

# Summary Part 1: Probabilistic Methods for Non-Asymptotic Analysis

Tianpei Xie

Jan. 26th., 2023

## Contents

<b>1</b>	<b>Basic Inequalities</b>	<b>3</b>
1.1	Arithmetic, Calculus and Algebra . . . . .	3
1.2	Function Space, Convexity and Duality . . . . .	4
1.3	Probability Theory . . . . .	5
1.4	Information Theory . . . . .	7
<b>2</b>	<b>Summary: General Proof Stratgy for Concentration Problem</b>	<b>11</b>
<b>3</b>	<b>Summary: Distribution-Free Concentration Inequality</b>	<b>16</b>
<b>4</b>	<b>Comparison: Gaussian Tail Bound vs. Poisson Tail Bound</b>	<b>18</b>
<b>5</b>	<b>The Cramér-Chernoff Method</b>	<b>19</b>
5.1	From Markov Inequality to Cramér-Chernoff Method . . . . .	19
5.2	Sub-Gaussian Random Variables . . . . .	21
5.3	Sub-Exponential and Sub-Gamma Random Variables . . . . .	23
5.4	Hoeffding's Inequality . . . . .	27
5.5	Bernstein's Inequality . . . . .	27
5.6	Bennett's Inequality . . . . .	29
5.7	The Johnson-Lindenstrauss Lemma . . . . .	30
<b>6</b>	<b>Martingale Method</b>	<b>30</b>
6.1	Martingale and Martingale Difference Sequence . . . . .	30
6.2	Bernstein Inequality for Martingale Difference Sequence . . . . .	33
6.3	Azuma-Hoeffding Inequality . . . . .	33
6.4	Bounded Difference Inequality . . . . .	33
<b>7</b>	<b>Bounding Variance</b>	<b>34</b>
7.1	Mean-Median Deviation . . . . .	34
7.2	The Efron-Stein Inequality and Jackknife Estimation . . . . .	35
7.3	Functions with Bounded Differences . . . . .	37
7.4	Convex Poincaré Inequality . . . . .	37
7.5	Gaussian Poincaré Inequality . . . . .	37

<b>8</b>	<b>Entropy Method</b>	<b>37</b>
8.1	Entropy Functional and $\Phi$ -Entropy . . . . .	37
8.2	Dual Formulation . . . . .	38
8.3	Tensorization Property . . . . .	40
8.4	Herbst's Argument . . . . .	40
8.5	Association Inequalities . . . . .	41
8.6	Connection to Variance Bounds . . . . .	41
<b>9</b>	<b>Transportation Method</b>	<b>42</b>
9.1	Optimal Transport, Wasserstein Distance and its Dual . . . . .	42
9.2	Concentration via Transportation Cost . . . . .	46
9.3	Tensorization for Transportation Cost . . . . .	47
9.4	Induction Lemma . . . . .	47
9.5	Marton's Transportation Inequality . . . . .	47
9.6	Talagrand's Gaussian Transportation Inequality . . . . .	49
<b>10</b>	<b>Proofs of Bounded Difference Inequality</b>	<b>49</b>
10.1	Martingale Method . . . . .	49
10.2	Entropy Method . . . . .	50
10.3	Isoperimetric Inequality on Binary Hypercube . . . . .	51
10.4	Transportation Method . . . . .	53

# 1 Basic Inequalities

## 1.1 Arithmetic, Calculus and Algebra

- Remark (*Basic Inequalities*)

1. *Arithmetic Mean-Geometric Mean Inequality:*

$$\sum_{i=1}^n \alpha_i x_i \geq \prod_{i=1}^n x_i^{\alpha_i}$$

where  $\sum_{i=1}^n \alpha_i = 1$  and  $\alpha_i \geq 0$ . Simple case,  $\frac{a+b}{2} \geq \sqrt{ab}$

2.  $a^2 + b^2 \geq \pm 2ab$ ; Also  $a + b \geq 2\sqrt{ab}$

3.  $(a + b)^2 \leq 2(a^2 + b^2)$ ; Also  $(\sqrt{a} + \sqrt{b})^2 \leq 2(a + b)$

4.  $x + 1 \leq e^x$

5.  $\log x \leq x - 1$  for  $x > 0$

6.  $e^s - e^t \leq e^t(s - t)$  for  $s \geq t$

- 7.

$$h(x) := (x + 1) \log(x + 1) - x \geq \frac{x^2}{2(1 + x/3)}, \quad \text{for } x > 0$$

- 8.

$$-\log(1 - x) - x \leq \frac{x^2}{2(1 - x)}, \quad \text{for } x \in (0, 1)$$

- 9.

$$h_1(x) := 1 + x - \sqrt{1 + 2x} \geq \frac{x^2}{2(1 + x)}, \quad \text{for } x > 0.$$

10. *Log-Sum Inequality:*

For non-negative numbers  $a_1, \dots, a_n$  and  $b_1, \dots, b_n$ ,

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left( \sum_{i=1}^n a_i \right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}$$

with equality if and only if  $\frac{a_i}{b_i}$  is constant.

11. Taylor series:

$$\begin{aligned}
e^x &:= \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \dots, \quad \forall x \\
\log(1+x) &:= \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n = x - \frac{x^2}{2} + \dots, \quad |x| < 1 \\
\frac{1}{1-x} &:= \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + \dots, \quad |x| < 1 \\
(1+x)^\alpha &:= \sum_{n=0}^{\infty} \binom{\alpha}{n} x^n
\end{aligned}$$

## 1.2 Function Space, Convexity and Duality

- **Proposition 1.1 (*Jensen's inequality*)** [Vershynin, 2018]

Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space. Let  $f : \Omega \rightarrow \mathbb{R}$  be a  $\mathbb{P}$ -measurable function and  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  be **convex function**. Then

$$\varphi(\mathbb{E}[X]) := \varphi\left(\int X d\mathbb{P}\right) \leq \int \varphi \circ X d\mathbb{P} := \mathbb{E}[\varphi(X)]. \quad (1)$$

- **Remark** As a simple consequence of Jensen's inequality,  $\|X\|_{L^p}$  is an **increasing function in  $p$** , that is

$$\|X\|_{L^p} \leq \|X\|_{L^q} \quad \text{for any } 1 \leq p \leq q \leq \infty \quad (2)$$

This inequality follows since  $\varphi(x) = x^{q/p}$  is a *convex function* if  $q/p \geq 1$ .

- **Proposition 1.2 (*Minkowski's inequality*)** [Vershynin, 2018]

For any  $p \in [1, \infty]$ ,  $X, Y \in L^p(\Omega, \mathbb{P})$ ,

$$\|X + Y\|_{L^p} \leq \|X\|_{L^p} + \|Y\|_{L^p}, \quad (3)$$

which implies that  $\|\cdot\|_{L^p}$  is a norm.

- **Proposition 1.3 (*Cauchy-Schwarz inequality*)** [Vershynin, 2018]

For any random variables  $X, Y \in L^2(\Omega, \mathbb{P})$ , the following inequality is satisfied:

$$|\langle X, Y \rangle_{L^2}| := |\mathbb{E}[XY]| \leq \|X\|_{L^2} \|Y\|_{L^2}. \quad (4)$$

This inequalities can be extended to *conjugate spaces*  $L^p$  and  $L^q$

- **Proposition 1.4 (*Hölder's inequality*)** [Vershynin, 2018]

For  $p, q \in (1, \infty)$ ,  $1/p + 1/q = 1$ , then the random variables  $X \in L^p(\Omega, \mathbb{P})$ ,  $Y \in L^q(\Omega, \mathbb{P})$  satisfy

$$|\langle X, Y \rangle_{L^2}| := |\mathbb{E}[XY]| \leq \|X\|_{L^p} \|Y\|_{L^q}. \quad (5)$$

- **Definition (*Legendre Transform*)**

Let  $\mathcal{X} \subset \mathbb{R}^n$  be a **convex set**, and  $f : \mathcal{X} \rightarrow \mathbb{R}$  a **convex** function; then its **Legendre transform** is the function  $f^* : \mathcal{X}^* \rightarrow \mathbb{R}$  defined by

$$f^*(x^*) = \sup_{x \in \mathcal{X}} (\langle x^*, x \rangle - f(x)), \quad x^* \in \mathcal{X}^*$$

where  $\sup$  denotes the supremum, and the domain  $\mathcal{X}^*$  is

$$\mathcal{X}^* = \left\{ x^* \in \mathbb{R}^n : \sup_{x \in \mathcal{X}} (\langle x^*, x \rangle - f(x)) < \infty \right\}.$$

The function  $f^*$  is called the **convex conjugate function** of  $f$ .

- **Theorem 1.5 (*Fenchel's Inequality / Fenchel-Young Inequality*)**

Suppose  $f^* : \mathcal{X}^* \rightarrow \mathbb{R}$  is the convex conjugate of function  $f : \mathcal{X} \rightarrow \mathbb{R}$ . For every  $x \in \mathcal{X}$  and  $p \in \mathcal{X}^*$ , i.e., independent  $(x, p)$  pairs,

$$\langle p, x \rangle \leq f(x) + f^*(p). \quad (6)$$

- **Theorem 1.6 (*Young's Convolution Inequality*)**

Suppose  $f$  is in the Lebesgue space  $L^p(\mathbb{R}^d)$  and  $g$  is in  $L^q(\mathbb{R}^d)$  and  $1/p + 1/q = 1/r + 1$  with  $1 \leq p, q, r \leq \infty$ . Then

$$\|f * g\|_r \leq \|f\|_p \|g\|_q. \quad (7)$$

Here the star denotes **convolution**:

$$(f * g)(t) = \int_{\mathbb{R}^d} f(t - \tau)g(\tau)d\tau$$

### 1.3 Probability Theory

- Assume a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  and a random variable  $X : \Omega \rightarrow \mathbb{R}$  is a real-valued measurable function on  $\Omega$ .
- For a random variable  $X$ , the **expectation** and **variance** are denoted as

$$\begin{aligned} \mathbb{E}[X] &= \int X d\mathbb{P} \\ \text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \end{aligned}$$

- The **moment generating function** of  $X$  and its **logarithm** are denoted as

$$\begin{aligned} M_X(\lambda) &:= \mathbb{E}[e^{\lambda X}] \\ \psi_X(\lambda) &:= \log \mathbb{E}[e^{\lambda X}] \end{aligned}$$

- For  $p > 0$ , the  **$p$ -th moment** of  $X$  is defined as  $\mathbb{E}[X^p]$ , and the  **$p$ -th absolute moment** is  $\mathbb{E}[|X|^p]$ .

- The  $L^p$  *norm* of  $X$  is

$$\|X\|_{L^p} := \mathbb{E} [|X|^p]^{1/p}$$

where  $1 \leq p < \infty$ . Note that the  $L^p$  space is a *Banach space*, which is defined as

$$L^p(\Omega, \mathbb{P}) := \{X : \|X\|_{L^p} < \infty\}.$$

- The *essential supremum* of  $|X|$  is the  $L^\infty$  *norm* of  $X$

$$\|X\|_{L^\infty} := \text{ess sup } |X|$$

Similarly,  $L^\infty$  is a Banach space as well

$$L^\infty(\Omega, \mathbb{P}) := \{X : \|X\|_{L^\infty} < \infty\}.$$

- For  $p = 2$ ,  $L^2$  space is a *Hilbert space* with inner product between random variables  $X, Y \in L^2(\Omega, \mathbb{P})$

$$\langle X, Y \rangle_{L^2} := \mathbb{E} [XY] = \int XY d\mathbb{P}$$

The *standard deviation* is

$$\sigma(X) = (\text{Var}(X))^{1/2} = \|X - \mathbb{E}[X]\|_{L^2}.$$

The *covariance* is defined as

$$\begin{aligned} \text{cov}(X, Y) &:= \langle X - \mathbb{E}[X], Y - \mathbb{E}[Y] \rangle \\ &= \mathbb{E} [(X - \mathbb{E}[X]) (Y - \mathbb{E}[Y])] \end{aligned}$$

When we consider random variables as vectors in the Hilbert space  $L^2$ , the identity above gives a *geometric interpretation of the notion of covariance*. The more the vectors  $X - \mathbb{E}[X]$  and  $Y - \mathbb{E}[Y]$  are aligned with each other, the bigger their inner product and covariance are.

- The *cumulative distribution function (CDF)* is defined as

$$F_X(t) := \mathbb{P}[X \leq t], \quad t \in \mathbb{R}.$$

The following result is important

**Lemma 1.7 (Integral Identity).** [Vershynin, 2018]

Let  $X$  be a *non-negative* random variable. Then

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > t] dt. \tag{8}$$

The two sides of this identity are either finite or infinite simultaneously.

- **Theorem 1.8 (Central Limit Theorem, Linderberg-Levy)**

Let  $X_1, \dots, X_n$  be **independent identically distributed** random variables with mean  $\mathbb{E}[X_i] = 0$  and variance  $\text{Var}(X_i) = 1$ . Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \xrightarrow{d} N(0, 1) \quad (9)$$

$$\text{i.e. } \lim_{n \rightarrow \infty} \sup_{t \in \mathbb{R}} \left| \mathbb{P} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \leq t \right\} - \Phi(t) \right| = 0$$

where  $\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = \mathbb{P}\{g \leq t\}$  for some Gaussian variable  $g$ .

- **Theorem 1.9 (Central Limit Theorem, Nonasymptotic, Berry-Esseen)** [Vershynin, 2018]

Let  $X_1, \dots, X_n$  be **independent identically distributed** random variables with mean  $\mathbb{E}[X_i] = 0$ , variance  $\text{Var}(X_i) = \sigma^2$  and  $\rho := \mathbb{E}[|X_i|^3] < \infty$ . Then with some constant  $C > 0$ ,

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P} \left\{ \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n X_i \leq t \right\} - \Phi(t) \right| \leq \frac{C}{\sigma^3\sqrt{n}} \rho \quad (10)$$

where  $\Phi(t) = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du = \mathbb{P}\{g \leq t\}$  for some Gaussian variable  $g$ .

- **Remark** The *Berry-Esseen* version of central limit theorem is **non-asymptotic** and it has a bound

$$\mathbb{P} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \leq t \right\} \leq \mathbb{P}\{g \leq t\} + \frac{C}{\sqrt{n}} \rho = \int_{-\infty}^t \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du + \frac{C}{\sqrt{n}} \rho$$

This bound is **sharp**, i.e. the equality is attained when  $X_i \sim \text{Bernoulli}(1/2)$ .

- **Theorem 1.10 (Poisson Limit Theorem)**. [Vershynin, 2018]

Let  $X_{N,i}$ ,  $1 \leq i \leq N$ , be independent random variables  $X_{N,i} \sim \text{Ber}(p_{N,i})$ , and let  $S_N = \sum_{i=1}^N X_{N,i}$ . Assume that, as  $N \rightarrow \infty$

$$\max_{i \leq N} p_{N,i} \rightarrow 0 \quad \text{and} \quad \mathbb{E}[S_N] = \sum_{i=1}^N p_{N,i} \rightarrow \lambda < \infty,$$

Then, as  $N \rightarrow \infty$ ,

$$S_N = \sum_{i=1}^N X_{N,i} \xrightarrow{d} \text{Pois}(\lambda)$$

## 1.4 Information Theory

- **Definition (Shannon Entropy)** [Cover and Thomas, 2006]

Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a probability space and  $X : \mathbb{R} \rightarrow \mathcal{X}$  be a random variable. Define  $p(x)$  as the

probability density function of  $X$  with respect to a base measure  $\mu$  on  $\mathcal{X}$ . **The Shannon Entropy** is defined as

$$\begin{aligned} H(X) &:= \mathbb{E}_p [-\log p(X)] \\ &= \int_{\Omega} -\log p(X(\omega)) d\mathbb{P}(\omega) \\ &= - \int_{\mathcal{X}} p(x) \log p(x) d\mu(x) \end{aligned}$$

- **Definition (Conditional Entropy)** [Cover and Thomas, 2006]

If a pair of random variables  $(X, Y)$  follows the joint probability density function  $p(x, y)$  with respect to a base product measure  $\mu$  on  $\mathcal{X} \times \mathcal{Y}$ . Then **the joint entropy** of  $(X, Y)$ , denoted as  $H(X, Y)$ , is defined as

$$H(X, Y) := \mathbb{E}_{X, Y} [-\log p(X, Y)] = - \int_{\mathcal{X} \times \mathcal{Y}} p(x, y) \log p(x, y) d\mu(x, y)$$

Then **the conditional entropy**  $H(Y|X)$  is defined as

$$\begin{aligned} H(Y|X) &:= \mathbb{E}_{X, Y} [-\log p(Y|X)] = - \int_{\mathcal{X} \times \mathcal{Y}} p(x, y) \log p(y|x) d\mu(x, y) \\ &= \mathbb{E}_X [\mathbb{E}_Y [-\log p(Y|X)]] = \int_{\mathcal{X}} p(x) \left( - \int_{\mathcal{Y}} p(y|x) \log p(y|x) d\mu(y) \right) d\mu(x) \end{aligned}$$

- **Proposition 1.11 (Properties of Shannon Entropy)** [Cover and Thomas, 2006]

Let  $X, Y, Z$  be random variables.

1. (**Non-negativity**)  $H(X) \geq 0$ ;
2. (**Concavity**)  $H(p) := \mathbb{E}_p [-\log p(X)]$  is a concave function in terms of p.d.f.  $p$ , i.e.

$$H(\lambda p_1 + (1 - \lambda)p_2) \geq \lambda H(p_1) + (1 - \lambda)H(p_2)$$

for any two p.d.fs  $p_1, p_2$  on  $\mathcal{X}$  and any  $\lambda \in [0, 1]$ .

- **Definition (Relative Entropy / Kullback-Leibler Divergence)** [Cover and Thomas, 2006]

Suppose that  $P$  and  $Q$  are probability measures on a measurable space  $\mathcal{X}$ , and  $P$  is absolutely continuous with respect to  $Q$ , then **the relative entropy** or **the Kullback-Leibler divergence** is defined as

$$\text{KL}(P \parallel Q) := \mathbb{E}_P \left[ \log \left( \frac{dP}{dQ} \right) \right] = \int_{\mathcal{X}} \log \left( \frac{dP(x)}{dQ(x)} \right) dP(x)$$

where  $\frac{dP}{dQ}$  is the Radon-Nikodym derivative of  $P$  with respect to  $Q$ . Equivalently, the KL-divergence can be written as

$$\text{KL}(P \parallel Q) = \int_{\mathcal{X}} \left( \frac{dP(x)}{dQ(x)} \right) \log \left( \frac{dP(x)}{dQ(x)} \right) dQ(x)$$

which is the entropy of  $P$  relative to  $Q$ . Furthermore, if  $\mu$  is a base measure on  $\mathcal{X}$  for which densities  $p$  and  $q$  with  $dP = p(x)d\mu$  and  $dQ = q(x)d\mu$  exist, then

$$\text{KL}(P \parallel Q) = \int_{\mathcal{X}} p(x) \log \left( \frac{p(x)}{q(x)} \right) d\mu(x)$$



- **Definition (Mutual Information)** [Cover and Thomas, 2006]

Consider two random variables  $X, Y$  on  $\mathcal{X} \times \mathcal{Y}$  with joint probability distribution  $P_{(X,Y)}$  and marginal distribution  $P_X$  and  $P_Y$ . **The mutual information**  $I(X; Y)$  is the relative entropy between the joint distribution  $P_{(X,Y)}$  and the product distribution  $P_X \otimes P_Y$ :

$$I(X; Y) = \mathbb{KL}(P_{(X,Y)} \parallel P_X \otimes P_Y) = \mathbb{E}_{P_{(X,Y)}} \left[ \log \frac{dP_{(X,Y)}}{dP_X \otimes dP_Y} \right]$$

If  $P_{(X,Y)}$  has a probability density function  $p(x, y)$  with respect to a base measure  $\mu$  on  $\mathcal{X} \times \mathcal{Y}$ , then

$$I(X; Y) = \int_{\mathcal{X} \times \mathcal{Y}} p(x, y) \log \left( \frac{p(x, y)}{p_X(x)p_Y(y)} \right) d\mu(x, y)$$

- **Proposition 1.12 (Properties of Relative Entropy and Mutual Information)** [Cover and Thomas, 2006]

Let  $X, Y$  be random variables.

