Summary: Part 1

Tianpei Xie

Jan. 26th., 2023

Contents

1 Basic Inequalities

1.1 Arithmetic, Calculus and Algebra

-

1.2 Function Space, Convexity and Duality

- **Proposition 1.1** ***(Jensen's inequality)*** *[Vershynin, 2018]*
  Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. Let $f : \Omega \to \mathbb{R}$ be a $\mathbb{P}$-measurable function and $\varphi : \mathbb{R} \to \mathbb{R}$ be ***convex function****. Then*

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

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

$$\left\|X\right\|_{L^p} \le \left\|X\right\|_{L^q} \quad \text{for any } 1 \le p \le q \le \infty \tag{2}$$

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

- **Proposition 1.2** ***(Minkowski's inequality)*** *[Vershynin, 2018]*
  For any $p \in [1, \infty]$, $X, Y \in L^p(\Omega, \mathbb{P})$,

$$\left\|X + Y\right\|_{L^p} \le \left\|X\right\|_{L^p} + \left\|Y\right\|_{L^p}, \tag{3}$$

  which implies that $\left\|\cdot\right\|_{L^p}$ 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:

$$\left|\langle X\,,\,Y\rangle_{L^2}\right| := \left|\mathbb{E}\left[XY\right]\right| \le \left\|X\right\|_{L^2}\,\left\|Y\right\|_{L^2}. \tag{4}$$

  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

$$\left|\langle X\,,\,Y\rangle_{L^2}\right| := \left|\mathbb{E}\left[XY\right]\right| \le \left\|X\right\|_{L^p}\,\left\|Y\right\|_{L^q}. \tag{5}$$

1.3 Probability Theory

- Assume a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ and a random variable $X : \Omega \to \mathbb{R}$ is a real-valued measurable function on Ω.

- For a random variable X, the ***expectation*** and ***variance*** are denoted as

$$\mathbb{E}\left[X\right] = \int X d\mathbb{P}$$
$$Var(X) = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right]$$

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

$$M_X(\lambda) := \mathbb{E}\left[e^{\lambda X}\right]$$
$$\psi_X(\lambda) := \log \mathbb{E}\left[e^{\lambda X}\right]$$

- 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^∞ ***norm*** of X

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

Similarly, L^∞ 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

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

The ***covariance*** is defined as

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

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 1.5 ***(Integral Identity)***. *[Vershynin, 2018]*
Let X be a ***non-negative*** *random variable. Then*

$$\mathbb{E}\left[X\right] = \int_0^\infty \mathbb{P}\left[X > t\right] dt. \tag{6}$$

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

- **Theorem 1.6** ***(Central Limit Theorem, Linderberg-Leváy)***
  Let $X_1, \ldots, X_n$ be ***independent identically distributed*** *random variables with mean $\mathbb{E}\left[X_i\right] = 0$ and variance $Var(X_i) = 1$. Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{d} N(0,1) \tag{7}$$

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

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

- **Theorem 1.7** ***(Central Limit Theorem, Nonasymptotic, Berry-Esseen)*** *[Vershynin, 2018]*
  Let $X_1, \ldots, X_n$ be ***independent identically distributed*** *random variables with mean $\mathbb{E}\left[X_i\right] = 0$, variance $Var(X_i) = \sigma^2$ and $\rho := \mathbb{E}\left[|X_i|^3\right] < \infty$. Then with some constant $C > 0$,*

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}\left\{ \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i \le t \right\} - \Phi(t) \right| \le \frac{C}{\sigma^3\sqrt{n}}\rho \tag{8}$$

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

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

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

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

- **Theorem 1.8** ***(Poisson Limit Theorem)***. *[Vershynin, 2018]*
  Let $X_{N,i}$, $1 \le i \le N$, be independent random variables $X_{N,i} \sim Ber(p_{N,i})$, and let $S_N = \sum_{i=1}^{N} X_{N,i}$. Assume that, as $N \to \infty$

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

  Then, as $N \to \infty$,

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

1.4 Information Theory

- **Definition** (***Shannon Entropy***) [Cover and Thomas, 2006]
  Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $X : \mathbb{R} \to \mathcal{X}$ be a random variable. Define $p(x)$ as *the probability density function* of X with respect to a base measure μ on $\mathcal{X}$. ***The Shannon Entropy*** is defined as

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

- **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 μ on $\mathcal{X} \times \mathcal{Y}$. Then ***the joint entropy*** of (X, Y), denoted as $H(X, Y)$, is defined as

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

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

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

- **Proposition 1.9** *(**Properties of Shannon Entropy**) [Cover and Thomas, 2006]*
  Let X, Y, Z be random variables.

