of random variables such that $\mathbb{E}[X_i|X_1, \ldots, X_{i-1}] = X_{i-1}$ $\forall i$.*

more generally,:

Definition 2 *Let $(X_i)_{i=1}^n$ and $(Z_i)_{i=1}^n$ be sequences of random variables on a common probability space such that:*

$$\begin{aligned}
\mathbb{E}\big[X_i|Z_{i-1}, \ldots, X_1\big] &= \mathbb{E}\big[X_i|g_i(Z_{i-1}, \ldots, X_1)\big] \\
&= Z_{i-1} \qquad \forall i
\end{aligned} \tag{74}$$

let (Z_i) is called a **martingale** *with respect to (X_i). In addition, sequence $Y_i = Z_i - Z_{i-1}$ is called a* **martingale difference** *sequence. By definition, $\mathbb{E}\big[Y_i|X_1, \ldots, X_{i-1}\big] = 0$ $\forall i$.*

6.2 Azuma inequality bound

Theorem 11 *Let (X_i) be a martingale and let $Y_i = X_i - X_{i-1}$ be corresponding difference sequence. If $c_i > 0$ such that $|Y_i| \leq c_i \quad \forall i$, then:*

$$\begin{aligned}
\Pr\Big(X_n - X_0 \geq +\epsilon\Big) &\leq \exp\Big(-\frac{\epsilon^2}{2\sum_{i=1}^{n} c_i^2}\Big) \\
\Pr\Big(X_n - X_0 \leq -\epsilon\Big) &\leq \exp\Big(-\frac{\epsilon^2}{2\sum_{i=1}^{n} c_i^2}\Big)
\end{aligned} \tag{75}$$

in case you wonder why we bound $|Y_n|$ here. It is because of:

$$\begin{aligned}
X_n - X_0 &= \sum_{i=1}^{n} X_i - X_{i-1} \quad \text{telescope sum} \\
&= \sum_{i=1}^{n} Y_i
\end{aligned} \tag{76}$$

we starting working with X_n:

$$\begin{aligned}
\Pr\left(X_n - X_0 \geq \epsilon\right) &= \Pr\left[\exp^{\lambda(X_n - X_0)} \geq \exp^{\lambda\epsilon}\right] \\
&\leq \exp^{-\lambda\epsilon}\,\mathbb{E}_{X_{1:n}}\left[\exp^{\lambda(X_n - X_0)}\right] \\
&= \exp^{-\lambda\epsilon}\,\mathbb{E}_{X_{1:n}}\left[\exp^{\lambda\left(\sum_{0\leq i\leq n} X_i - X_{i-1}\right)}\right] \quad \text{telescope sum } X_n - X_0 = \sum_{0\leq i\leq n} X_i - X_{i-1}
\end{aligned} \tag{77}$$

before we go on further, let's have a look at filtration in Martingale:

6.2.1 what is filtration?

A Filtration is a growing sequence of sigma algebras

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \ldots \subseteq \mathcal{F}_n \tag{78}$$

whenever we write:

$$\mathbb{E}[Y_n | X_1, X_2, \ldots, X_n] \tag{79}$$

we can alternatively write it as

$$\mathbb{E}[Y_{n+1} | \mathcal{F}_n] \tag{80}$$

where $\mathcal{F}_n$ is a σ-algebra that makes random variables $X_1, \ldots, X_n$ measurable.

6.2.2 back to proof

let's just be looking at the term $\mathbb{E}_{X_{1:n}}\left[\exp^{\lambda\left(X_n - X_0\right)}\right]$:

$$\begin{aligned}
\mathbb{E}_{X_{1:n}}\left[\exp^{\lambda\left(X_n - X_0\right)}\right] &\equiv \mathbb{E}_{X_{1:n}}\left[\exp^{\lambda\left(\sum_{0<i\leq n} X_i - X_{i-1}\right)}\right] = \mathbb{E}_{X_{1:n}}\left[\exp^{\lambda\left(\sum_{0<i\leq n} Y_i\right)}\right] \\
&= \mathbb{E}_{X_{1:n-1}}\left[\mathbb{E}_{X_n}\left[\exp^{\lambda\left(\sum_{0\leq i\leq n}(X_i - X_{i-1})\right)} \;|\mathcal{F}_{n-1}\right]\right] \\
&= \mathbb{E}_{X_{1:n-1}}\left[\exp^{\lambda\left(\sum_{0<i\leq n-1}(X_i - X_{i-1})\right)}\,\mathbb{E}\left[\exp^{\lambda(X_n - X_{n-1})} \;|\mathcal{F}_{n-1}\right]\right] \\
&= \mathbb{E}_{X_{1:n-1}}\left[\exp^{\lambda\left(\sum_{0<i\leq n-1}(X_i - X_{i-1})\right)}\,\underbrace{\mathbb{E}\left[\exp^{\lambda Y_n} \;|\mathcal{F}_{n-1}\right]}\right]
\end{aligned} \tag{81}$$

here we may take two different routes:

1. first one is to apply Eq.(40) to bound $\mathbb{E}\left[\exp^{\lambda Y_n} \;|\mathcal{F}_{n-1}\right]$:

$$\begin{aligned}
\mathbb{E}\left[\exp^{\lambda Y_n} \;|\mathcal{F}_{n-1}\right] &= \mathbb{E}\left[\exp^{\lambda c_n \frac{Y_n}{c_n}} \;|\mathcal{F}_{n-1}\right]\Big] \quad \text{note that } \left|\frac{Y_n}{c_n}\right| \leq 1 \\
&\leq \exp\frac{c_n^2\lambda^2}{2} \quad \text{apply Eq.(40)} \quad \lambda \to \lambda c_n
\end{aligned} \tag{82}$$

