Instructor : Avishek Ghosh

Course website : figure it out (content, assignments @ website, moodle)

Tue / Fri — 5.30 – 7 pm

Extra classes — Saturday 3.30 – 5 pm

Grading :

Homeworks	20%	(2–3)
Midsem	30%	
Endsem	30%	
Scribes	15%	
Class Participation	5%	

What is this course about?

1. Analyzing ML algorithms from a statistical point of view
2. Involves a large of statistical tools / techniques that can be used independently

NOT about

1. Theory of Deep Learning
2. Not particularly algorithmic

Reference :

1. High dimensional Statistics — Martin Wainwright
2. Asymptotic Stats — A. W. Vanderbaart

Classification (Binary)

Given data points $(X_1, Y_1)\ (X_2, Y_2) \dots (X_n, Y_n)$ where $X_i \in \mathcal{X} \subseteq \mathbb{R}^d$ (features)

$Y_i \in \{-1, +1\}$

Goal : Find a classifier $g : \mathcal{X} \to \{-1, +1\}$

How to obtain this? — Notion of loss function

Binary loss (error) : $\mathbb{1}[g(x) \neq Y] = \begin{cases} 1, & \text{if } g(x) \neq Y \\ 0, & \text{if } g(x) = Y \end{cases}$

Construct "Empirical loss" $L_n(g) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{g(x_i) \neq y_i\}$

We select classifier for which $L_n(g)$ is minimized

$$\hat{g}_n = \arg\min_{g \in \mathcal{C}} L_n(g)$$, where $\mathcal{C}$ denotes family of classifier

<u>Problem</u> : 1. Performance on "unseen" data is not considered

2. $\mathcal{C}$ can be complicated, n can be much smaller

Upto this point, purely empirical (no statistics)

<u>Statistical Model</u>

We assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d samples from a joint distribution $\mathcal{D}$ having same distribution as (X, Y)

Then, for a classifier $g: \mathcal{X} \to \{\pm 1\}$, we can write,

$$\underbrace{L(g)}_{\text{Avg loss / expected loss}} = \mathbb{E}_{(X,Y)\sim\mathcal{D}} \left[\mathbb{1}\{g(X) \neq Y\}\right] = \mathbb{P}(g(X) \neq Y)$$

It is a good idea to study $L(\hat{g}_n)$

2 Questions :

1. Is $L(\hat{g}_n)$ comparable to $\overbrace{\inf_{g \in \mathcal{C}} L(g)}^{\text{Naive Bayes classifier}}$?

whether $\hat{g}_n$ is comparable with the best classifier in $\mathcal{C}$

2. Is $L(\hat{g}_n)$ comparable to $L_n(\hat{g}_n)$?

Comparison between "in-sample" error and average error for $\hat{g}_n$

Assume $g^* = \arg\min_{g \in \mathcal{C}} L(g)$ {Naive Bayes}

We write $L(\hat{g}_n) = L(g^*) + L(\hat{g}_n) - L_n(\hat{g}_n) + \overbrace{L_n(\hat{g}_n)}^{\text{lesser loss}} - L(g^*)$

$$\leq L(g^*) + \underbrace{L(\hat{g}_n) - L_n(\hat{g}_n)}_{\wedge} + \underbrace{L_n(g^*) - L(g^*)}$$

$$\Rightarrow L(\hat{g}_n) - L(g^*) \leq \sup_{g \in \mathcal{C}} |L(g) - L_n(g)| + \sup_{g \in \mathcal{C}} |L_n(g) - L(g)|$$

$$\Rightarrow \quad L(\hat{g}_n) - L(g^*) \le 2 \sup_{g \in \mathcal{C}} |L_n(g) - L(g)| \qquad - (*)$$

Remark : ① is controlled by $(*)$

② $$L(\hat{g}_n) - L_n(\hat{g}_n) \le \sup_{g \in \mathcal{C}} |L_n(g) - L(g)|$$

Remark : Performance of $\hat{g}_n$ is governed by $\sup_{g \in \mathcal{C}} |L_n(g) - L(g)|$.