  1. *(**Non-negativity**) $H(X) \geq 0$;*
  2. *(**Concavity**) $H(p) := \mathbb{E}_p\left[-\log p(X)\right]$ 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

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

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

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

which is *the entropy of P relative to Q*. Furthermore, if μ 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

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

- **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* $P_X \otimes P_Y$:

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

  If $P_{(X,Y)}$ has a probability density function $p(x,y)$ with respect to a base measure μ 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 1.10** ***(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.*

  $$\mathbb{KL}\left(P \,\|\, Q\right) \geq 0$$

  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)*

  $$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned} \tag{9}$$

  4. ***(Shannon Entropy as Self-Information)*** $I(X;X) = H(X)$

  5. ***(Joint Convexity of Relative Entropy)*** *The relative entropy $\mathbb{KL}\left(p \,\|\, q\right)$ 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 $\lambda \in [0, 1]$,*

  $$\mathbb{KL}\left(\lambda p_1 + (1-\lambda)p_2 \,\|\, \lambda q_1 + (1-\lambda)q_2\right) \leq \lambda \mathbb{KL}\left(p_1 \,\|\, q_1\right) + (1-\lambda)\mathbb{KL}\left(p_2 \,\|\, q_2\right) \tag{10}$$

- **Proposition 1.11** ***(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

$$H(X|Y) \leq H(X) \tag{11}$$

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

- **Proposition 1.12** ***(Chain Rule for Entropy)*** *[Cover and Thomas, 2006]*
  Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1) \tag{12}$$

- **Proposition 1.13** ***(Sub-Additivity of Entropy)*** *[Cover and Thomas, 2006]*
  Let $X_1, X_2, \ldots, X_n$ be drawn according to $p(x_1, x_2, \ldots, x_n)$. Then

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i) \tag{13}$$

  with equality if and only if the X_i are independent.

- **Proposition 1.14** ***(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*

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

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

$$\mathbb{KL} \left(P_{(X,Y)} \parallel Q_{(X,Y)} \right) = \mathbb{KL} \left(P_X \parallel Q_X \right) + \mathbb{E}_X \left[\mathbb{KL} \left(P_{Y|X} \parallel Q_{Y|X} \right) \right] \tag{14}$$

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

$$\mathbb{KL} \left(P \parallel Q \right) = \sum_{i=1}^{n} \mathbb{E}_{P_{1:i-1}} \left[\mathbb{KL} \left(P_{X_i|1\ldots i-1} \parallel Q_{X_i|1\ldots i-1} \right) \right] \tag{15}$$

- **Proposition 1.15** ***(Han's Inequality)*** *[Cover and Thomas, 2006, Boucheron et al., 2013]*
  Let $X_1, X_2, \ldots, X_n$ be random variables. Then

$$H(X_1, X_2, \ldots, X_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \tag{16}$$

$$\Leftrightarrow H(X) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(X_{(-i)})$$

2 Summary: General Proof Stratgy for Concentration Problem

3 Summary: Distribution-Free Concentration Inequality

4 The Cramér-Chernoff Method

4.1 From Markov Inequality to Cramér-Chernoff Method

- **Proposition 4.1** ***(Markov's Inequality)***. *[Vershynin, 2018]*
  *For any **non-negative** random variable X and $t > 0$, we have*

$$\mathbb{P}\left\{X \geq t\right\} \leq \frac{\mathbb{E}\left[X\right]}{t} \tag{17}$$

- **Proposition 4.2** ***(Chebyshev's Inequality)***. *[Vershynin, 2018]*
  Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$, we have

$$\mathbb{P}\left\{|X - \mu| \geq t\right\} \leq \frac{\sigma^2}{t^2}. \tag{18}$$

- **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

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

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

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

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

  In other word, in order to prove concentration around mean

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

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

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

- **Proposition 4.3** ***(Chernoff's inequality)*** *[Boucheron et al., 2013]*
  Let X be a real-valued random variable. For $\lambda \ge 0$, $\psi_X(\lambda)$ is the ***the logarithm of moment generating function*** *of X and $\psi_X^*(t)$ is its* ***Legendre (Cramér) transform****. Then*

$$
\mathbb{P}\left\{X \ge t\right\} \le \exp\left(-\psi_X^*(t)\right). \tag{19}
$$