2. the second is to apply Hoeffding Lemma (strong), i.e., Eq.(54):

$$\begin{aligned}
&\mathbb{E}\big[\exp^{\lambda(X-\mathbb{E}[X])}\big] \leq \exp\Big(\frac{\lambda^2(b-a)^2}{8}\Big) \\
\Longrightarrow\ &\mathbb{E}\big[\exp^{\lambda Y_n}\ |\mathcal{F}_{n-1}\big] \leq \exp\Big(\frac{\lambda^2\big(c_n-(-c_n)\big)^2}{8}\Big) \\
&= \exp\Big(\frac{c_n^2\lambda^2}{2}\Big)
\end{aligned} \tag{83}$$

doesn't matter which route to take, we have:

$$\begin{aligned}
\mathbb{E}_{X_{1:n}}\big[\exp^{\lambda\left(X_n-X_0\right)}\big] &\leq \mathbb{E}_{X_{1:n-1}}\Big[\exp^{\lambda\left(\sum_{0<i\leq n-1}(X_i-X_{i-1})\right)}\exp\Big(\frac{c_n^2\lambda^2}{2}\Big)\Big] \\
&= \exp\Big(\frac{c_n^2\lambda^2}{2}\Big)\mathbb{E}_{X_{1:n-1}}\Big[\exp^{\lambda\left(\sum_{0<i\leq n-1}(X_i-X_{i-1})\right)}\Big] \\
&= \exp\Big(\frac{c_n^2\lambda^2}{2}\Big)\mathbb{E}_{X_{1:n-1}}\Big[\exp^{\lambda\left(X_{n-1}-X_0\right)}\Big] \\
&= \exp^{\frac{1}{2}\lambda^2\sum_{i=1}^n c_i^2}\,\mathbb{E}_{X_0}\Big[\exp^{\lambda\left(X_0-X_0\right)}\Big] \quad \text{apply recursion} \\
&= \exp^{\frac{1}{2}\lambda^2\sum_{i=1}^n c_i^2}
\end{aligned} \tag{84}$$

substitute it back to Eq.(77):

$$\begin{aligned}
\Pr\big(X_n-X_0\geq\epsilon\big) &\leq \exp^{-\lambda\epsilon}\,\mathbb{E}_{X_{1:n}}\big[\exp^{\lambda(X_n-X_0)}\big] \\
&\leq \exp^{-\lambda\epsilon}\ \exp^{\frac{1}{2}\lambda^2\sum_{i=1}^n c_i^2} \\
&= \exp^{-\lambda\epsilon+\frac{1}{2}\lambda^2\sum_{i=1}^n c_i^2}
\end{aligned} \tag{85}$$

minimizing λ:

$$\begin{aligned}
&\nabla_\lambda(-\lambda\epsilon+\frac{1}{2}\lambda^2 A) = -\epsilon + A\lambda = 0 \\
\Longrightarrow\ &\lambda = \frac{\epsilon}{A} = \frac{\epsilon}{\sum_{i=1}^n c_i^2}
\end{aligned} \tag{86}$$

$$\begin{aligned}
\Pr\big(X_n-X_0\geq\epsilon\big) &\leq \exp^{-\frac{\epsilon}{\sum_{i=1}^n c_i^2}\epsilon+\frac{1}{2}\left(\frac{\epsilon}{\sum_{i=1}^n c_i^2}\right)^2\sum_{i=1}^n c_i^2} \\
&= \exp^{-\frac{\epsilon^2}{\sum_{i=1}^n c_i^2}+\frac{\epsilon^2}{2\sum_{i=1}^n c_i^2}} \\
&= \exp^{-\frac{\epsilon^2}{2\sum_{i=1}^n c_i^2}}
\end{aligned} \tag{87}$$

6.3 Application of Azuma inequality: Stochastic convex optimization

objective function $F(w)$ is defined as:

$$F(w) = \mathbb{E}_{Z\sim\mathcal{D}}[h(w,Z)] \tag{88}$$

and if we minimize this objective function:

$$\begin{aligned} F(w^*) &= \min_{w \in \mathcal{W}} F(w) \\ &= \min_{w \in \mathcal{W}} \mathbb{E}_{Z \sim \mathcal{D}}[h(w, Z)] \end{aligned} \tag{89}$$

First, observe that for each $t \in \{1 : T\}$, we may define the cost function $c_t(\boldsymbol{w}, Z_t) = f(w, Z_t)$. We may thus use **online gradient descent** algorithm to obtain low regret, i.e., to ensure that:

$$R_T = \sum_{t=1}^{T} h(w_t, Z_t) - \inf_{w \in \mathcal{W}} \sum_{t=1}^{T} h(w, Z_t) \tag{90}$$

Note that because of gradient descend:

$$w_t = w_{t-1} + \eta \nabla h(w_{t-1}, Z_{t-1}) \tag{91}$$

So, $\sum_{t=1}^{T} h(w_t, Z_t)$ the accumulated function values (not minimized, and each may have different function values) when w_t is computed with one data at the time in SGD fashion.

On the other hand, $w^* = \inf_{w \in \mathcal{W}} \sum_{t=1}^{T} h(w, Z_t)$ is the batch optimized result combining T data together. (it may use some arbitrary optimizer other than gradient descend)

6.3.1 unbiased estimator

now we have Z_t are i.i.d., then $h(w, Z_t)$ are also i.i.d. when we assume w to be fixed:

$$\begin{aligned} \mathbb{E}_{Z_t \sim \mathcal{D}}[\nabla_{\boldsymbol{w}} h(w, Z_t)] &= \nabla_w \mathbb{E}_{Z_t \sim \mathcal{D}}[h(w, Z_t)] \\ &= \nabla_w \mathbb{E}_{Z \sim \mathcal{D}}[h(w, Z)] \\ &= \nabla_w F(w) \end{aligned} \tag{92}$$

looking at the last line:

$$\nabla_w F(w) = \mathbb{E}_Z[\nabla_w h(w, Z)] \tag{93}$$

for **fixed** action w, the stochastic gradient $\nabla_w h(w, Z_t)$ is an unbiased estimator of the gradient $\nabla_w F(w)$. Taking a step in $-\nabla_w h(w, Z_t)$ should in expectation move us towards optimum w^*.