1. **(Non-negativity)** Let  $p(x), q(x)$  be probability density function of  $P, Q$ .

$$\mathbb{KL}(P \parallel Q) \geq 0$$

with equality if and only if  $p(x) = q(x)$  almost surely. Therefore, the mutual information is non-negative as well:

$$I(X; Y) \geq 0$$

with equality if and only if  $X$  and  $Y$  are independent.

2. **(Symmetry)**  $I(X; Y) = I(Y; X)$
3. **(Information Gain via Conditioning)** The mutual information  $I(X; Y)$  is the reduction in the uncertainty of  $X$  due to the knowledge of  $Y$  (and vice versa)

$$\begin{aligned} I(X; Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X, Y) \end{aligned} \tag{11}$$

4. **(Shannon Entropy as Self-Information)**  $I(X; X) = H(X)$
5. **(Joint Convexity of Relative Entropy)** The relative entropy  $\mathbb{KL}(p \parallel q)$  is **convex** in the pair  $(p, q)$ ; that is, if  $(p_1, q_1)$  and  $(p_2, q_2)$  are two pairs of probability density functions, then for  $\lambda \in [0, 1]$ ,

$$\mathbb{KL}(\lambda p_1 + (1 - \lambda)p_2 \parallel \lambda q_1 + (1 - \lambda)q_2) \leq \lambda \mathbb{KL}(p_1 \parallel q_1) + (1 - \lambda) \mathbb{KL}(p_2 \parallel q_2) \tag{12}$$

- **Proposition 1.13 (Conditioning Reduces Entropy)** [Cover and Thomas, 2006]

From non-negativity of mutual information, we see that the entropy of  $X$  is non-increasing when conditioning on  $Y$

$$H(X|Y) \leq H(X) \tag{13}$$

where equality holds if and only if  $X$  and  $Y$  are independent.

- **Proposition 1.14** (*Chain Rule for Entropy*) [Cover and Thomas, 2006]  
Let  $X_1, X_2, \dots, X_n$  be drawn according to  $p(x_1, x_2, \dots, x_n)$ . Then

$$H(X_1, X_2, \dots, X_n) = \sum_{i=1}^n H(X_i | X_{i-1}, \dots, X_1) \quad (14)$$

- **Proposition 1.15** (*Sub-Additivity of Entropy*) [Cover and Thomas, 2006]  
Let  $X_1, X_2, \dots, X_n$  be drawn according to  $p(x_1, x_2, \dots, x_n)$ . Then

$$H(X_1, X_2, \dots, X_n) \leq \sum_{i=1}^n H(X_i) \quad (15)$$

with equality if and only if the  $X_i$  are independent.

- **Proposition 1.16** (*Chain Rule for Relative Entropy*) [Cover and Thomas, 2006]  
Let  $P_{(X,Y)}$  and  $Q_{(X,Y)}$  be two probability measures on product space  $\mathcal{X} \times \mathcal{Y}$  and  $P \ll Q$ . Denote the marginal distributions  $P_X, Q_X$  and  $P_Y, Q_Y$  on  $\mathcal{X}$  and  $\mathcal{Y}$ , respectively.  $P_{Y|X}$  and  $Q_{Y|X}$  are conditional distributions (Note that  $P_{Y|X} \ll Q_{Y|X}$ ). Define **the conditional relative entropy** as

$$\mathbb{E}_X [\text{KL}(P_{Y|X} \parallel Q_{Y|X})] := \mathbb{E}_X \left[ \mathbb{E}_{P_{Y|X}} \left[ \log \left( \frac{dP_{Y|X}}{dQ_{Y|X}} \right) \right] \right].$$

Then the relative entropy of joint distribution  $P_{(X,Y)}$  with respect to  $Q_{(X,Y)}$  is

$$\text{KL}(P_{(X,Y)} \parallel Q_{(X,Y)}) = \text{KL}(P_X \parallel Q_X) + \mathbb{E}_X [\text{KL}(P_{Y|X} \parallel Q_{Y|X})] \quad (16)$$

In addition, let  $P$  and  $Q$  denote two joint distributions for  $X_1, X_2, \dots, X_n$ , let  $P_{1:i}$  and  $Q_{1:i}$  denote the marginal distributions of  $X_1, X_2, \dots, X_i$  under  $P$  and  $Q$ , respectively. Let  $P_{X_i|1\dots i-1}$  and  $Q_{X_i|1\dots i-1}$  denote the conditional distribution of  $X_i$  with respect to  $X_1, X_2, \dots, X_{i-1}$  under  $P$  and under  $Q$ .

$$\text{KL}(P \parallel Q) = \sum_{i=1}^n \mathbb{E}_{P_{1:i-1}} [\text{KL}(P_{X_i|1\dots i-1} \parallel Q_{X_i|1\dots i-1})] \quad (17)$$

- **Proposition 1.17** (*Han's Inequality*) [Cover and Thomas, 2006, Boucheron et al., 2013]  
Let  $X_1, X_2, \dots, X_n$  be random variables. Then

$$\begin{aligned} H(X_1, X_2, \dots, X_n) &\leq \frac{1}{n-1} \sum_{i=1}^n H(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n) \\ &\Leftrightarrow H(X) \leq \frac{1}{n-1} \sum_{i=1}^n H(X_{(-i)}) \end{aligned} \quad (18)$$

## 2 Summary: General Proof Stratgy for Concentration Problem

There are many proof techniques introduced. We can summarize them as follows:

### 1. *The Cramér-Chernoff Method:*

This class of methods essentially apply *the Markov inequality* on *exponential transform*  $e^{\lambda X}$  with parameter  $\lambda$ . The key is to *bound* the *log-moment generating function* from above and then use *the Legendre transform* to find the concentration bound.

Specifically, for a real-valued random variable  $X$ , any  $\lambda \geq 0$ , the following inequality holds

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

where  $\psi_X(\lambda) := \log \mathbb{E}[e^{\lambda X}]$ . One can choose optimal  $\lambda^*$  that *minimizes the upper bound above*. Since  $\psi_X(\lambda)$  is a *convex function*, we can define its *Legendre transform*

$$\psi_X^*(t) := \sup_{\lambda \in \mathbb{R}} \{\lambda t - \psi_X(\lambda)\}.$$

The expression of the right-hand side is known as the *convex conjugate* of  $\psi_X$ . The Legendre transform of log-moment generating function is also its *convex conjugate*. Thus we have

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

The lower bound can be found by applying above formula to  $-X$ .

In other word, in order to prove concentration around mean

$$\mathbb{P}\{f(X) \geq \mathbb{E}[f(X)] + t\} \text{ or } \mathbb{P}\{f(X) \leq \mathbb{E}[f(X)] - t\}$$

using *the Cramér-Chernoff Method*, we just need to find *the upper bound*  $\phi(\lambda)$  of the *logarithmic moment generating function*  $\psi(\lambda)$

$$\psi(\lambda) := \log \mathbb{E}[e^{\lambda(f(X) - \mathbb{E}[f(X)])}] \leq \phi(\lambda).$$

### Remark (*Advantages and Disadvantages of Cramér-Chernoff Method*)

There are several advantages for this method:

- (a) The derivation is *distribution-free*, since *Markov inequality* is based on fundamental properties of *measure and integration theory*. Moreover, *the bounds on logarithmic moment generating function*  $\psi(\lambda)$  can be used to *characterize different distributions* in terms of *their concentration behavior*.
- (b) This method is *widely applicable*. Most of techniques we learned here is to compute *the upper bound* for  $\psi(\lambda)$  and then apply the *Cramér-Chernoff method*.
- (c) The formula is *easy to compute* if the *simple bounds* on *logarithmic moment generating function* is *computed*. Then it will compute the rate via *Legendre transform of upper bound of*  $\psi(\lambda)$ .
- (d) The function  $\psi(\lambda)$  easily handles *product measures*  $\mathbb{P} = \otimes_{k=1}^n \mathbb{P}_k$  (i.e. *independent variables*).

$$\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda \sum_{i=1}^n X_i}] = \log \prod_{i=1}^n \mathbb{E}[e^{\lambda X_i}] = n \psi_X(\lambda)$$

and consequently,

$$\psi_Z^*(t) = n \psi_X^* \left( \frac{t}{n} \right).$$

For ***martingale difference sequence***, we see that by conditioning on previous input

$$\begin{aligned} \mathbb{E} \left[ \exp \left\{ \lambda \left( \sum_{k=1}^n D_k \right) \right\} \right] &= \mathbb{E} \left[ \mathbb{E} \left[ \exp \left\{ \lambda \left( \sum_{k=1}^n D_k \right) \right\} \mid \mathcal{B}_{n-1} \right] \right] \\ &= \mathbb{E} \left[ \exp \left\{ \lambda \left( \sum_{k=1}^{n-1} D_k \right) \right\} \mathbb{E} \left[ \exp \{ \lambda D_n \} \mid \mathcal{B}_{n-1} \right] \right] \end{aligned}$$

If we can control each martingale difference by

$$\log \mathbb{E} \left[ \exp \{ \lambda D_n \} \mid \mathcal{B}_{n-1} \right] \leq \phi(\lambda)$$

then we have

$$\begin{aligned} \psi_Z(\lambda) &\leq \log \mathbb{E} \left[ \exp \left\{ \lambda \left( \sum_{k=1}^{n-1} D_k \right) \right\} \right] + \phi(\lambda) \\ &\leq \dots \\ &\leq n\phi(\lambda). \end{aligned}$$

***The main disadvantage*** is that the Chernoff bound is ***not necessarily sharp***, since the Markov inequality is not necessarily sharp.

## 2. ***Entropy Method:***

The entropy method focus on the ***tensorization property*** of the ***entropy functional***  $\text{Ent}(X)$

$$\text{Ent}(X) := \mathbb{E} [X \log X] - \mathbb{E} [X] \log (\mathbb{E} [X]).$$

Specifically, let  $Z_1, Z_2, \dots, Z_n$  be *independent random variables* taking values in  $\mathcal{X}$ , and let  $f : \mathcal{X}^n \rightarrow [0, \infty)$  be a measurable function. Letting  $X = f(Z_1, Z_2, \dots, Z_n)$  such that  $\mathbb{E} [X \log X] < \infty$ , we have

$$\text{Ent}(X) \leq \mathbb{E} \left[ \sum_{i=1}^n \text{Ent}_{(-i)}(X) \right].$$

where  $\mathbb{E}_{(-i)}[\cdot]$  is the conditional expectation operator conditioning on  $Z_{(-i)}$ , which is equal to  $Z$  after dropping  $i$ -component. In other word, the ***key strategy*** for proving concentration using entropy method is to find ***the upper bound*** for each ***single variable entropy functional***

$$\text{Ent}_{(-i)}(X) := \mathbb{E}_{(-i)} [X \log X] - \mathbb{E}_{(-i)} [X] \log (\mathbb{E}_{(-i)} [X]) \equiv H_\Phi(\mathbb{P}_i).$$

Note that for independent random variables  $Z$ , this term ***depends only on distribution of***  $Z_i$ , since the rest  $Z_{(-i)}$  are ***controlled*** by the conditioning. For distributions such as *Gaussian, Bernoulli and Poisson*, one can use ***the logarithmic Sobolev inequalities*** to derive ***the upper bound of the entropy functional*** via ***norm of gradients***.

To obtain *the concentraion bound*, we use **the Herbst's argument**; that is, the find the bound

$$\text{Ent}(e^{\lambda X}) \leq \mathbb{E} \left[ e^{\lambda X} \right] \phi(\lambda)$$

and using the differential equation for the log-moment generating function  $\psi$

$$\frac{\text{Ent}(e^{\lambda Z})}{\mathbb{E} [e^{\lambda Z}]} = \lambda \psi'(\lambda) - \psi(\lambda) = \lambda^2 \left( \frac{\psi(\lambda)}{\lambda} \right)',$$

we can obtain the upper bound for  $\psi(\lambda)$ :

$$\begin{aligned} \left( \frac{\psi(\lambda)}{\lambda} \right)' &\leq \lambda^{-2} \phi(\lambda) \\ \left( \frac{\psi(\lambda)}{\lambda} \right) &\leq \lim_{\lambda \rightarrow 0} \left( \frac{\psi(\lambda)}{\lambda} \right) + \int_0^\lambda s^{-2} \phi(s) ds \\ \psi(\lambda) &\leq \lambda \left( \mathbb{E} [X] + \int_0^\lambda s^{-2} \phi(s) ds \right). \end{aligned}$$

Finally, we apply *the Chernoff bound*.

In general, **the key advantage** of the **entropy method** is that the tensorization property allows us to **generalize the concentration result from 1-dimensional distribution to n-dimensional product distribution**.

The main effort is to find a concentration inequality for **entropy of single variable distribution**. One way to find such concentration is to use **the logarithmic Sobolev inequalities**.

### 3. **Transportation Method:**

*The transportation method* is closed related to various *statistical divergence* esp. **the Kullback-Leibler divergence** and **the information inequality**. The central part of the proof is to show that for *given distribution*  $\mathbb{P}$  of concern, **the transportation cost inequality** holds:

$$\mathcal{W}_1^d(\mathbb{Q}, \mathbb{P}) := \min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \mathbb{E}_\gamma [d(Y, X)] \leq \phi^{*-1}(\text{KL}(\mathbb{Q} \parallel \mathbb{P})) \quad \forall \text{ distribution } \mathbb{Q}$$

where  $\Pi(\mathbb{Q}, \mathbb{P}) = \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) : Y_\# \gamma = \mathbb{Q}, X_\# \gamma = \mathbb{P}\}$  i.e.  $\gamma$  is a **coupling** of marginal distribution  $\mathbb{Q}$  and  $\mathbb{P}$ . And, for every  $s \geq 0$ ,

$$\phi^{*-1}(s) = \inf\{t \in \text{dom}(\phi^*) : \phi^*(t) > s\}$$

is defined as the **the generalized inverse** of the Legendre transform  $\phi^* = \sup_{\lambda \in (0, b)} (\lambda x - \phi(\lambda))$ .

There are *two ways to proceed*:

- (a) Based on *the duality of 1-Wasserstein distance*, this *transportation cost inequality* implies that for any 1-Lipschitz function  $f : \mathcal{X} \rightarrow \mathbb{R}$  with respect to metric  $d$

$$\mathbb{E}_\mathbb{Q} [f(Y)] - \mathbb{E}_\mathbb{P} [f(X)] = \mathbb{E}_\gamma [f(Y) - f(X)] \leq \mathcal{W}_1^d(\mathbb{Q}, \mathbb{P}) \leq \phi^{*-1}(\text{KL}(\mathbb{Q} \parallel \mathbb{P})).$$

(b) Or, we use *the Cauchy-Schwartz inequality*

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[f(Y)] - \mathbb{E}_{\mathbb{P}}[f(X)] &= \mathbb{E}_{\gamma}[f(Y) - f(X)] \leq \sum_{i=1}^n \alpha_i \mathbb{E}_{\gamma}[d(Y_i, X_i)] \\ &\leq \left( \sum_{i=1}^n \alpha_i^2 \right)^{1/2} \left( \sum_{i=1}^n (\mathbb{E}_{\gamma}[d(Y_i, X_i)])^2 \right)^{1/2}\end{aligned}$$

If we can show that the quadratic of transportation cost

$$\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^n (\mathbb{E}_{\gamma}[d(Y_i, X_i)])^2 \leq \varphi(\text{KL}(\mathbb{Q} \parallel \mathbb{P}))$$

Then

$$\mathbb{E}_{\mathbb{Q}}[f(Y)] - \mathbb{E}_{\mathbb{P}}[f(X)] \leq \left( \left( \sum_{i=1}^n \alpha_i^2 \right) \varphi(\text{KL}(\mathbb{Q} \parallel \mathbb{P})) \right)^{1/2}$$

Finally by the **transportation lemma**, we can show that

$$\psi_{f(X)}(\lambda) := \mathbb{E}_{\mathbb{P}} \left[ e^{\lambda(f(X) - \mathbb{E}[f(X)])} \right] \leq \phi(\lambda).$$

The concentration follows from *Chernoff bound* with rate function  $\phi^*(t)$ .

Note that the transportation cost inequality has **the tensorization property** as well. This allows us to generalize the the inequality from 1-dimension distribution to product distributions.

**Remark (*Advantages and Disadvantages of Transportation Method*)**

There are several advantages for this method:

- (a) *The optimal transport problem and the Wasserstein distance* is closely related to **the information geometry** of probability space  $\mathcal{P}(\mathcal{X})$ . In particular, the transportation cost inequality relates the optimal transport cost to the relative entropy:

$$\mathcal{W}_p^d(\mathbb{Q}, \mathbb{P}) \leq \varphi(\text{KL}(\mathbb{Q} \parallel \mathbb{P})).$$

This provides an alternative **information theoretical interpretation** of the concentration behavior of independent random variables.

- (b) *The low optimal transportation cost* is closely associated with **the concentration of measure** in  $\mathbb{P} \in \mathcal{P}(\mathcal{X})$ . In fact, we can bound the concentration function  $\alpha_{\mathbb{P},(\mathcal{X},d)}(t)$  from above by the upper bound of optimal transport cost.
- (c) **The dual formulation** naturally leads to *the concentration of Lipschitz function* or other *strong uniform continuous functions*.
- (d) The concept of **coupling**  $\gamma \in \Pi(\mathbb{Q}, \mathbb{P})$  allows us to extend the concentration results to **dependent variables**, such as *Markov chains*, *Markov random field* etc. In those cases, we can separate the conditional distribution  $\mathbb{P}(X_i | X_{1:i-1})$  and the marginal distributions  $\mathbb{P}(X_{1:i-1})$ .

4. **Concentration of Measure and Isoperimetric Inequalities:**

The applicability of **isoperimetric Inequalities** are *limited* to a few cases such as **Gaussian measure**, **Bernoulli measure** (or the uniform distribution on **binary hypercube**, compact manifolds, Lebesgue measure on  $\mathbb{R}^n$ , graph vertex and edge boundaries etc.).

The key is to derive the upper bound for **the concentration function**:

$$\alpha_{\mathbb{P},(\mathcal{X},d)} := \sup \left\{ \mathbb{P} \{A_t^c\} : A \subset \mathcal{X}, \mathbb{P}(A) \geq \frac{1}{2} \right\}.$$

Note that for  $d(x, A) := \inf_{y \in A} d(x, y)$ , the  $t$ -blowup of  $A$  is defined as

$$\mathbb{P} \{A_t^c\} := \mathbb{P} \{d(X, A) \geq t\}.$$

Then the goal is to find **the isoperimetric inequality**

$$\alpha_{\mathbb{P},(\mathcal{X},d)} \leq \exp(-\phi(t)) \quad \Leftrightarrow \quad \mathbb{P}(A)\mathbb{P}(A_t^c) := \mathbb{P}(A)\mathbb{P}\{d(X, A) \geq t\} \leq \exp(-\phi(t)).$$

By Levy's inequality, for *Lipschitz function*  $f : \mathcal{X} \rightarrow \mathbb{R}$ , let  $A := \{x : f(x) \leq \text{Med}(f(X))\}$ , so that  $\mathbb{P}(A) \geq 1/2$ , and the complement of  $t$ -blowup of  $A$  becomes

$$\mathbb{P}(A_t^c) = \mathbb{P}\{f(X) \geq \text{Med}(f(X)) + t\}$$

Then the above isoperimetric inequality is equivalent to

$$\begin{aligned} \mathbb{P}\{f(X) \geq \text{Med}(f(X)) + t\} &\leq 2 \exp(-\phi(t)) \\ \mathbb{P}\{f(X) \leq \text{Med}(f(X)) - t\} &\leq 2 \exp(-\phi(t)). \end{aligned}$$

Besides the existing result for *Gaussian and Bernoulli random variables*, the Talagrand's **convex distance inequality** is very useful to derive such isoperimetric inequality based on *weighted Hamming distance*.

$$\mathbb{P}(A)\mathbb{P}\left\{\sup_{\alpha \in \mathbb{R}_+^n : \|\alpha\|_2=1} \inf_{y \in A} \sum_{i=1}^n \alpha_i \mathbb{1}\{x_i \neq y_i\} \geq t\right\} \leq \exp\left(-\frac{t^2}{4}\right).$$

Note that if  $A$  is a convex set,

$$d(x, A) := \inf_{y \in A} \|x - y\|_2 \leq d_T(x, A) := \sup_{\alpha \in \mathbb{R}_+^n : \|\alpha\|_2=1} \inf_{y \in A} \sum_{i: x_i \neq y_i} \alpha_i$$

If **isoperimetric theorem** exists **for given distribution or space**, then the derived concentration bound is known to be **sharp** due to **the concentration of measure phenomenon**. This is the main **advantage** of using isoperimetric inequalities. However, proving **isoperimetric theorem** is extremely hard and thus is not widely available.

