# Group Study and Seminar Series (Summer 20)

## Statistics, Optimization and Reinforcement Learning

Yingru Li

The Chinese University of Hong Kong, Shenzhen, China

June 24, 2020

Mainly based on:
Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint (Vol. 48). Cambridge University Press.
Selected paper on Mathematical and Algorithmic issues in Reinforcement Learning

# Outline

- Introduction to the seminar series
  - Why Non-Asymptotic?
- Basic tail and concentration bounds
  - Classical bounds
    - Sub-Gaussian
    - Sub-Exponential
  - Martingale based methods
  - Lipschitz functions of Gaussian variables

# Arrangements

- 14:00 - 17:00 in Administrative Building 603
- Every Thursday from 6/18 to 8/20
  - One exception: the seminar will be held in Wednesday (6/24) next week due to the Dragon boat festival.
  - 10 weeks in total!
- Enjoy the journey!

# What to discuss and why important?

- From (theoretical) computer scientists perspective:
  - “While traditional areas of computer science remain highly important, increasingly researchers of the future will be involved with using computers to understand and extract usable information from massive data arising in applications, not just how to make computers useful on specific well-defined problems. With this in mind we have written this book to cover the theory we expect to be useful in the next 40 years, just as an understanding of automata theory, algorithms, and related topics gave students an advantage in the last 40 years. One of the major changes is an increase in emphasis on probability, statistics, and numerical methods.”

    — Blum, A., Hopcroft, J., & Kannan, R. (2020). Foundations of data science. Cambridge University Press.
  - First chapter of the book is “High-Dimensional Space”.
  - Early draft of this book with free access was used in undergrad-level and graduate-level lectures by Cornell, Princeton, Stanford and many other US and Chinese Universities.

# What to discuss and why important?

- Fundamental tools in “High-Dimensional Statistics: A Non-Asymptotic Viewpoint” (used for grad-level courses in UCB, MIT, Princeton, etc.)
  - Basic tail and concentration bounds (week 1 and 2)
  - Minimax Lower bounds (week 3)
  - Uniform Law of Large Number (week 5)
  - Metric Entropy and its uses (week 7)
  - Random matrices (week 9)
- Potential Applications:
  - Analysis of Randomized and Online Algorithms
  - Statistical Learning Theory
  - Statistical Signal Processing
  - Differential Privacy, Random Graph Theory, (Big) Data Science
  - Any subject where randomness plays an important role ...

# What to discuss and why important?

- Mathematical and Algorithmic foundations of Reinforcement Learning
  - Sample complexity lower and upper bound for tabular RL with generative oracle (week 4)
  - Optimality and approximation issues on policy gradient methods (week 6)
  - Regret lower and upper bound for tabular RL problem (week 8)
  - Model-based RL with function approximation (week 10)
  - New problems and settings
- Decision is always a higher-level of intelligence.
- Understanding a general paradigm for solving (online) sequential decision problem
- Towards understanding the fundamental limits of Reinforcement Learning
- Towards designing better general decision-making agents

# Outline

- Introduction to the seminar series
  - Why Non-Asymptotic?
- Basic tail and concentration bounds
  - Classical bounds
    - Sub-Gaussian
    - Sub-Exponential
  - Martingale based methods
  - Lipschitz functions of Gaussian variables

# Data matrix

- A good example to keep in mind is a dataset organized as an $n$ by $d$ matrix $\mathbf{X}$ where,
  - for example, the rows correspond to patients and,
  - the columns correspond to measurements on each patient (height, weight, heart rate, $\ldots$ ).
- Row $i$ is a random vector $X_i^\top \in \mathbb{R}^d$ of the measurements performed on patient $i$.

$$\mathbf{X} = \begin{pmatrix} X_1^\top \\ X_2^\top \\ \vdots \\ X_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}$$

- Modern applications in science and engineering:
  - large-scale problems: both $d$ and $n$ may be large (possibly $d \gg n$)
  - need for high-dimensional theory that provides non-asymptotic results for $(n, d)$

# Asymptotic v.s. Non-Asymptotic

- Mean estimation problem: Here $d = 1$ and we observe $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} P$ where $P$ is a distribution over $\mathbb{R}$ with mean $\mu$ and variance $\sigma^2$. We consider the sample mean:

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

- Asymptotic statements:
  - Weak Law of Large Numbers (WLLN): $\overline{X}_n \xrightarrow{\mathbb{P}} \mu$, as $n \to \infty$.
  - Central limit Theorem (CLT): $\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{(d)} \mathcal{N}\left(0, \sigma^2\right)$, as $n \to \infty$.
  - Asymptotic confidence interval for $\mu$: direct result of the CLT (<u>Classical asymptotic regime</u>)

$$\mathbb{P}\left(\mu \in \left[\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]\right) \underset{n\to\infty}{\longrightarrow} .95$$

- What will happen for finite fixed sample size $n$? (<u>Non-asymptotic regime</u>: fixed pair $(n, d)$)

# Asymptotic v.s. Non-Asymptotic

- Asymptotic statements: Asymptotic confidence interval for $\mu$

$$\mathbb{P}\left(\mu \in \left[\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right]\right) \underset{n\to\infty}{\longrightarrow} .95$$

- Non-asymptotic statements:
  - Quadratic risk: $\mathbb{E}\left[\left(\overline{X}_n - \mu\right)^2\right] = \frac{\sigma^2}{n}$
  - Tail bounds: $\mathbb{P}\left(\left|\overline{X}_n - \mu\right| > t\right) \leq 2e^{-Cnt^2}$, where $C > 0$ is a constant that depends on further assumptions on $P$, such as having a bounded support (Hoeffding's inequality)
  - Non-Asymptotic confidence interval for $\mu$: result of Hoeffding's inequality

$$\mathbb{P}\left(\mu \in \left[\overline{X}_n - C\frac{\sigma}{\sqrt{n}}, \overline{X}_n + C\frac{\sigma}{\sqrt{n}}\right]\right) \geq .95$$

where $C$ is a constant typically larger than $1.96$.

# Asymptotic v.s. Non-Asymptotic

- Covariance matrix estimation: assume now that we observe $X_1, \ldots, X_n \overset{\text{i.i.d}}{\sim} P$
  - where $P$ is a distribution over $\mathbb{R}^d$ with zero mean and covariance matrix $\mathbf{\Sigma} = \operatorname{cov}(X_1) = \mathbb{E}\left[X_1 X_1^\top\right]$. The sample covariance matrix $\widehat{\mathbf{\Sigma}}$ is defined as

$$\widehat{\mathbf{\Sigma}} := \overbrace{\frac{1}{n}\sum_{k=1}^{n} X_k X_k^\top}^{\text{average of } d \times d \text{ rank one matrices}}, \quad \text{with} \quad \mathbb{E}\left(\widehat{\mathbf{\Sigma}}\right) = \mathbf{\Sigma},$$

$$\operatorname{Var}\left(\widehat{\mathbf{\Sigma}}_{i,j}\right) = \mathbb{E}\left(\widehat{\mathbf{\Sigma}}_{i,j} - \mathbf{\Sigma}_{i,j}\right)^2 = \frac{1}{n^2}\sum_{k=1}^{n} \mathbb{E}\left(\mathbf{X}_{k,i}\mathbf{X}_{k,j} - \mathbf{\Sigma}_{i,j}\right)^2 = \frac{1}{n}\operatorname{Var}\left(\mathbf{X}_{1,i}\mathbf{X}_{1,j}\right)$$

- Asymptotic statements:
  - WLLN: $\widehat{\mathbf{\Sigma}}_{i,j} \xrightarrow{\mathbb{P}} \mathbf{\Sigma}_{i,j}$, for all $i, j = 1, \ldots, d$ as $n \to \infty$
  - CLT: $\sqrt{n}\left(\widehat{\mathbf{\Sigma}}_{i,j} - \mathbf{\Sigma}_{i,j}\right) \xrightarrow{(d)} \mathcal{N}\left(0, \operatorname{Var}\left(\mathbf{X}_{1,i}\mathbf{X}_{1,j}\right)\right)$ as $n \to \infty$
  - In this case, letting $n \to \infty$ implicitly assumes that $n \gg d$. But what if $n$ is of the order of $d$ or even if $n \ll d$?