Online gradient descent uses gradients **evaluated** at played action w_t, which can depend on $Z_{1:t-1}$ and hence w_t is stochastic (having different $Z_{1:t-1}$, online gradient descend may result to different w_t).

However, conditioning on $z_{1:t-1}$, w_t becomes fixed this is because:

$$w_t = w_{t-1} + \eta \nabla h(w_{t-1}, Z_{t-1}) \tag{94}$$

Note that w_t does not depend on Z_t yet. Therefore, same as the case of arbitrary w in Eq.(93). However, we now need to make it depend on $Z_{1:t-1}$:

$$\mathbb{E}_{Z_t \sim \mathcal{D}}\big[h(w_t, Z_t) | Z_{1:t-1}\big] = F(w_t) \tag{95}$$

same happen to gradient descend:

$$\begin{aligned} \mathbb{E}_{Z_t \sim \mathcal{D}}\big[\nabla_w h(w_t, Z_t) \; | Z_{1:t-1}\big] &= \nabla_w \mathbb{E}_{Z_t \sim \mathcal{D}}\big[h(w_t, Z_t) \; | Z_{1:t-1}\big] \\ &= \nabla_w F(w_t) \end{aligned} \tag{96}$$

6.4 online to batch conversion

Suppose an online learning algorithm that plays $w_1, \dots, w_T$ **against** sequence $Z_1, \dots, Z_T$ obtains regret R_T. We will prove that the simple average:

$$\bar{w}_T := \frac{1}{T}\sum_{t=1}^{T} w_t \tag{97}$$

obtains low excess risk whenever R_T is small. Note $\{w_1, \dots, w_T\}$ are fixed, and there is only one Z:

$$\begin{aligned}
F(\bar{w}_T) &= \mathbb{E}_{Z\sim\mathcal{D}}[h(\bar{w}_T, Z)] \\
&= \mathbb{E}_{Z\sim\mathcal{D}}\Big[h\Big(\frac{1}{T}\sum_{t=1}^{T} w_t, Z\Big)\Big] \\
&\le \mathbb{E}_{Z\sim\mathcal{D}}\Big[\frac{1}{T}\sum_{t=1}^{T} h(w_t, Z)\Big] \quad \text{Jensen's inequality}
\end{aligned} \tag{98}$$

$$\begin{aligned}
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z\sim\mathcal{D}}\Big[h(w_t, Z)\Big] \\
&= \frac{1}{T}\sum_{t=1}^{T} F(w_t)
\end{aligned}$$

so we can try to bound the R.H.S $\frac{1}{T}\sum_{t=1}^{T} F(w_t)$:

6.5 in expectation bound

this is to bound the expectation of $\frac{1}{T}\sum_{t=1}^{T} F(w_t)$:

$$\begin{aligned}
\mathbb{E}_{w_1,\dots,w_T}\Big[\frac{1}{T}\sum_{t=1}^{T} F(w_t)\Big] &= \mathbb{E}_{w_1,\dots,w_T}\Big[\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z\sim\mathcal{D}}[h(w_t, Z)]\Big] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{w_1,\dots,w_T,Z}[h(w_t, Z)] \quad \text{basically just bound itself: } \frac{1}{T}\sum_{t=1}^{T} F(w_t) \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{Z_1,\dots,Z_T\sim\mathcal{D}}[h(w_t, Z_t)] \quad w_t \text{ depends on } Z_{1:t-1} \\
&= \frac{1}{T}\mathbb{E}_{Z_1,\dots,Z_T\sim\mathcal{D}}\Big[\inf_{w\in\mathcal{W}}\sum_{t=1}^{T} h(w, Z_t)\Big] + \frac{1}{T}\mathbb{E}[R_T] \\
&\qquad \text{from Eq.(91): } R_T = \sum_{t=1}^{T} h(w_t, Z_t) - \inf_{w\in\mathcal{W}}\sum_{t=1}^{T} h(w, Z_t) \text{ then add } \mathbb{E}[\cdot] \\
&\le \inf_{w\in\mathcal{W}}\mathbb{E}_{Z_1,\dots,Z_T\sim\mathcal{D}}\Big[\frac{1}{T}\sum_{t=1}^{T} h(w, Z_t)\Big] + \frac{1}{T}\mathbb{E}[R_T] \\
&= F(w^*) + \frac{1}{T}\mathbb{E}[R_T]
\end{aligned} \tag{99}$$

So, an upper bound on expected regret implies an upper bound on the expected excess risk of $\bar{w}_T$. With a little more work, we can establish a similar guarantee that holds with high probability with respect to $Z_1, \ldots, Z_T$

6.6 probability bound

We now begin proving a high probability excess risk bound for $\bar{w}_T$. Define for each $t \in [T]$:

$$X_t = f(w^*, Z_t) - f(w_t, Z_t) - \mathbb{E}\big[f(w^*, Z_t) - f(w_t, Z_t)|Z_{1:t-1}\big] \tag{100}$$

the reason why this **definition** is useful, is that we can show:

$$\begin{aligned} \mathbb{E}[|X_t|] &\leq \infty \\ \mathbb{E}[X_t|\mathcal{F}_{t-1}] &= 0 \end{aligned} \tag{101}$$