We use uniform law of large numbers to handle this

Empirical Process Theory

1. Uniform law of large numbers
2. Uniform central limit theorem

Uniform Law of large Nos.

Suppose $X_1, X_2, \dots, X_n$ i.i.d random objects taking value in $\mathcal{X}$. Let $\mathcal{F}$ be class of real-valued function on $\mathcal{X}$, what can we say about

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right| = Z$$

In particular,

① whether $Z \to 0$ when n is large ?

② Can we obtain non-asymptotic guarantees ? i.e. guarantees for every n

③ Can we provide conditions on f s.t. Z converges to 0 ?

Connection to ML and statistics

① Binary Classification —

$$X_i \mapsto (X_i, Y_i)$$

$$\mathcal{F} \mapsto \{ \mathbb{1}\{g(x) \neq y\} : g \in \mathcal{C} \}$$

② M - estimation

$$\hat{\theta}_n = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m_\theta(x_i)$$

think as −ve of loss fn.

where $x_1 \cdots x_n$ are i.i.d observations, Θ is parameter space

m_θ are real valued fn parametrized by θ.

Examples, 1. $m_\theta(x) = \log p_\theta(x)$ ← Maximum likelihood estimator (MLE)

2. $m_\theta(x) = -(x-\theta)^2$ ← sample mean (Mean estimator)

3. $m_\theta(x) = -|x-\theta|$ ← Median estimator

In mean estimation, target quantity for $\hat{\theta}_n$ is

$$\theta^* = \arg\max_{\theta \in \Theta} \mathbb{E}\, m_\theta(x)$$

Similar to binary classification example we want $\underbrace{d(\hat{\theta}_n, \theta^*)}_{\text{distance b/w } \hat{\theta}_n \text{ and } \theta^*}$ to be small

It turns out $d(\hat{\theta}_n, \theta^*)$ is governed by $2 \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} m_\theta(x_i) - \mathbb{E}\, m_\theta(x) \right|$

which is an instance of uniform law of large numbers.

Strategy to Control Z

① Key observation : Z concentrates around $\mathbb{E} Z$ i.e. $\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \mathbb{E} f(x) \right|$

(concentration of measure)

② We control $\mathbb{E} Z$ through techniques like symmetrization (Rademacher Complexity)

or chaining (vc dimension)

Remark : (Asymptotic result) $\mathcal{F}$ is called "Gliven Ko – Cantelli" if $Z \to 0$

almost - surely as $n \to \infty$

Assumption : $\sup_x |f(x)| \le B \quad \forall f \in \mathcal{F}$

McDiarmid's inequality

Suppose $X_1, X_2, \dots X_n$ and $g : \mathcal{X} \times \dots \times \mathcal{X}_n \to \mathbb{R}$ satisfies "bounded difference".

$$|g(x_1, \dots, x_n) - g(x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \le c_i$$

$$\forall x_1 \dots x_n \quad \forall i \in [n]$$

Then we have

$$\mathbb{P}\{ g(x_1 \dots x_n) - \mathbb{E}\, g(x_1 \dots x_n) \ge t) \le \exp\left(- \frac{2t^2}{\sum_{i=1}^n c_i^2} \right)$$

$$\le -t \qquad \le \quad " \quad - \quad "$$

Rmk : The bounded difference says that a f^n that is not too sensitive on any of its argument concentrates.

Apply Mcdiarmid's to Z

$$Z = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E} f(X) \right|$$

we construct

$$g(x_1 \dots x_n) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(x_i) - \mathbb{E} f(x) \right|$$

$$g(x_1 \dots x_i' \dots x_n) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j \ne i} f(x_j) + \frac{f(x_i')}{n} - \mathbb{E} f(x) \right|$$

$$= \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{j=1}^n f(x_j) - \mathbb{E} f(x) + \frac{f(x_i')}{n} - \frac{f(x_i)}{n} \right|$$

$$\le g(x_1 \dots x_n) + \sup_{f \in \mathcal{F}} \left| \frac{f(x_i')}{n} \right| + \sup_{f \in \mathcal{F}} \left| \frac{f(x_i)}{n} \right|$$

$$\therefore \underbrace{|g(x_1 \dots x_n) - g(x_1 \dots x_i' \dots x_n)|}_{\text{by switching } x_i, x_i' \text{ to get both side}} \le \boxed{\frac{2B}{n} = c_i}$$