*Transportation methods* and *logarithmic Sobolev inequalities* can also use to show *isoperimetric inequalities*.

### 3 Summary: Distribution-Free Concentration Inequality

- **Remark (*Distribution-Free Concentration Inequality*)**

Some concentration results are based on **assumption on specific underlying distributions** such as *Gaussian, Bernoulli, Poisson, sub-Gaussian, sub-Gamma* etc. On the other hand, some concentration results are based on assumption on specific function class such as *bounded (actually is sub-Gaussian), Lipschitz function, bounded difference, convex function* etc. The latter results do not rely on specific distribution assumption, so it is called **the distribution-free concentration inequality**.

We list out several important inequalities:

1. **Theorem 3.1 (*Markov's Inequality*)**. [Vershynin, 2018]

For any **non-negative** random variable  $X$  and  $t > 0$ , we have

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}[X]}{t}$$

2. **Theorem 3.2 (*Chebyshev's Inequality*)**. [Vershynin, 2018]

Let  $X$  be a random variable with mean  $\mu$  and variance  $\sigma^2$ . Then, for any  $t > 0$ , we have

$$\mathbb{P}\{|X - \mu| \geq t\} \leq \frac{\sigma^2}{t^2}.$$

3. **Theorem 3.3 (*Chernoff's inequality*)** [Boucheron et al., 2013]

Let  $X$  be a real-valued random variable. For  $\lambda \geq 0$ ,  $\psi_X(\lambda)$  is the **the logarithm of moment generating function** of  $X$  and  $\psi_X^*(t)$  is its **Legendre (Cramér) transform**. Then

$$\mathbb{P}\{X \geq t\} \leq \exp(-\psi_X^*(t)).$$

4. **Theorem 3.4 (*Hoeffding's inequality*)** [Boucheron et al., 2013]

Let  $X_1, \dots, X_n$  be independent random variables such that  $X_i$  takes its values in  $[a_i, b_i]$  **almost surely** for all  $i \leq n$ . Then for every  $t > 0$ ,

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

5. **Corollary 3.5 (*Azuma-Hoeffding Inequality*)** [Wainwright, 2019]

Let  $\{(D_k, \mathcal{B}_k), k \geq 1\}$  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  $t \geq 0$ ,

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

6. **Theorem 3.6 (*McDiarmid's Inequality / Bounded Differences Inequality*)** [Boucheron et al., 2013, Wainwright, 2019]



Suppose that  $f$  satisfies **the bounded difference property** (50) with parameters  $(L_1, \dots, L_n)$  i.e. for each index  $k = 1, 2, \dots, n$ ,

$$|f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq L_k, \quad \text{for all } x, x' \in \mathcal{X}^n.$$

Assume that the random vector  $X = (X_1, X_2, \dots, X_n)$  has **independent** components. Then

$$\mathbb{P}\{|f(X) - \mathbb{E}[f(X)]| \geq t\} \leq 2 \exp\left(-\frac{2t^2}{\sum_{k=1}^n L_k^2}\right).$$

Note that functions with bounded difference property are **Lipschitz function** with respect to **Hamming distance**.

**7. Theorem 3.7 (Concentration of Separately Convex Lipschitz Functions)** [Boucheron et al., 2013]

Let  $Z := (Z_1, \dots, Z_n)$  be independent random variables, each taking values in the interval  $[a_i, b_i]$  and let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a **separately convex function** (i.e.  $f$  is **convex in each coordinate** while the **others** are **fixed**) such that

$$|f(x) - f(y)| \leq L \|x - y\| \quad \text{for all } x, y \in [0, 1]^n.$$

Then  $X = f(Z_1, \dots, Z_n)$  satisfies, for all  $t > 0$ ,

$$\mathbb{P}\{f(Z) - \mathbb{E}[f(Z)] \geq t\} \leq \exp\left(-\frac{t^2}{2L^2 \sum_{k=1}^n (b_k - a_k)^2}\right).$$

Convex Lipschitz assumption is **stronger** than bounded difference assumption.

**8. Theorem 3.8 (Concentration of Quasi-Convex Lipschitz Functions)** [Boucheron et al., 2013]

Let  $Z := (Z_1, \dots, Z_n)$  be independent random variables taking values in the interval  $[0, 1]$  and let  $f : [0, 1]^n \rightarrow \mathbb{R}$  be a **quasi-convex function**; that is

$$\{z : f(z) \leq s\} \text{ is convex set for all } s \in \mathbb{R}.$$

Moreover,  $f$  is Lipschitz function satisfying

$$|f(x) - f(y)| \leq \|x - y\| \quad \text{for all } x, y \in [0, 1]^n.$$

Then  $X = f(Z_1, \dots, Z_n)$  satisfies, for all  $t > 0$ ,

$$\begin{aligned} \mathbb{P}\{f(Z) \geq \text{Med}(f(Z)) + t\} &\leq 2 \exp\left(-\frac{t^2}{4}\right), \\ \mathbb{P}\{f(Z) \leq \text{Med}(f(Z)) - t\} &\leq 2 \exp\left(-\frac{t^2}{4}\right). \end{aligned}$$

## 4 Comparison: Gaussian Tail Bound vs. Poisson Tail Bound

- **Remark** (*Gaussian Tail Bound vs. Poisson Tail Bound*)

Based on Chernoff bound, we can derive the tail bound for two important class of distributions:

1. **Gaussian tail bound:** for any  $t > 0$ ,

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

where  $\nu > 0$ .

2. **Poisson tail bound:** for any  $t > \mathbb{E}[X] = \nu$ ,

$$\mathbb{P}\{X \geq t\} \leq \exp\left(-\nu h\left(\frac{t}{\nu}\right)\right) = e^{-\nu} \left(\frac{e\nu}{t+\nu}\right)^{t+\nu}$$

where  $h(x) = (1+x)\log(1+x) - x$  for all  $x \geq -1$  and  $\nu > 0$ .

Note that for **small deviation** around the mean, the tail of Poisson distribution  $\text{Pois}(\nu)$  behaves like *Gaussian*  $\mathcal{N}(\nu, \nu)$ :

$$\mathbb{P}\{|X - \nu| \geq t\} \leq 2 \exp\left(-\frac{ct^2}{\nu}\right)$$

But in the **large deviation** regime, the Poisson tail is **heavier** than Gaussian. Such distribution is a **sub-Gamma distribution**.

- **Remark** *The Bennet inequality* captures **the Poisson tail behavior**: for sum of  $n$  independent random variables  $X_i$  such that  $X_i \leq b$  almost surely with zero mean, finite variance  $\nu = \sum_{i=1}^n \mathbb{E}[X_i^2]$ . The Bennet inequality provides a tail bound as

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

In small deviation regime, where  $u := bt/\nu \ll 1$ ,  $h(u) \approx u^2$ , the Bennet inequality gives a Gaussian tail bound  $\approx \exp(-t^2/\nu)$ .

In large deviation regime,  $u \gg bt/\nu \geq 2$ ,  $h(u) \geq \frac{1}{2}u \log(u)$ , thus the Bennet inequality gives a Poisson tail bound  $\approx (\nu/b^2 t)^{t/(2b^2)}$

## 5 The Cramér-Chernoff Method

### 5.1 From Markov Inequality to Cramér-Chernoff Method

- **Proposition 5.1 (Markov's Inequality).** [Vershynin, 2018]  
For any **non-negative** random variable  $X$  and  $t > 0$ , we have

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}[X]}{t} \quad (19)$$

- **Proposition 5.2 (Chebyshev's Inequality).** [Vershynin, 2018]  
Let  $X$  be a random variable with mean  $\mu$  and variance  $\sigma^2$ . Then, for any  $t > 0$ , we have

$$\mathbb{P}\{|X - \mu| \geq t\} \leq \frac{\sigma^2}{t^2}. \quad (20)$$

- **Remark (Cramér-Chernoff Method)**

In this section we describe and formalize the Cramér-Chernoff bounding method. This method determines *the best possible bound* for a **tail probability** that one can possibly obtain using *Markov's inequality* with an exponential function  $\phi(t) = e^{\lambda t}$ .

Recall that for a real-valued random variable  $X$ , any  $\lambda \geq 0$ , the following inequality holds

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

where  $\psi_X(\lambda) := \log \mathbb{E}[e^{\lambda X}]$ . One can choose optimal  $\lambda^*$  that **minimizes** the upper bound above. Since  $\psi_X(\lambda)$  is a **convex function**, we can define its **Legendre transform**

$$\psi_X^*(t) := \sup_{\lambda \in \mathbb{R}} \{\lambda t - \psi_X(\lambda)\}.$$

The expression of the right-hand side is known as the **Fenchel-Legendre dual function** (or the **convex conjugate**) of  $\psi_X$ . The Legendre transform of log-moment generating function is also its convex conjugate.

In other word, in order to prove concentration around mean

$$\mathbb{P}\{f(X) \geq \mathbb{E}[f(X)] + t\} \text{ or } \mathbb{P}\{f(X) \leq \mathbb{E}[f(X)] - t\}$$

using **the Cramér-Chernoff Method**, we just need to find **the upper bound** of the **logarithmic moment generating function**

$$\psi(\lambda) := \log \mathbb{E}[e^{\lambda(f(X) - \mathbb{E}[f(X)])}] \leq \phi(\lambda)$$

- **Proposition 5.3 (Chernoff's inequality)** [Boucheron et al., 2013]  
Let  $X$  be a real-valued random variable. For  $\lambda \geq 0$ ,  $\psi_X(\lambda)$  is the **the logarithm of moment generating function** of  $X$  and  $\psi_X^*(t)$  is its **Legendre (Cramér) transform**. Then

$$\mathbb{P}\{X \geq t\} \leq \exp(-\psi_X^*(t)). \quad (21)$$

- **Remark** The *Legendre transform* is also called *the Cramér transform* [Boucheron et al., 2013].

Since  $\psi_X(0) = 0$ , its *Legendre transform*  $\psi_X^*(t)$  is *nonnegative*.

- **Definition** (*The Rate Function*)

The rate function is defined as *the Legendre transformation* of the logarithm of the moment generating function of a random variable. That is,

$$\psi_X^*(t) := \sup_{\lambda \in \mathbb{R}} \{ \lambda t - \psi_X(\lambda) \}, \quad (22)$$

where  $\psi_X(\lambda) := \log \mathbb{E} [e^{\lambda X}]$ . Thus, by *Chernoff's inequality*, we can bound the tail probabilities of random variables via its rate function.

- **Remark** (*Sums of independent random variables*)

The reason why Chernoff's inequality became popular is that it is very simple to use when applied to a sum of independent random variables. As an illustration, assume that  $Z := X_1 + \dots + X_n$  where  $X_1, \dots, X_n$  are *independent and identically distributed* real-valued random variables. Denote the logarithm of the moment-generating function of the  $X_i$  by  $\psi_X(\lambda) = \log \mathbb{E} [e^{\lambda X_i}]$ , and the corresponding *Legendre transform* by  $\psi_X^*(t)$ . Then, by independence, for all  $\lambda$  for which  $\psi_X(\lambda) < \infty$ ,

$$\psi_Z(\lambda) = \log \mathbb{E} [e^{\lambda \sum_{i=1}^n X_i}] = \log \prod_{i=1}^n \mathbb{E} [e^{\lambda X_i}] = n \psi_X(\lambda)$$

and consequently,

$$\psi_Z^*(t) = n \psi_X^*\left(\frac{t}{n}\right).$$

Thus the *Chernoff's inequality* states that

$$\mathbb{P} \{ Z \geq t \} \leq \exp(-\psi_Z^*(t)) = \exp\left(-n \psi_X^*\left(\frac{t}{n}\right)\right).$$

- **Example** (*Normal Distribution*)

Let  $X$  be a *centered normal random variable* with variance  $\sigma^2$ . Then

$$\psi_X(\lambda) = \frac{\lambda^2 \sigma^2}{2}, \quad \lambda_t = \frac{t}{\sigma^2}$$

and, therefore for every  $t > 0$ ,

$$\psi_X^*(t) = \frac{t^2}{2\sigma^2}.$$

Hence, *Chernoff's inequality* implies, for all  $t > 0$ ,

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

*Chernoff's inequality* appears to be quite sharp in this case. In fact, one can show that it cannot be improved uniformly by more than a factor of 1/2. ■

- **Example (*Poisson Distribution*)**

Let  $X$  be a **Poisson random variable** with parameter  $\nu$ , that is,  $\mathbb{P}\{X = k\} = \frac{1}{k!}e^{-\nu}\nu^k$  for all  $k = 0, 1, 2, \dots$ . Let  $Z = X - \nu$  be the *corresponding centered variable*. Then by direct calculation,

$$\psi_Z(\lambda) = \nu(e^\lambda - \lambda - 1), \quad \lambda_t = \log\left(1 + \frac{t}{\nu}\right)$$

Therefore the Legendre transform equals, for every  $t > 0$ ,

$$\psi_Z^*(t) = \nu h\left(\frac{t}{\nu}\right).$$

where the function  $h$  is defined, for all  $x \geq -1$ , by  $h(x) = (1+x)\log(1+x) - x$ . Similarly, for every  $t \leq \nu$ ,

$$\psi_{-Z}^*(t) = \nu h\left(-\frac{t}{\nu}\right).$$

- **Example (*Bernoulli Distribution*)**

Let  $X$  be a **Bernoulli random variable** with probability of success  $p$ , that is,  $\mathbb{P}\{X = 1\} = p$  and  $\mathbb{P}\{X = 0\} = 1 - p$ . Let  $Z = X - p$  be the *corresponding centered variable*. If  $0 < t < 1 - p$ , we have

$$\psi_Z(\lambda) = \log(pe^\lambda + 1 - p) - p\lambda, \quad \lambda_t = \log\frac{(1-p)(p+t)}{p(1-p-t)}$$

and therefore, for every  $t \in (0, 1 - p)$ ,

$$\psi_Z^*(t) = (1 - p - t) \log \frac{1 - p - t}{1 - p} + (p + t) \log \frac{p + t}{p}.$$

Equivalently, setting  $a = t + p$  for every  $a \in (p, 1)$ ,

$$\psi_Z^*(t) = h_p(a) = (1 - a) \log \frac{1 - a}{1 - p} + a \log \frac{a}{p}.$$

We note here that  $h_p(a)$  is just the **Kullback-Leibler divergence**  $\text{KL}(\mathbb{P}_a \parallel \mathbb{P}_p)$  between a Bernoulli distribution  $\mathbb{P}_a$  of parameter  $a$  and a Bernoulli distribution  $\mathbb{P}_p$  of parameter  $p$ .

$$\mathbb{P}\{X \geq t\} \leq \exp(-\text{KL}(\mathbb{P}_{p+t} \parallel \mathbb{P}_p))$$

## 5.2 Sub-Gaussian Random Variables

- **Definition (*Sub-Gaussian Random Variable*)**

A **centered** random variable  $X$  is said to be **sub-Gaussian with variance factor  $\nu$**  if

$$\psi_X(\lambda) \leq \frac{\lambda^2 \nu}{2}, \quad \text{for every } \lambda \in \mathbb{R}. \quad (23)$$

We denote the collection of such random variables by  $\mathcal{G}(\nu)$ .

- **Proposition 5.4 (Moment Characterization of Sub-Gaussian Random Variables)**  
[Boucheron et al., 2013]

Let  $X$  be a random variable with  $\mathbb{E}[X] = 0$ . If for some  $\nu > 0$

$$\mathbb{P}\{X > t\} \vee \mathbb{P}\{-X > t\} \leq \exp\left(-\frac{t^2}{2\nu}\right), \quad \text{for all } t > 0 \quad (24)$$

then for every integer  $q \geq 1$ ,

$$\mathbb{E}[X^{2q}] \leq 2q!(2\nu)^q \leq q!(4\nu)^q. \quad (25)$$

**Conversely**, if for some positive constant  $C$

$$\mathbb{E}[X^{2q}] \leq q!C^q,$$

then  $X \in \mathcal{G}(4C)$  (and therefore (25) holds with  $\nu = 4C$ ).

- **Proposition 5.5 (Equivalent Definitions for Sub-Gaussian Random Variables)**.  
[Vershynin, 2018]

Let  $X$  be a random variable. Then the following properties are **equivalent**; the parameters  $K_i > 0$  appearing in these properties differ from each other by at most an absolute constant factor.

1. The **tails** of  $X$  satisfy

$$\mathbb{P}\{|X| \geq t\} \leq 2 \exp(-t^2/K_1^2) \quad \text{for all } t \geq 0.$$

2. The **moments** of  $X$  satisfy

$$\|X\|_{L^p} = (\mathbb{E}[|X|^p])^{1/p} \leq K_2 \sqrt{p} \quad \text{for all } p \geq 1.$$

3. The **moment-generating function (MGF)** of  $X^2$  satisfies

$$\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(K_3^2 \lambda^2) \quad \text{for all } \lambda \text{ such that } |\lambda| \leq \frac{1}{K_3}$$

4. The **MGF** of  $X^2$  is **bounded** at some point, namely

$$\mathbb{E}[\exp(X^2/K_4^2)] \leq 2.$$

Moreover, if  $\mathbb{E}[X] = 0$  then properties (1)-(4) are also **equivalent** to the following one.

5. The **MGF** of  $X$  satisfies

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(K_5^2 \lambda^2) \quad \text{for all } \lambda \in \mathbb{R}.$$

- **Definition (Sub-Gaussian Norm)**

The **sub-gaussian norm** of  $X$ , denoted  $\|X\|_{\psi_2}$ , is defined to be the **smallest**  $K_4$  that satisfies

$$\mathbb{E}[\exp(X^2/K_4^2)] \leq 2.$$

In other words, we define

$$\|X\|_{\psi_2} = \inf \{t > 0 : \mathbb{E}[\exp(X^2/t^2)] \leq 2\}. \quad (26)$$

- **Remark** (*Sub-Gaussian Characterizations via Sub-Gaussian Norm*)

We can restate the properties of sub-gaussian random variables in terms of sub-gaussian norm:

$$\begin{aligned}\mathbb{P}\{|X| \geq t\} &\leq 2 \exp\left(-ct^2 / \|X\|_{\psi_2}^2\right) \quad \text{for all } t \geq 0; \\ \|X\|_{L^p} &\leq C \|X\|_{\psi_2} \sqrt{p} \quad \text{for all } p \geq 1; \\ \mathbb{E}\left[\exp(X^2 / \|X\|_{\psi_2}^2)\right] &\leq 2; \\ \text{if } \mathbb{E}[X] = 0, \quad \text{then } \mathbb{E}[\exp(\lambda X)] &\leq \exp(C\lambda^2 \|X\|_{\psi_2}^2) \quad \text{for all } \lambda \in \mathbb{R}.\end{aligned}$$

- **Example** Here are some classical examples of sub-gaussian distributions.

1. (**Gaussian**): As we already noted,  $X \sim N(0, 1)$  is a sub-gaussian random variable with  $\|X\|_{\psi_2} \leq C$ , where  $C$  is an absolute constant. More generally, if  $X \sim N(0, \sigma^2)$  then  $X$  is sub-gaussian with

$$\|X\|_{\psi_2} \leq C\sigma \quad (27)$$

2. (**Bernoulli**): Let  $X$  be a random variable with *symmetric Bernoulli distribution*. Since  $|X| = 1$ , it follows that  $X$  is a sub-gaussian random variable with

$$\|X\|_{\psi_2} \leq \frac{1}{\sqrt{\log 2}} \quad (28)$$

3. (**Bounded**): More generally, any *bounded random variable*  $X$  is sub-gaussian with

$$\|X\|_{\psi_2} \leq C \|X\|_{\infty} \quad (29)$$

where  $C = 1/\sqrt{\log 2}$ .