If we define:

$$\begin{aligned} X_t &= f(w^*, Z_t) - f(w_t, Z_t) - \mathbb{E}\big[f(w^*, Z_t) - f(w_t, Z_t)|Z_{1:t-1}\big] \\ &= f(w^*, Z_t) - f(w_t, Z_t) - \big(F(w^*) - F(w_t)\big) \end{aligned} \tag{102}$$

it's obvious that $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$, or $\mathbb{E}[X_t|g(Z_{t-1}, \ldots Z_1)] = 0$ (see definition at Eq.(74):

$$\begin{aligned} X_t &= f(w^*, Z_t) - f(w_t, Z_t) - \big(F(w^*) - F(w_t)\big) \\ \implies F(w_t) &= F(w^*) + f(w_t, Z_t) - f(w^*, Z_t) + X_t \\ \frac{1}{T}\sum_{t=1}^{T} F(w_t) &= F(w^*) + \frac{1}{T}\sum_{t=1}^{T} f(w_t, Z_t) - \frac{1}{T}\sum_{t=1}^{T} f(w^*, Z_t) + \frac{1}{T}\sum_{t=1}^{T} X_t \\ &= F(w^*) + \frac{1}{T}\sum_{t=1}^{T} f(w_t, Z_t) - \inf_{w \in \mathcal{W}} \Big(\frac{1}{T}\sum_{t=1}^{T} f(w, Z_t)\Big) + \frac{1}{T}\sum_{t=1}^{T} X_t \end{aligned} \tag{103}$$

so the remaining is to bound $\sum_{t=1}^{T} X_t$

7 McDiarmid's Inequality

Theorem 12 *Let $X_1, \ldots, X_m$ be i.i.d random variables. let $f : \mathcal{X}^m \to \mathbb{R}$ be a function of $X_1, \ldots, X_m$ that satisfies $\forall x_1, \ldots, x_m, x_i' \in \mathcal{X}$*

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i \tag{104}$$

then for all $\epsilon > 0$:

$$\Pr\big(f(x_1, \ldots, x_m) - \mathbb{E}[f(x_1, \ldots, x_m)] \geq \epsilon\big) \leq \exp\Big(\frac{-2\epsilon^2}{\sum_{i=1}^{m} c_i^2}\Big) \tag{105}$$

7.1 proof

7.1.1 $Z_i = \mathbb{E}\big[f(X_{1:m}) \mid X_{1:i}\big]$

Define random variables:

$$
\begin{aligned}
Z_i &= \mathbb{E}\big[f(X_{1:m}) \mid X_{1:i}\big] \\
&\equiv \mathbb{E}_{1:m}\big[f(X_{1:m}) \mid X_{1:i}\big] \\
&= \mathbb{E}_{X_{i+1:m}}\big[f(X_{1:m}) \mid X_{1:i}\big] \\
&= g(X_{1:i})
\end{aligned}
\tag{106}
$$

1. (Z_i) form a Martingale difference sequence (we need to prove), in which case we can use Azuma inequality to prove the bound $Z_m - Z_0$
2. This is a peculiar way of defining random variable, as $X_{1:i}$ is a random variable, but they are used in the condition. Therefore remaining variables $X_{i+1:m}$ gets integrated out. So one can consider Z_i as some function $g \in \mathbb{R}$ of random variable $X_{1:i}$, i.e., $g(X_{1:i})$.
3. When we have random variable of the form $f(Y)|Y$. Ordinarily, $\mathbb{E}[f(Y)|Y] = \mathbb{E}[f(Y)]$ if integral is performed on entire Y. However, when we condition on part of dimension of Y, the integral only applies to the remainder dimension of Y.

It has a few properties, where the two "end" terms are:

$$
\begin{aligned}
Z_0 &= \mathbb{E}\big[f(X_{1:m})\big] \quad \text{constant value} \\
Z_m &= \mathbb{E}\big[f(X_{1:m}) \mid X_{1:m}\big] \\
&= f(X_{1:m}) \quad \text{all arguments are random variables}
\end{aligned}
\tag{107}
$$

This tells us that we can use them for telescope.

7.1.2 $\mathbb{E}\big[Z_i - Z_{i-1} \mid X_{1:i-1}\big]$

by definition, we have:

$$
\mathbb{E}\big[Z_i - Z_{i-1} \mid X_{1:i-1}\big] = \mathbb{E}_{X_{1:m}}\Big[\mathbb{E}_{i+1:m}\big[f(X_{1:m}) \mid X_{1:i}\big] - \mathbb{E}_{i:m}\big[f(X_{1:m}) \mid X_{1:i-1}\big] \,\Big|\, X_{1:i-1}\Big]
\tag{108}
$$

for L.H.S:

$$
\begin{aligned}
\mathbb{E}\big[Z_i \mid X_{1:i-1}\big] &\equiv \mathbb{E}_{X_{i:m}}\Big[\underbrace{\mathbb{E}_{X_{i+1:m}}\big[f(X_{1:m}) \mid X_{1:i}\big]}_{g(X_{1:i})} \,\Big|\, X_{1:i-1}\Big] \quad \text{we use appropriate integrand index} \\
&\equiv \mathbb{E}_{X_i}\Big[\mathbb{E}_{X_{i+1:m}}\big[f(X_{1:m}) \mid X_{1:i}\big] \,\Big|\, X_{1:i-1}\Big] \quad \text{for outer integral } X_{i+1:m} \text{ are integrated out} \\
&= \mathbb{E}_{X_{i:m}}\big[f(X_{1:m}) \,\big|\, X_{1:i-1}\big]
\end{aligned}
\tag{109}
$$

for R.H.S:

$$
\begin{aligned}
\mathbb{E}\big[Z_{i-1} \mid X_{1:i-1}\big] &\equiv \mathbb{E}_{X_{i:m}}\Big[\underbrace{\mathbb{E}_{X_{i:m}}\big[f(X_{1:m}) \mid X_{1:i-1}\big]}_{g(X_{1:i-1})} \,\Big|\, X_{1:i-1}\Big] \\
&= \mathbb{E}_{X_{i:m}}\big[f(X_{1:m}) \,\big|\, X_{1:i-1}\big] \quad \text{outer integral has no effect}
\end{aligned}
\tag{110}
$$

since LHS and RHS are the same, then we have:

$$
\mathbb{E}\big[Z_i - Z_{i-1} \,\big|\, X_{1:i-1}\big] = 0
\tag{111}
$$

7.1.3 bounds for $Z_i - Z_{i-1} \mid X_{1:i-1}$

now know the Random variable $Z_i - Z_{i-1} \mid X_{1:i-1}$ has mean of zero, then what may be its bound?