# Asymptotic v.s. Non-Asymptotic

- Can we make similar statements simultaneously for all the entries of $\widehat{\mathbf{\Sigma}}$? In other words, can we guarantee that the matrix $\widehat{\mathbf{\Sigma}}$ converges to the matrix $\mathbf{\Sigma}$?
  - For example, let's say that we are interested in understanding the random variable:

$$|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty := \max_{i,j} \left|\widehat{\mathbf{\Sigma}}_{i,j} - \mathbf{\Sigma}_{i,j}\right|$$

  - An attempt using a union bound with Chebyshev's inequality assuming $\mathrm{Var}\left(\mathbf{X}_{1,i}\mathbf{X}_{1,j}\right) \leq C$

$$\mathbb{P}\left(\left|\widehat{\mathbf{\Sigma}}_{i,j} - \mathbf{\Sigma}_{i,j}\right| > t\right) \leq \frac{\mathrm{Var}\left(\widehat{\mathbf{\Sigma}}_{i,j}\right)}{t^2} = \frac{\mathrm{Var}\left(\mathbf{X}_{1,i}\mathbf{X}_{1,j}\right)}{nt^2} \leq \frac{C}{nt^2}$$

$$\mathbb{P}\left(|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty > t\right) = \mathbb{P}\left(\exists (i,j) : \left|\widehat{\mathbf{\Sigma}}_{i,j} - \mathbf{\Sigma}_{i,j}\right| > t\right) \leq \sum_{1 \leq i,j \leq d} \mathbb{P}\left(\left|\widehat{\mathbf{\Sigma}}_{i,j} - \mathbf{\Sigma}_{i,j}\right| > t\right) \leq C\frac{d^2}{nt^2}$$

# Asymptotic v.s. Non-Asymptotic

- Can we make similar statements simultaneously for all the entries of $\widehat{\mathbf{\Sigma}}$? In other words, can we guarantee that the matrix $\widehat{\mathbf{\Sigma}}$ converges to the matrix $\mathbf{\Sigma}$?
  - For example, let's say that we are interested in understanding the random variable:

$$|\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty := \max_{i,j} \left| \widehat{\mathbf{\Sigma}}_{i,j} - \mathbf{\Sigma}_{i,j} \right|$$

  - An attempt using a union bound

$$\mathbb{P}\left( |\widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}|_\infty > t \right) \leq C \frac{d^2}{nt^2}$$

  - This attempt fails if $d$ grows faster than $\sqrt{n}$, that is, $d = \omega(\sqrt{n})$
  - While this attempt uses loose arguments, one can show that convergence of $\widehat{\mathbf{\Sigma}}$ to $\mathbf{\Sigma}$ fails if $d/n \to \gamma \in (0, 1]$ (High-dimensional asymptotic regime: asymptotically fixed aspect ratio).
  - Spoiler! Jiancong will lead a detailed lecture on covariance estimation as well as random matrices that could provide concentration bounds for all fixed pairs $(n, d)$.

# Asymptotic v.s. Non-Asymptotic

- Classical asymptotic or high-dimensional asymptotic regimes:
  - preferred for statistical inference tasks such as confidence intervals and hypothesis testing
- Non-asymptotic regime:
  - preferred to produce a qualitative description of the performance of a possibly complicated and high-dimensional method such as the ones arising in machine learning.
  - E.g. double descent phenomenon in overparameterized neural networks or even linear models.
  - E.g. the true shape of the regret in bandit and RL problems.

# Outline

Introduction to the seminar series

- Why Non-Asymptotic?

Basic tail and concentration bounds

- Classical bounds
    - Sub-Gaussian
    - Sub-Exponential
- Martingale based methods
- Lipschitz functions of Gaussian variables

# Outline

Introduction to the seminar series

- Why Non-Asymptotic?

Basic tail and concentration bounds

- Classical bounds
  - Sub-Gaussian
  - Sub-Exponential
- Martingale based methods
- Lipschitz functions of Gaussian variables

# Mills Ratio Inequality

- Consider a standard normal random variable $Z \sim \mathcal{N}(0,1)$ with p.d.f:

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

- **Proposition (Mills Ratio Inequality)**: $\forall t > 0$

$$\mathbb{P}[|Z| > t] \leq \sqrt{\frac{2}{\pi}} \frac{e^{-\frac{t^2}{2}}}{t}$$

# Proof of Mills Ratio

- Integral the tail distribution:

$$\mathbb{P}[Z > t] = \int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x$$

- On the interval $[t, \infty), x \geq t$, i.e., $\frac{x}{t} \geq 1$,

$$\int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x \leq \int_t^\infty \left(\frac{x}{t}\right) \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \mathrm{d}x = \frac{1}{t\sqrt{2\pi}} \int_t^\infty x \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x = \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{t^2}{2}}}{t}$$

- Since $Z$ is symmetric ($Z \overset{(d)}{=} -Z$),

$$\mathbb{P}[|Z| > t] = \mathbb{P}[Z > t] + \mathbb{P}[Z < -t] = 2\mathbb{P}[X > t] = 2\frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{t^2}{2}}}{t}$$

# Motivation

- Immediate results of Mills Ratio for the tail bound for sample mean $\overline{X}_n$ with $X_1, \ldots, X_n \overset{i.i.d}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$

$$\mathbb{P}\left[\sqrt{n}\frac{\left|\bar{X}_n - \mu\right|}{\sigma} > t\right] \leq \sqrt{\frac{2}{\pi}}\frac{e^{-\frac{t^2}{2}}}{t}$$

- Set $t = \frac{\sqrt{n}}{\sigma}\epsilon$, we rewrite the bound $\mathbb{P}\left[\left|\bar{X}_n - \mu\right| > \epsilon\right] \leq \sqrt{\frac{2\sigma^2}{\pi n \epsilon^2}} \exp\left(-\frac{n\epsilon^2}{2\sigma^2}\right)$
- Thus, the probability that the sample average $\overline{X}_n$ deviates away from $\mu$ decays rapidly (exponential decay on sample size $n$)
- We want to replicate this type of behavior for other random variables in a manner that allows us to (1) obtain finite samples guarantees (i.e. for every $n$ ), and (2) circumvent the need for too many distributional assumptions on $X_1, \ldots, X_n$.

# From Markov to Chernoff

- **Markov's inequality**: given a non-negative random variable $X$ with finite mean, we have $\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$ for all $t > 0$. (Proof hints: $X \geq t\mathbb{1}_{X \geq t}$ always holds)
- **Chebyshev's inequality**: for a random variable $X$ that also has a finite variance, we have$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\mathrm{Var}(X)}{t^2}$ for all $t > 0$. (Proof hints: apply Markov's to the new non-negative random variable $Y = (X - \mu)^2$)
- If there is some constant $b > 0$ s.t. the $\underline{\text{moment generating function } \varphi_X(\lambda) = \mathbb{E}\left[e^{\lambda(X-\mu)}\right]}$ exists for all $\lambda \leq b$. In this case, for any $\lambda \in [0, b]$, we may $\underline{\text{apply Markov's inequality to the random variable } Y = e^{\lambda(X-\mu)}}$, thereby obtaining the upper bound

$$\mathbb{P}[(X - \mu) \geq t] = \mathbb{P}\left[e^{\lambda(X-\mu)} \geq e^{\lambda t}\right] \leq \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}}$$

# From Markov to Chernoff

- **Markov's inequality**: given a non-negative random variable $X$ with finite mean, we have $\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$ for all $t > 0$. (Proof hints: $X \geq t\mathbb{1}_{X \geq t}$ always holds)
- Define the $\underline{\text{cumulant generating function } \psi_X(\lambda) := \log\left(\mathbb{E}\left[e^{\lambda(X-\mu)}\right]\right).}$
- For any $\lambda \in [0, b]$,

$$\mathbb{P}[(X-\mu) \geq t] = \mathbb{P}\left[e^{\lambda(X-\mu)} \geq e^{\lambda t}\right] \leq \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}} = \exp\left(\psi_X(\lambda) - \lambda t\right)$$

- **Chernoff's inequality**: Let $X$ be a random variable and $\mathbb{E}[X] = \mu$ with well-defined moment generating function $\varphi_X(\lambda)$ for $|\lambda| \leq b$, then

$$\mathbb{P}(X - \mu \geq t) \leq \exp\left(-\psi_X^*(t)\right),$$

where $\psi_X^*(t) := \sup_{\lambda \in [0,b]} \left(\lambda t - \psi_X(\lambda)\right)$ is the conjugate function of $\psi_X(\lambda)$.

# Chernoff bound for Gaussian

- Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, then $\mathbb{E}\left[e^{\lambda X}\right] = e^{\mu\lambda + \sigma^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$. So, we have

$$\sup_{\lambda \geq 0}\left(\lambda t - \log\left(\mathbb{E}\left[e^{\lambda(X-\mu)}\right]\right)\right) = \sup_{\lambda \geq 0}\left(\lambda t - \frac{\lambda^2\sigma^2}{2}\right) = \frac{t^2}{2\sigma^2}$$

  which yields the bound for all $t > 0$,

$$P(X - \mu \geq t) \leq \exp\left(\frac{-t^2}{2\sigma^2}\right)$$

- This bound is sharp up to polynomial-factor corrections, compared to the Mills Ratio.
- Deriving a Chernoff bound does not require less knowledge about a distribution than a Markov-based bound since we need the moment generating function of $X - \mu$. (Assume the existence of infinity many moments)
- A main advantage is that these moments do not have to be painstakingly calculated.