- **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,

$$
\psi_X^*(t) := \sup_{\lambda \in \mathbb{R}} \left\{\lambda\, t - \psi_X(\lambda)\right\}, \tag{20}
$$

  where $\psi_X(\lambda) := \log \mathbb{E}\left[e^{\lambda X}\right]$. 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 $Z := X_1 + \ldots + X_n$ where $X_1, \ldots, X_n$ are ***independent*** *and* ***identically distributed*** *real-valued random variables*. Denote the logarithm of the moment-generating function of the X_i by $\psi_X(\lambda) = \log \mathbb{E}\left[e^{\lambda X_i}\right]$, and the corresponding *Legendre transform* by $\psi_X^*(t)$. Then, by independence, for all λ for which $\psi_X(\lambda) < \infty$,

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

  and consequently,

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

  Thus *the Chernoff's inequality* states that

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

- **Example** (***Normal Distribution***)
  Let X be a ***centered normal random variable*** with variance σ^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 $1/2$. $\blacksquare$

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

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

  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, $\mathbb{P}\{X = 1\} = 1 - \mathbb{P}\{X = 0\} = p$. Let $Z = X - p$ be the *corresponding centered variable*. If $0 < t < 1 - p$, we have

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

  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*** $\mathbb{KL}\left(\mathbb{P}_a \,\|\, \mathbb{P}_p\right)$ between a Bernoulli distribution $\mathbb{P}_a$ of parameter a and a Bernoulli distribution $\mathbb{P}_p$ of parameter p.

$$\mathbb{P}\{X \geq t\} \leq \exp\left(-\mathbb{KL}\left(\mathbb{P}_{p+t} \,\|\, \mathbb{P}_p\right)\right)$$

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

4.2 Sub-Gaussian Random Variables

- **Definition** (***Sub-Gaussian Random Variable***)
  A ***centered*** random variable X is said to be ***sub-Gaussian*** *with* ***variance factor*** ν if

$$\psi_X(\lambda) \le \frac{\lambda^2 \nu}{2}, \qquad \text{for every } \lambda \in \mathbb{R}. \tag{21}$$

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

- **Proposition 4.4** ***(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 $\nu > 0$

$$\mathbb{P}\{X > t\} \vee \mathbb{P}\{-X > t\} \le \exp\left(-\frac{t^2}{2\nu}\right), \qquad \text{for all } t > 0 \tag{22}$$

  then for every integer $q \ge 1$,

$$\mathbb{E}\left[X^{2q}\right] \le 2q!(2\nu)^q \le q!(4\nu)^q. \tag{23}$$

  Conversely, *if for some positive constant C*

$$\mathbb{E}\left[X^{2q}\right] \le q!C^q,$$

  then $X \in \mathcal{G}(4C)$ (and therefore (23) holds with $\nu = 4C$).

- **Proposition 4.5** ***(Sub-Gaussian Characterizations)****. [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*

$$\mathbb{P}\{|X| \ge t\} \le 2\exp\left(-t^2/K_1^2\right) \qquad \text{for all } t \ge 0.$$

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

$$\|X\|_{L^p} = \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \le K_2\sqrt{p} \qquad \text{for all } p \ge 1.$$

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

$$\mathbb{E}\left[\exp(\lambda^2 X^2)\right] \le \exp(K_3^2\,\lambda^2) \qquad \text{for all } \lambda \text{ such that } |\lambda| \le \frac{1}{K_3}$$

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

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

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

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

$$\mathbb{E}\left[\exp(\lambda X)\right] \le \exp(K_5^2\,\lambda^2) \qquad \text{for all } \lambda \in \mathbb{R}.$$

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

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

  In other words, we define

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\left[\exp(X^2/t^2)\right] \le 2\right\}. \tag{24}$$

- **Remark** (***Sub-Gaussian Characterizations via Sub-Gaussian Norm***)
  We can restate the properties of sub-gaussian random variables in terms of sub-gaussian norm:

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

- **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 $\|X\|_{\psi_2} \le C$, where C is an absolute constant. More generally, if $X \sim N(0,\sigma^2)$ then X is sub-gaussian with

$$\|X\|_{\psi_2} \le C\sigma \tag{25}$$

  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

$$\|X\|_{\psi_2} \le \frac{1}{\sqrt{\log 2}} \tag{26}$$

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

$$\|X\|_{\psi_2} \le C\,\|X\|_{\infty} \tag{27}$$

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

4.3 Sub-Exponential and Sub-Gamma Random Variables

4.4 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 4.6** ***(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 $\psi_X(\lambda) := \log \mathbb{E}\left[e^{\lambda X}\right]$. Then*

$$\psi''_X(\lambda) \le \frac{(b-a)^2}{4}$$

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

- **Proposition 4.7** ***(Hoeffding's inequality)*** *[Boucheron et al., 2013]*
  Let $X_1, \ldots, X_n$ be independent random variables such that X_i takes its values in $[a_i, b_i]$ ***almost surely*** *for all $i \le n$. Let*

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

  Then for every $t > 0$,

$$\mathbb{P}\{S \ge t\} \le \exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right). \tag{28}$$

- **Proposition 4.8** ***(General Hoeffding's inequality)*** *[Vershynin, 2018]*
  Let $X_1, \ldots, X_n$ be ***independent sub-gaussian*** *random variables. Let*

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

  Then for every $t > 0$,

$$\mathbb{P}\{S \ge t\} \le \exp\left(-\frac{c\,t^2}{\sum_{i=1}^{n} \|X_i\|_{\psi_2}}\right). \tag{29}$$

4.5 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 ν, c* holds if the *variance* $\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mu^2 \le \nu$, and

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

  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 4.9** ***(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 ν and c. If $S = \sum_{i=1}^n (X_i - \mathbb{E}[X_i])$, then for all $\lambda \in (0, 1/c)$ and $t > 0$*

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

and

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

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

$$\mathbb{P}\left\{S \ge \sqrt{2\nu t} + ct\right\} \le e^{-t}. \tag{30}$$

- **Proposition 4.10** ***(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 $t > 0$,*

$$\mathbb{P}\{S \ge t\} \le \exp\left(-\frac{t^2}{2(\nu + ct)}\right). \tag{31}$$

- **Corollary 4.11** ***(Bernstein's Inequality for Bounded Distributions)****. [Vershynin, 2018]*
  Let $X_1, \dots, X_n$ be ***independent, mean zero*** *random variables, such that $|X_i| \le b$ all i. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \ge t\right\} \le 2\exp\left(-\frac{t^2}{2(\nu + bt/3)}\right). \tag{32}$$

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

- **Corollary 4.12** ***(Bernstein's Inequality)****. [Vershynin, 2018]*
  Let $X_1, \dots, X_n$ be ***independent, mean zero, sub-exponential random variables****. Then, for every $t \ge 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n X_i\right| \ge t\right\} \le 2\exp\left[-c\min\left\{\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}}\right\}\right] \tag{33}$$

  where $c > 0$ is an absolute constant.

- **Proposition 4.13** ***(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 $t \ge 0$, we have*

$$\mathbb{P}\left\{\left|\sum_{i=1}^n a_i X_i\right| \ge t\right\} \le 2\exp\left[-c\min\left\{\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right\}\right] \tag{34}$$

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

- **Corollary 4.14** *(**Bernstein's Inequality, Average Form**). [Vershynin, 2018]*
*Let $X_1, \ldots, X_n$ be **independent, mean zero, sub-exponential random variables**. Then, for every $t \geq 0$, we have*
$$
\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\right\} \leq 2\exp\left[-c\min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}n\right] \tag{35}
$$
where $K = \max_i \|X_i\|_{\psi_1}$.

4.6 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,
$$
\psi_S(\lambda) = \sum_{i=1}^{n}\left(\log \mathbb{E}\left[e^{\lambda X_i}\right] - \lambda \mathbb{E}\left[X_i\right]\right).
$$
Using $\log u \leq u - 1$ for $u > 0$,
$$
\psi_S(\lambda) \leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda X_i} - \lambda X_i - 1\right]. \tag{36}
$$
Both Bennett's and Bernstein's inequalities may be derived from this bound, under different integrability conditions for the X_i.

- **Proposition 4.15** *(**Bennett's Inequality**) [Boucheron et al., 2013]*
*Let $X_1, \ldots, X_n$ 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*
$$
S = \sum_{i=1}^{n}\left(X_i - \mathbb{E}\left[X_i\right]\right)
$$
and $\nu = \sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right]$. 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 $t > 0$,
$$
\mathbb{P}\left\{S \geq t\right\} \leq \exp\left(-\frac{\nu}{b^2} h\left(\frac{b\,t}{\nu}\right)\right) \tag{37}
$$
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*** $(\nu/bt)^{t/2b}$.