now the raw random variable $Z_i - Z_{i-1} \mid X_{1:i-1}$ may be a bit confusing at first. However, we can treat $X_{1:i-1}$ as a fixed value.

Also $Z_i - Z_{i-1} \mid X_{1:i-1} \equiv g(X_{1:i}) - g(X_{1:i-1})$. Important thing is that both are dependent on the **same** $X_{1:i-1}$. That's the reason why we need to have $Z_i - Z_{i-1} \mid X_{1:i-1}$ instead of just $Z_i - Z_{i-1}$

Compare $Z_i|X_{1:i-1}$ with $Z_{i-1}|X_{1:i-1}$, LHS has one random variable Z_i so we can maximize/minimize. This is because:

1. $(X_{i+1:m}$ are used to compute the expectation, so it's the same LHS and RHS.
2. $X_{1:i-1}$ are the same, so they cancel out

therefore, we condition on $X_{1:i-1}$:

$$\begin{aligned} U_i &= \sup_{u} \left\{ \mathbb{E}[f(X_{1:m}) \mid \underbrace{X_{1:i-1}}_{\text{same}}, X_i = u] - \mathbb{E}[f(X_{1:m}) \mid \underbrace{X_{1:i-1}}_{\text{same}}] \right\} \\ L_i &= \inf_{l} \left\{ \mathbb{E}[f(X_{1:m}) \mid \underbrace{X_{1:i-1}}_{\text{same}}, X_i = l] - \mathbb{E}[f(X_{1:m}) \mid \underbrace{X_{1:i-1}}_{\text{same}}] \right\} \end{aligned} \tag{112}$$

then:

$$\begin{aligned} U_i - L_i = &\sup_{u} \left\{ \mathbb{E}[f(X_{1:m}) \mid X_{1:i-1}, X_i = u] - \mathbb{E}[f(X_{1:m}) \mid X_{1:i-1}] \right\} \\ &\qquad - \inf_{l} \left\{ \mathbb{E}[f(X_{1:m}) \mid X_{1:i-1}, X_i = l] - \mathbb{E}[f(X_{1:m}) \mid X_{1:i-1}] \right\} \\ = &\sup_{u} \left\{ \mathbb{E}[f(X_{1:m}) \mid X_{1:i-1}, X_i = u] \right\} - \inf_{l} \left\{ \mathbb{E}[f(X_{1:m}) \mid X_{1:i-1}, X_i = l] \right\} \end{aligned} \tag{113}$$

By taking expectation of its argument, $\mathbb{E}[f(X_{1:m})]$ is still in the range of $f(X_{1:m})$. So the assumption in **theorem 12** still applies. So by our assumption:

$$\begin{aligned} |f(x_1, \dots, x_i, \dots, x_m) - f(x_1, ..., x_i', \dots, x_m)| &\leq c_i \\ \implies U_i - L_i &\leq c_i \end{aligned} \tag{114}$$

Eq.(114) can be understood by by the fact that the condition:

$$\underbrace{U_i - L_i}_{\text{LHS}} \leq \underbrace{\max\left(|f(x_1, \dots, x_i, \dots, x_m) - f(x_1, ..., x_i', \dots, x_m)|\right)}_{\text{RHS}} \leq c_1 \tag{115}$$

as for the R.H.S, one may choose any $\{x_{j\neq i}\}$, whereas when we compute LHS, the set of $\{x_{j\neq i}\}$ are more constrained to their expectations, hence the difference is smaller.

it becomes clear that:

$$\begin{aligned} &\mathbb{E}_{Z_{1:m}}\left[\exp^{\lambda\left(z_i - z_{i-1}\right)} |X_{1:i-1}\right] \\ &\qquad\qquad \text{note we do \textbf{not} express as } \mathbb{E}_{Z_{1:m}}\left[\exp^{\lambda\left(z_i - z_{i-1}|X_{1:i-1}\right)}\right] \\ &\equiv \mathbb{E}_{Z_m}\left[\exp^{\lambda\left(z_i - z_{i-1}\right)} |X_{1:i-1}\right] \\ &\leq \exp\left(\frac{c_i^2\lambda^2}{8}\right) \quad \text{by strong Hoeffding lemma} \end{aligned} \tag{116}$$

the rest then becomes easier to deal with, one need to recognize:

$$
\begin{aligned}
f(X_{1:m}) - \mathbb{E}[f(X_{1:m})] = Z_m - Z_0 \quad & \text{Eq.(107)} \\
= \sum_{i=1}^{m} Z_i - Z_{i-1} \quad & \text{telescope sum}
\end{aligned}
\tag{117}
$$