Hence, g satisfies bounded difference $\Rightarrow$ apply mcdiarmid's inequality.

$$\mathbb{P}(Z - \mathbb{E}Z \geq t) \leq \exp\left(\frac{-2t^2}{\sum_i \frac{4B^2}{n^2}}\right) = \exp\left(-\frac{nt^2}{2B^2}\right)$$

$$\text{lly} \quad \mathbb{P}(Z - \mathbb{E}Z \leq -t) \leq \underbrace{\exp\left(-\frac{nt^2}{2B^2}\right)}_{=\delta}$$

Then we say w.p. $\geq 1-\delta$,

$$Z \leq \mathbb{E}Z + \underbrace{B\sqrt{\frac{2}{n}\log\frac{1}{\delta}}}_{t \text{ (very small bcz } \propto \frac{1}{\sqrt{n}})}$$

Remark : We need to control $\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X) \right|$

To control Z ① $Z \to \mathbb{E}Z$ ✓

② control $\mathbb{E}Z$?

Concentration Inequality (Hoeffding)

Suppose indep. $X_1 \dots X_n$ r.v. such that $a_i \leq X_i \leq b_i$ $\overbrace{\text{almost surely}}^{\text{w.p. } 1}$

where $a_1, \dots a_n, b_1, \dots, b_n$ are real numbers. Then for any $t \geq 0$

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

and

$$\mathbb{P}\left\{\sum_{i=1}^n (X_i - \mathbb{E}X_i) \leq -t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Proof : Let $S = \sum_{i=1}^n (X_i - \mathbb{E}X_i)$. Fix $\lambda \geq 0$. we have

$$\mathbb{P}(S \geq t) = \mathbb{P}(e^{\lambda S} \geq e^{\lambda t}) \underset{\text{Markov}}{\leq} \frac{\mathbb{E}(e^{\lambda S})}{e^{\lambda t}} = e^{-\lambda t}\,\mathbb{E}e^{\lambda t} = \exp(-\lambda t + \psi_S(t)) \quad \text{— ①}$$

HW: Find Ω for which $e^{\lambda S} \geq e^{\lambda t}$ is true but $S \not\geq t$ $\exists x \cdot t$

$\psi_S(t) = \log \mathbb{E} e^{\lambda S}$ (log-mgf / cumulant fn)

$$\psi_S(\lambda) = \log \mathbb{E}\, e^{\lambda \sum_i (X_i - \mathbb{E}X_i)} \xrightarrow{\text{independence}} = \log \prod_{i=1}^{n} \mathbb{E}\, e^{\lambda(X_i - \mathbb{E}X_i)}$$

$$= \sum_{i=1}^{n} \underbrace{\log \mathbb{E}\, e^{\lambda(X_i - \mathbb{E}X_i)}}$$

$$= \sum_{i=1}^{n} \psi_{X_i - \mathbb{E}X_i}(\lambda) \quad \text{— (1)}$$

Fix i, we analyze

$\psi_{\underbrace{X_i - \mathbb{E}X_i}_{U}}(\lambda)$. Need to bound ψ_U

$b_i - \mathbb{E}X_i \le U \le a_i - \mathbb{E}X_i$ almost surely.

Taylor series,

$$\psi_U(\lambda) = \cancel{\psi_U(0)}^{0} + \cancel{\psi_U'(0)}^{0}(\lambda) + \overbrace{\psi_U''(c)\left(\frac{\lambda^2}{2}\right)}^{\text{key term.}}, \quad 0 < c < \lambda$$

* $\psi_U(0) = \log \mathbb{E}\, e^{0(X_i - \mathbb{E}X_i)} = 0$

* $\psi_U'(0) = \dfrac{d}{d\lambda} \log \mathbb{E}\, e^{\lambda U} = \dfrac{1}{\mathbb{E}e^{\lambda U}} \mathbb{E}(e^{\lambda U} \cdot U)$