### 5.3 Sub-Exponential and Sub-Gamma Random Variables

- **Remark** For *exponential distribution*  $X \sim \exp(a)$  with rate  $a$  (inverse of scale parameter), the p.d.f. and moment generating function

$$\begin{aligned}f_X(x) &= ae^{-ax}, \quad x > 0 \\ M_X(\lambda) &= \frac{1}{1 - \lambda/a}, \quad 0 < \lambda < a\end{aligned}$$

For *Gamma distribution*  $X \sim \Gamma(a, 1/b)$  with shape parameter  $a$  and scale parameter  $b$ , the p.d.f. and the moment generating function

$$\begin{aligned}f_X(x) &= \frac{1}{\Gamma(a) b^a} x^{a-1} e^{-x/b}, \quad x > 0 \\ M_X(\lambda) &= \left(\frac{1}{1 - b\lambda}\right)^a, \quad 0 < \lambda < 1/b\end{aligned}$$

Also  $\mathbb{E}[X] = ab$  and  $\text{Var}(X) = ab^2$ .

- **Definition (*Sub-Exponential Random Variables*)**

A **nonnegative** random variable  $X$  has a **sub-exponential distribution** if there exists a constant  $a > 0$  such that

$$\mathbb{E} \left[ e^{\lambda X} \right] \leq \frac{1}{1 - \lambda/a} \quad \text{for every } \lambda \text{ such that } 0 < \lambda < a$$

or  $\psi_X(\lambda) \leq \log \left( \frac{1}{1 - \lambda/a} \right)$

- **Definition (*Sub-Gamma Random Variables*)**

A real-valued **centered** random variable  $X$  is said to be **sub-gamma on the right tail** with **variance factor**  $\nu$  and **scale parameter**  $c$  if

$$\psi_X(\lambda) \leq \frac{\lambda^2 \nu}{2(1 - c\lambda)} \quad \text{for every } \lambda \text{ such that } 0 < \lambda < 1/c$$

We denote the collection of such random variables by  $\Gamma_+(\nu, c)$ .

Similarly, a real-valued centered random variable  $X$  is said to be **sub-gamma on the left tail** with **variance factor**  $\nu$  and **scale parameter**  $c$  if  $-X$  is **sub-gamma on the right tail** with variance factor  $\nu$  and tail parameter  $c$ . We denote the collection of such random variables by  $\Gamma_-(\nu, c)$ .

Finally,  $X$  is simply said to be **sub-gamma with variance factor**  $\nu$  and **scale parameter**  $c$  if  $X$  is **sub-gamma both on the right and left tails** with **the same** variance factor  $\nu$  and scale parameter  $c$ . The collection of such random variables is denoted by  $\Gamma(\nu, c)$ .

Observe that  $\Gamma(\nu, 0) = \mathcal{G}(\nu)$ .

- **Remark** To derive the definition for sub-gamma distribution, we see that *the variance factor*  $\nu := ab^2$  and  $c := b$ . Also  $\mathbb{E}[X] = ab$ . The logarithmic moment generating function of Gamma distribution  $\Gamma(a, 1/b) = \Gamma(\nu/c^2, 1/c)$  is

$$\psi_{X - \mathbb{E}[X]}(\lambda) = a \log \left( \frac{1}{1 - b\lambda} \right) - \lambda ab \leq \frac{\lambda^2 b^2 a}{2(1 - b\lambda)} \equiv \frac{\lambda^2 \nu}{2(1 - c\lambda)}$$

The last inequality is due to

$$\log \left( \frac{1}{1 - u} \right) - u \leq \frac{u^2}{2(1 - u)}$$

- **Remark** Note that the sum of  $n$  i.i.d. random variables with exponential distribution  $\exp(1/b)$  have the Gamma distribution  $\Gamma(n, 1/b)$ . So *the sub-gamma distributed* random variable follows *the sub-exponential distribution* as well (with shape parameter = 1).

- **Proposition 5.6 (*Equivalent Definitions for Sub-Exponential Random Variables*).**  
[Vershynin, 2018]

Let  $X$  be a random variable. Then the following properties are **equivalent**; the parameters  $K_i > 0$  appearing in these properties differ from each other by at most an absolute constant factor.

1. The **tails** of  $X$  satisfy

$$\mathbb{P} \{|X| \geq t\} \leq 2 \exp(-t/K_1) \quad \text{for all } t \geq 0.$$



2. The **moments** of  $X$  satisfy

$$\|X\|_{L^p} = (\mathbb{E}[|X|^p])^{1/p} \leq K_2 p \quad \text{for all } p \geq 1.$$

3. The **moment-generating function (MGF)** of  $|X|$  satisfies

$$\mathbb{E}[\exp(\lambda |X|)] \leq \exp(K_3 \lambda) \quad \text{for all } \lambda \text{ such that } 0 \leq \lambda \leq \frac{1}{K_3}$$

4. The **MGF** of  $|X|$  is **bounded** at some point, namely

$$\mathbb{E}[\exp(|X|/K_4)] \leq 2.$$

Moreover, if  $\mathbb{E}[X] = 0$  then properties (1)-(4) are also **equivalent** to the following one.

5. The **MGF** of  $X$  satisfies

$$\mathbb{E}[\exp(\lambda X)] \leq \exp(K_5^2 \lambda^2) \quad \text{for all } \lambda \text{ such that } |\lambda| \leq \frac{1}{K_5}.$$

- **Definition (Sub-Exponential Norm)**

The **sub-exponential norm** of  $X$ , denoted  $\|X\|_{\psi_1}$ , is defined to be the **smallest**  $K_4$  that satisfies

$$\mathbb{E}[\exp(|X|/K_4)] \leq 2.$$

In other words, we define

$$\|X\|_{\psi_1} = \inf \{t > 0 : \mathbb{E}[\exp(|X|/t)] \leq 2\}. \quad (30)$$

- **Remark** Sub-gaussian and sub-exponential distributions are closely related.

1. First, *any sub-gaussian distribution is clearly sub-exponential.*
2. Second, *the square of a sub-gaussian random variable is sub-exponential:*

**Lemma 5.7 (Sub-exponential is Sub-gaussian Squared).** [Vershynin, 2018]  
A random variable  $X$  is **sub-gaussian** if and only if  $X^2$  is **sub-exponential**. Moreover,

$$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$$

More generally, *the product of two sub-gaussian random variables is sub-exponential:*

**Lemma 5.8 (Product of Sub-Gaussians is Sub-Exponential).** [Vershynin, 2018]  
Let  $X$  and  $Y$  be **sub-gaussian** random variables. Then  $XY$  is **sub-exponential**. Moreover,

$$\|XY\|_{\psi_1} \leq \|X\|_{\psi_2} \|Y\|_{\psi_2}.$$

- **Proposition 5.9 (Moment Characterization of Sub-Exponential Random Variables)**  
[Boucheron et al., 2013]

Let  $X$  be a nonnegative random variable. If  $X$  is sub-exponential distributed with parameter  $a > 0$  then for every integer  $q \geq 1$ ,

$$\mathbb{E}[X^q] \leq 2^{q+1} \frac{q!}{a^q}. \quad (31)$$

**Conversely**, if there exists a constant  $a > 0$  in order that for every positive integer  $q$ ,

$$\mathbb{E}[X^q] \leq \frac{q!}{a^q},$$

then  $X$  is sub-exponential. More precisely, for any  $0 < \lambda < a$ ,

$$\mathbb{E}[e^{\lambda X}] \leq \frac{1}{1 - \lambda/a}.$$

• **Remark (Concentration Inequalities for Sub-Gamma Distribution)**

Similarly to the *sub-Gaussian property*, the **sub-gamma property** can be characterized in terms of *tail or moment conditions*. We start by computing **the Fenchel-Legendre dual function** of

$$\psi(\lambda) = \frac{\lambda^2 \nu}{2(1 - c\lambda)}.$$

Setting

$$h_1(u) = 1 + u - \sqrt{1 + 2u} \text{ for } u > 0,$$

it follows by elementary calculation that for every  $t > 0$ ,

$$\psi^*(t) = \sup_{\lambda \in (0, 1/c)} \left\{ t\lambda - \frac{\lambda^2 \nu}{2(1 - c\lambda)} \right\} = \frac{\nu}{c^2} h_1\left(\frac{ct}{\nu}\right).$$

Since  $h_1$  is an increasing function from  $(0, \infty)$  onto  $(0, \infty)$  with **inverse function**

$$h^{-1}(u) = u + \sqrt{2u} \text{ for } u > 0,$$

we finally get

$$\psi^{*-1}(u) = \sqrt{2\nu u} + cu.$$

Hence, *Chernoff's inequality* implies that whenever  $X$  is a *sub-gamma random variable on the right tail* with *variance factor*  $\nu$  and *scale parameter*  $c$ , for every  $t > 0$ , we have

$$\mathbb{P}\{X > t\} \leq \exp\left(\frac{\nu}{c^2} h_1\left(\frac{ct}{\nu}\right)\right), \quad (32)$$

or equivalently, for every  $t > 0$ ,

$$\mathbb{P}\left\{X > \sqrt{2\nu t} + ct\right\} \leq e^{-t}. \quad (33)$$

Therefore, if  $X$  belongs to  $\Gamma(\nu, c)$ , then for every  $t > 0$ ,

$$\mathbb{P}\left\{X > \sqrt{2\nu t} + ct\right\} \vee \mathbb{P}\left\{-X > \sqrt{2\nu t} + ct\right\} \leq e^{-t}. \quad \blacksquare$$

## 5.4 Hoeffding's Inequality

- **Remark (*Bounded Variables*)**

Bounded variables are an important class of *sub-Gaussian random variables*. The *sub-Gaussian property* of *bounded random variables* is established by the following lemma:

- **Lemma 5.10 (*Hoeffding's Lemma*)** [Boucheron et al., 2013]

Let  $X$  be a random variable with  $\mathbb{E}[X] = 0$ , taking values in a **bounded interval**  $[a, b]$  and let  $\psi_X(\lambda) := \log \mathbb{E}[e^{\lambda X}]$ . Then

$$\psi_X''(\lambda) \leq \frac{(b-a)^2}{4}$$

and  $X \in \mathcal{G}((b-a)^2/4)$ .

- **Proposition 5.11 (*Hoeffding's inequality*)** [Boucheron et al., 2013]

Let  $X_1, \dots, X_n$  be independent random variables such that  $X_i$  takes its values in  $[a_i, b_i]$  **almost surely** for all  $i \leq n$ . Let

$$S = \sum_{i=1}^n (X_i - \mathbb{E}[X_i]).$$

Then for every  $t > 0$ ,

$$\mathbb{P}\{S \geq t\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \quad (34)$$

- **Proposition 5.12 (*General Hoeffding's inequality*)** [Vershynin, 2018]

Let  $X_1, \dots, X_n$  be **independent sub-gaussian** random variables. Let

$$S = \sum_{i=1}^n (X_i - \mathbb{E}[X_i]).$$

Then for every  $t > 0$ ,

$$\mathbb{P}\{S \geq t\} \leq \exp\left(-\frac{ct^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}\right). \quad (35)$$

## 5.5 Bernstein's Inequality

- **Definition (*Bernstein's Condition*)**

Given a random variable  $X$  with mean  $\mu = \mathbb{E}[X]$  we say that **Bernstein's condition** with parameter  $\nu, c$  holds if the *variance*  $\text{Var}(X) = \mathbb{E}[X^2] - \mu^2 \leq \nu$ , and

$$\sum_{i=1}^n \mathbb{E}[(X - \mu)_+^q] \leq \frac{q!}{2} \nu c^{q-2}, \quad \text{for all integers } q \geq 2,$$

where  $(x)_+ = \max\{x, 0\}$ .

- **Remark** If  $X$  is bounded, then it satisfies the Bernstein's condition.

If  $X$  satisfies the Bernstein's condition,  $X$  follows a **sub-gamma distribution**.

- **Proposition 5.13** (**Bernstein's Condition  $\Rightarrow$  Sub-Gamma Distribution**). [Boucheron et al., 2013]

Let  $X_1, \dots, X_n$  be independent real-valued random variables and each  $X_i$  satisfies **the Bernstein's condition** with parameter  $\nu$  and  $c$ . If  $S = \sum_{i=1}^n (X_i - \mathbb{E}[X_i])$ , then for all  $\lambda \in (0, 1/c)$  and  $t > 0$

$$\psi_S(\lambda) \leq \frac{\lambda^2 \nu}{2(1 - c\lambda)}$$

and

$$\psi_S^*(t) \geq \frac{\nu}{c^2} h_1\left(\frac{ct}{\nu}\right),$$

where  $h_1(u) = 1 + u - \sqrt{1 + 2u}$  for  $u > 0$ . In particular, for all  $t > 0$ ,

$$\mathbb{P}\{S \geq \sqrt{2\nu t} + ct\} \leq e^{-t}. \quad (36)$$

- **Proposition 5.14** (**Bernstein's Inequality**). [Boucheron et al., 2013]

Let  $X_1, \dots, X_n$  be independent real-valued random variables satisfying **the Bernstein's conditions** above and let  $S = \sum_{i=1}^n (X_i - \mathbb{E}[X_i])$ . Then for all  $t > 0$ ,

$$\mathbb{P}\{S \geq t\} \leq \exp\left(-\frac{t^2}{2(\nu + ct)}\right). \quad (37)$$

- **Corollary 5.15** (**Bernstein's Inequality for Bounded Distributions**). [Vershynin, 2018]

Let  $X_1, \dots, X_n$  be **independent, mean zero** random variables, such that  $|X_i| \leq b$  all  $i$ . Then, for every  $t \geq 0$ , we have

$$\mathbb{P}\left\{\left|\frac{1}{n} \sum_{i=1}^n X_i\right| \geq t\right\} \leq 2 \exp\left(-\frac{t^2}{2(\nu + bt/3)}\right). \quad (38)$$

Here  $\nu = \sum_{i=1}^n \mathbb{E}[X_i^2]$  is the variance of the sum.

- **Corollary 5.16** (**Bernstein's Inequality**). [Vershynin, 2018]

Let  $X_1, \dots, X_n$  be **independent, mean zero, sub-exponential** random variables. Then, for every  $t \geq 0$ , we have

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i\right| \geq t\right\} \leq 2 \exp\left[-c \min\left\{\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right\}\right] \quad (39)$$

where  $c > 0$  is an absolute constant.

- **Proposition 5.17** (**Bernstein's Inequality, Linear Combination Form**). [Vershynin, 2018]

Let  $X_1, \dots, X_n$  be **independent, mean zero, sub-exponential random variables**, and  $a = (a_1, \dots, a_n) \in \mathbb{R}^n$ . Then, for every  $t \geq 0$ , we have

$$\mathbb{P} \left\{ \left| \sum_{i=1}^n a_i X_i \right| \geq t \right\} \leq 2 \exp \left[ -c \min \left\{ \frac{t^2}{K^2 \|a\|_2^2}, \frac{t}{K \|a\|_\infty} \right\} \right] \quad (40)$$

where  $c > 0$  is an absolute constant and  $K = \max_i \|X_i\|_{\psi_1}$ .

- **Corollary 5.18 (Bernstein's Inequality, Average Form).** [Vershynin, 2018]  
Let  $X_1, \dots, X_n$  be **independent, mean zero, sub-exponential random variables**. Then, for every  $t \geq 0$ , we have

$$\mathbb{P} \left\{ \left| \frac{1}{n} \sum_{i=1}^n X_i \right| \geq t \right\} \leq 2 \exp \left[ -c \min \left\{ \frac{t^2}{K^2}, \frac{t}{K} \right\} n \right] \quad (41)$$

where  $K = \max_i \|X_i\|_{\psi_1}$ .

## 5.6 Bennett's Inequality

- **Remark** Our starting point is the fact that *the logarithmic moment-generating function of an independent sum equals the sum of the logarithmic moment-generating functions of the centered summands*, that is,

$$\psi_S(\lambda) = \sum_{i=1}^n \left( \log \mathbb{E} \left[ e^{\lambda X_i} \right] - \lambda \mathbb{E} [X_i] \right).$$

Using  $\log u \leq u - 1$  for  $u > 0$ ,

$$\psi_S(\lambda) \leq \sum_{i=1}^n \mathbb{E} \left[ e^{\lambda X_i} - \lambda X_i - 1 \right]. \quad (42)$$

Both Bennett's and Bernstein's inequalities may be derived from this bound, under different integrability conditions for the  $X_i$ .

- **Proposition 5.19 (Bennett's Inequality)** [Boucheron et al., 2013]  
Let  $X_1, \dots, X_n$  be independent random variables with **finite variance** such that  $X_i \leq b$  for some  $b > 0$  **almost surely** for all  $i \leq n$ . Let

$$S = \sum_{i=1}^n (X_i - \mathbb{E} [X_i])$$

and  $\nu = \sum_{i=1}^n \mathbb{E} [X_i^2]$ . If we write  $\phi(u) = e^u - u - 1$  for  $u \in \mathbb{R}$ , then, for all  $\lambda > 0$ ,

$$\log \mathbb{E} \left[ e^{\lambda S} \right] \leq n \log \left( 1 + \frac{\nu}{nb^2} \phi(b\lambda) \right) \leq \frac{\nu}{b^2} \phi(b\lambda),$$

and for any  $t > 0$ ,

$$\mathbb{P} \{ S \geq t \} \leq \exp \left( -\frac{\nu}{b^2} h \left( \frac{bt}{\nu} \right) \right) \quad (43)$$

where  $h(u) = (1 + u) \log(1 + u) - u$  for  $u > 0$ .

- **Remark** This bound can be analyzed in two different regimes:

1. In the **small deviation regime**, where  $u := bt/\nu \ll 1$ , we have asymptotically  $h(u) \approx u^2$  and Bennett's inequality gives approximately the Gaussian tail bound  $\approx \exp(-t^2/\nu)$ .
2. In the **large deviations regime**, say where  $u := bt/\nu \geq 2$ , we have  $h(u) \geq \frac{1}{2}u \log u$ , and Bennett's inequality gives a **Poisson-like tail**  $(\nu/bt)^{t/2b}$ .

## 5.7 The Johnson-Lindenstrauss Lemma

- **Remark (Overview of The Johnson-Lindenstrauss Lemma)**

The celebrated **Johnson-Lindenstrauss lemma** states roughly that, given an arbitrary set of  $n$  points in a (high-dimensional) Euclidean space, there exists a **linear embedding** of these points in a  $d$ -dimensional Euclidean space such that **all pairwise distances are preserved** within a factor of  $1 \pm \epsilon$  if  $d$  is proportional to  $(\log n)/\epsilon^2$ . It is remarkable that this result does not involve the dimension of the space to which the  $n$  points belong. In fact, the dimension of this space may even be *infinite*.

- **Definition ( $\epsilon$ -Isometry)**

Consider an arbitrary set  $A \subset \mathbb{R}^D$  or  $A \subset \mathcal{H}$  for separable Hilbert space  $\mathcal{H}$ . Given  $\epsilon \in (0, 1)$ , a map  $f : \mathbb{R}^D \rightarrow \mathbb{R}^d$  is called an  $\epsilon$ -isometry on  $A$  if for every pair  $a, a' \in A$ , we have

$$(1 - \epsilon) \|a - a'\|_2^2 \leq \|f(a) - f(a')\|_2^2 \leq (1 + \epsilon) \|a - a'\|_2^2.$$

## 6 Martingale Method

### 6.1 Martingale and Martingale Difference Sequence

- **Definition (Martingale)** [Resnick, 2013]

Let  $\{X_n, n \geq 0\}$  be a stochastic process on  $(\Omega, \mathcal{F})$  and  $\{\mathcal{F}_n, n \geq 0\}$  be a **filtration**; that is,  $\{\mathcal{F}_n, n \geq 0\}$  is an *increasing sub  $\sigma$ -fields* of  $\mathcal{F}$

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \dots \subseteq \mathcal{F}.$$

Then  $\{(X_n, \mathcal{F}_n), n \geq 0\}$  is a **martingale (mg)** if

1.  $X_n$  is **adapted** in the sense that for each  $n$ ,  $X_n \in \mathcal{F}_n$ ; that is,  $X_n$  is  $\mathcal{F}_n$ -measurable.
2.  $X_n \in L_1$ ; that is  $\mathbb{E}[|X_n|] < \infty$  for  $n \geq 0$ .
3. For  $0 \leq m < n$

$$\mathbb{E}[X_n | \mathcal{F}_m] = X_m, \quad \text{a.s.} \tag{44}$$

If the equality of (44) is replaced by  $\geq$ ; that is, things are getting better on the average:

$$\mathbb{E}[X_n | \mathcal{F}_m] \geq X_m, \quad \text{a.s.} \tag{45}$$

then  $\{X_n\}$  is called a **sub-martingale (submg)** while if things are getting worse on the average

$$\mathbb{E}[X_n | \mathcal{F}_m] \leq X_m, \quad \text{a.s.} \quad (46)$$

$\{X_n\}$  is called a **super-martingale (supermg)**.