$$
\begin{aligned}
&\Pr(f(X_{1:m}) - \mathbb{E}[f(X_{1:m})] \geq \epsilon) \\
&\quad \leq \min_{\lambda} \Big\{ \exp^{-\lambda\epsilon} \, \mathbb{E}_{X_{1:m}} \big[\exp^{\lambda\big(f(X_{1:m}) - \mathbb{E}[f(X_{1:m})]\big)} \big] \Big\} \\
&\quad = \min_{\lambda} \Big\{ \exp^{-\lambda\epsilon} \, \mathbb{E}_{\underbrace{X_{1:m}}_{U}} \big[\exp^{\lambda \sum_{i=1}^{m} Z_i - Z_{i-1}} \big] \Big\} \\
&\quad = \min_{\lambda} \Big\{ \exp^{-\lambda\epsilon} \, \mathbb{E}_{\underbrace{X_{1:m-1}}_{V}} \Big[\mathbb{E}_{\underbrace{X_{1:m}}_{U}} \big[\exp^{\lambda \sum_{i=1}^{m} Z_i - Z_{i-1}} \mid \underbrace{X_{1:m-1}}_{V} \big] \Big] \Big\} \quad \text{law of total expectation} \\
&\quad = \min_{\lambda} \Big\{ \exp^{-\lambda\epsilon} \, \mathbb{E}_{X_{1:m-1}} \Big[\mathbb{E}_{X_{1:m}} \big[\exp^{\lambda \sum_{i=1}^{m-1} Z_i - Z_{i-1}} \times \exp^{\lambda Z_m - Z_{m-1}} \mid X_{1:m-1} \big] \Big] \Big\} \\
&\quad = \min_{\lambda} \Big\{ \exp^{-\lambda\epsilon} \, \mathbb{E}_{X_{1:m-1}} \Big[\mathbb{E}_{X_{1:m-1}} \big[\exp^{\lambda \sum_{i=1}^{m-1} Z_i - Z_{i-1}} \big] \times \mathbb{E}_{Z_m} \big[\exp^{\lambda Z_m - Z_{m-1}} \mid X_{1:m-1} \big] \Big] \Big\} \\
&\quad \leq \min_{\lambda} \Big\{ \exp^{-\lambda\epsilon} \, \mathbb{E}_{X_{1:m-1}} \Big[\mathbb{E}_{X_{1:m-1}} \big[\exp^{\lambda \sum_{i=1}^{m-1} Z_i - Z_{i-1}} \big] \exp\Big(\frac{c_i^2 \lambda^2}{8}\Big) \Big] \Big\} \\
&\quad \leq \min_{\lambda} \Big\{ \exp\big(-\lambda\epsilon\big) \exp\Big(\frac{c_i^2\lambda^2}{8}\Big) \mathbb{E}_{X_{1:m-1}} \Big[\exp^{\lambda \sum_{i=1}^{m-1} Z_i - Z_{i-1}} \Big] \Big\} \quad \text{two } \mathbb{E}_{X_{1:m-1}} \big[\mathbb{E}_{X_{1:m-1}}[\cdot]\big] \text{ can be combined} \\
&\quad \leq \min_{\lambda} \Big\{ \exp\Big(-\lambda\epsilon + \frac{\lambda^2}{8} \sum_{i=1}^{m} c_i^2 \Big) \Big\}
\end{aligned}
\tag{118}
$$

the rest of the proof are the same as Eq.(87) of Azuma inequality, except we have $|[a,b]| = 2c_i \rightarrow [a,b] = c_i$, so instead of $\leq \exp\left(\frac{-\epsilon^2}{2\sum_{i=1}^{m} c_i^2}\right)$, we need to multiply by 2^2 in the exponent.

$$
\Pr\big(f(x_1,\ldots,x_m) - \mathbb{E}[f(x_1,\ldots,x_m)] \geq \epsilon\big) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m} c_i^2}\right) \tag{119}
$$

7.1.4 law of total expectation

$$
\begin{aligned}
\mathbb{E}_V\big[\mathbb{E}_U[U|V]\big] &= \int_v \Big[\int_u uP(U=u|V=v) \Big] P(V=v) \\
&= \int_u u \int_v P(U=u|V=v)P(V=v) \\
&= \int_u u \int_v P(U=u\,,\,V=v) \\
&= \int_u uP(U=u) \\
&= \mathbb{E}[U]
\end{aligned}
\tag{120}
$$

7.2 relationship with Heoffding's inequality

Let $f(X_1,\ldots,X_m) = \frac{1}{m}\sum_{i=1}^{m} X_i$, then we get back Hoeffding's inequality and let $|X_i - X_i'| < c_i$

$$
\begin{aligned}
|f(X_1,\dots,X_i,\dots,X_m)-f(X_1,\dots,X_i',\dots,X_m)| &= \Big|\frac{1}{m}\sum_{i=1}^{m}X_i - \big(\frac{1}{m}\sum_{j=1,j\neq i}^{m}X_i + X_i'\big)\Big| \\
&= \frac{|X_i-X_i'|}{m} \\
&\leq \frac{c_i}{m}
\end{aligned}
\tag{121}
$$

then:

$$
\begin{aligned}
\Pr\big(f(x_1,\dots,x_m)-\mathbb{E}[f(x_1,\dots,x_m)]\geq\epsilon\big) &\leq \exp\Big(\frac{-2\epsilon^2}{\sum_{i=1}^{m}c_i^2}\Big) \\
\Pr\big(\overline{X}-\mu\geq\epsilon\big) &\leq \exp\Big(\frac{-2\epsilon^2}{\sum_{i=1}^{m}\big(\frac{c_i}{m}\big)^2}\Big) \\
&= \exp\Big(\frac{-2m^2\epsilon^2}{\sum_{i=1}^{m}c_i^2}\Big) \\
&= \exp\Big(-\frac{2m^2\epsilon^2}{\sum_{i=1}^{m}(b_i-a_i)^2}\Big)
\end{aligned}
\tag{122}
$$

8 Bernstein inequality

all Heoffding's inequality derivatives only needing input X to be with a certain range. However, if Var$[X]$ is also known, then we can have an even tighter bound.