4.7 The Johnson-Lindenstrauss Lemma

5 Martingale Method

5.1 Martingale and Martingale Difference Sequence

- **Definition** (***Martingale***) [Resnick, 2013]
  Let $\{X_n, n \geq 0\}$ be a stochastic process on $(\Omega, \mathscr{F})$ and $\{\mathscr{F}_n, n \geq 0\}$ be a ***filtration***; that is, $\{\mathscr{F}_n, n \geq 0\}$ is an *increasing sub σ-fields* of $\mathscr{F}$

  $$\mathscr{F}_0 \subseteq \mathscr{F}_1 \subseteq \mathscr{F}_2 \subseteq \ldots \subseteq \mathscr{F}.$$

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

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

  $$\mathbb{E}\left[X_n \mid \mathscr{F}_m\right] = X_m, \quad \text{a.s.} \tag{38}$$

  If the equality of (38) is replaced by $\geq$; that is, things are getting better on the average:

  $$\mathbb{E}\left[X_n \mid \mathscr{F}_m\right] \geq X_m, \quad \text{a.s.} \tag{39}$$

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

  $$\mathbb{E}\left[X_n \mid \mathscr{F}_m\right] \leq X_m, \quad \text{a.s.} \tag{40}$$

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

- **Remark** $\{X_n\}$ is ***martingale*** if it is *both* a ***sub*** and ***supermartingale***. $\{X_n\}$ is a ***supermartingale*** if and only if $\{-X_n\}$ is a ***submartingale***.

- **Remark** If $\{X_n\}$ is a ***martingale***, then $\mathbb{E}\left[X_n\right]$ is *constant*. In the case of a ***submartingale***, *the mean increases* and for a ***supermartingale***, *the mean decreases.*

- **Proposition 5.1** *[Resnick, 2013]*
  If $\{(X_n, \mathscr{F}_n), n \geq 0\}$ is a ***(sub, super) martingale****, then*

  $$\{(X_n, \sigma\left(X_0, X_1, \ldots, X_n\right)), n \geq 0\}$$

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

- **Definition** (***Martingale Differences***). [Resnick, 2013]
  $\{(d_j, \mathscr{B}_j), j \geq 0\}$ is a ***(sub, super) martingale difference sequence*** or a ***(sub, super) fair sequence*** if

  1. For $j \geq 0$, $\mathscr{B}_j \subset \mathscr{B}_{j+1}$.
  2. For $j \geq 0$, $d_j \in L_1$, $d_j \in \mathscr{B}_j$; that is, d_j is *absolutely integrable* and *$\mathscr{B}_j$-measurable.*

3. For $j \geq 0$,

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

- **Proposition 5.2** ***(Construction of Martingale From Martingale Difference)****[Resnick, 2013]*
  If $\{(d_j, \mathscr{B}_j), j \geq 0\}$ is ***(sub, super) martingale difference sequence****, and*

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

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

- **Proposition 5.3** ***(Construction of Martingale Difference From Martingale)*** *[Resnick, 2013]*
  Suppose $\{(X_n, \mathscr{B}_n), n \geq 0\}$ is a ***(sub, super) martingale****. Define*

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

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

- **Proposition 5.4** ***(Orthogonality of Martingale Differences)****. [Resnick, 2013]*
  If $\{(X_n, \mathscr{B}_n), n \geq 0\}$ is a ***martingale*** *where X_n can be decomposed as*

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

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

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

- **Example** (***Smoothing as Martingale***)
  Suppose $X \in L_1$ and $\{\mathscr{B}_n, n \geq 0\}$ is an increasing family of sub σ-algebra of $\mathscr{B}$. Define for $n \geq 0$

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

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

  $$d_n := \mathbb{E}\left[X | \mathscr{B}_n\right] - \mathbb{E}\left[X | \mathscr{B}_{n-1}\right], \quad n \geq 1. \tag{41}$$

- **Example** (***Sums of Independent Random Variables***)
  Suppose that $\{Z_n, n \geq 0\}$ is an ***independent*** *sequence of integrable random variables* satisfying for $n \geq 0$, $\mathbb{E}\left[Z_n\right] = 0$. Set