# Sub-Gaussian random variables

- **Definition (Sub-Gaussian random variables):** A random variable $X$ is Sub-Gaussian with parameter $\sigma$ if
$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$$
for all $\lambda \in \mathbb{R}$. In that case, we write $X \in \mathrm{SG}\left(\sigma^2\right)$.
- First simple observation: $X \in \mathrm{SG}\left(\sigma^2\right)$ iff $-X \in \mathrm{SG}\left(\sigma^2\right)$.
- Second observation: if $X \in \mathrm{SG}\left(\sigma^2\right)$ , then the MGF of $X$ can be bounded by the Gaussian MGF which yields the same Chernoff bound, i. e.
$$\mathbb{P}(|X-\mu| \geq t) \leq 2\exp\left(\frac{-t^2}{2\sigma^2}\right)$$

# Properties of Sub-Gaussian random variables

- Let $X \in \mathrm{SG}\left(\sigma^2\right)$, then $\mathrm{Var}[X] \leq \sigma^2$ with $\mathrm{Var}[X] = \sigma^2$ if $X$ is Gaussian.
- If there are $a, b \in \mathbb{R}$ such that $a \leq X - \mu \leq b$ almost everywhere, then $X \in \mathrm{SG}\left(\left(\frac{b-a}{2}\right)^2\right)$
- Let $X \in \mathrm{SG}\left(\sigma^2\right)$ and $Y \in \mathrm{SG}\left(\tau^2\right)$, then
  - $X\alpha \in \mathrm{SG}\left(\sigma^2\alpha^2\right)$ for all $\alpha \in \mathbb{R}$ with $\alpha \neq 0$
  - $X + Y \in \mathrm{SG}\left((\tau + \sigma)^2\right)$, and
  - if $X \perp Y, X + Y \in \mathrm{SG}\left(\tau^2 + \sigma^2\right)$

# Proof of Sub-Gaussian property 1

- **Property 1**: Let $X \in \mathrm{SG}\left(\sigma^2\right)$, then $\operatorname{Var}[X] \leq \sigma^2$ with $\operatorname{Var}[X] = \sigma^2$ if $X$ is Gaussian.
- It holds by assumption that $\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$ for all $\lambda \in \mathbb{R}$, and hence by taylor expansion

$$1 + \lambda \underbrace{\mathbb{E}[X-\mu]}_{=0} + \lambda^2 \frac{\mathbb{E}\left[(X-\mu)^2\right]}{2} + o\left(\lambda^2\right) \leq 1 + \frac{\lambda^2\sigma^2}{2} + o\left(\lambda^2\right)$$

- We divide both sides of this inequality by $\lambda^2$ (and assume $\lambda \neq 0$ ), and let $\lambda \to 0$.

# Proof of Sub-Gaussian property 2

- **Property 2:** If there are $a, b \in \mathbb{R}$ such that $a \leq X - \mu \leq b$ almost everywhere, then $X \in \mathrm{SG}\left(\left(\frac{b-a}{2}\right)^2\right)$
- W.L.O.G, let $\mu = 0$. We need to show that $\log\left(\mathbb{E}\left[e^{\lambda X}\right]\right) \leq \frac{(b-a)^2\lambda^2}{8}$ for all $\lambda \in \mathbb{R}$.
- First, notice that for any distribution supported on $[a, b]$, the variance $\mathrm{Var}[X] = \underbrace{\mathbb{E}(X - \mathbb{E}(X))^2 \leq \mathbb{E}(X - (a+b)/2)^2}_{\mathbb{E}(X) = \arg\min_s g(s) := \mathbb{E}(X-s)^2} = \frac{1}{4}\mathbb{E}(\underbrace{(X-a)}_{\geq 0} + \underbrace{(X-b)}_{\leq 0})^2 \leq \left(\frac{b-a}{2}\right)^2$.
- Remind the CGF $\psi_X(\lambda) = \log \mathbb{E}\left[e^{\lambda X}\right]$, and $\psi_X(0) = 0$. Define a new prob. measure:

$$\widetilde{\mathbb{P}}(A) = \frac{\int_A e^{\lambda x}\mathbb{P}(\mathrm{d}x)}{\int e^{\lambda y}\mathbb{P}(\mathrm{d}y)}$$

- We have the derivatives: $\psi_X'(\lambda) = \frac{\mathbb{E}\left[Xe^{\lambda X}\right]}{\mathbb{E}[e^{\lambda X}]} = \mathbb{E}_{\widetilde{\mathbb{P}}}(X)$. Note that $\psi_X'(0) = \mathbb{E}_{\widetilde{\mathbb{P}}}(0) = 0$.

# Proof of Sub-Gaussian property 2

- Remind the CGF $\psi_X(\lambda) = \log \mathbb{E}\left[e^{\lambda X}\right]$, define a new prob. measure

$$\widetilde{\mathbb{P}}(A) = \frac{\int_A e^{\lambda x}\mathbb{P}(\mathrm{d}x)}{\int e^{\lambda y}\mathbb{P}(\mathrm{d}y)},$$

- We also have the second-order derivatives:

$$\begin{aligned}
\psi_X''(\lambda) &= \frac{\mathbb{E}\left[X^2 e^{\lambda X}\right]\mathbb{E}\left[e^{\lambda X}\right] - \mathbb{E}\left[X e^{\lambda X}\right]\mathbb{E}\left[X e^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]^2} \\
&= \frac{\mathbb{E}\left[X^2 e^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]} - \left(\frac{\mathbb{E}\left[X e^{\lambda X}\right]}{\mathbb{E}\left[e^{\lambda X}\right]}\right)^2 \\
&= \mathbb{E}_{\widetilde{\mathbb{P}}}\left[X^2\right] - \mathbb{E}_{\widetilde{\mathbb{P}}}[X]^2 \\
&= \mathrm{Var}_{\widetilde{\mathbb{P}}}(X) \leq (b-a)^2/4
\end{aligned}$$

# Proof of Sub-Gaussian property 2

- **Property 2:** If there are $a, b \in \mathbb{R}$ such that $a \leq X - \mu \leq b$ almost everywhere, then $X \in \mathrm{SG}\left(\left(\frac{b-a}{2}\right)^2\right)$
- $\psi_X(0) = \psi'_X(0) = 0$ and $\psi''_X(\lambda) \leq (b-a)^2/4$ for all $\lambda$.
- The fundamental theorem of calculus:

$$\psi_X(\lambda) = \int_0^\lambda \psi'_X(u)du = \int_0^\lambda \int_0^u \psi''_X(w)dwdu \leq \int_0^\lambda \int_0^u \frac{(b-a)^2}{4} dwdu = \lambda^2 \frac{(b-a)^2}{8}$$

- This result can be used to prove Hoeffding's bound.

- **Property 3**: Let $X \in \mathrm{SG}\left(\sigma^2\right)$ and $Y \in \mathrm{SG}\left(\tau^2\right)$ , then
  i $X\alpha \in \mathrm{SG}\left(\sigma^2\alpha^2\right)$ for all $\alpha \in \mathbb{R}$ with $\alpha \neq 0$
  ii $X + Y \in \mathrm{SG}\left((\tau+\sigma)^2\right)$ , and
  iii if $X \perp Y, X + Y \in \mathrm{SG}\left(\tau^2 + \sigma^2\right)$
- We prove (ii) and (iii) and assume that $\mathbb{E}[X] = \mathbb{E}[Y]$. If $X \perp Y$, the proof is immediate. If not, it holds for every $\lambda \in \mathbb{R}$ that $\mathbb{E}\left[e^{\lambda(X+Y)}\right] = \mathbb{E}\left[e^{\lambda X} e^{\lambda Y}\right]$ . By Hölder's inequality,

$$
\begin{aligned}
\mathbb{E}\left[e^{\lambda(X+Y)}\right] &= \mathbb{E}\left[e^{\lambda X} e^{\lambda Y}\right] \leq \left(\mathbb{E}\left[e^{\lambda p X}\right]\right)^{1/p} \left(\mathbb{E}\left[e^{\lambda q Y}\right]\right)^{1/q} \\
&\overset{\mathrm{SG}}{\leq} \exp\left(\frac{\lambda^2 p^2 \sigma^2}{2} \frac{1}{p} + \frac{\lambda^2 q^2 \tau^2}{2} \frac{1}{q}\right) \\
&= \exp\left(\frac{\lambda^2}{2}\left(p\sigma^2 + q\tau^2\right)\right) = \exp\left(\frac{\lambda^2}{2}(\sigma+\tau)^2\right)
\end{aligned}
$$

where we set $p = (\tau+\sigma)/\sigma$ in the last step.

# Hölder's inequality

- If $p, q > 0$ with $\frac{1}{p} + \frac{1}{q} = 1$, it holds that

$$\mathbb{E}\left[|X_1 X_2|\right] \leq \left(\mathbb{E}\left[|X_1|^p\right]\right)^{1/p} \left(\mathbb{E}\left[|X_2|^q\right]\right)^{1/q}$$

- The special case with $p = q = 2$ is referred to as Cauchy-Schwartz inequality. The Cauchy-Schwartz inequality can, for example, be used to show that

$$\left| \frac{\mathrm{Cov}\left[X_1, X_2\right]}{\sqrt{\mathrm{Var}\left[X_1\right]}\sqrt{\mathrm{Var}\left[X_2\right]}} \right| \leq 1$$

# Hoeffding inequality

- **Theorem (Hoeffding inequality)**: Let$X_1, \ldots, X_n$ be independent random variables such that $X_i \in \mathrm{SG}\left(\sigma_i^2\right)$ for all i, then

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

- Usually, we have $\sigma_i^2 = \sigma^2$ for all $i$. In this case, it holds that

$$2\exp\left(\frac{-n^2 t^2}{2\sum_{i=1}^{n} \sigma_i^2}\right) = 2\exp\left(\frac{-nt^2}{2\sigma^2}\right)$$

- The proof is trivial using the properties of Sub-Gaussian.