8.1 Bernstein Lemma

Just like the hoeffding lemma (which bounds the MGF almost surely) used to help the proof of hoeffding inequality, we also need **Bernstein Lemma** to bound the MGF a.s.:

Lemma 13 *Suppose that $|X|\leq c$ and $\mathbb{E}[X]=0$ For any $\lambda>0$:*

$$
\mathbb{E}\big[\exp^{\lambda X}\big] \leq \exp\Big(\lambda^2\sigma^2\Big(\frac{\exp^{\lambda c}-1-\lambda c}{(\lambda c)^2}\Big)\Big)
\tag{123}
$$

where $\sigma^2 = Var(X)$

basically Lemma (13) is to express MGF in a way that to make Var$[X]=\sigma^2$ explicit in the MGF bound. Note that this applies to all random variable X with condition set out in the Lemma (**??**):

$$
\begin{aligned}
\mathbb{E}\big[\exp^{\lambda X}\big] &= \mathbb{E}\Big[1+\lambda\mathbf{x}+\sum_{r=2}^{\infty}\frac{\lambda^r X^r}{r!}\Big] \\
&= 1+\lambda\mathbb{E}[\mathbf{x}]+\sum_{r=2}^{\infty}\frac{\lambda^r\mathbb{E}[X^r]}{r!} \\
&= 1+\lambda^2\sigma^2\underbrace{\sum_{r=2}^{\infty}\frac{\lambda^{r-2}\mathbb{E}[X^r]}{r!\sigma^2}}_{F} \\
&= 1+\lambda^2\sigma^2 F \\
&\leq \exp^{\lambda^2\sigma^2 F} \quad \because 1+x\leq\exp^x \quad \text{for } x>0
\end{aligned}
\tag{124}
$$

so now we can obtain the expression of $\mathbb{E}[X^r]$, so we need to somehow put $\mathbb{E}[X^2] = \text{Var}[X] = \sigma^2$ (when $\mathbb{E}[X] = 0$) into this:

for $r \geq 2$:

$$\mathbb{E}[X^r] = \mathbb{E}[X^{r-2}X^2] \tag{125}$$

Lemma 14

$$\mathbb{E}[fg] \leq \mathbb{E}[f] \max_x \big(|g(x)|\big) \tag{126}$$

this is because if $f \geq 0$, then $fg \leq f \max_x \big(|g(x)|\big)$:

although we have no control over the sign of X^{r-2}, but we do know that $X^2 \geq 0$, so by letting $f \equiv X^2$, we have:

$$\begin{aligned} \mathbb{E}[X^r] &\leq \mathbb{E}[X^2] \max\big(|X^{r-2}|\big) \\ &\leq \sigma^2 c^{r-2} \end{aligned} \tag{127}$$

let's go back to the defintion of F:

$$\begin{aligned} F &= \sum_{r=2}^{\infty} \frac{\lambda^{r-2}\mathbb{E}[X^r]}{r!\sigma^2} \\ &\leq \sum_{r=2}^{\infty} \frac{\lambda^{r-2}\sigma^2 c^{r-2}}{r!\sigma^2} \\ &= \sum_{r=2}^{\infty} \frac{\lambda^{r-2}c^{r-2}}{r!} \times \frac{\lambda^2 c^2}{\lambda^2 c^2} \\ &= \frac{1}{(\lambda c)^2} \sum_{r=2}^{\infty} \frac{(\lambda c)^r}{r!} \end{aligned} \tag{128}$$

the term $\sum_{r=2}^{\infty} \frac{(\lambda c)^r}{r!}$ looks suspiciously familiar, as we have:

$$\begin{aligned} \exp^{\lambda c} &= \sum_{r=1}^{\infty} \frac{(\lambda c)^r}{r!} \\ \implies \sum_{r=2}^{\infty} \frac{(\lambda c)^r}{r!} &= \exp^{\lambda c} -1 - \lambda c \end{aligned} \tag{129}$$

Substituting Eq.(130) into the above into Eq.(128), we have:

$$F \leq \frac{\exp^{\lambda c} -1 - \lambda c}{(\lambda c)^2} \tag{130}$$

so looking at Eq.(124), we have:

$$\begin{aligned} \mathbb{E}\big[\exp^{\lambda X}\big] &\leq \exp^{\lambda^2\sigma^2 F} \\ &\leq \exp^{\lambda^2\sigma^2 \frac{\exp^{\lambda c} -1-\lambda c}{(\lambda c)^2}} \end{aligned} \tag{131}$$

8.2 Bernstein Inequality

Theorem 15 *If $|X_i| \leq c$ and $\mathbb{E}[X_i] = \mu$, then for any $\epsilon > 0$:*

$$\Pr(|\bar{X}_n - \mu| > \epsilon) \leq 2\exp\Big(- \frac{n\epsilon^2}{2\sigma^2 + \frac{2c\epsilon}{3}}\Big) \tag{132}$$

where $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} Var(X_i)$

obviously, the smaller the σ^2, the smaller the bound. Making things simple, we let $\mu = 0$, then we apply Lemma 13:

$$\begin{aligned}
\Pr(\bar{X}_n > \epsilon) &= \Pr\big(\sum_{i=1}^{n} X_i > n\epsilon\big) \\
&= \Pr\Big(\exp^{\lambda \sum_{i=1}^{n} X_i} > \exp^{\lambda n\epsilon}\Big) \\
&\leq \frac{\mathbb{E}\big[\exp^{\lambda \sum_{i=1}^{n} X_i}\big]}{\exp^{\lambda n\epsilon}} = \frac{\mathbb{E}\big[\prod_{i=1}^{n} \exp^{\lambda X_i}\big]}{\exp^{\lambda n\epsilon}} \\
&\leq \frac{\prod_{i=1}^{n} \mathbb{E}\big[\exp^{\lambda X_i}\big]}{\exp^{\lambda n\epsilon}} \quad X_i \text{ are independent} \\
&\leq \frac{\prod_{i=1}^{n} \exp\Big(\lambda^2\sigma^2\Big(\frac{\exp^{\lambda c} - 1 - \lambda c}{(\lambda c)^2}\Big)\Big)}{\exp^{\lambda n\epsilon}} \quad \because \text{Bernstein lemma 13} \\
&\leq \frac{\exp\Big(n\lambda^2\sigma^2\Big(\frac{\exp^{\lambda c} - 1 - \lambda c}{(\lambda c)^2}\Big)\Big)}{\exp^{\lambda n\epsilon}}
\end{aligned} \tag{133}$$