Put $\lambda = 0$ $\quad = \dfrac{\mathbb{E}[U]}{(\ )} = 0 ..$

$$\psi_U(\lambda) = \psi_U''(c)\, \frac{\lambda^2}{2}$$

$$\psi_U''(\lambda) = \frac{d}{d\lambda} \frac{1}{\mathbb{E}e^{\lambda U}} \mathbb{E}(U e^{\lambda U}) = \frac{\mathbb{E}(e^{\lambda U})\,\mathbb{E}(U^2 e^{\lambda U}) - \mathbb{E}(U e^{\lambda U})^2}{\mathbb{E}(e^{\lambda U})^2}$$

We will show : $\psi_U''(\lambda) \ge 0$ for any λ

Consider a r.v. V whose density w.r.t. density of U is $e^{\lambda U} / \mathbb{E} e^{\lambda U}$

$$\underbrace{P_V}_{\text{density fn for random var. } V} = P_u \frac{e^{\lambda u}}{\mathbb{E}e^{\lambda u}} \qquad (u = 0 \Rightarrow P_u = 0 \Rightarrow P_V = 0)$$

$$\psi_U''(\lambda) = \mathbb{E}(V^2) - (\mathbb{E}V)^2$$ ← variance of V

≥ 0.

Rmk : Support of U is $a_i - \mathbb{E}X_i$, $b_i - \mathbb{E}X_i$

support of V is " ——— " ($P_U = 0 \Rightarrow P_V = 0$)

Exercise : For any random variable V bound,

$$\mathrm{Var}(V) \leq \left(\frac{b_i - a_i}{4}\right)^2$$ ← $1/2$ pr a_i, $1/2$ b_i

$$\Rightarrow \psi_U(\lambda') = \mathrm{Var}(V) \leq \frac{(b_i - a_i)^2}{4}$$

we have,

$$\boxed{\psi_U(\alpha) = \frac{\lambda^2}{2}\psi_U''(x) \leq \frac{\lambda^2}{8}(b_i - a_i)^2}$$

Now, substituting (ii)

$$\psi_S(\lambda) = \sum_{i=1}^{n} \psi_{X_i - \mathbb{E}X_i}(\lambda) \leq \sum_{i=1}^{n} \frac{\lambda^2}{8}(b_i - a_i)^2$$

Substituting this in (i),

$$\Pr(S \geq t) \leq \exp\left(-\lambda t + \sum_{i=1}^{n} \frac{\lambda^2}{8}(b_i - a_i)^2\right)$$

optimizing over λ, put $\lambda^* = \dfrac{4t}{\sum (b_i - a_i)^2}$

Putting $\lambda = \lambda^*$,

$$\Pr(S \geq t) \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

To get other side of concentration put $Y_i = -X_i$

Reading assignment:

1. Go through proof of lec 2

2. Revise martingales

Central Limit Theorem

$X_1, X_2, \ldots, X_n$ independent with mean μ, variance σ^2

$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

(Reading: Convergence of r.v. convergence in distribution)

CLT: $\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \xrightarrow[\text{distribution}]{n\to\infty} N(0,1)$

Suppose n is very large (for $t \geq 0$)

$$\mathbb{P}\left(\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma}\right) \geq t\right) \approx \mathbb{P}\left(N(0,1) \geq t\right)$$

Scaling r.v. by σ, variance scales by σ^2.

$$\mathbb{P}\left(\sqrt{n}\left(\overline{X}_n - \mu\right) \geq t\right) \approx \mathbb{P}\left(N(0,\sigma^2) \geq t\right)$$

We use MGF based method to upperbound

$$\mathbb{P}\left(N(0,\sigma^2) \geq t\right) \leq \mathbb{P}\left(e^{\lambda N(0,\sigma^2)} \geq e^{\lambda t}\right)$$

$$\leq e^{-\lambda t}\, \mathbb{E}\, e^{\lambda N(0,\sigma^2)}$$

$$= \exp\left(-\lambda t + \underbrace{\psi_{N(0,\sigma^2)}}_{\text{log-mgf / cumulant}}(\lambda)\right)$$