- **Example (Hoeffding for Bernoulli R.V. ):** Let$X_1, \ldots, X_n$ be independent R.V. with $X_i \sim$ Bernoulli$(p_i)$ for some $p_i \in (0,1)$. Then, $X_i \in \mathrm{SG}(1/4)$ and thus

$$P\left(\left|\bar{X}_n - \bar{p}_n\right| \geq t\right) \leq 2\exp\left(-2nt^2\right)$$

where $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ and $\bar{p}_n := \frac{1}{n}\sum_{i=1}^n p_i$. Thus, we have that

$$\left|\bar{X}_n - \bar{p}_n\right| \leq \sqrt{\frac{1}{2n}\log\left(\frac{1}{\delta}\right)}$$

with probability at least $1-\delta$.

- In the case $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Bernoulli $(p)$. Set $\delta = \frac{1}{n^c}$ for some $c > 0$. For example, we have that with probability at least $1 - \frac{1}{n}$,

$$\left|\bar{X}_n - p\right| \leq \sqrt{\frac{1}{2n}\log(2n)} = \mathcal{O}\left(\sqrt{\frac{\log n}{n}}\right) = \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{n}}\right)$$

## Comparing Hoeffding and Chernoff Bounds

- Hoeffding's inequality is not the sharpest concentration inequality in general. Note that the above calculations would similarly hold for any bounded random variable.
- **Chernoff's multiplicative inequality** yields an improvement on Hoeffding's inequality ' whenever $p$ is small. Generally for $X_i \sim$ Bernoulli $(p_i)$ for some $p_i \in (0,1)$,

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i \geq (1+\epsilon)\sum_{i=1}^{n} p_i\right\} \leq \left\{\begin{array}{ll} \exp\left\{-\frac{\epsilon^2 \sum_{i=1}^{n} p_i}{3}\right\}, & \epsilon \in (0,1] \\ \exp\left\{-\frac{\epsilon^2 \sum_{i=1}^{n} p_i}{2+\epsilon}\right\}, & \epsilon > 1 \end{array}\right. ,$$

and

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i \leq (1-\epsilon)\sum_{i=1}^{n} p_i\right\} \leq \exp\left\{-\frac{\epsilon^2 \sum_{i=1}^{n} p_i}{2}\right\}, \quad \forall \epsilon \in (0,1)$$

# Comparing Hoeffding and Chernoff Bounds

- Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Bernoulli(p). Hoeffding's inequality gave, w. p. at least $1 - \delta$. $p - \bar{X}_n \leq \sqrt{\frac{1}{2n} \log(1/\delta)}$
- On the other hand, Chernoff's multiplicative inequality yields

$$\mathbb{P}\left\{p - \bar{X}_n \geq \epsilon p\right\} \leq \exp\left(-\frac{np\epsilon^2}{2}\right), \quad \forall \epsilon \in (0, 1)$$

- Provided $p \geq \frac{2}{n} \log(1/\delta)$, we have $p - \bar{X}_n \leq \sqrt{\frac{2p}{n} \log(1/\delta)}$ w.p. $1 - \delta$
- If we let $p \equiv p_n$ so that $p_n \to 0$, then Chernoff's provides a significant improvement upon Hoeffding's.
  - Because the variance is upper bounded by $p(1-p)$, which is shrinking as the sample size grows. Chernoff's multiplicative inequality incorporates this information.

# Equivalent Definitions of Sub-Gaussian Random Variables

- **Proposition.** Let $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ be the Gamma function. If $X \in \mathrm{SG}\left(\sigma^2\right)$,

  $$\mathbb{E}\left[|X|^p\right] \leq p 2^{p/2} \sigma^p \Gamma(p/2), \quad \forall p > 0$$

  In particular, there exists $C > 0$ not depending on $p$ such that

  $$\left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}} \leq C\sigma\sqrt{p}.$$

- **Remark.** For example, if $X \sim \mathcal{N}\left(0, \sigma^2\right)$, we have

  $$\mathbb{E}\left[|X|^p\right] = \frac{\sigma^p 2^{\frac{p}{2}} \Gamma\left(\frac{p+1}{2}\right)}{\sqrt{\pi}}$$

# Proof of the equivalent definition

- If $x \geq 0$ then $x = \int_0^x \mathrm{d}t = \int_0^\infty \mathbb{I}[t \leq x]\mathrm{d}t$. Using Fubini's theorem, ($p$-th order moment exists due to Sub-Gaussianess)

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right] &= \int_0^\infty \mathbb{P}\left(|X|^p \geq u\right)\mathrm{d}u = \int_0^\infty \mathbb{P}\left(|X| \geq u^{\frac{1}{p}}\right)\mathrm{d}u \leq 2\int_0^\infty \exp\left\{-\frac{u^{2/p}}{2\sigma^2}\right\}\mathrm{d}u \\
&\leq 2\left(2\sigma^2\right)^{\frac{p}{2}}\frac{p}{2}\underbrace{\int_0^\infty e^{-t}t^{\frac{p}{2}-1}\mathrm{d}t}_{\Gamma\left(\frac{p}{2}\right)} \qquad \left(\text{where } t = \frac{u^{\frac{2}{p}}}{2\sigma^2}\right) \\
&= \left(2\sigma^2\right)^{\frac{p}{2}} p\Gamma\left(\frac{p}{2}\right) \leq \left(2\sigma^2\right)^{\frac{p}{2}} p\left(\frac{p}{2}\right)^{p/2}
\end{aligned}$$

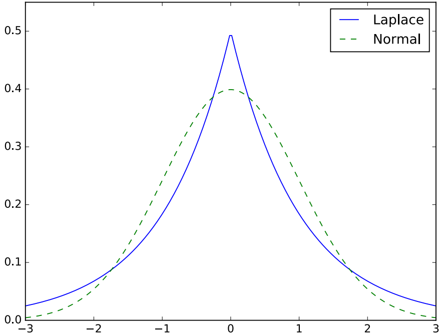
- $\|X\|_p := \left(\mathbb{E}\left[|X|^p\right]\right)^{\frac{1}{p}} \leq p^{1/p}\sigma\sqrt{p} \leq C\sigma\sqrt{p}$, since we can optimize fn. $p^{1/p}$ over $p$.

# Laplace distribution

- **Example** Let $X \sim$ Laplace $(0, b)$ for $b > 0$. Then it can be shown that

$$\mathbb{P}(|X| \geq t) \leq \exp(-tb), \quad \forall t > 0$$

- The Laplace distribution has fatter tails than the normal distribution



# Laplace Distribution

- **Example.** Let $X \sim$ Laplace $(b)$ for $b > 0$. Then it can be shown that

$$\mathbb{P}(|X| \geq t) \leq \exp(-tb), \quad \forall t > 0$$

- The Laplace distribution has fatter tails than the normal distribution
- $X \notin \mathrm{SG}\left(\sigma^2\right)$ since its MGF is only defined on a subset of the real line:

$$\mathbb{E}\left[e^{\lambda X}\right] = \frac{1}{1 - b^2\lambda^2}, \quad \forall |\lambda| < \frac{1}{b}$$

- The notion of Sub-Gaussianess is fairly restrictive, consider relaxations.
- Now turn to the class of sub-exponential variables, which are defined by a slightly milder condition on the moment generating function:

# Sub-Exponential Random Variable

- **Definition (Sub-Exponential Random Variable).** We say that a random variable $X$ is Sub-Exponential with parameters $\nu, \alpha > 0$, and we write $X \in \mathrm{SE}\left(\nu^2, \alpha\right)$, if

$$\mathbb{E}\left[e^{\lambda(X-\mathbb{E}(X))}\right] \leq e^{\frac{\lambda^2\nu^2}{2}}, \quad \forall|\lambda| < \frac{1}{\alpha}$$

- **Remark.** An immediate consequence of the definition is that $\mathrm{SG}\left(\sigma^2\right) \subseteq \mathrm{SE}\left(\sigma^2, 0\right)$. Thus, all Sub-Gaussian random variables are also Sub-Exponential.