- **Remark**  $\{X_n\}$  is **martingale** if it is both a **sub** and **supermartingale**.  $\{X_n\}$  is a **super-martingale** if and only if  $\{-X_n\}$  is a **submartingale**.
- **Remark** If  $\{X_n\}$  is a **martingale**, then  $\mathbb{E}[X_n]$  is *constant*. In the case of a **submartingale**, the mean increases and for a **supermartingale**, the mean decreases.
- **Proposition 6.1** [Resnick, 2013]  
If  $\{(X_n, \mathcal{F}_n), n \geq 0\}$  is a **(sub, super) martingale**, then

$$\{(X_n, \sigma(X_0, X_1, \dots, X_n)), n \geq 0\}$$

is also a **(sub, super) martingale**.

- **Definition (Martingale Differences)**. [Resnick, 2013]  
 $\{(d_j, \mathcal{B}_j), j \geq 0\}$  is a **(sub, super) martingale difference sequence** or a **(sub, super) fair sequence** if
  1. For  $j \geq 0$ ,  $\mathcal{B}_j \subset \mathcal{B}_{j+1}$ .
  2. For  $j \geq 0$ ,  $d_j \in L_1$ ,  $d_j \in \mathcal{B}_j$ ; that is,  $d_j$  is absolutely integrable and  $\mathcal{B}_j$ -measurable.
  3. For  $j \geq 0$ ,

$$\begin{aligned} \mathbb{E}[d_{j+1} | \mathcal{B}_j] &= 0, & (\text{martingale difference / fair sequence}); \\ &\geq 0, & (\text{submartingale difference / subfair sequence}); \\ &\leq 0, & (\text{supermartingale difference / supfair sequence}) \end{aligned}$$

- **Proposition 6.2 (Construction of Martingale From Martingale Difference)** [Resnick, 2013]  
If  $\{(d_j, \mathcal{B}_j), j \geq 0\}$  is **(sub, super) martingale difference sequence**, and

$$X_n = \sum_{j=0}^n d_j,$$

then  $\{(X_n, \mathcal{B}_n), n \geq 0\}$  is a **(sub, super) martingale**.

- **Proposition 6.3 (Construction of Martingale Difference From Martingale)** [Resnick, 2013]  
Suppose  $\{(X_n, \mathcal{B}_n), n \geq 0\}$  is a **(sub, super) martingale**. Define

$$\begin{aligned} d_0 &:= X_0 - \mathbb{E}[X_0] \\ d_j &:= X_j - X_{j-1}, \quad j \geq 1. \end{aligned}$$

Then  $\{(d_j, \mathcal{B}_j), j \geq 0\}$  is a **(sub, super) martingale difference sequence**.

- **Proposition 6.4 (Orthogonality of Martingale Differences).** [Resnick, 2013]  
If  $\{(X_n, \mathcal{B}_n), n \geq 0\}$  is a **martingale** where  $X_n$  can be decomposed as

$$X_n = \sum_{j=0}^n d_j,$$

$d_j$  is  $\mathcal{B}_j$ -measurable and  $\mathbb{E}[d_j^2] < \infty$  for  $j \geq 0$ , then  $\{d_j\}$  are **orthogonal**:

$$\mathbb{E}[d_i d_j] = 0 \quad i \neq j.$$

- **Example (Smoothing as Martingale)**

Suppose  $X \in L_1$  and  $\{\mathcal{B}_n, n \geq 0\}$  is an increasing family of sub  $\sigma$ -algebra of  $\mathcal{B}$ . Define for  $n \geq 0$

$$X_n := \mathbb{E}[X | \mathcal{B}_n].$$

Then  $(X_n, \mathcal{B}_n)$  is a **martingale**. From this result, we see that  $\{(d_n, \mathcal{B}_n), n \geq 0\}$  is a **martingale difference sequence** when

$$d_n := \mathbb{E}[X | \mathcal{B}_n] - \mathbb{E}[X | \mathcal{B}_{n-1}], \quad n \geq 1. \quad (47)$$

- **Example (Sums of Independent Random Variables)**

Suppose that  $\{Z_n, n \geq 0\}$  is an **independent** sequence of integrable random variables satisfying for  $n \geq 0$ ,  $\mathbb{E}[Z_n] = 0$ . Set

$$\begin{aligned} X_0 &:= 0, \\ X_n &:= \sum_{i=1}^n Z_i, \quad n \geq 1 \\ \mathcal{B}_n &:= \sigma(Z_0, \dots, Z_n). \end{aligned}$$

Then  $\{(X_n, \mathcal{B}_n), n \geq 0\}$  is a **martingale** since  $\{(Z_n, \mathcal{B}_n), n \geq 0\}$  is a **martingale difference sequence**.

- **Example (Likelihood Ratios).**

Suppose  $\{Y_n, n \geq 0\}$  are **independent identically distributed** random variables and suppose the true density of  $Y_n$  is  $f_0$  (The word “density” can be understood with respect to some fixed reference measure  $\mu$ .) Let  $f_1$  be some other probability density. For simplicity suppose  $f_0(y) > 0$ , for all  $y$ . For  $n \geq 0$ , define the likelihood ratio

$$\begin{aligned} X_n &:= \frac{\prod_{i=0}^n f_1(Y_i)}{\prod_{i=0}^n f_0(Y_i)} \\ \mathcal{B}_n &:= \sigma(Y_0, \dots, Y_n) \end{aligned}$$

Then  $(X_n, \mathcal{B}_n)$  is a **martingale**.



## 6.2 Bernstein Inequality for Martingale Difference Sequence

- **Proposition 6.5** (*Bernstein Inequality, Martingale Difference Sequence Version*) [Wainwright, 2019]

Let  $\{(D_k, \mathcal{B}_k), k \geq 1\}$  be a **martingale difference sequence**, and suppose that

$$\mathbb{E} [\exp(\lambda D_k) | \mathcal{B}_{k-1}] \leq \exp\left(\frac{\lambda^2 \nu_k^2}{2}\right)$$

almost surely for any  $|\lambda| < 1/\alpha_k$ . Then the following hold:

1. The sum  $\sum_{k=1}^n D_k$  is **sub-exponential** with **parameters**  $(\sqrt{\sum_{k=1}^n \nu_k^2}, \alpha_*)$  where  $\alpha_* := \max_{k=1, \dots, n} \alpha_k$ . That is, for any  $|\lambda| < 1/\alpha_*$ ,

$$\mathbb{E} \left[ \exp \left\{ \lambda \left( \sum_{k=1}^n D_k \right) \right\} \right] \leq \exp \left( \frac{\lambda^2 \sum_{k=1}^n \nu_k^2}{2} \right)$$

2. The sum satisfies **the concentration inequality**

$$\mathbb{P} \left\{ \left| \sum_{k=1}^n D_k \right| \geq t \right\} \leq \begin{cases} 2 \exp \left( -\frac{t^2}{2 \sum_{k=1}^n \nu_k^2} \right) & \text{if } 0 \leq t \leq \frac{\sum_{k=1}^n \nu_k^2}{\alpha_*} \\ 2 \exp \left( -\frac{t}{\alpha_*} \right) & \text{if } t > \frac{\sum_{k=1}^n \nu_k^2}{\alpha_*}. \end{cases} \quad (48)$$

## 6.3 Azuma-Hoeffding Inequality

- **Corollary 6.6** (*Azuma-Hoeffding Inequality*) [Wainwright, 2019]

Let  $\{(D_k, \mathcal{B}_k), k \geq 1\}$  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  $t \geq 0$ ,

$$\mathbb{P} \left\{ \left| \sum_{k=1}^n D_k \right| \geq t \right\} \leq 2 \exp \left( -\frac{2t^2}{\sum_{k=1}^n (b_k - a_k)^2} \right) \quad (49)$$

## 6.4 Bounded Difference Inequality

- An important application of *Azuma-Hoeffding Inequality* concerns functions that satisfy a **bounded difference property**.

**Definition** (*Functions with Bounded Difference Property*)

Given vectors  $x, x' \in \mathcal{X}^n$  and an index  $k \in \{1, 2, \dots, n\}$ , we define a new vector  $x^{(-k)} \in \mathcal{X}^n$  via

$$x_j^{(-k)} = \begin{cases} x_j & j \neq k \\ x'_k & j = k \end{cases}$$

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

$$\left| f(x) - f(x^{(-k)}) \right| \leq L_k, \quad \text{for all } x, x' \in \mathcal{X}^n. \quad (50)$$

- **Corollary 6.7** (*McDiarmid's Inequality / Bounded Differences Inequality*) [Wainwright, 2019]  
Suppose that  $f$  satisfies **the bounded difference property** (50) with parameters  $(L_1, \dots, L_n)$  and that the random vector  $X = (X_1, X_2, \dots, X_n)$  has **independent** components. Then

$$\mathbb{P}\{|f(X) - \mathbb{E}[f(X)]| \geq t\} \leq 2 \exp\left(-\frac{2t^2}{\sum_{k=1}^n L_k^2}\right). \quad (51)$$

## 7 Bounding Variance

### 7.1 Mean-Median Deviation

- **Definition** (*Median of Random Variable*)

The **median** of a random variable  $X \in \mathcal{X}$  with distribution  $\mathbb{P}$  is a constant  $m$  such that

$$\mathbb{P}\{X \geq m\} \geq \frac{1}{2} \quad \wedge \quad \mathbb{P}\{X \leq m\} \geq \frac{1}{2}$$

- **Proposition 7.1** (*Mean-Median Deviation, Variance Bound*) [Boucheron et al., 2013]  
Let  $X \in \mathcal{X}$  be a random variable with distribution  $\mathbb{P}$ ,  $m$  be the **median** of  $X$  and  $\mu = \mathbb{E}[X]$  be the **mean** of  $X$ . If  $\text{Var}(X) = \sigma^2 < \infty$ , then

$$|m - \mu| \leq \sqrt{\text{Var}(X)} = \sigma \quad (52)$$

(proof by Jensen's inequality  $|m - \mu| = |\mathbb{E}[X - m]| \leq \mathbb{E}[|X - m|] \leq \mathbb{E}[|X - \mu|] \leq \sqrt{\mathbb{E}[|X - \mu|^2]}$ )

- **Exercise 7.2** (*Mean-Median Deviation via Concentration Inequality*) [Boucheron et al., 2013]  
Let  $X$  be a random variable with **median**  $m$  such that positive constants  $a$  and  $b$  exist so that for all  $t > 0$ ,

$$\mathbb{P}\{|X - m| \geq t\} \leq a \exp\left(-\frac{t^2}{b}\right)$$

Show that

$$|m - \mu| \leq \min\left\{\sqrt{ab}, \frac{a}{2}\sqrt{b\pi}\right\}.$$

- **Exercise 7.3** (*Concentration Inequality Around Medians and Means*) [Wainwright, 2019]  
Given a scalar random variable  $X$ , suppose that there are positive constants  $c_1, c_2$  such that for all  $t \geq 0$ ,

$$\mathbb{P}\{|X - \mathbb{E}[X]| \geq t\} \leq c_1 \exp(-c_2 t^2) \quad (53)$$

1. Prove that  $\text{Var}(X) \leq \frac{c_1}{c_2}$

2. Let  $m_X$  be the a **median** of  $X$ . Show that **whenever the mean concentration bound (53) holds**, then for **any median**  $m_X$ , we have, for all  $t \geq 0$ , **the median concentration**

$$\mathbb{P}\{|X - m_X| \geq t\} \leq c_3 \exp(-c_4 t^2) \quad (54)$$

where  $c_3 := 4c_1$  and  $c_4 := \frac{c_2}{8}$ .

3. Conversely, show that **whenever the median concentration bound (54) holds**, then **mean concentration (53) holds** with  $c_1 = 2c_3$  and  $c_2 = \frac{c_4}{4}$ .

## 7.2 The Efron-Stein Inequality and Jackknife Estimation

- **Remark (Variance of Smoothing Martingale Difference Sequence)**

Suppose  $X \in L_1$  and  $\{\mathcal{B}_n, n \geq 0\}$  is an increasing family of sub  $\sigma$ -algebra of  $\mathcal{B}$  formed by

$$\mathcal{B}_n := \sigma(Z_1, \dots, Z_n).$$

For  $n \geq 1$ , define

$$\begin{aligned} d_0 &:= \mathbb{E}[X] \\ d_n &:= \mathbb{E}[X|\mathcal{B}_n] - \mathbb{E}[X|\mathcal{B}_{n-1}] \\ &= \mathbb{E}[X|Z_1, \dots, Z_n] - \mathbb{E}[X|Z_1, \dots, Z_{n-1}]. \end{aligned}$$

From (47) we see that  $(d_n, \mathcal{B}_n)$  is a martingale difference sequence. By *orthogonality of martingale difference*, we see that

$$\mathbb{E}[d_i d_j] = 0 \quad i \neq j.$$

Therefore, based on the decomposition

$$X - EX = \sum_{i=1}^n d_i$$

we have

$$\begin{aligned} \text{Var}(X) &= \mathbb{E}\left[\left(\sum_{i=1}^n d_i\right)^2\right] = \sum_{i=1}^n \mathbb{E}[d_i^2] + 2 \sum_{i>j} \mathbb{E}[d_i d_j] \\ &= \sum_{i=1}^n \mathbb{E}[d_i^2]. \end{aligned} \quad (55)$$

- **Remark (Variance of General Functions of Independent Random Variables)**

Then above formula (55) holds when  $X = f(Z_1, \dots, Z_n)$  for general function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  with  $n$  independent random variables  $(Z_1, \dots, Z_n)$ . By *Fubini's theorem*,

$$\mathbb{E}[X|Z_1, \dots, Z_i] = \int_{\mathcal{Z}^{n-i}} f(Z_1, \dots, Z_i, z_{i+1}, \dots, z_n) d\mu_{i+1}(z_{i+1}) \dots d\mu_n(z_n)$$

where  $\mu_j$  is the probability distribution of  $Z_j$  for  $j \geq 1$ .

Let  $Z_{(-i)} := (Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n)$  be all random variables  $(Z_1, \dots, Z_n)$  **except for**  $Z_i$ . Denote  $\mathbb{E}_{(-i)}[\cdot]$  as the conditional expectation of  $X$  given  $Z_{(-i)}$

$$\begin{aligned}\mathbb{E}_{(-i)}[X] &:= \mathbb{E}[X | Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n] \\ &= \int_{\mathcal{Z}} f(Z_1, \dots, Z_{i-1}, z_i, Z_{i+1}, \dots, Z_n) d\mu_i(z_i).\end{aligned}$$

Then, again by *Fubini's theorem (smoothing properties of conditional expectation)*,

$$\mathbb{E}[\mathbb{E}_{(-i)}[X] | Z_1, \dots, Z_i] = \mathbb{E}[X | Z_1, \dots, Z_{i-1}] \quad (56)$$

• **Proposition 7.4 (Efron-Stein Inequality)** [Boucheron et al., 2013]

Let  $Z_1, \dots, Z_n$  be **independent random variables** and let  $X = f(Z)$  be a square-integrable function of  $Z = (Z_1, \dots, Z_n)$ . Then

$$\text{Var}(X) \leq \sum_{i=1}^n \mathbb{E}[(X - \mathbb{E}_{(-i)}[X])^2] := \nu. \quad (57)$$

Moreover, if  $Z'_1, \dots, Z'_n$  are **independent** copies of  $Z_1, \dots, Z_n$  and if we define, for every  $i = 1, \dots, n$ ,

$$X'_i := f(Z_1, \dots, Z_{i-1}, Z'_i, Z_{i+1}, \dots, Z_n),$$

then

$$\nu = \frac{1}{2} \sum_{i=1}^n \mathbb{E}[(X - X'_i)^2] = \sum_{i=1}^n \mathbb{E}[(X - X'_i)_+^2] = \sum_{i=1}^n \mathbb{E}[(X - X'_i)_-^2]$$

where  $x_+ = \max\{x, 0\}$  and  $x_- = \max\{-x, 0\}$  denote the **positive** and **negative** parts of a real number  $x$ . Also,

$$\nu = \inf_{X_i} \sum_{i=1}^n \mathbb{E}[(X - X_i)^2],$$

where the infimum is taken over the class of all  $Z_{(-i)}$ -measurable and square-integrable variables  $X_i$ ,  $i = 1, \dots, n$ .

• **Example (The Jackknife Estimate)**

We should note here that the Efron-Stein inequality was first motivated by the study of the so-called **jackknife estimate of statistics**.

To describe this estimate, assume that  $Z_1, \dots, Z_n$  are i.i.d. random variables and one wishes to *estimate a functional  $\theta$  of the distribution* of the  $Z_i$  by a function  $X = f(Z_1, \dots, Z_n)$  of the data. The quality of the estimate is often measured by its bias  $\mathbb{E}[X] - \theta$  and its variance  $\text{Var}(X)$ . Since the distribution of the  $Z_i$ 's is unknown, one needs to *estimate* the bias and variance **from the same sample**. The jackknife estimate of the bias is defined by

$$(n-1) \left( \frac{1}{n} \sum_{i=1}^n X_i - X \right) \quad (58)$$

where  $X_i$  is an appropriately defined function of  $Z_{(-i)} := (Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n)$ .  $Z_{(-i)}$  is often called *the  $i$ -th jackknife sample* while  $X_i$  is the so-called *jackknife replication* of  $X$ . In an analogous way, the jackknife estimate of the variance is defined by

$$\sum_{i=1}^n (X - X_i)^2 \quad (59)$$

Using this language, *the Efron-Stein inequality* simply states that *the jackknife estimate of the variance is always positively biased*. In fact, this is how Efron and Stein originally formulated their inequality.