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

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

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

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

5.2 Bernstein Inequality for Martingale Difference Sequence

- **Proposition 5.5** ***(Bernstein Inequality, Martingale Difference Sequence Version)*** *[Wainwright, 2019]*
  Let $\{(D_k, \mathscr{B}_k), k \geq 1\}$ be a ***martingale difference sequence****, and suppose that*

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

  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*** $\left(\sqrt{\sum_{k=1}^{n} \nu_k^2}\;,\ \alpha_*\right)$ *where $\alpha_* := \max_{k=1\,,\ldots,\,n} \alpha_k$. 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***

$$\mathbb{P}\left\{\left|\sum_{k=1}^{n} D_k\right| \geq t\right\} \leq \begin{cases} 2\exp\left(-\frac{t^2}{2\sum_{k=1}^{n} \nu_k^2}\right) & \text{if } 0 \leq t \leq \frac{\sum_{k=1}^{n} \nu_k^2}{\alpha_*} \\ 2\exp\left(-\frac{t}{\alpha_*}\right) & \text{if } t > \frac{\sum_{k=1}^{n} \nu_k^2}{\alpha_*}. \end{cases} \tag{42}$$

5.3 Azuma-Hoeffding Inequality

- **Corollary 5.6** ***(Azuma-Hoeffding Inequality)****[Wainwright, 2019]*
  Let $\{(D_k, \mathscr{B}_k), k \geq 1\}$ 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 $k = 1\,, \ldots, n$. Then, for all $t \geq 0$,*

$$\mathbb{P}\left\{\left|\sum_{k=1}^{n} D_k\right| \geq t\right\} \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right) \tag{43}$$

5.4 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 $k \in \{1, 2, \ldots, n\}$, 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 $f : \mathcal{X}^n \to \mathbb{R}$ satisfies ***the bounded difference inequality*** with parameters $(L_1, \ldots, L_n)$ if, for each index $k = 1, 2, \ldots, n$,

  $$\left| f(x) - f(x^{(-k)}) \right| \le L_k, \quad \text{for all } x, x' \in \mathcal{X}^n. \tag{44}$$

- **Corollary 5.7** *(**McDiarmid's Inequality / Bounded Differences Inequality**)[Wainwright, 2019]*
  *Suppose that f satisfies **the bounded difference property** (44) with parameters $(L_1, \ldots, L_n)$ and that the random vector $X = (X_1, X_2, \ldots, X_n)$ has **independent** components. Then*

  $$\mathbb{P}\left\{ |f(X) - \mathbb{E}[f(X)]| \ge t \right\} \le 2 \exp\left(-\frac{2t^2}{\sum_{k=1}^{n} L_k^2} \right). \tag{45}$$

6 Bounding Variance

6.1 Mean-Median Deviation

6.2 The Efron-Stein Inequality and Jackknife Estimation

- **Remark** (***Variance of Smoothing Martingale Difference Sequence***)
  Suppose $X \in L_1$ and $\{\mathscr{B}_n, n \ge 0\}$ is an increasing family of sub σ-algebra of $\mathscr{B}$ formed by

  $$\mathscr{B}_n := \sigma(Z_1, \ldots, Z_n).$$

  For $n \ge 1$, define

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

  From (41) we see that $(d_n, \mathscr{B}_n)$ 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

$$\begin{aligned}\mathrm{Var}(X) = \mathbb{E}\left[\left(\sum_{i=1}^{n} d_i\right)^2\right] &= \sum_{i=1}^{n} \mathbb{E}\left[d_i^2\right] + 2\sum_{i>j} \mathbb{E}\left[d_i\, d_j\right] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[d_i^2\right]. \end{aligned} \tag{46}$$

- **Remark** (***Variance of General Functions of Independent Random Variables***)
Then above formula (46) holds when $X = f(Z_1, \ldots, Z_n)$ for general function $f : \mathbb{R}^n \to \mathbb{R}$ with n independent random variables $(Z_1, \ldots, Z_n)$. By *Fubini's theorem*,

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

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

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

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

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

$$\mathbb{E}\left[\mathbb{E}_{(-i)}[X]\,|Z_1, \ldots, Z_i\right] = \mathbb{E}\left[X|Z_1, \ldots, Z_{i-1}\right] \tag{47}$$