In the last expression the terms inside $\prod$ are replaced by their bound, therefore, even if they were not identical before, now their bound are identical (hence there is no index):

let's choose $\lambda = \frac{1}{c}\log\big(1 + \frac{\epsilon c}{\sigma^2}\big)$:

$$\begin{aligned}
\Pr(\bar{X}_n > \epsilon) &\leq \frac{\exp\Big(n\lambda^2\sigma^2\Big(\frac{\exp^{\lambda c} - 1 - \lambda c}{(\lambda c)^2}\Big)\Big)}{\exp^{\lambda n\epsilon}} \\
&= \frac{\exp\Big(n\sigma^2\Big(\frac{\exp^{\lambda c} - 1 - \lambda c}{c^2}\Big)\Big)}{\exp^{\lambda n\epsilon}} \\
&= \frac{\exp\Big(n\sigma^2\Big(\frac{\exp^{\left[\frac{1}{c}\log\left(1 + \frac{\epsilon c}{\sigma^2}\right)\right]c} - 1 - \left[\frac{1}{c}\log\left(1 + \frac{\epsilon c}{\sigma^2}\right)\right]c}{c^2}\Big)\Big)}{\exp^{\left[\frac{1}{c}\log\left(1 + \frac{\epsilon c}{\sigma^2}\right)\right]n\epsilon}} \\
&= \frac{\exp\Big(n\sigma^2\Big(\frac{\left(1 + \frac{\epsilon c}{\sigma^2}\right) - 1 - \log\left(1 + \frac{\epsilon c}{\sigma^2}\right)}{c^2}\Big)\Big)}{\exp^{\left[\frac{1}{c}\log\left(1 + \frac{\epsilon c}{\sigma^2}\right)\right]n\epsilon}} \\
&= \frac{\exp\Big(\frac{n\sigma^2}{c^2}\Big(\big(1 + \frac{\epsilon c}{\sigma^2}\big) - 1 - \log\big(1 + \frac{\epsilon c}{\sigma^2}\big)\Big)\Big)}{\exp^{\left[\frac{1}{c}\log\left(1 + \frac{\epsilon c}{\sigma^2}\right)\right]n\epsilon}} \\
&= \frac{\exp\Big(\frac{n\sigma^2}{c^2}\Big(\frac{\epsilon c}{\sigma^2} - \log\big(1 + \frac{\epsilon c}{\sigma^2}\big)\Big)\Big)}{\exp^{\left[\frac{1}{c}\log\left(1 + \frac{\epsilon c}{\sigma^2}\right)\right]n\epsilon}}
\end{aligned} \tag{134}$$

$$
\begin{aligned}
&= \exp\Big(\frac{n\sigma^2}{c^2}\Big(\frac{\epsilon c}{\sigma^2} - \log\big(1+\frac{\epsilon c}{\sigma^2})\Big)\Big) - \big[\frac{1}{c}\log\big(1+\frac{\epsilon c}{\sigma^2})\big]n\epsilon\Big) \\
&= \exp\Big(-\frac{n\sigma^2}{c^2}\Big(\log\big(1+\frac{\epsilon c}{\sigma^2}) - \frac{\epsilon c}{\sigma^2}\Big)\Big) - \frac{n\epsilon}{c}\log\big(1+\frac{\epsilon c}{\sigma^2})\Big) \\
&= \exp\Big(-\frac{n\sigma^2}{c^2}\Big(\log\big(1+\frac{\epsilon c}{\sigma^2}) - \frac{\epsilon c}{\sigma^2}\Big)\Big) - \frac{n\sigma^2}{c^2}\frac{\epsilon c}{\sigma^2}\log\big(1+\frac{\epsilon c}{\sigma^2})\Big) \\
&= \exp\Big(-\frac{n\sigma^2}{c^2}\Big[\big(1+\frac{\epsilon c}{\sigma^2})\log\big(1+\frac{\epsilon c}{\sigma^2}) - \frac{\epsilon c}{\sigma^2}\Big]\Big) \\
&= \exp\Big(-\frac{n\sigma^2}{c^2}\Big[\big(1+u\log\big(1+u) - u\Big]\Big) \qquad \text{let } u = \frac{\epsilon c}{\sigma^2} \\
&\leq \exp\Big(-\frac{n\sigma^2}{c^2}\Big[\frac{u^2}{2+2u/3}\Big]\Big) \quad \because 1+u\log\big(1+u\big) - u \geq \frac{u^2}{2+2u/3} \\
&= \exp\Big(-\frac{n\sigma^2}{c^2}\left[\frac{\big(\frac{\epsilon c}{\sigma^2}\big)^2}{2+2\big(\frac{\epsilon c}{\sigma^2}\big)/3}\right]\Big) \\
&= \exp\Big(-n\left[\frac{\frac{\epsilon^2}{\sigma^2}}{2+2\big(\frac{\epsilon c}{\sigma^2}\big)/3}\right]\Big) \\
&= \exp\Big(-\frac{n\epsilon^2}{2\sigma^2 + \frac{2\epsilon c}{3}}\Big)
\end{aligned}
\tag{135}
$$