$\psi_{N(0,\sigma^2)}(\lambda) = \log \mathbb{E}\, e^{\lambda N(0,\sigma^2)}$

$$\mathbb{E}\, e^{\lambda N(0,\sigma^2)} = e^{\frac{\sigma^2\lambda^2}{2}} \quad \text{(Exercise)}$$

Using this,

$$\mathbb{P}\left(N(0,\sigma^2) \geq t\right) \leq \exp\left(\underbrace{-\lambda t + \frac{\sigma^2\lambda^2}{2}}_{\text{optimal } \lambda}\right)$$

$$\boxed{\mathbb{P}\left(N(0,\sigma^2) \geq t\right) \leq \exp\left(\frac{-t^2}{2\sigma^2}\right)}$$

* CLT: $\boxed{\mathbb{P}\left(\sqrt{n}\left(\overline{X}_n - \mu\right) \geq t\right) \lessapprox \exp\left(-\frac{t^2}{2\sigma^2}\right)}$ — (I) (CLT)

Using hoeffding's we get,

$X_1 \dots X_n$, $\mathbb{E}[X] = \mu$, $Var(X) = \sigma^2$, $a \leq X \leq b$ a.s.

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}(X) \geq t'\right) \leq \exp\left(\frac{-2t'^2}{n(b-a)^2}\right)$$

let $t' = \sqrt{n} \cdot t$

$$\Rightarrow \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X] \geq \frac{\sqrt{n} \cdot t}{n}\right) \leq \exp\left(-\frac{2t}{(b-a)^2}\right)$$

$$\Rightarrow \boxed{\mathbb{P}\left(\sqrt{n}\left(\bar{X}_n - \mu\right) \geq t\right) \leq \exp\left(\frac{-2t}{(b-a)^2}\right)} \quad \text{II} \quad \text{(Hoeffding's)}$$

Comparison :

If $a \leq X_i \leq b$ a.s., then

$$\sigma^2 \leq \frac{(b-a)^2}{4}$$

CLT : $\exp\left(\frac{-t^2}{2\sigma^2}\right)$

Hoeffding's : $\exp\left(\frac{-2t^2}{(b-a)^2}\right)$

With this, CLT bound is $\underset{\text{(CLT)}}{\exp\left(\frac{-t^2}{2\sigma^2}\right)} \leq \underset{\text{(Hoeffding)}}{\exp\left(\frac{-2t^2}{(b-a)^2}\right)}$.

Remarks : * Error upper bound matches b/w CLT, hoeffding

* σ^2 can be much smaller than range, makes CLT bound sharper/stronger

* CLT holds asymptotically and is an approximation. On the other hand, hoeffding's is exact, non-asymptotic (holds for any n)

Berry Esses Thm : How close apx w.r.t CLT (approximation scales $\sim 1/\sqrt{n}$) (v. deep thm — not in course)

Note : MGF based techniques we use — (Cramer - chernoff Method.)

Confidence Interval

Suppose $X_1 \dots X_n$ i.i.d random variables with $\mathbb{E}X = \mu$, $\operatorname{Var} X = \sigma^2$

$a \le X_i \le b$ almost surely.

Problem: We know (a, b, σ^2) and want to estimate μ.

We obtain a set (interval), also known as confidence interval (CI), where μ lies w.h.p.

CLT based bound (classical) (for $t \ge 0$)

By CLT, $\mathbb{P}\left(\left|\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right)\right| \le t\right) \xrightarrow{n \to \infty} \mathbb{P}\left(|N(0,1)| \le t\right)$

(union of two events)

Put $t = z_{\alpha/2}$ (α-quantile) s.t. $\mathbb{P}\left(|N(0,1)| \le z_{\alpha/2}\right) = 1 - \alpha$

w.p. $1-\alpha$,

$$-z_{\alpha/2} \le \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \le z_{\alpha/2}$$

$$\Rightarrow \boxed{\bar{X}_n - \frac{\sigma}{\sqrt{n}} z_{\alpha/2} \le \mu \le \bar{X}_n + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}} \quad \text{C.I}$$