# Chi-square distribution

- **Example.** Let $Z \sim \mathcal{N}(0,1)$, and $X = Z^2 \sim \chi^2_{(1)}, \mathbb{E}(X) = 1$. Let $\lambda < \frac{1}{2}$. Then

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda(X-1)}\right] &= \frac{1}{\sqrt{2\pi}} e^{-\lambda} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}(1-2\lambda)} dz \\
&= e^{-\lambda} \frac{1}{\sqrt{1-2\lambda}} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dy \qquad \left(y = z\sqrt{1-2\lambda}\right) \\
&= \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \overset{(a)}{\leq} \exp\left\{\frac{\lambda^2}{1-2\lambda}\right\} \overset{(b)}{\leq} \exp\left\{\frac{4\lambda^2}{2}\right\} \quad \left(\lambda < \frac{1}{4}\right)
\end{aligned}$$

- Thus, $X \in \mathrm{SE}(4,4)$. The inequality $(a)$ follows from the following elementary inequality $-\log(1-u) - u \leq \frac{u^2}{2(1-u)}, \forall u \in (0,1)$ with $u = 2\lambda$. And $(b)$ is due to the chosen $\lambda < 1/4$.
- Also, it is possible to show that $\mathbb{P}(X - 1 > 2t + 2\sqrt{t}) \leq e^{-t}, \forall t > 0$

## Tail Bounds for Sub-Exponential Random Variables

- **Theorem.** Let $X \in \mathrm{SE}\left(\nu^2, \alpha\right)$, and $t > 0$. Then

$$\mathbb{P}\{X - \mathbb{E}(X) \geq t\} \leq \begin{cases} \exp\left\{-\frac{t^2}{2\nu^2}\right\}, & t \leq \frac{\nu^2}{\alpha} \\ \exp\left\{-\frac{t}{2\alpha}\right\}, & t > \frac{\nu^2}{\alpha} \end{cases}$$

(Sub-Gaussian behavior)

Equivalently,

$$\mathbb{P}\{X - \mathbb{E}(X) \geq t\} \leq \exp\left\{-\frac{1}{2}\min\left(\frac{t^2}{\nu^2}, \frac{t}{\alpha}\right)\right\}$$

# Proof of tail bounds for Sub-Exponential R.V.

- Assume $\mu = 0$. Then repeating Chernoff argument, one obtains:

$$\mathbb{P}(X \geq t) \leq e^{-\lambda t + \frac{\lambda^2 \nu^2}{2}} = e^{g(\lambda, t)}, \quad \forall \lambda \in [0, 1/\alpha)$$

- To obtain the tightest bound one needs to find: $g^*(t) = \inf_{\lambda \in [0,1/\alpha)} g(\lambda, t)$
- To do so, notice, firstly, that unconstrained infimum occurs at $\lambda^* = t/\nu^2 > 0$. Consider two cases:
  1. If $\lambda^* < 1/\alpha \Leftrightarrow t \leq \frac{\nu^2}{\alpha}$, then sub-Gaussian behavior
  2. If $\lambda^* > 1/\alpha \Leftrightarrow t > \nu^2/\alpha$, then notice that the function $g(\lambda, t)$ is decreasing in $\lambda$ in the interval $\lambda \in [0, 1/\alpha)$ . Thus, the constrained infimum occurs at the boundary:

$$\lambda^*_{\text{constrained}} = \frac{1}{\alpha} \quad \text{and} \quad g\left(\lambda^*_{\text{constrained}}, t\right) = -\frac{t}{\alpha} + \frac{1}{2\alpha}\frac{\nu^2}{\alpha} \leq -\frac{t}{2\alpha}$$

# Bernstein condition

- Recall that sufficient conditions for a random variable to be a Sub-Gaussian include:
  - Boundedness of a random variable
  - Condition on the moments $\left(\mathbb{E}|X|^k\right)^{1/k}$
- Bernstein condition is one similar condition allowing unbounded random variables to behave sub-exponentially.
- **Definition (Bernstein condition).** Let $X$ be a random variable with mean $\mu$ and variance $\sigma^2$. Assume that $\exists b > 0$,

$$\mathbb{E}|X-\mu|^k \leq \frac{1}{2}k!\sigma^2 b^{k-2}, \quad k = 3, 4, \ldots$$

Then one says that $X$ satisfies Bernstein condition.

# Bernstein lemma

- **Lemma.** If random variable $X$ satisfies Bernstein condition with parameter $b$, then:

$$\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\frac{\lambda^2\sigma^2}{2}\frac{1}{1-b|\lambda|}}, \quad \forall|\lambda| < \frac{1}{b}$$

Additionally, from the bound on the moment generating function one can obtain the following tail bound (also known as Bernstein inequality):

$$\mathbb{P}(|X-\mu| \geq t) \leq 2\exp\left(-\frac{t^2}{2\left(\sigma^2+bt\right)}\right), \forall t > 0$$

## Proof of Bernstein lemma

- Pick $\lambda : |\lambda| < \frac{1}{b}$ (allowing interchanging summation and taking expectation) and expand the MGF in a Taylor series:

$$\mathbb{E}e^{\lambda(X-\mu)} = 1 + \frac{\lambda^2\sigma^2}{2} + \sum_{k=3}^{\infty} \frac{\mathbb{E}|X-\mu|^k}{k!}\lambda^k \leq 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}(|\lambda|b)^{k-2}$$

$$= 1 + \frac{\lambda^2\sigma^2}{2}\frac{1}{1-b|\lambda|} \leq e^{\frac{\lambda^2\sigma^2}{2}\frac{1}{1-b|\lambda|}}, \qquad (1 + x \leq e^x)$$

- To show the final bound, take $\lambda : |\lambda| < \frac{1}{2b}$. Then the bound becomes:

$$e^{\frac{\lambda^2\sigma^2}{2}\frac{1}{1-b|\lambda|}} \leq e^{\lambda^2\sigma^2} = e^{\frac{\lambda^2\left(2\sigma^2\right)}{2}} \quad \Longrightarrow \quad X \in \mathrm{SE}\left(2\sigma^2, 2b\right).$$

- The concentration result follows by taking $\lambda = \frac{t}{bt+\sigma^2} \in [0, \frac{1}{b})$ in the Chernoff bound.

# Proof of the Bernstein lemma

- The concentration result follows by taking $\lambda = \frac{t}{bt+\sigma^2} \in [0, \frac{1}{b})$ in the Chernoff bound.

$$\begin{aligned}
\mathbb{P}[X - \mu \geq t] &\leq \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)} - \lambda t\right) \\
&= \exp\left(\frac{\left(\frac{t}{bt+\sigma^2}\right)^2 \sigma^2}{\frac{2\sigma^2}{bt+\sigma^2}} - \frac{t^2}{bt+\sigma^2}\right) \\
&= \exp\left(-\frac{t^2}{2(bt+\sigma^2)}\right)
\end{aligned}$$

# Composition property of Sub-Exponential random variables

- **Property.** Let $X_1, \ldots, X_n$ be independent random variables such that $\mathbb{E}X_i = \mu_i$ and $X_i \in \mathrm{SE}\left(\nu_i^2, \alpha_i\right)$. Then $\sum_{i=1}^n \left(X_i - \mu_i\right) \in \mathrm{SE}\left(\sum_{i=1}^n \nu_i^2, \max_i \alpha_i\right)$
- In particular, denote $\nu_*^2 = \sum_{i=1}^n \nu_i^2, \alpha_* = \max_i \alpha_i$. Then:

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n \left(X_i - \mu_i\right)\right| \geq t\right) \leq \begin{cases} 2\exp\left(-\frac{nt^2}{2\nu_*^2/n}\right), & 0 < t \leq \frac{\nu_*^2/n}{\alpha_*} \\ 2\exp\left(-\frac{nt}{2\alpha_*}\right), & \text{otherwise} \end{cases}$$

  or, equivalently, $\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n \left(X_i - \mu_i\right)\right| \geq t\right) \leq \exp\left(-\frac{n}{2}\min\left\{\frac{t^2}{\nu_*^2/n}, \frac{t}{\alpha_*}\right\}\right)$
- **Remark**: Notice that that the range over which one obtains sub-Gaussian tail behavior gets smaller after composition. $\left(0, \frac{\nu_*^2/n}{\alpha_*}\right)$.

# Orlicz norm

- **Definition** ($\psi$**-Orlicz norm).** Let function $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ satisfy the following properties:
  - $\psi(x)$ is strictly increasing function
  - $\psi(x)$ is a convex function
  - $\psi(0) = 0$

  Then the $\psi$-Orlicz norm of a random variable $X$ is defined as

$$\|X\|_\psi = \inf \left\{ t > 0 : \mathbb{E}\psi\left(\frac{|X|}{t}\right) \leq 1 \right\}$$

- **Example 1.** Let $\psi(x) = x^p, p \geq 1$. Then:

$$\|X\|_\psi = \|X\|_p = (\mathbb{E}|X|^p)^{\frac{1}{p}}$$

# Orlicz norm

- **Example 2.** Let $\psi_p(x) = e^{x^p} - 1, p \geq 1.$ The corresponding Orlicz has two properties:
  (a) $p = 1:$ then $\|X\|_{\psi_1} < \infty$ is equivalent to $X$ belonging to the class of Sub-Exponential random variables
  (b) $p = 2:$ then $\|X\|_{\psi_2} < \infty$ is equivalent to $X$ belonging to the class of Sub-Gaussian random variables
- It is easy to show that (by definition):

$$\left\|X^2\right\|_{\psi_1} = \left(\|X\|_{\psi_2}\right)^2, \quad \|XY\|_{\psi_1} \leq \|X\|_{\psi_2}\|Y\|_{\psi_2}$$

- Using Orlicz norms allows to straightforwardly implies the following facts:
  1 Squared Sub-Gaussian random variable is Sub-Exponential.
  2 Product of two Sub-Gaussian random variables is Sub-Exponential.