### 7.3 Functions with Bounded Differences

- **Corollary 7.5** [Boucheron et al., 2013]  
If  $f$  has the **bounded differences property** with parameters  $(L_1, \dots, L_n)$ , then

$$\text{Var}(f(Z)) \leq \frac{1}{4} \sum_{i=1}^n L_i^2.$$

### 7.4 Convex Poincaré Inequality

- **Theorem 7.6 (Convex Poincaré Inequality)** [Boucheron et al., 2013]  
Let  $Z_1, \dots, Z_n$  be **independent** random variables taking values in the interval  $[0, 1]$  and let  $f : [0, 1]^n \rightarrow \mathbb{R}$  be a **separately convex function** whose partial derivatives exist; that is, for every  $i = 1, \dots, n$  and fixed  $z_1, \dots, z_{i-1}, z_{i+1}, \dots, z_n$ ,  $f$  is a convex function of its  $i$ -th variable. Then  $f(Z) = f(Z_1, \dots, Z_n)$  satisfies

$$\text{Var}(f(Z)) \leq \mathbb{E} \left[ \|\nabla f(Z)\|_2^2 \right]. \quad (60)$$

### 7.5 Gaussian Poincaré Inequality

- **Theorem 7.7 (Gaussian Poincaré Inequality)** [Boucheron et al., 2013]  
Let  $Z = (Z_1, \dots, Z_n)$  be a vector of **i.i.d. standard Gaussian** random variables (i.e.  $Z$  is a Gaussian vector with **zero mean** vector and **identity covariance matrix**). Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be any **continuously differentiable** function. Then

$$\text{Var}(f(Z)) \leq \mathbb{E} \left[ \|\nabla f(Z)\|_2^2 \right]. \quad (61)$$

## 8 Entropy Method

### 8.1 Entropy Functional and $\Phi$ -Entropy

- **Definition ( $\Phi$ -Entropy)** [Boucheron et al., 2013]  
Let  $\Phi : [0, \infty) \rightarrow \mathbb{R}$  be a **convex** function, and assign, to every **non-negative integrable**

random variable  $X$ , the  $\Phi$ -entropy of  $X$  is defined as

$$H_\Phi(X) = \mathbb{E} [\Phi(X)] - \Phi(\mathbb{E} [X]). \quad (62)$$

- **Remark** The  $\Phi$ -entropy is a **functional** of *distribution*  $P_X$  instead of a function of  $X$ .
- **Remark** By Jensen's inequality, the  $\Phi$ -entropy is *non-negative*

$$\begin{aligned} \Phi(\mathbb{E} [X]) &\leq \mathbb{E} [\Phi(X)] \\ \Rightarrow H_\Phi(X) &= \mathbb{E} [\Phi(X)] - \Phi(\mathbb{E} [X]) \geq 0. \end{aligned}$$

- **Example** (*Special Examples for  $\Phi$ -Entropy*)

1. For  $\Phi(x) = x^2$ , the  $\Phi$ -entropy of  $X$  is the **variance** of  $X$ :

$$H_\Phi(X) = \mathbb{E} [X^2] - (\mathbb{E} [X])^2 = \text{Var}(X).$$

2. For  $\Phi(x) = -\log(x)$ , the  $\Phi$ -entropy of  $Y = e^{\lambda X}$  is the **logarithm of moment generating function** of  $X - \mathbb{E} [X]$ :

$$H_\Phi(e^{\lambda X}) = -\lambda \mathbb{E} [X] + \log \left( \mathbb{E} [e^{\lambda X}] \right) = \log \mathbb{E} [e^{\lambda(X - \mathbb{E} [X])}] := \psi_{X - \mathbb{E} [X]}(\lambda). \quad (63)$$

3. For  $\Phi(x) = x \log x$ , the  $\Phi$ -entropy of  $X$  is defined as the **entropy functional** of  $X$

$$H_\Phi(X) = \text{Ent}(X) := \mathbb{E} [X \log X] - \mathbb{E} [X] \log (\mathbb{E} [X]). \quad (64)$$

Let  $(\Omega, \mathcal{B})$  be measurable space, and  $P$  and  $Q$  are probability measures on  $\Omega$  with  $P \ll Q$ . Define a random variable  $X$  by the *Radon-Nikodym derivative* of  $P$  with respect to  $Q$ ; that is,

$$X(\omega) := \begin{cases} \frac{dP}{dQ}(\omega) & Q(\omega) > 0 \\ 0 & \text{o.w.} \end{cases}.$$

We see that  $X$  is  $Q$ -measurable and  $dP = X dQ$  with  $\mathbb{E}_Q [X] = 1$ . Then the entropy of  $X$  is the relative entropy of  $P$  with respect to  $Q$ .

$$\text{Ent}(X) = \text{KL}(P \parallel Q) \quad (65)$$

## 8.2 Dual Formulation

- **Lemma 8.1** The **Legendre transform** (or **convex conjugate**) of  $\Phi(x) = x \log(x)$  is  $e^{u-1}$ . That is,

$$\sup_{x>0} \{u x - x \log(x)\} = e^{u-1}$$

- **Proposition 8.2** (**Duality Formula of Entropy**) [Boucheron et al., 2013]  
Let  $X$  be a non-negative random variable defined on a probability space  $(\Omega, \mathcal{A}, P)$  such that  $\mathbb{E} [\Phi(X)] < \infty$ . Then we have **the duality formula**

$$\text{Ent}(X) = \sup_{U \in \mathcal{U}} \mathbb{E} [U X] \quad (66)$$

where the supremum is taken over the set  $\mathcal{U}$  of all random variables  $U : \Omega \rightarrow \mathbb{R} \cup \{\infty\}$  with  $\mathbb{E}[e^U] = 1$ . Moreover, if  $U$  is such that  $\mathbb{E}[UX] \leq \text{Ent}(X)$  for all non-negative random variable  $X$  such that  $\Phi(X)$  is integrable and  $\mathbb{E}[X] = 1$ , then  $\mathbb{E}[e^U] \leq 1$ .

- **Corollary 8.3 (Alternative Duality Formula of Entropy)** [Boucheron et al., 2013]

$$\text{Ent}(X) = \sup_T \mathbb{E}[X(\log(T) - \log(\mathbb{E}[T]))] \quad (67)$$

where the supremum is taken over all non-negative and integrable random variables.

- **Corollary 8.4 (Duality Formula of Log-MGF)** [Cover and Thomas, 2006, Boucheron et al., 2013]

Let  $X$  be a real-valued integrable random variable. Then for every  $\lambda \in \mathbb{R}$ ,

$$\log \mathbb{E}_{\mathbb{P}}[e^{\lambda(X - \mathbb{E}[X])}] = \sup_{\mathbb{Q} \ll \mathbb{P}} \{\lambda(\mathbb{E}_{\mathbb{Q}}[X] - \mathbb{E}_{\mathbb{P}}[X]) - \text{KL}(\mathbb{Q} \parallel \mathbb{P})\}, \quad (68)$$

where the supremum is taken over all probability measures  $\mathbb{Q}$  absolutely continuous with respect to  $\mathbb{P}$ .

- **Corollary 8.5 (Duality Formula of Kullback-Leibler Divergence)** [Cover and Thomas, 2006, Boucheron et al., 2013]

Let  $\mathbb{P}$  and  $\mathbb{Q}$  be two probability distributions on the same space. Then

$$\text{KL}(\mathbb{Q} \parallel \mathbb{P}) = \sup_X \{\mathbb{E}_{\mathbb{Q}}[X] - \log \mathbb{E}_{\mathbb{P}}[e^X]\}, \quad (69)$$

where the supremum is taken over all random variables such that  $\mathbb{E}_{\mathbb{P}}[\exp(X)] < \infty$ .

- **Definition (Bregman Divergence)**

Let  $F : \mathcal{X} \rightarrow \mathbb{R}$  be a continuously-differentiable, **strictly convex** function defined on a convex set  $\mathcal{X}$ . The **Bregman divergence** associated with  $F$  for points  $p, q \in \mathcal{X}$  is the difference between the value of  $F$  at point  $p$  and the value of the *first-order Taylor expansion* of  $F$  around point  $q$  evaluated at point  $p$ :

$$\mathbb{D}^F(p \parallel q) = F(p) - F(q) - \langle \nabla F(q), p - q \rangle \quad (70)$$

- **Theorem 8.6 (The Expected Value Minimizes Expected Bregman Divergence)** [Boucheron et al., 2013]

Let  $I \subseteq \mathbb{R}$  be an open interval and let  $f : I \rightarrow \mathbb{R}$  be **convex** and **differentiable**. For any  $x, y \in I$ , the **Bregman divergence** of  $f$  from  $x$  to  $y$  is  $f(y) - f(x) - f'(x)(y - x)$ . Let  $X$  be an  $I$ -valued random variable. Then

$$\mathbb{E}[f(X) - f(\mathbb{E}[X])] = \inf_{a \in I} \mathbb{E}[f(X) - f(a) - f'(a)(X - a)] \quad (71)$$

- **Corollary 8.7 (Duality Formula of Entropy via Bregman Divergence)** [Boucheron et al., 2013]

Let  $X$  be a non-negative random variable such that  $\mathbb{E}[\Phi(X)] < \infty$ . Then

$$\text{Ent}(X) = \inf_{u > 0} \mathbb{E}[X(\log(X) - \log(u)) - (X - u)] \quad (72)$$

### 8.3 Tensorization Property

- **Proposition 8.8** (*Sub-Additivity of The Entropy / Tensorization Property*) [Boucheron et al., 2013]

Let  $\Phi(x) = x \log x$ , for  $x > 0$  and  $\Phi(0) = 0$ . Let  $Z_1, Z_2, \dots, Z_n$  be independent random variables taking values in  $\mathcal{X}$ , and let  $f : \mathcal{X}^n \rightarrow [0, \infty)$  be a measurable function. Letting  $X = f(Z_1, Z_2, \dots, Z_n)$  such that  $\mathbb{E}[X \log X] < \infty$ , we have

$$\text{Ent}(X) := \mathbb{E}[\Phi(X)] - \Phi(\mathbb{E}[X]) \leq \sum_{i=1}^n \mathbb{E}[\mathbb{E}_{(-i)}[\Phi(X)] - \Phi(\mathbb{E}_{(-i)}[X])], \quad (73)$$

where  $\mathbb{E}_{(-i)}[\cdot]$  is the conditional expectation operator conditioning on  $Z_{(-i)}$ . Introducing the notation  $\text{Ent}_{(-i)}(X) = \mathbb{E}_{(-i)}[\Phi(X)] - \Phi(\mathbb{E}_{(-i)}[X])$ , this can be re-written as

$$\text{Ent}(X) \leq \mathbb{E} \left[ \sum_{i=1}^n \text{Ent}_{(-i)}(X) \right]. \quad (74)$$

### 8.4 Herbst's Argument

- **Remark** (*Entropy Functional for Moment Generating Function*)

Let  $X = e^{\lambda Z}$  where  $Z$  is a random variable. The entropy function of  $X$  becomes

$$\text{Ent}(e^{\lambda Z}) = \mathbb{E}[\lambda Z e^{\lambda Z}] - \mathbb{E}[e^{\lambda Z}] \log(\mathbb{E}[e^{\lambda Z}])$$

Denote  $\psi_{Z-\mathbb{E}[Z]}(\lambda) := \log \mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}]$ . Then we have

$$\frac{\text{Ent}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} = \lambda \psi'_{Z-\mathbb{E}[Z]}(\lambda) - \psi_{Z-\mathbb{E}[Z]}(\lambda). \quad (75)$$

Our strategy is based on using (75) *the sub-additivity of entropy* and then univariate calculus to derive *upper bounds for the derivative of  $\psi(\lambda)$* . By solving the obtained *differential inequality*, we obtain tail bounds via *Chernoff's bounding*.

For example, if

$$\begin{aligned} \frac{\text{Ent}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} &\leq \frac{\nu \lambda^2}{2} \\ \Leftrightarrow \lambda \psi'_{Z-\mathbb{E}[Z]}(\lambda) - \psi_{Z-\mathbb{E}[Z]}(\lambda) &\leq \frac{\nu \lambda^2}{2}, \\ \Leftrightarrow \frac{1}{\lambda} \psi'_{Z-\mathbb{E}[Z]}(\lambda) - \frac{1}{\lambda^2} \psi_{Z-\mathbb{E}[Z]}(\lambda) &\leq \frac{\nu}{2}. \end{aligned}$$

Setting  $G(\lambda) = \lambda^{-1} \psi_{Z-\mathbb{E}[Z]}(\lambda)$ , we see that the differential inequality becomes

$$G'(\lambda) \leq \frac{\nu}{2}.$$

Since  $G(\lambda) \rightarrow 0$  as  $\lambda \rightarrow 0$ , which implies that

$$G(\lambda) \leq \frac{\nu \lambda}{2},$$

and the result follows.



- **Proposition 8.9 (*Herbst's Argument*)** [Boucheron et al., 2013, Wainwright, 2019]  
Let  $Z$  be an integrable random variable such that for some  $\nu > 0$ , we have, for every  $\lambda > 0$ ,

$$\frac{\text{Ent}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} \leq \frac{\nu \lambda^2}{2} \quad (76)$$

Then, for every  $\lambda > 0$ , the logarithmic moment generating function of centered random variable  $(Z - \mathbb{E}[Z])$  satisfies

$$\psi_{Z - \mathbb{E}[Z]}(\lambda) := \log \mathbb{E} \left[ e^{\lambda(Z - \mathbb{E}[Z])} \right] \leq \frac{\nu \lambda^2}{2}.$$

## 8.5 Association Inequalities

- **Proposition 8.10 (*Chebyshev's Association Inequalities*)** [Boucheron et al., 2013]  
Let  $f$  and  $g$  be **nondecreasing** real-valued functions defined on the real line. If  $X$  is a real-valued random variable and  $Y$  is a **nonnegative** random variable, then

$$\mathbb{E}[Y] \mathbb{E}[Y f(X) g(X)] \geq \mathbb{E}[Y f(X)] \mathbb{E}[Y g(X)] \quad (77)$$

If  $f$  is **nonincreasing** and  $g$  is **nondecreasing** then

$$\mathbb{E}[Y] \mathbb{E}[Y f(X) g(X)] \leq \mathbb{E}[Y f(X)] \mathbb{E}[Y g(X)] \quad (78)$$

- **Proposition 8.11 (*Harris's Inequalities*)** [Boucheron et al., 2013]  
Let  $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$  be **nondecreasing** functions. Let  $Z_1, \dots, Z_n$  be **independent** real-valued random variables and define the random vector  $Z = (Z_1, \dots, Z_n)$  taking values in  $\mathbb{R}^n$ . Then

$$\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)] \mathbb{E}[g(X)] \quad (79)$$

If  $f$  is **nonincreasing** and  $g$  is **nondecreasing** then

$$\mathbb{E}[f(X)g(X)] \leq \mathbb{E}[f(X)] \mathbb{E}[g(X)] \quad (80)$$

## 8.6 Connection to Variance Bounds

- **Proposition 8.12 (*A Modified Logarithmic Sobolev Inequalities for Moment Generating Function*)** [Boucheron et al., 2013]  
Consider independent random variables  $Z_1, \dots, Z_n$  taking values in  $\mathcal{X}$ , a real-valued function  $f : \mathcal{X}^n \rightarrow \mathbb{R}$  and the random variable  $X = f(Z_1, \dots, Z_n)$ . Also denote  $Z_{(-i)} = (Z_1, \dots, Z_{i-1}, Z_{i+1}, \dots, Z_n)$  and  $X_{(-i)} = f_i(Z_{(-i)})$  where  $f_i : \mathcal{X}^{n-1} \rightarrow \mathbb{R}$  is an arbitrary function. Let  $\phi(x) = e^x - x - 1$ . Then for all  $\lambda \in \mathbb{R}$ ,

$$\text{Ent}(e^{\lambda X}) := \mathbb{E}[\lambda X e^{\lambda X}] - \mathbb{E}[e^{\lambda X}] \log \mathbb{E}[e^{\lambda X}] \leq \sum_{i=1}^n \mathbb{E} \left[ e^{\lambda X} \phi(-\lambda(X - X_{(-i)})) \right] \quad (81)$$

- **Proposition 8.13 (*Symmetrized Modified Logarithmic Sobolev Inequalities*)** [Boucheron et al., 2013]

Consider independent random variables  $Z_1, \dots, Z_n$  taking values in  $\mathcal{X}$ , a real-valued function  $f : \mathcal{X}^n \rightarrow \mathbb{R}$  and the random variable  $X = f(Z_1, \dots, Z_n)$ . Also denote  $\tilde{X}^{(i)} = f(Z_1, \dots, Z_{i-1}, Z'_i, Z_{i+1}, \dots, Z_n)$ . Let  $\phi(x) = e^x - x - 1$ . Then for all  $\lambda \in \mathbb{R}$ ,

$$\lambda \mathbb{E} [X e^{\lambda X}] - \mathbb{E} [e^{\lambda X}] \log \mathbb{E} [e^{\lambda X}] \leq \sum_{i=1}^n \mathbb{E} [e^{\lambda X} \phi(-\lambda(X - \tilde{X}^{(i)}))] \quad (82)$$

Moreover, denoting  $\tau(x) = x(e^x - 1)$ , for all  $\lambda \in \mathbb{R}$ ,

$$\begin{aligned} \lambda \mathbb{E} [X e^{\lambda X}] - \mathbb{E} [e^{\lambda X}] \log \mathbb{E} [e^{\lambda X}] &\leq \sum_{i=1}^n \mathbb{E} [e^{\lambda X} \tau(-\lambda(X - \tilde{X}^{(i)})_+)] , \\ \lambda \mathbb{E} [X e^{\lambda X}] - \mathbb{E} [e^{\lambda X}] \log \mathbb{E} [e^{\lambda X}] &\leq \sum_{i=1}^n \mathbb{E} [e^{\lambda X} \tau(\lambda(\tilde{X}^{(i)} - X)_-)] . \end{aligned}$$

• **Remark** In the proof, we have

$$\begin{aligned} \text{Ent}_{\mu_i}(e^{\lambda X}) &\leq \mathbb{E}_{\mu_i} [e^{\lambda X} (\log e^{\lambda X} - \log e^{\lambda X'_i}) - (e^{\lambda X} - e^{\lambda X'_i})] \\ &= \mathbb{E}_{\mu_i} [e^{\lambda X} (\lambda(X - X'_i) - (e^{\lambda X} - e^{\lambda X'_i}))] \\ &\leq \mathbb{E}_{\mu_i} [(e^{\lambda X} - e^{\lambda X'_i})(\lambda(X - X'_i)_+)] \\ &\leq \lambda^2 \mathbb{E}_{\mu_i} [(X - X'_i)_+^2] \end{aligned}$$

Using the convexity of  $e^x$ , we have  $e^s - e^t \leq e^t(s - t)$  if  $s > t$ . Thus

$$\text{Ent}(e^{\lambda X}) \leq \lambda^2 \sum_{i=1}^n \mathbb{E} [(X - X'_i)_+^2] .$$

From Efron-Stein inequality, if we can bound

$$\sum_{i=1}^n \mathbb{E} [(X - X'_i)_+^2] \leq \nu ,$$

then we can bound both the variance  $\text{Var}(X)$  and entropy  $\text{Ent}(e^{\lambda X})$ .

## 9 Transportation Method

### 9.1 Optimal Transport, Wasserstein Distance and its Dual

• **Definition (*Pushforward Measure*)** [Peyr and Cuturi, 2019]

Let  $(\mathcal{X}, \mathcal{B}_X)$  and  $(\mathcal{Y}, \mathcal{B}_Y)$  be two topological measurable spaces. Denote the spaces of *general (Radon) measures* on  $\mathcal{X}, \mathcal{Y}$  as  $\mathcal{M}(\mathcal{X})$  and  $\mathcal{M}(\mathcal{Y})$ . Also let  $\mathcal{C}(\mathcal{X})$  be space of continuous functions on  $\mathcal{X}$ . For a *continous* map  $T : \mathcal{X} \rightarrow \mathcal{Y}$ , the **push-forward operator** is defined as  $T_{\#} : \mathcal{M}(\mathcal{X}) \rightarrow \mathcal{M}(\mathcal{Y})$  that satisfies

$$\forall h \in \mathcal{C}(\mathcal{X}), \quad \int_{\mathcal{Y}} h(y) d(T_{\#}\alpha)(y) = \int_{\mathcal{X}} h(T(x)) d\alpha(x). \quad (83)$$

$$\text{or equivalently,} \quad (T_{\#}\alpha)(B) := \alpha(\{x : T(x) \in B \subset \mathcal{Y}\}) = \alpha(T^{-1}(B)) \quad (84)$$

where the *push-forward measure*  $\beta := T_{\#}\alpha \in \mathcal{M}(\mathcal{Y})$  of some  $\alpha \in \mathcal{M}(\mathcal{X})$ ,  $T^{-1}(\cdot)$  is the pre-image of  $T$ .

- **Remark (*Density Function of Pushforward Measure*)**

Assume that  $(\alpha, \beta)$  have densities  $(\rho_{\alpha}, \rho_{\beta})$  with respect to a fixed measure, and  $\beta = T_{\#}\alpha$ . We see that  $T_{\#}$  acts on a density  $\rho_{\alpha}$  linearly to a density  $\rho_{\beta}$  as a change of variable, i.e.

$$\begin{aligned} \rho_{\alpha}(\mathbf{x}) &= |\det(T'(\mathbf{x}))| \rho_{\beta}(T(\mathbf{x})) \\ |\det(T'(\mathbf{x}))| &= \frac{\rho_{\alpha}(\mathbf{x})}{\rho_{\beta}(T(\mathbf{x}))} \end{aligned} \tag{85}$$

- **Definition (*Optimal Transport Problem, Monge Problem*)** [Villani, 2009, Santambrogio, 2015, Peyr and Cuturi, 2019]

Let  $(\mathcal{X}, \mathcal{B}_X)$  and  $(\mathcal{Y}, \mathcal{B}_Y)$  be two measurable spaces, where  $\mathcal{X}$  and  $\mathcal{Y}$  are *complete separable metric spaces*. Denote  $\mathcal{P}(\mathcal{X})$  and  $\mathcal{P}(\mathcal{Y})$  as the space of probability measures on  $\mathcal{X}$  and  $\mathcal{Y}$ . Define a *cost function*  $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$  as non-negative real-valued measurable functions on  $\mathcal{X} \times \mathcal{Y}$ . The optimal transport problem by Monge (i.e. ***Monge Problem***) is defined as follows: given two probability measures  $\mathbb{P} \in \mathcal{P}(\mathcal{X})$  and  $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ , find a *continuous measurable map*  $T : \mathcal{X} \rightarrow \mathcal{Y}$  so that

$$\begin{aligned} \inf_T \int_{\mathcal{X}} c(x, T(x)) d\mathbb{P}(x) \\ \text{s.t. } \mathbb{Q} = T_{\#}\mathbb{P} \end{aligned}$$

The optimal solution  $T$  is also called an *optimal transportation plan*.