Length of interval $\frac{2\sigma}{\sqrt{n}} z_{\alpha/2}$ shrinks with n ($\to 0$ as $n \to \infty$)

Hoeffding's for C.I.

$$\mathbb{P}\left(\left|\sqrt{n}(\bar{X}_n - \mu)\right| \ge t\right) \le \underbrace{2\exp\left(\frac{-2t^2}{(b-a)^2}\right)}_{\alpha}$$

i.e. $t = (b-a)\sqrt{\frac{1}{2}\log\frac{\alpha}{2}}$

So, w.p. $1-\alpha$,

$$\boxed{|\bar{X}_n - \mu| \le \frac{1}{\sqrt{n}}(b-a)\sqrt{\frac{1}{2}\log\frac{\alpha}{2}}}$$

μ lies in $\left[\bar{X}_n - \text{''}, \bar{X}_n + \text{''}\right]$ C.I.

Length of interval,

$\sigma \le \frac{b-a}{2}$

so, CLT is better (asymptotic)

Sub - Gaussian

Assume $X \sim N(\mu, \sigma^2)$. We know that

$$\mathbb{E}\, e^{\lambda(X-\mu)} = e^{\lambda^2 \sigma^2 / 2} \quad \text{for any } \lambda.$$

This motivates us to define class of r.v. exhibiting similar properties.

<u>Def</u>: A r.v. X w/ mean μ is called sub-gaussian if there exists a positive number σ such that

$$\mathbb{E}\, e^{\lambda(X-\mu)} \leq e^{\lambda^2\sigma^2/2} \quad \text{for all } \lambda.$$

It is denoted as $X \sim \text{sub}G(\sigma)$. σ is called the parameter of sub-Gaussian r.v.; σ^2 is a proxy for variance.

<u>Examples</u> ① Gaussian

② <u>Rademacher r.v.</u> :

$\epsilon \in \{-1, +1\}$ with equal probability

We want to show that ϵ is 1 sub gaussian

$\ell! \sim \ell^\ell$

$(2k)!$ much bigger than $2^k k!$

$$\mu = 0, \quad \sigma = 1 \qquad \mathbb{E}\, e^{\lambda\epsilon} = \frac{e^\lambda + e^{-\lambda}}{2} = \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \leq \sum_{k=0}^{\infty} \frac{(\lambda^2)^k}{k!\, 2^k}$$

$$\leq e^{\lambda^2/2}$$

③ <u>Bounded r.v.</u>

Suppose $a \leq X \leq b$ a.s.

Show that $X \sim \text{Sub}G\left(\frac{b-a}{2}\right)$

<u>Last class</u>: $\log \mathbb{E}\, e^{\lambda X} \leq \frac{(b-a)^2}{8}\lambda^2$

$$\Rightarrow \mathbb{E}\, e^{\lambda X} \leq \exp\left(\lambda^2 \frac{(b-a)^2}{8}\right) \leftarrow \text{exactly what we want.}$$

Hoeffding's inequality for sub Gaussian [not just for bounded r.v.]

Suppose $X_1 \dots X_n$ are $\text{sub}\,\mathcal{G}(\sigma_i)$ and $\mathbb{E} X_i = \mu_i$ then, for all $t \geq 0$

$$\mathbb{P}\left(\sum_{i=1}^{n} (X_i - \mathbb{E} X_i) \geq t\right) \leq \exp\left(\frac{-t^2}{2\sum_{i=1}^{n} \sigma_i^2}\right)$$

$$\mathbb{P}\left(\sum_{i=1}^{n} (\;\text{"}\;) \leq -t\right) \leq \;\text{"}$$

Remark : $\sigma_i = \frac{b_i - a_i}{2}$, we get back hoeffding's for bounded r.v. (proved in lec2)

Proof : Use Cramer - chernoff technique, bound on $\mathbb{E} e^{\lambda X}$ comes from defn of sub-gaussian directly!

Reading exercise :

Chap 2 from HDS (Martin) — Sub-gaussian (bigger class : sub-exponential)

— Martingales.