# Concentraiton of a sub-gaussian random vector

- **Lemma (Concentration of a sub-gaussian random vector)** Let $X = (X_1, \ldots, X_d)^\top \in \mathbb{R}^d$ be such that: $\mathbb{E}X_i = 0, \mathbb{V}(X_i) = 1$ and assume that $X_i \in \mathrm{SG}(1)$. Then we can show that $\|X\|_2$ concentrates around $\sqrt{d}$.
- **Proof:**
  - Consider $\|X\|_2^2 = \sum_{i=1}^n X_i^2$. Then $X_i^2 - 1 \in \mathrm{SE}(\nu^2, \alpha)$ Thus, we have $\mathbb{P}\left(\left|\frac{\|X\|^2}{d} - 1\right| \geq t\right) \leq 2\exp\left(-\frac{d}{2}\min\left\{\frac{t^2}{\nu^2}, \frac{t}{\alpha}\right\}\right), \forall t > 0$
  - We will need to use the following fact: fix $c > 0$. Then for any numbers $z > 0$, $|z-1| \geq c \overset{\text{implies}}{\Longrightarrow} z^2 - 1 \geq \max\{c, c^2\}$.

$$\mathbb{P}\left(\left|\frac{\|X\|}{\sqrt{d}} - 1\right| \geq u\right) \leq \mathbb{P}\left(\left|\frac{\|X\|^2}{d} - 1\right| \geq \max\{u, u^2\}\right) \leq 2\exp\left(-\frac{du^2}{2C}\right)$$

- See more modern discussions on concentration of random vector in [Jin, Netrapalli, Ge, Kakade, and Jordan, 2019].

# Hoeffding vs. Bernstein

- One would like to compare two type of bounds/inequalities: Hoeffding's and Bernstein's. Denote $\mu = \mathbb{E}X$ and $\sigma^2 = \mathbb{V}(X)$. Assume that $|X - \mu| \leq b$ a.e. Then:

$$\mathbb{P}(|X - \mu| \geq t) \leq \begin{cases} 2\exp\left(-\frac{t^2}{2b^2}\right), & \text{Hoeffding} \\ 2\exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right), & \text{Bernstein} \end{cases}$$

- For small $t$ (meaning $bt \ll \sigma^2$ ) Bernstein's inequality gives rise to a bound of the order: $\mathbb{P}(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{c\sigma^2}}$ while Hoeffding's gives: $\mathbb{P}(|X - \mu| \geq t) \leq 2e^{-\frac{t^2}{cb^2}}$
- But $\sigma^2 \leq b^2$ and, thus, Bernstein's bound is tighter. Substantially tighter when $\sigma^2 \ll b^2$.
  - The case for a random variable that occasionally takes on large values.
  - (Further strengthen) For bounded R.V., Bennett's inequality is sharper then Bernstein's.

- **Theorem (Classic Bernstein inequality).** Let$X_1, \dots, X_n$ be independent random variables such that $|X_i - \mathbb{E}X_i| \leq b$ a.e. and $\max_i \mathbb{V}\left(X_i\right) \leq \sigma^2$. Then:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mathbb{E}X_i\right)\right| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2\left(\sigma^2 + \frac{bt}{3}\right)}\right)$$

- **Theorem (Laurent-Massart bounds for $\chi^2$).** Let $Z_1, \dots, Z_d \sim \mathcal{N}(0,1)$ and $a = (a_1, \dots, a_d)$ with $a_i \geq 0, \forall i \in \{1, \dots, n\}$. Let $X = \sum_{i=1}^{n} a_i \left(X_i^2 - 1\right)$. Then for right-tail behavior is described by:

$$\mathbb{P}\left(X \geq 2\|a\|_2\sqrt{t} + 2\|a\|_\infty t\right) \leq e^{-t}, \forall t > 0$$

and for left-tail behavior:

$$\mathbb{P}(X \leq -2\|a\|_2\sqrt{t}) \leq e^{-t}, \forall t > 0$$

Reduce to classic $\chi^2$ when $a_1 = a_2 = \dots = a_d = 1$.

# Johnson–Lindenstrauss embedding

- Suppose that we are given $N \geq 2$ distinct vectors $\{u^1, \ldots, u^N\}$, with each vector in $\mathbb{R}^d$.
- The data dimension $d$ is large. Expensive to store and manipulate the data set.
- The idea of dimensionality reduction is to construct a mapping $F : \mathbb{R}^d \to \mathbb{R}^m$
- Want mapping $F$ satisfies
  1. the projected dimension
$$m \ll d$$
  2. mapping $F$ preserves **pairwise distances, or equivalently norms and inner products**.
- Many ML algorithms are based on such pairwise quantities, including (1) linear regression, (2) methods for principal components, (3) the $k$ -means algorithm for clustering, and (4) nearest-neighbor algorithms for density estimation.

# Johnson–Lindenstrauss embedding

- Suppose that we are given $N \geq 2$ distinct vectors $\left\{u^1, \ldots, u^N\right\}$, with each vector in $\mathbb{R}^d$.
- The data dimension $d$ is large. Expensive to store and manipulate the data set.
- The idea of dimensionality reduction is to construct a mapping $F : \mathbb{R}^d \to \mathbb{R}^m$
- **more precisely**, given some tolerance $\delta \in (0, 1)$, we might be interested in a mapping $F$ with the guarantee that

$$(1-\delta) \leq \frac{\left\|F\left(u^i\right) - F\left(u^j\right)\right\|_2^2}{\left\|u^i - u^j\right\|_2^2} \leq (1+\delta) \quad \text{for all pairs } u^i \neq u^j \tag{1}$$

# Johnson–Lindenstrauss embedding

- An interesting and simple construction of the mapping that satisfies the condition 1 is probabilistic (known as Johnson–Lindenstrauss embedding or **Random Projection**):
  - Form a random matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ filled with independent $\mathcal{N}(0,1)$ entries
  - and use it to define a linear mapping

$$F : \mathbb{R}^d \to \mathbb{R}^m \quad \text{via} \quad u \mapsto \mathbf{X}u/\sqrt{m}$$

- Random projection satisfies the condition 1 with high probability as long as the projected dimension is lower bounded as

$$m \gtrsim \frac{1}{\delta^2} \log N.$$

  - The projected dimension $m$ is independent of the ambient dimension $d$, and scales only logarithmically with the number of data points $N$.

# Johnson–Lindenstrauss embedding

- Now verify that $F$ satisfies condition 1 with high probability.
- Let $x_i \in \mathbb{R}^d$ denote the $i$ th row of $\mathbf{X}$, and consider some fixed $u \neq 0$.
- Since $x_i$ is a standard normal vector, the variable $\langle x_i, u/\|u\|_2 \rangle$ follows a $\mathcal{N}(0,1)$ distribution, and hence the quantity

$$Y := \frac{\|\mathbf{X}u\|_2^2}{\|u\|_2^2} = \sum_{i=1}^{m} \langle x_i, u/\|u\|_2 \rangle^2$$

  follows a $\chi^2$ distribution with $m$ degrees of freedom, using the independence of the rows.
- Specifically $Y \in \mathrm{SE}(4m, 4)$, applying the sub-exponential tail bound, we find that

$$\mathbb{P}\left[ \left| \frac{\|\mathbf{X}u\|_2^2}{m\|u\|_2^2} - 1 \right| \geq \delta \right] \leq 2e^{-m\delta^2/8} \quad \text{for all } \delta \in (0,1)$$

## Johnson–Lindenstrauss embedding

- Rearranging and recalling the definition of $F$ yields the bound

$$\mathbb{P}\left[\frac{\|F(u)\|_2^2}{\|u\|_2^2} \notin [(1-\delta),(1+\delta)]\right] \le 2e^{-m\delta^2/8}, \qquad \text{for any fixed } 0 \neq u \in \mathbb{R}^d$$

- Noting that there are $\binom{N}{2}$ distinct pairs of data points, then applying the union bound:

$$\mathbb{P}\left[\frac{\left\|F\left(u^i-u^j\right)\right\|_2^2}{\|u^i-u^j\|_2^2} \notin [(1-\delta),(1+\delta)] \text{ for some } u^i \neq u^j\right] \le 2\binom{N}{2}e^{-m\delta^2/8}$$

- For any $\epsilon \in (0,1)$, this probability can be driven below $\epsilon$ by choosing $m > \frac{16}{\delta^2}\log(N/\epsilon)$.

# Application on Maxima

- **Theorem.** Let$X_1, \ldots, X_n$ be centered random variables that are not necessarily independent such that for all $\lambda \in [0, b), b \leq \infty, \mathbb{E}\left[e^{\lambda X_i}\right] \leq \psi(\lambda)$, where $\psi(\cdot)$ is convex on $[0, b)$. Then,
$$\mathbb{E}\left[\max_i X_i\right] \leq \inf_{\lambda \in [0,b)} \left\{\frac{\log(n) + \psi(\lambda)}{\lambda}\right\}$$
- By Jensen's inequality, for any $\lambda \in [0, b)$,
$$\exp\left\{\lambda \mathbb{E}\left[\max_i X_i\right]\right\} \leq \mathbb{E}\left[e^{\lambda \max_i X_i}\right] = \mathbb{E}\left[\max_i e^{\lambda X_i}\right] \leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i}\right] \leq n e^{\psi(\lambda)}$$
- Furthermore, if $\psi$ is convex, continuously differentiable on $[0, b)$, and $\psi(0) = \psi'(0) = 0$, then $\forall \mu > 0, \inf_{\lambda \in [0,b)} \left\{\frac{\mu + \psi(\lambda)}{\lambda}\right\} = \inf\{t \geq 0 : \psi^*(t) \geq \mu\}$ where $\psi^*(t) = \sup_{\lambda \in [0,b)} \lambda t - \psi(\lambda)$. (Hint: let $w = \inf_{\lambda \in [0,b)} \left\{\frac{\mu + \psi(\lambda)}{\lambda}\right\}$ and use definition.)