- **Definition (*Optimal Transport Problem, Kantorovich Relaxation*)** [Villani, 2009, Santambrogio, 2015, Peyr and Cuturi, 2019]

The optimal transport problem by Kantorovich (i.e. ***Kantorovich Relaxation***) is defined as follows: given two probability measures  $\mathbb{P} \in \mathcal{P}(\mathcal{X})$  and  $\mathbb{Q} \in \mathcal{P}(\mathcal{Y})$ , find a *joint probability measure*  $\gamma \in \Pi(\mathbb{P}, \mathbb{Q})$  so that

$$\begin{aligned} \inf_{\gamma} \int_{\mathcal{X} \times \mathcal{Y}} c(x, y) d\gamma(x, y) \\ \text{s.t. } \gamma \in \Pi(\mathbb{P}, \mathbb{Q}) := \{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y}) : \pi_{\mathcal{X}, \#}\gamma = \mathbb{P}, \pi_{\mathcal{Y}, \#}\gamma = \mathbb{Q}\} \end{aligned}$$

where  $\mathcal{P}(\mathcal{X} \times \mathcal{Y})$  is the space of joint probability measure on  $\mathcal{X} \times \mathcal{Y}$ ,  $\pi_{\mathcal{X}}$  and  $\pi_{\mathcal{Y}}$  are the coordinate projection onto  $\mathcal{X}$  and  $\mathcal{Y}$ .  $\pi_{\mathcal{X}, \#}\gamma = \mathbb{P}$  means that  $\mathbb{P}$  is the marginal distribution of  $\gamma$  on  $\mathcal{X}$ . Similarly  $\mathbb{Q}$  is the marginal distribution of  $\gamma$  on  $\mathcal{Y}$ .

Equivalently, let  $X$  and  $Y$  are *random variables* taking values in  $\mathcal{X}$  and  $\mathcal{Y}$ . The *joint distribution* of  $(X, Y)$  is  $\gamma$  with marginal distribution of  $X$  and  $Y$  being  $\mathbb{P}$  and  $\mathbb{Q}$ . Then the problem is

$$\min_{\gamma \in \Pi(\mathbb{P}, \mathbb{Q})} \mathbb{E}_{\gamma} [c(X, Y)]$$

The joint distribution  $\gamma \in \Pi(\mathbb{P}, \mathbb{Q})$  such that  $X_{\#}\gamma = \mathbb{P}$  and  $Y_{\#}\gamma = \mathbb{Q}$  is called a *coupling*.

- **Definition (*Dual Problem of Kantorovich Problem*)** [Villani, 2009, Santambrogio, 2015, Peyr and Cuturi, 2019]

The **dual problem** of *Kantorovich problem* is described as below:

$$\begin{aligned} \mathcal{L}_c(\mathbb{P}, \mathbb{Q}) = & \max_{(\varphi, \psi) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y})} \int_{\mathcal{X}} \varphi(x) d\mathbb{P}(x) + \int_{\mathcal{Y}} \psi(y) d\mathbb{Q}(y) \\ \text{s.t. } & \varphi(x) + \psi(y) \leq c(x, y), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}, \end{aligned}$$

Here,  $(\varphi, \psi)$  is a pair of *continuous functions* on  $\mathcal{X}$  and  $\mathcal{Y}$  respectively and they are also the **Kantorovich potentials**. The feasible region is

$$\mathcal{R}(c) := \{(\varphi, \psi) \in \mathcal{C}(\mathcal{X}) \times \mathcal{C}(\mathcal{Y}) : \varphi \oplus \psi \leq c\}$$

where  $(\varphi \oplus \psi)(x, y) = \varphi(x) + \psi(y)$ .

In other words, the dual optimization problem is

$$\max_{(\varphi, \psi) \in \mathcal{R}(c)} \mathbb{E}_{\mathbb{P}}[\varphi(X)] + \mathbb{E}_{\mathbb{Q}}[\psi(Y)]$$

• **Proposition 9.1 (Strong Duality)** [Santambrogio, 2015]

Let  $\mathcal{X}, \mathcal{Y}$  be **complete separable spaces**, and  $c : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$  be **lower semi-continuous and bounded from below**. Then the optimal value of primal and dual problems are the same

$$\min_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} \mathbb{E}[c(X, Y)] = \mathcal{L}_c(\mathbb{P}, \mathbb{Q}) = \max_{(\varphi, \psi) \in \mathcal{R}(c)} \mathbb{E}_{\mathbb{P}}[\varphi(X)] + \mathbb{E}_{\mathbb{Q}}[\psi(Y)].$$

• **Definition (Wasserstein Distance)**

Let  $((\mathcal{X}, d), \mathcal{B})$  be a *metric measurable space* with *Borel  $\sigma$ -algebra* induced by metric  $d$ . Let  $X, Y$  be two random variables taking values in  $\mathcal{X}$  with distribution  $\mathbb{P}$  and  $\mathbb{Q}$ . **The Wasserstein distance** between probability distributions  $\mathbb{P}$  and  $\mathbb{Q}$  induced by  $d$  is defined as

$$\mathcal{W}_1(\mathbb{P}, \mathbb{Q}) \equiv \mathcal{W}_d(\mathbb{P}, \mathbb{Q}) := \min_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} \mathbb{E}[d(X, Y)] \quad (86)$$

In general, for  $p \in [1, \infty)$ , we can define **Wasserstein  $p$ -distance** as

$$\mathcal{W}_p(\mathbb{P}, \mathbb{Q}) \equiv \mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) := \left( \min_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} \mathbb{E}[(d(X, Y))^p] \right)^{1/p}. \quad (87)$$

• **Remark** Not to confuse the **2-Wasserstein distance** with **the Wasserstein distance induced by  $L_2$  norm**:

$$\begin{aligned} \mathcal{W}_{\|\cdot\|_2}(\mathbb{P}, \mathbb{Q}) &\equiv \mathcal{W}_{1,\|\cdot\|_2}(\mathbb{P}, \mathbb{Q}) := \min_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} \mathbb{E}[\|X - Y\|_2] \\ \mathcal{W}_2(\mathbb{P}, \mathbb{Q}) &\equiv \mathcal{W}_{2,d}(\mathbb{P}, \mathbb{Q}) := \sqrt{\min_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} \mathbb{E}[d(X, Y)^2]} \end{aligned}$$

• **Remark (Wasserstein  $p$ -Distance is a Metric in  $\mathcal{P}(\mathcal{X})$ )**

The **Wasserstein  $p$ -distance**  $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) := (\min_{X \sim \mathbb{P}, Y \sim \mathbb{Q}} \mathbb{E}[(d(X, Y))^p])^{1/p}$  is a well-defined metric in  $\mathcal{P}(\mathcal{X})$ : for all  $\mathbb{P}, \mathbb{Q}, \mathbb{M} \in \mathcal{P}(\mathcal{X})$ ,

1. (Non-Negativity):  $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) \geq 0$ .

2. (*Definiteness*):  $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) = 0$  iff  $\mathbb{P} = \mathbb{Q}$
3. (*Symmetric*):  $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) = \mathcal{W}_{p,d}(\mathbb{Q}, \mathbb{P})$
4. (*Triangular inequality*):  $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) \leq \mathcal{W}_{p,d}(\mathbb{P}, \mathbb{M}) + \mathcal{W}_{p,d}(\mathbb{M}, \mathbb{Q})$

• **Definition (*Total Variation / Variational Distance*)**

Let  $P, Q$  be two probability measures on measurable space  $(\Omega, \mathcal{F})$ . The ***total variation*** or ***variational distance*** between  $P$  and  $Q$  is defined by

$$V(P, Q) := \sup_{A \in \mathcal{F}} |P(A) - Q(A)| \quad (88)$$

• **Remark (*Equivalent Formulation of Total Variation*)**

It is a well-known and simple fact that the total variation is half the  $L_1$ -distance, that is, if  $\mu$  is a *common dominating measure* of  $P$  and  $Q$  and  $p(x) = dP/d\mu$  and  $q(x) = dQ/d\mu$  denote their respective densities, then

$$V(P, Q) := P(A^*) - Q(A^*) = \frac{1}{2} \int_{\Omega} |p(x) - q(x)| d\mu(x), \quad (89)$$

where  $A^* = \{x : p(x) \geq q(x)\}$ .

• **Remark (*Total Variation via Optimal Coupling of Two Measures*)**

We note that another important interpretation of the *variational distance* is related to the *best coupling of the two measures*

$$V(P, Q) = \min P\{X \neq Y\} \quad (90)$$

where the minimum is taken over *all pairs of joint distributions* for the random variables  $(X, Y)$  whose marginal distributions are  $X \sim P$  and  $Y \sim Q$ .

• **Proposition 9.2 (*Pinsker's Inequality*)** [Cover and Thomas, 2006, Boucheron et al., 2013]

Let  $P, Q$  be two probability distributions on measurable space  $(\Omega, \mathcal{F})$  such that  $P \ll Q$ . Then

$$V(P, Q)^2 \leq \frac{1}{2} \text{KL}(P \parallel Q). \quad (91)$$

• **Remark (*Total Variation as 1-Wasserstein Distance*)**

The total variation between  $P$  and  $Q$  is the ***Wasserstein distance*** induced by the ***Hamming distance***  $d(x, y) = \# \{i : x_i \neq y_i\}$ .

$$V(P, Q) = \mathcal{W}_1(P, Q).$$

Thus the *Pinsker's inequality* (91) is the special case of *transportation cost inequality* (93).

• **Theorem 9.3 (*Kantorovich-Rubenstein Duality*)** [Villani, 2009]

Let  $\mathcal{X}$  be a ***Polish space***, i.e.  $\mathcal{X}$  a ***complete separable metric space*** equipped with a Borel  $\sigma$ -algebra induced by metric  $d$ , and  $\mathbb{P}$  and  $\mathbb{Q}$  be probability measures on  $\mathcal{X}$ . For fixed  $p \in [1, \infty)$ , let  $\text{Lip}_1$  be the space of all 1-***Lipschitz*** function with respect to metric  $d$  such that

$$\|f\|_L := \sup_{x, y \in \mathcal{X}} \left\{ \frac{|f(x) - f(y)|}{d(x, y)} \right\} \leq 1.$$

Then

$$\mathcal{W}_d(\mathbb{P}, \mathbb{Q}) \equiv \mathcal{W}_{1,d}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \text{Lip}_1} \{ \mathbb{E}_{\mathbb{P}}[f(X)] - \mathbb{E}_{\mathbb{Q}}[f(Y)] \}. \quad (92)$$

## 9.2 Concentration via Transportation Cost

- **Lemma 9.4 (Transportation Lemma)** [Boucheron et al., 2013]

Let  $X$  be a real-valued integrable random variable. Let  $\phi$  be a **convex** and **continuously differentiable** function on a (possibly unbounded) interval  $[0, b)$  and assume that  $\phi(0) = \phi'(0) = 0$ . Define, for every  $x \geq 0$ , **the Legendre transform**  $\phi^*(x) = \sup_{\lambda \in (0, b)} (\lambda x - \phi(\lambda))$ , and let, for every  $t \geq 0$ ,  $\phi^{*-1}(t) = \inf\{x \geq 0 : \phi^*(x) > t\}$ , i.e. **the generalized inverse** of  $\phi^*$ . Then the following two statements are equivalent:

1. for every  $\lambda \in (0, b)$ ,

$$\psi_{X - \mathbb{E}[X]}(\lambda) \leq \phi(\lambda)$$

where  $\psi_X(\lambda) := \log \mathbb{E}_Q [e^{\lambda X}]$  is the logarithm of moment generating function;

2. for any probability measure  $P$  absolutely continuous with respect to  $Q$  such that  $\text{KL}(P \parallel Q) < \infty$ ,

$$\mathbb{E}_P[X] - \mathbb{E}_Q[X] \leq \phi^{*-1}(\text{KL}(P \parallel Q)). \quad (93)$$

In particular, given  $\nu > 0$ ,  $X$  follows a *sub-Gaussian distribution*, i.e.

$$\psi_{X - \mathbb{E}[X]}(\lambda) \leq \frac{\nu \lambda^2}{2}$$

for every  $\lambda > 0$  **if and only if** for any probability measure  $P$  absolutely continuous with respect to  $Q$  and such that  $\text{KL}(P \parallel Q) < \infty$ ,

$$\mathbb{E}_P[X] - \mathbb{E}_Q[X] \leq \sqrt{2\nu \text{KL}(P \parallel Q)}. \quad (94)$$

- **Remark (Transportation Method)**

Let  $\mathbb{P} = \otimes_{i=1}^n \mathbb{P}_i$  be the product measure for  $Z := (Z_1, \dots, Z_n)$  on  $\mathcal{X}^n$  and  $f : \mathcal{X}^n \rightarrow \mathbb{R}$  be 1-Lipschitz function. Consider a probability measure  $\mathbb{Q}$  on  $\mathcal{X}^n$ , absolutely continuous with respect to  $\mathbb{P}$  and let  $Y$  be a random variable (defined on the same probability space as  $\mathcal{X}$ ) such that  $Y$  has distribution  $\mathbb{Q}$ .

The lemma above suggests that one may prove *sub-Gaussian concentration inequalities* for  $X = f(Z_1, \dots, Z_n)$  by proving a “*transportation*” inequality as above. The key to achieving this relies on *coupling*. In particular, the *Kantorovich-Rubenstein duality* for  $\mathcal{W}_{1,d}$  suggests that

$$\mathbb{E}_{\mathbb{Q}}[f(Y)] - \mathbb{E}_{\mathbb{P}}[f(Z)] \leq \min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \mathbb{E}_{\gamma}[d(Y, Z)] := \mathcal{W}_{1,d}(\mathbb{Q}, \mathbb{P})$$

Thus, it suffices to *upper bound* the 1-Wasserstein distance between  $\mathbb{Q}$  and  $\mathbb{P}$ .

- **Definition (*d-Transportation Cost Inequality*)** [Wainwright, 2019]

Let  $(\mathcal{X}, d)$  be a *metric space* with metric  $d$ , and  $(\mathcal{X}, \mathcal{B})$  be a *measurable space*, where  $\mathcal{B}$  is the *Borel  $\sigma$ -algebra* induced by metric  $d$ , **the probability measure**  $\mathbb{P}$  is said to satisfy a ***d-transportation cost inequality*** with parameter  $\nu > 0$  if

$$\mathcal{W}_{1,d}(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\nu \text{KL}(\mathbb{Q} \parallel \mathbb{P})} \quad (95)$$

for all probability measure  $\mathbb{Q} \ll \mathbb{P}$  on  $\mathcal{B}$ .

- **Theorem 9.5 (Isoperimetric Inequality via Transportation Cost)** [Wainwright, 2019]  
Consider a metric measure space  $(\mathcal{X}, \mathcal{B}, \mathbb{P})$  with metric  $d$ , and suppose that  $\mathbb{P}$  satisfies the  $d$ -transportation cost inequality in (95). Then its **concentration function** satisfies the bound

$$\alpha_{\mathbb{P},(\mathcal{X},d)}(t) \leq \exp\left(-\frac{(t-t_0)_+^2}{2\nu}\right), \text{ for } t \geq t_0 \quad (96)$$

where  $t_0 := \sqrt{2\nu \log 2}$ . Moreover, for any  $Z \sim \mathbb{P}$  and any  $L$ -Lipschitz function  $f : \mathcal{X} \rightarrow \mathbb{R}$ , we have the **concentration inequality**

$$\mathbb{P}\{|f(Z) - \mathbb{E}[f(Z)]| \geq t\} \leq 2 \exp\left(-\frac{t^2}{2\nu L^2}\right). \quad (97)$$

### 9.3 Tensorization for Transportation Cost

- **Proposition 9.6 (Tensorization for Transportation Cost)** [Boucheron et al., 2013]  
Suppose that, for each  $k = 1, 2, \dots, n$ , the univariate distribution  $\mathbb{P}_k$  satisfies a  $d_k$ -transportation cost inequality with parameter  $\nu_k$ . Then the **product distribution**  $\mathbb{P} = \otimes_{k=1}^n \mathbb{P}_k$  satisfies the transportation cost inequality

$$\mathcal{W}_{1,d}(\mathbb{Q}, \mathbb{P}) = \sqrt{2 \left( \sum_{k=1}^n \nu_k \right) \text{KL}(\mathbb{Q} \parallel \mathbb{P})}, \quad \text{for all distributions } \mathbb{Q} \ll \mathbb{P} \quad (98)$$

where the Wasserstein metric is defined using the distance  $d(x, y) := \sum_{k=1}^n d_k(x_k, y_k)$ .

### 9.4 Induction Lemma

### 9.5 Marton's Transportation Inequality

- **Theorem 9.7 (Marton's Transportation Inequality)** [Boucheron et al., 2013]  
Let  $\mathbb{P} = \otimes_{k=1}^n \mathbb{P}_k$  be a product probability measure on  $\mathcal{X}^n$ , and let  $\mathbb{Q}$  be a probability measure absolutely continuous with respect to  $\mathbb{P}$ . Define two random vectors  $X = (X_1, \dots, X_n), Y = (Y_1, \dots, Y_n)$  in  $\mathcal{X}^n$  with distribution  $\mathbb{P}$  and  $\mathbb{Q}$  respectively. Then

$$\begin{aligned} \mathcal{W}_{2,d_H}(\mathbb{Q}, \mathbb{P}) &:= \sqrt{\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^n \gamma^2 \{X_i \neq Y_i\}} \leq \sqrt{\frac{1}{2} \text{KL}(\mathbb{Q} \parallel \mathbb{P})} \\ &\Leftrightarrow \min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^n \gamma^2 \{X_i \neq Y_i\} \leq \frac{1}{2} \text{KL}(\mathbb{Q} \parallel \mathbb{P}) \end{aligned} \quad (99)$$

- **Theorem 9.8 (Marton's Conditional Transportation Inequality)** [Boucheron et al., 2013]  
Let  $\mathbb{P} = \otimes_{k=1}^n \mathbb{P}_k$  be a product probability measure on  $\mathcal{X}^n$ , and let  $\mathbb{Q}$  be a probability measure absolutely continuous with respect to  $\mathbb{P}$ . Define two random vectors  $X = (X_1, \dots, X_n), Y = (Y_1, \dots, Y_n)$  in  $\mathcal{X}^n$  with distribution  $\mathbb{P}$  and  $\mathbb{Q}$  respectively. Then

$$\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \mathbb{E}_\gamma \left[ \sum_{i=1}^n (\gamma^2 \{X_i \neq Y_i | X_i\} + \gamma^2 \{X_i \neq Y_i | Y_i\}) \right] \leq 2 \text{KL}(\mathbb{Q} \parallel \mathbb{P}) \quad (100)$$

- **Proposition 9.9 (Concentration of Lipschitz Function with Function Weighted Hamming Distance)** [Boucheron et al., 2013]

Let  $f : \mathcal{X}^n \rightarrow \mathbb{R}$  be a measurable function and let  $Z_1, \dots, Z_n$  be independent random variables taking their values in  $\mathcal{X}$ . Define  $X = f(Z_1, \dots, Z_n)$ . Assume that there exist **measurable functions**  $c_i : \mathcal{X}^n \rightarrow [0, \infty)$  such that for all  $x, y \in \mathcal{X}^n$ ,

$$f(y) - f(z) \leq \sum_{i=1}^n c_i(z) \mathbb{1}\{y_i \neq z_i\}.$$

Setting

$$\nu = \mathbb{E} \left[ \sum_{i=1}^n c_i^2(Z) \right] \quad \text{and} \quad \nu_\infty = \sup_{z \in \mathcal{X}^n} \sum_{i=1}^n c_i^2(z)$$

for all  $\lambda > 0$ , we have

$$\psi_{X - \mathbb{E}[X]}(\lambda) \leq \frac{\nu \lambda^2}{2} \quad \text{and} \quad \psi_{-X + \mathbb{E}[X]}(\lambda) \leq \frac{\nu_\infty \lambda^2}{2}$$

In particular, for all  $t > 0$ ,

$$\begin{aligned} \mathbb{P}\{X \geq \mathbb{E}[X] + t\} &\leq \exp\left(-\frac{t^2}{2\nu}\right) \\ \mathbb{P}\{X \leq \mathbb{E}[X] - t\} &\leq \exp\left(-\frac{t^2}{2\nu_\infty}\right). \end{aligned} \tag{101}$$

- **Remark** The condition in above proposition covers

1. *Lipschitz functions* such as *functions with bounded difference*,
2. **self-bounding functions** including **configuration functions**: Let  $f$  be such a configuration function. For any  $z \in \mathcal{X}^n$ , fix a *maximal sub-sequence*  $(z_{i,1}, \dots, z_{i,m})$  satisfying property  $\Pi$  (so that  $f(z) = m$ ). Let  $c_i(z)$  denote the indicator that  $z_i$  belongs to the sub-sequence  $(z_{i,1}, \dots, z_{i,m})$ . Thus,

$$\sum_{i=1}^n c_i^2(z) = \sum_{i=1}^n c_i(z) = f(z).$$

It follows from the definition of a configuration function that for all  $z, y \in \mathcal{X}^n$ ,

$$f(y) \geq f(z) - \sum_{i=1}^n c_i(z) \mathbb{1}\{z_i \neq y_i\}$$

So  $g = -f$  satisfies the condition in above proposition.