## Application on Maxima

$$\mathbb{E}\left[\max_i X_i\right] \leq \inf_{\lambda \in [0,b)} \left\{ \frac{\log(n) + \psi(\lambda)}{\lambda} \right\}$$

- **Example (Expectation for maxima of $n$ Sub-Gaussian R.V.).** Given all $X_i \in \mathrm{SG}\left(\sigma^2\right)$, then $\psi(\lambda) = \frac{\lambda^2\sigma^2}{2}$. The bound is $\inf_{\lambda>0} \frac{\log(n)}{\lambda} + \frac{\lambda^2\sigma^2}{2\lambda}$. The infimum is achieved at $\lambda = \sqrt{\frac{2\log(n)}{\sigma^2}}$ and we obtain

$$\mathbb{E}\left[\max_i X_i\right] \leq \sqrt{2\sigma^2 \log(n)}.$$

- This yields an important result: the expected value of the maximum of sub-Gaussian random variables with the same parameter grows at a rate of $\sqrt{\log(n)}$.

# Application on Maxima

- **Example (Sub-Exponential).** If $\psi(\lambda) = \frac{\lambda^2 \nu^2}{2(1-\lambda b)}$ for $\lambda \in \left(0, \frac{1}{b}\right)$, then we have

$$\mathbb{E}\left[\max_i X_i\right] \leq \sqrt{2\nu^2 \log(n)} + b \log(n).$$

- Note that this bound looks similar to the one for sub-Gaussian random variables, except that it includes an additional $b \log(n)$ term.
- Special case: if $X_i \sim \chi_p^2, \mathbb{E}\left[\max X_i\right] \leq 2\sqrt{p \log(n)} + 2 \log(n)$
- More detailed discussions in Sec. 2.4 and 2.5 in [Boucheron, Lugosi, and Massart, 2013].

# Outline

Introduction to the seminar series

- Why Non-Asymptotic?

Basic tail and concentration bounds

- Classical bounds
  - Sub-Gaussian
  - Sub-Exponential
- Martingale based methods
- Lipschitz functions of Gaussian variables

# Motivation

- Up until this point, these techniques have provided various types of bounds on sums of independent random variables. Many problems require bounds on more general functions of random variables.
- **Picture an arbitrary function of independent random variables. Can we create a concentration inequality for this arbitrary function?**
- One classical approach is based on martingale decompositions. In this section, we describe some of the results in this area along with some examples.

# Background

- $\mathbb{E}(Y|X)$ is a random variable because it is a function of the random variable, $X$ .
- $\mathbb{E}(Y|X=x)$ is not a random variable because it is a function of the fixed $x$.
- Let $Z = f\left(X_1, \ldots, X_n\right)$ with $f : \mathbb{R}^n \rightarrow \mathbb{R}$.
- We are interested in the concentration inequality for $Z - \mathbb{E}(Z)$
- **How to bound it?**

# Filtration and adaptation

- **Filtration.** Let $\{\mathcal{F}_k\}_{k=1}^{\infty}$ be a sequence of $\sigma$ -fields that are **nested**, meaning that

  $$\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$$

  for all $k \geq 1$ such a sequence is known as a filtration.

- **Adapted sequence.** Let $\{Y_k\}_{k=1}^{\infty}$ be a sequence of random variables such that $Y_k$ is measurable with respect to the $\sigma$-field $\mathcal{F}_k$. In this case, we say that $\{Y_k\}_{k=1}^{\infty}$ is adapted to the filtration $\{\mathcal{F}_k\}_{k=1}^{\infty}$ .

# Martingales

- **Definition (martingale).** Given a sequence $\{Y_k\}_{k=1}^{\infty}$ of random variables adapted to a filtration $\{\mathcal{F}_k\}_{k=1}^{\infty}$, the pair $\{(Y_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ is a martingale if, for all $k \geq 1$,

$$\mathbb{E}\left[|Y_k|\right] < \infty \quad \text{and} \quad \mathbb{E}\left[Y_{k+1} | \mathcal{F}_k\right] = Y_k$$

- **Doob construction.** Given a sequence of random variables $\{X_k\}_{k=1}^{n}$, and consider the random variable $f(X) = f\left(X_1 \ldots, X_n\right)$. Define the sequence

$$Y_k = \mathbb{E}\left[f(X) | X_1, \ldots, X_k\right], \quad \text{with } Y_n = f(X), Y_0 = \mathbb{E}(f(X))$$

and suppose that $\mathbb{E}[|f(X)|] < \infty$.

# Justification of Doob construction

- **Doob construction.**

$$Y_k = \mathbb{E}\left[f(X)|X_1, \ldots, X_k\right], \qquad \text{with } Y_n = f(X), Y_0 = \mathbb{E}(f(X))$$

- We claim that $\{Y_k\}_{k=0}^n$ is a martingale with respect to $\{X_k\}_{k=1}^n$. Indeed, in terms of the shorthand $X_1^k = (X_1, X_2, \ldots, X_k)$, we have

  1. follows from Jensen's inequality,

  $$\mathbb{E}\left[|Y_k|\right] = \mathbb{E}\left[\left|\mathbb{E}\left[f(X)|X_1^k\right]\right|\right] \le \mathbb{E}[|f(X)|] < \infty$$

  2. the tower property of conditional expectation in step (i),

  $$\mathbb{E}\left[Y_{k+1}|X_1^k\right] = \mathbb{E}\left[\mathbb{E}\left[f(X)|X_1^{k+1}\right]|X_1^k\right] \overset{\text{(i)}}{=} \mathbb{E}\left[f(X)|X_1^k\right] = Y_k$$

# Likelihood ratio construction

- **Likelihood ratio.** Let $f$ and $g$ be two mutually absolutely continuous densities, and let $\{X_k\}_{k=1}^{\infty}$ be a sequence of random variables drawn i.i.d. according to $f$. For each $k \geq 1$, let $Y_k := \prod_{\ell=1}^{k} \frac{g(X_t)}{f(X_\ell)}$ be the likelihood ratio based on the first $k$ samples. Then the sequence $\{Y_k\}_{k=1}^{\infty}$ is a martingale with respect to $\{X_k\}_{k=1}^{\infty}$. Indeed, we have

  $$\mathbb{E}\left[Y_{n+1} | X_1, \ldots, X_n\right] = \mathbb{E}\left[\frac{g\left(X_{n+1}\right)}{f\left(X_{n+1}\right)}\right] \prod_{k=1}^{n} \frac{g\left(X_k\right)}{f\left(X_k\right)} = Y_n$$

  using the fact that $\mathbb{E}\left[\frac{g(X_{n+1})}{f(X_{n+1})}\right] = 1$.
- Important in analyzing stopping rules for sequential hypothesis tests.

# Martingale difference sequence

- **Definition (martingale difference sequence).** A closely related notion is that of martingale difference sequence, meaning an adapted sequence $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ such that, for all $k \geq 1$,

$$\mathbb{E}\left[|D_k|\right] < \infty \quad \text{and} \quad \mathbb{E}\left[D_{k+1}|\mathcal{F}_k\right] = 0$$

- As suggested by their name, such difference sequences arise in a natural way from martingales. In particular, given a martingale $\{(Y_k, \mathcal{F}_k)\}_{k=0}^{\infty}$, let us define for $k \geq 1$

$$D_k = Y_k - Y_{k-1}$$

- We then have

$$\mathbb{E}\left[D_{k+1}|\mathcal{F}_k\right] = \mathbb{E}\left[Y_{k+1}|\mathcal{F}_k\right] - \mathbb{E}\left[Y_k|\mathcal{F}_k\right] = \mathbb{E}\left[Y_{k+1}|\mathcal{F}_k\right] - Y_k = 0$$

# Telescoping decomposition

- Using the martingale property and the fact that $Y_k$ is measurable with respect to $\mathcal{F}_k$ Thus, for any martingale sequence $\{Y_k\}_{k=0}^{\infty}$, we have the telescoping decomposition

$$f(X) - \mathbb{E}(f(X)) = Y_n - Y_0 = \sum_{k=1}^{n} (Y_k - Y_{k-1}) = \sum_{k=1}^{n} D_k$$

  where $\{D_k\}_{k=1}^{\infty}$ is a martingale difference sequence.
- This decomposition plays an important role in our development of concentration inequalities to follow.

# Concentration bounds for martingale difference sequences

- **Theorem (Bernstein-Freedman).** Let $\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty}$ be a martingale difference sequence, and suppose that $\mathbb{E}\left[e^{\lambda D_k} | \mathcal{F}_{k-1}\right] \leq e^{\lambda^2 v_k^2/2}$ almost surely for any $|\lambda| < 1/\alpha_k$. Then the following hold:
  (a) $\sum_{k=1}^{n} D_k \in \mathrm{SE}\left(\sum_{k=1}^{n} v_k^2, \alpha_*\right)$ where $\alpha_* := \max_k \alpha_k$. (same as if they were independent)
  (b) The sum satisfies the concentration inequality

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} D_k\right| \geq t\right] \leq \begin{cases} 2e^{-\frac{t^2}{2\sum_{k=1}^{n} v_k^2}} & \text{if } 0 \leq t \leq \frac{\sum_{k=1}^{n} v_k^2}{\alpha_*} \\ 2e^{-\frac{t}{2\alpha_*}} & \text{if } t > \frac{\sum_{k=1}^{n} v_k^2}{\alpha_*} \end{cases}$$

# Proof of Bernstein-Freedman

- For any scalar $\lambda$ such that $|\lambda| < \frac{1}{\alpha_*}$, conditioning on $\mathcal{F}_{n-1}$ and applying iterated expectation yields

$$\mathbb{E}\left[e^{\lambda\left(\sum_{k=1}^{n} D_k\right)}\right] = \mathbb{E}\left[e^{\lambda\left(\sum_{k=1}^{n-1} D_k\right)} \underbrace{\mathbb{E}\left[e^{\lambda D_n} | \mathcal{F}_{n-1}\right]}_{\text{see condition}}\right] \leq \mathbb{E}\left[e^{\lambda \sum_{k=1}^{n-1} D_k}\right] e^{\lambda^2 v_n^2/2}$$

- Iterating this procedure yields the bound $\mathbb{E}\left[e^{\lambda \sum_{k=1}^{n} D_k}\right] \leq e^{\lambda^2 \sum_{k=1}^{n} v_k^2/2}$, valid for all $|\lambda| < \frac{1}{\alpha_*}$. The tail bound follows by applying sub-exponential tail bound.
- Need sufficient and easily checkable conditions for the differences $D_k$.

## Azuma-Hoeffding

- **Corollary (Azuma-Hoeffding).** Let $(\{(D_k, \mathcal{F}_k)\}_{k=1}^{\infty})$ be a martingale difference sequence for which there are constants $\{(a_k, b_k)\}_{k=1}^{n}$ such that $D_k \in [a_k, b_k]$ almost surely for all $k = 1, \dots, n$. Then, for all $t \geq 0$

$$\mathbb{P}\left[\left|\sum_{k=1}^{n} D_k\right| \geq t\right] \leq 2e^{-\frac{2t^2}{\Sigma_{k-1}^{n}(b_k - a_k)^2}}$$

- **Proof Hint.**
  - Since $D_k \in [a_k, b_k]$ almost surely,
  - the conditioned variable $(D_k | \mathcal{F}_{k-1}) \in [a_k, b_k]$ almost surely,
  - and hence applying the result of property of sub-Gaussian R.V. (Hoeffding's lemma).

# Application: bounded difference function

- Given vectors $x, x' \in \mathbb{R}^n$ and an index $k \in \{1, 2, \ldots, n\}$, define a new vector $x^{\setminus k} \in \mathbb{R}^n$ via

$$x_j^{\setminus k} := \begin{cases} x_j & \text{if } j \neq k \\ x'_k & \text{if } j = k \end{cases}$$

- With this notation, we say that $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference inequality with parameters $(L_1, \ldots, L_n)$ if, for each index $k = 1, 2, \ldots, n$

$$\left| f(x) - f\left(x^{\setminus k}\right) \right| \leq L_k \quad \text{for all } x, x' \in \mathbb{R}^n \tag{2}$$

- For instance, if the function $f$ is $L$-Lipschitz with respect to the Hamming norm $d_H(x, y) = \sum_{i=1}^{n} \mathbb{I}\left[x_i \neq y_i\right]$, then the bounded difference inequality holds with parameter $L$ uniformly across all coordinates. (Recall the definition of $L$-Lipschitz)

# Bounded difference inequality

- **Corollary (Bounded differences inequality).** Suppose that $f$ satisfies the bounded difference property 2 with parameters $(L_1, \ldots, L_n)$ and that the random vector $X = (X_1, X_2, \ldots, X_n)$ has independent components. Then

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{2t^2}{\sum_{k=1}^{n} L_k^2}} \qquad \text{for all } t \geq 0$$

- Recalling the Doob martingale and constructing two corresponding auxiliary R.V.

$$\begin{aligned} D_k &= \mathbb{E}\left[f(X) \mid X_1, \ldots, X_k\right] - \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{k-1}\right] \\ A_k &:= \inf_x \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{k-1}, x\right] - \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{k-1}\right] \\ B_k &:= \sup_x \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{k-1}, x\right] - \mathbb{E}\left[f(X) \mid X_1, \ldots, X_{k-1}\right] \end{aligned}$$

Find something? $D_k \leq B_k - A_k$ a.s. and recall the definition $B_k - A_k \leq L_k$ a.s.

# Application of Bounded difference inequality

- Example 2.22 (Classical Hoeffding from bounded differences)
- Example 2.23 (U-statistics)
- Example 2.24 (Clique number in random graphs)
- Example 2.25 (Rademacher complexity)

# Outline

Introduction to the seminar series

- Why Non-Asymptotic?

Basic tail and concentration bounds

- Classical bounds
  - Sub-Gaussian
  - Sub-Exponential
- Martingale based methods
- Lipschitz functions of Gaussian variables

## Lipschitz functions of Gaussian variables

- A function $f : \mathbb{R}^n \to \mathbb{R}$ is $L$ -Lipschitz with respect to the Euclidean norm $\|\cdot\|_2$ if

$$|f(x) - f(y)| \leq L\|x - y\|_2 \quad \text{for all } x, y \in \mathbb{R}^n$$

- **Theorem.** Let $(X_1, \ldots, X_n)$ be a vector of i.i.d. standard Gaussian variables, and let $f : \mathbb{R}^n \to$ R be L-Lipschitz with respect to the Euclidean norm. Then the variable $f(X) - \mathbb{E}(f(X)) \in \mathrm{SG}(L^2)$ and hence

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq 2e^{-\frac{t^2}{2L^2}} \quad \text{for all } t \geq 0$$

- **Remark.** Guarantees that any $L$-Lipschitz function of a standard Gaussian random vector, **regardless of the dimension**, exhibits concentration like a scalar Gaussian variable with variance $L^2$.

# Proof

- Assume $f$ is also differentiable, we prove a slightly weak version with loose constant.
- **Lemma.** Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Then for any convex function $\phi : \mathbb{R} \to \mathbb{R}$, we have

$$\mathbb{E}[\phi(f(X) - \mathbb{E}[f(X)])] \leq \mathbb{E}\left[\phi\left(\frac{\pi}{2}\langle \nabla f(X), Y\rangle\right)\right]$$

where $X, Y \sim \mathcal{N}(0, \mathbf{I}_n)$ are standard multivariate Gaussian, and independent

# Proof

- Using this lemma with convex fn. $\phi(x) = \exp(\lambda x)$ for any fixed $\lambda \in \mathbb{R}$,

$$\mathbb{E}_X[\exp(\lambda\{f(X) - \mathbb{E}[f(X)]\})] \leq \mathbb{E}_{X,Y}\left[\exp\left(\overbrace{\frac{\lambda\pi}{2}\langle Y, \nabla f(X)\rangle}^{\mathcal{N}(0,\frac{\lambda^2\pi^2}{4}\|\nabla f(X)\|_2^2)}\right) \mid X\right]$$

$$= \mathbb{E}_X\left[\exp\left(\frac{\lambda^2\pi^2}{8}\|\nabla f(X)\|_2^2\right)\right]$$

using the fact if $Z \sim \mathcal{N}(0, \sigma^2)$, then $\mathbb{E}(\exp(Z)) = \exp(\sigma^2/2)$

# Proof

- Using this lemma with convex fn. $\phi(x) = \exp(\lambda x)$ for any fixed $\lambda \in \mathbb{R}$,

$$\mathbb{E}_X[\exp(\lambda\{f(X) - \mathbb{E}[f(X)]\})] \leq \mathbb{E}_X\left[\exp\left(\frac{\lambda^2\pi^2}{8}\underbrace{\|\nabla f(X)\|_2^2}_{\leq L^2}\right)\right] \leq \exp\left(\frac{1}{8}\lambda^2\pi^2L^2\right)$$

- $f(X) - \mathbb{E}[f(X)] \in \mathrm{SG}((\frac{\pi L}{2})^2)$
- See proof of the lemma in the book.

# Applications

- Example 2.28 ($\chi^2$ concentration )
- Example 2.29 (Order statistics)
- Example 2.30 (Gaussian complexity)
- Example 2.31 (Gaussian chaos variables)
- Example 2.32 (singular values of Gaussian random matrices)

# References I

S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities: A nonasymptotic theory of independence. Oxford university press, 2013.

C. Jin, P. Netrapalli, R. Ge, S. M. Kakade, and M. I. Jordan. A short note on concentration inequalities for random vectors with subgaussian norm. arXiv preprint arXiv:1902.03736, 2019.

M. J. Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge University Press, 2019.