3. **weakly self-bounding functions**
4. **convex distance function**

$$d_T(z, A) := \sup_{\alpha \in \mathbb{R}_+^n : \|\alpha\|_2 = 1} \inf_{y \in A} \sum_{i=1}^n \alpha_i \mathbb{1}\{z_i \neq y_i\}$$



Denote by  $c(z) = (c_1(z), \dots, c_n(z)) = \alpha^*$  the vector of nonnegative components in the unit ball for which the supremum is achieved. Thus

$$\begin{aligned} d_T(z, A) - d_T(y, A) &\leq \inf_{z' \in A} \sum_{i=1}^n c_i(z) \mathbb{1}\{z_i \neq z'_i\} - \inf_{y' \in A} \sum_{i=1}^n c_i(z) \mathbb{1}\{y_i \neq y'_i\} \\ &\leq \sum_{i=1}^n c_i(z) \mathbb{1}\{z_i \neq y_i\} \end{aligned}$$

## 9.6 Talagrand's Gaussian Transportation Inequality

- **Theorem 9.10 (Talagrand's Gaussian Transportation Inequality)** [Boucheron et al., 2013]

Let  $\mathbb{P}$  be the standard Gaussian probability measure on  $\mathbb{R}^n$ , and let  $\mathbb{Q}$  be a probability measure absolutely continuous with respect to  $\mathbb{P}$ . Define two random vectors  $X = (X_1, \dots, X_n), Y = (Y_1, \dots, Y_n)$  in  $\mathcal{X}^n$  with distribution  $\mathbb{P}$  and  $\mathbb{Q}$  respectively. Then

$$\begin{aligned} \mathcal{W}_{2,d}(\mathbb{Q}, \mathbb{P}) &:= \sqrt{\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^n \mathbb{E}_\gamma [(X_i - Y_i)^2] \leq \sqrt{2\text{KL}(\mathbb{Q} \parallel \mathbb{P})}} \\ &\Leftrightarrow \min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^n \mathbb{E}_\gamma [(X_i - Y_i)^2] \leq 2\text{KL}(\mathbb{Q} \parallel \mathbb{P}) \end{aligned} \quad (102)$$

## 10 Proofs of Bounded Difference Inequality

- **Theorem 10.1 (McDiarmid's Inequality / Bounded Differences Inequality)** [Boucheron et al., 2013, Wainwright, 2019]

Suppose that  $f$  satisfies **the bounded difference property** (50) with parameters  $(L_1, \dots, L_n)$  i.e. for each index  $k = 1, 2, \dots, n$ ,

$$|f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, x'_i, x_{i+1}, \dots, x_n)| \leq L_k, \quad \text{for all } x, x' \in \mathcal{X}^n.$$

Assume that the random vector  $X = (X_1, X_2, \dots, X_n)$  has **independent** components. Then

$$\mathbb{P}\{|f(X) - \mathbb{E}[f(X)]| \geq t\} \leq 2 \exp\left(-\frac{2t^2}{\sum_{k=1}^n L_k^2}\right).$$

### 10.1 Martingale Method

- **Proof:** Consider the associated *martingale difference sequence*

$$D_k := \mathbb{E}[f(X)|X_1, \dots, X_k] - \mathbb{E}[f(X)|X_1, \dots, X_{k-1}].$$

We claim that  $D_k$  lies in an interval of length at most  $L_k$  almost surely. In order to prove this claim, define the random variables

$$\begin{aligned} A_k &:= \inf_x \{\mathbb{E}[f(X)|X_1, \dots, X_{k-1}, x]\} - \mathbb{E}[f(X)|X_1, \dots, X_{k-1}] \\ B_k &:= \sup_x \{\mathbb{E}[f(X)|X_1, \dots, X_{k-1}, x]\} - \mathbb{E}[f(X)|X_1, \dots, X_{k-1}]. \end{aligned}$$

On one hand, we have

$$D_k - A_k = \mathbb{E}[f(X)|X_1, \dots, X_k] - \inf_x \{\mathbb{E}[f(X)|X_1, \dots, X_{k-1}, x]\},$$

so that  $D_k \geq A_k$  *almost surely*. A similar argument shows that  $D_k \leq B_k$  *almost surely*. We now need to show that  $B_k - A_k \leq L_k$  *almost surely*. Observe that by the independence of  $\{X_k\}_{k=1}^n$ , we have

$$\mathbb{E}[f(X)|x_1, \dots, x_k] = \mathbb{E}_{(k+1)}[f(x_1, \dots, x_k, X_{k+1}, \dots, X_n)], \text{ for any } (x_1, \dots, x_k),$$

where  $\mathbb{E}_{(k+1)}[\cdot]$  denote the expectation over  $(X_{k+1}, \dots, X_n)$ . Consequently, we have

$$\begin{aligned} B_k - A_k &= \sup_x \mathbb{E}_{(k+1)}[f(X_1, \dots, X_{k-1}, x, X_{k+1}, \dots, X_n)] \\ &\quad - \inf_x \mathbb{E}_{(k+1)}[f(X_1, \dots, X_{k-1}, x, X_{k+1}, \dots, X_n)] \\ &\leq \sup_{x,y} \{\mathbb{E}_{(k+1)}[f(X_{1:k-1}, x, X_{k+1:n})] - \mathbb{E}_{(k+1)}[f(X_{1:k-1}, y, X_{k+1:n})]\} \\ &\leq L_k, \end{aligned}$$

using *the bounded differences assumption*. Thus, the variable  $D_k$  lies within an interval of length  $L_k$  at most surely, so that the claim follows as a corollary of *the Azuma-Hoeffding inequality*. ■

## 10.2 Entropy Method

- **Proof:** Recall that for a random variable  $Y$  taking its values in  $[a, b]$ , then we know from *Hoeffding's Lemma* that the logarithmic moment generating functions  $\psi(\lambda)$  satisfies

$$\psi(\lambda)'' = \text{Var}(Y) \leq \frac{(b-a)^2}{4}$$

for every  $\lambda \in \mathbb{R}$ . Hence, Hoeffding's inequality is obtained since

$$\frac{\text{Ent}(e^{\lambda Y})}{\mathbb{E}[e^{\lambda Y}]} = \lambda \psi'(\lambda) - \psi(\lambda) = \int_0^\lambda s \psi''(s) ds \leq \frac{(b-a)^2}{4} \int_0^\lambda s ds = \frac{(b-a)^2 \lambda^2}{8},$$

Note that by the bounded differences assumption, given  $X_{(-i)}$ ,  $f(X)$  is a random variable whose range is in an interval of length at most  $L_i$ , so

$$\frac{\text{Ent}_{(-i)}(e^{\lambda f(X)})}{\mathbb{E}_{(-i)}[e^{\lambda f(X)}]} \leq \frac{L_i^2 \lambda^2}{8}$$

From the tensorization property of entropy, we can bound the entropy of total function

$$\begin{aligned} \text{Ent}(e^{\lambda f(X)}) &\leq \mathbb{E} \left[ \sum_{i=1}^n \text{Ent}_{(-i)}(e^{\lambda f(X)}) \right] \leq \sum_{i=1}^n \frac{L_i^2 \lambda^2}{8} \mathbb{E} \left[ \mathbb{E}_{(-i)}[e^{\lambda f(X)}] \right] \\ \frac{\text{Ent}(e^{\lambda f(X)})}{\mathbb{E}[e^{\lambda f(X)}]} &\leq \frac{\sum_{i=1}^n L_i^2 \lambda^2}{8} \equiv \frac{\nu \lambda^2}{2}, \end{aligned}$$

where  $\nu := \frac{1}{4} \sum_{i=1}^n L_i^2$ . Using *Herbst's argument*, it leads to the bound of logarithmic moment generating function:

$$\psi_{f(X)}(\lambda) \leq \frac{\nu \lambda^2}{2}.$$

Finally, we apply *the Chernoff's inequality*

$$\mathbb{P} \{f(X) - \mathbb{E} [f(X)] \geq t\} \leq \inf_{\lambda > 0} \exp (\psi_{f(X)}(\lambda) - \lambda t) \leq \exp \left( -\frac{t^2}{2\nu} \right). \quad \blacksquare$$

### 10.3 Isoperimetric Inequality on Binary Hypercube

- **Definition (Vertex Boundary of Graph)** [Boucheron et al., 2013]

Consider a graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  and let  $\mathcal{A} \subset \mathcal{V}$  be a set of its vertices. The vertex boundary of  $\mathcal{A}$  is defined as *the set of those vertices, not in  $\mathcal{A}$ , which are connected to some vertex in  $\mathcal{V}$  by an edge*. We denote the vertex boundary of  $\mathcal{A}$  by  $\partial V(\mathcal{A})$ .

- **Remark (Binary Hypercube as Nearest Neighbor Graph with Hamming Distance)**  
Consider the graph as binary hypercube  $\{-1, +1\}^n$  in which two vertices are connected by an edge if and only if their **Hamming distance** equals 1. Define the *norm* as the *Hamming distance* to  $-1^n = (-1, \dots, -1)$

$$\|x\|_H := \sum_{i=1}^n \mathbb{1} \{x_i = 1\} = d_H(x, -1^n)$$

- **Definition (Simplicial Order)**

We define the so-called simplicial order of the elements of the binary hypercube. We say that  $x = (x_1, \dots, x_n) \in \{-1, 1\}^n$  **precedes**  $y = (y_1, \dots, y_n) \in \{-1, 1\}^n$  in the *simplicial order* if either  $\|x\|_H < \|y\|_H$  (where  $\|x\|_H := \sum_{i=1}^n \mathbb{1} \{x_i = 1\} = d_H(x, -1^n)$ ) or  $\|x\|_H = \|y\|_H$  and  $x_i = 1$  and  $y_i = -1$  for the smallest  $i$  for which  $x_i \neq y_i$ . That is

$$\begin{aligned} & x \prec y \\ \Leftrightarrow & \{(x, y) : \|x\|_H < \|y\|_H \vee (\|x\|_H = \|y\|_H \wedge (x_i = 1 \wedge y_i = -1, \text{ where } i = \min \{k : x_k \neq y_k\}))\} \end{aligned}$$

In other words, the vector with **less** 1's **precedes** the vector with more 1's. If the number of 1's are the same, then the first 1's on the leftmost position is preferred.

**Theorem 10.2 (Harp's Vertex Isoperimetric Theorem)** [Boucheron et al., 2013]  
For  $N = 1, \dots, 2^n$ , let  $S_N$  denote the set of **first  $N$  elements** of  $\{-1, +1\}^n$  in the **simplicial order**. For any subset  $A \subset \{-1, +1\}^n$ , where  $|A| = N$ ,

$$|\partial V(A)| \geq |\partial V(S_N)|$$

- **Remark** Note that if  $N = \sum_{i=0}^k \binom{n}{i}$ , for  $k = 0, \dots, n$ , then

$$S_N = \{x \in \{-1, +1\}^n : d_H(x, -1^n) \leq k\} = B_H(-1^n, k)$$

In other words,  $S_N$  is a **Hamming ball** centered at the vector  $-1^n = (-1, \dots, -1)$ .

- **Definition (*t*-Blowup of Set  $A$  in Binary Hypercube)**

For any  $A \subset \{-1, +1\}^n$  and  $x \in \{-1, +1\}^n$ , let  $d_H(x, A) = \min_{y \in A} d_H(x, y)$  be the *Hamming distance* of  $x$  to the set  $A$ . Also, denote by

$$A_t := \{x \in \{-1, +1\}^n : d_H(x, A) < t\}$$

the *t*-blowup of the set  $A$ , that is, the set of points whose Hamming distance from  $A$  is at most  $t$ .

- **Corollary 10.3 (*Isoperimetric Inequality in Binary Hypercube*)** [Boucheron et al., 2013]

Let  $A \subset \{-1, +1\}^n$  such that  $|A| \geq \sum_{i=0}^k \binom{n}{i}$ . Then for any  $t = 1, 2, \dots, n - k + 1$ ,

$$|A_t| \geq \sum_{i=0}^{k+1-t} \binom{n}{i}. \quad (103)$$

In particular, if  $|A|/2^n \geq 1/2$  then we may take  $k = \lfloor n/2 \rfloor$  in the corollary above and

$$\frac{|A_t|}{2^n} \geq \mathbb{P}\{X < \mathbb{E}[X] + t\} \geq 1 - \exp\left(-\frac{2t^2}{n}\right) \quad (104)$$

where  $X \sim \text{Ber}(1/2)$  is a symmetric Bernoulli random variable taking values in  $\{-1, +1\}$  with  $\mathbb{P}\{X = 1\} = \mathbb{P}\{X = -1\} = 1/2$ .

- **Proof: (*Proof of Bounded Difference Inequality on Binary Hypercube*)**

Note that any function with **bounded difference property** is **Lipschitz function** with respect to **Hamming distance**.

$$\begin{aligned} & \sup_{x \in \mathcal{X}^n, y_i \in \mathcal{X}} |f(x_1, \dots, x_n) - f(x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_n)| \\ & \leq c_i = c_i d_H((x_1, \dots, x_n), (x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_n)), \quad 1 \leq i \leq n \\ \Rightarrow |f(x) - f(y)| &= \left| \sum_{i=1}^n (f(x_{(i-1)}) - f(x_{(i)})) \right| \\ & \leq \sum_{i=1}^n |f(x_{(i-1)}) - f(x_{(i)})| \\ & \leq \sum_{i=1}^n L_i \mathbb{1}\{x_{(i-1)}[i] \neq x_{(i)}[i]\} \\ & \leq \left( \sum_i^n L_i^2 \right)^{1/2} d_H(x, y) \end{aligned}$$

where  $x_{(i)}$  is replicate of  $x_{(i-1)}$  except for  $i$ -th component, which is replaced by  $y_i$ . Note that  $x_{(0)} = x$  and  $x_{(n)} = y$ .

The Harp's isoperimetric theorem suggests that the concentration function

$$\alpha_{\mathbb{P}, (\{-1, +1\}^n, d_{H,L})}(t) := \sup_{A: \mathbb{P}\{A\} \geq 1/2} \mathbb{P}\{A_t\} \leq \exp\left(-\frac{2t^2}{\sum_i^n L_i^2}\right)$$

where  $\mathbb{P}$  is uniform distribution on  $\{-1, +1\}^n$ . Thus by *Levy's inequality*, we prove that for  $Z \in \{-1, 1\}^n$  and Lipschitz function  $f : \{-1, 1\}^n \rightarrow \mathbb{R}$  with respect to weighted Hamming distance  $d_{H,L}$ ,

$$\mathbb{P} \{|f(Z) - \text{Med}(f(Z))| \geq t\} \leq 2 \exp \left( -\frac{2t^2}{\sum_i^n L_i^2} \right). \quad \blacksquare$$

## 10.4 Transportation Method

- **Proof:** Any function with *bounded difference property* is *Lipschitz function* with respect to *Hamming distance*. This implies that for all  $x, y \in \mathcal{X}^n$ ,

$$f(y) - f(x) \leq \sum_{i=1}^n L_i \mathbb{1} \{x_i \neq y_i\} \equiv d_{H,L}(x, y).$$

Note that for coupling  $\gamma \in \Pi(\mathbb{Q}, \mathbb{P})$  where  $Y \sim \mathbb{Q}$  and  $X \sim \mathbb{P}$ ,

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}[f(Y)] - \mathbb{E}_{\mathbb{P}}[f(X)] &= \mathbb{E}_{\gamma}[f(Y) - f(X)] \\ &\leq \sum_{i=1}^n L_i \mathbb{E}_{\gamma}[\mathbb{1} \{X_i \neq Y_i\}] \\ &\leq \left( \sum_{i=1}^n L_i^2 \right)^{1/2} \left( \sum_{i=1}^n (\mathbb{E}_{\gamma}[\mathbb{1} \{X_i \neq Y_i\}])^2 \right)^{1/2} \end{aligned}$$

We want to prove the concentration using transportation cost inequality. That is, to bound the term

$$\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^n (\mathbb{E}_{\gamma}[\mathbb{1} \{X_i \neq Y_i\}])^2 = \min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^n \gamma^2 \{X_i \neq Y_i\}.$$

We have shown that

$$\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \gamma \{X \neq Y\} = \mathcal{W}_{1,d_H}(\mathbb{Q}, \mathbb{P}) = \sup_{A \in \mathcal{X}} |\mathbb{Q}(A) - \mathbb{P}(A)| \equiv \|\mathbb{Q} - \mathbb{P}\|_{TV}.$$

For each independent variable  $X_i, Y_i$ , and their marginal distribution  $\mathbb{P}_i, \mathbb{Q}_i$  where  $\mathbb{Q}_i \ll \mathbb{P}_i$ , by Pinsker's inequality,

$$\begin{aligned} \left[ \min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \gamma \{X_i \neq Y_i\} \right]^2 &= \|\mathbb{Q} - \mathbb{P}\|_{TV}^2 \leq \frac{1}{2} \text{KL}(\mathbb{Q}_i \| \mathbb{P}_i) \\ \min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \gamma^2 \{X_i \neq Y_i\} &\leq \left[ \min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \gamma \{X_i \neq Y_i\} \right]^2 \leq \frac{1}{2} \text{KL}(\mathbb{Q}_i \| \mathbb{P}_i) \end{aligned}$$

Thus by induction lemma,

$$\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^n \gamma^2 \{X_i \neq Y_i\} \leq \frac{1}{2} \text{KL}(\mathbb{Q} \| \mathbb{P})$$

which is the *Marton's transportation inequality*. Finally, we have

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[f(Y)] - \mathbb{E}_{\mathbb{P}}[f(X)] &\leq \left( \sum_{i=1}^n L_i^2 \right)^{1/2} \left( \sum_{i=1}^n (\mathbb{E}_{\gamma} [\mathbb{1}_{\{X_i \neq Y_i\}}])^2 \right)^{1/2} \\ &\leq \sqrt{\frac{(\sum_{i=1}^n L_i^2)}{2}} \text{KL}(\mathbb{Q} \parallel \mathbb{P}).\end{aligned}$$

Then we can apply the transportation lemma with  $\nu := \frac{1}{4} \sum_{i=1}^n L_i^2$ , which proves the bounded difference inequality.  $\blacksquare$

## References

- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.
- Thomas M. Cover and Joy A. Thomas. *Elements of information theory (2. ed.)*. Wiley, 2006. ISBN 978-0-471-24195-9.
- Gabriel Peyr and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019. ISSN 1935-8237.
- Sidney I Resnick. *A probability path*. Springer Science & Business Media, 2013.
- Filippo Santambrogio. *Optimal transport for applied mathematicians*, volume 55. Springer, 2015.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.