- **Proposition 6.1** *(**Efron-Stein Inequality**) [Boucheron et al., 2013]*
*Let $Z_1, \ldots, Z_n$ be **independent random variables** and let $X = f(Z)$ be a square-integrable function of $Z = (Z_1, \ldots, Z_n)$. Then*

$$Var(X) \leq \sum_{i=1}^{n} \mathbb{E}\left[\left(X - \mathbb{E}_{(-i)}[X]\right)^2\right] := \nu. \tag{48}$$

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

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

then

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

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

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

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

- **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 $Z_1, \ldots, Z_n$ are i.i.d. random variables and one wishes to *estimate a functional θ of the distribution* of the Z_i by a function $X = f(Z_1, \ldots, Z_n)$ 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

$$(n-1)\left(\frac{1}{n}\sum_{i=1}^{n} X_i - X\right) \tag{49}$$

  where X_i is an appropriately defined function of $Z_{(-i)} := (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$. $Z_{(-i)}$ 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

$$\sum_{i=1}^{n} (X - X_i)^2 \tag{50}$$

  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.

6.3 Functions with Bounded Differences

- **Corollary 6.2** *[Boucheron et al., 2013]*
  *If f has the **bounded differences property** with parameters $(L_1, \ldots, L_n)$, then*

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

6.4 Convex Poincaré Inequality

- **Theorem 6.3** ***(Convex Poincaré Inequality)*** *[Boucheron et al., 2013]*
  *Let $Z_1, \ldots, Z_n$ be **independent** random variables taking values in the interval $[0, 1]$ and let $f : [0, 1]^n \to \mathbb{R}$ be a **separately convex function** whose partial derivatives exist; that is, for every $i = 1, \ldots, n$ and fixed $z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n$, f is a convex function of its i-th variable. Then $f(Z) = f(Z_1, \ldots, Z_n)$ satisfies*

$$Var(f(Z)) \le \mathbb{E}\left[\|\nabla f(Z)\|_2^2\right]. \tag{51}$$

6.5 Gaussian Poincaré Inequality

- **Theorem 6.4** ***(Gaussian Poincaré Inequality)*** *[Boucheron et al., 2013]*
  Let $Z = (Z_1, \ldots, Z_n)$ 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 $f : \mathbb{R}^n \to \mathbb{R}$ be any* ***continuously differentiable*** *function. Then*
$$
Var(f(Z)) \le \mathbb{E}\left[\|\nabla f(Z)\|_2^2\right]. \tag{52}
$$

7 Entropy Method

7.1 Entropy Functional and Φ-Entropy

- **Definition** (***Φ-Entropy***)[Boucheron et al., 2013]
  Let $\Phi : [0, \infty) \to \mathbb{R}$ be a ***convex*** function, and assign, to every ***non-negative*** *integrable random variable* X, ***the Φ-entropy*** of X is defined as
$$
H_\Phi(X) = \mathbb{E}[\Phi(X)] - \Phi(\mathbb{E}[X]). \tag{53}
$$

- **Remark** The Φ-entropy is a ***functional*** of *distribution* P_X instead of a function of X.

- **Remark** By Jenson's inequality, the Φ-entropy is *non-negative*
$$
\begin{aligned}
&\Phi(\mathbb{E}[X]) \le \mathbb{E}[\Phi(X)] \\
&\Rightarrow H_\Phi(X) = \mathbb{E}[\Phi(X)] - \Phi(\mathbb{E}[X]) \ge 0.
\end{aligned}
$$

- **Example** (***Special Examples for Φ-Entropy***)
  1. For $\Phi(x) = x^2$, *the Φ-entropy* of X is *the* ***variance*** of X:
$$
H_\Phi(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \mathrm{Var}(X).
$$
  2. For $\Phi(x) = -\log(x)$, *the Φ-entropy* of $Y = e^{\lambda X}$ is *the* ***logarithm of moment generating function*** of $X - \mathbb{E}[X]$:
$$
H_\Phi(e^{\lambda X}) = -\lambda \mathbb{E}[X] + \log\left(\mathbb{E}\left[e^{\lambda X}\right]\right) = \log \mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right] := \psi_{X - \mathbb{E}[X]}(\lambda). \tag{54}
$$
  3. For $\Phi(x) = x \log x$, *the Φ-entropy* of X is defined as the ***entropy functional*** of X
$$
H_\Phi(X) = \mathrm{Ent}(X) := \mathbb{E}[X \log X] - \mathbb{E}[X] \log(\mathbb{E}[X]). \tag{55}
$$
     Let $(\Omega, \mathscr{B})$ be measurable space, and P and Q are probability measures on Ω 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.
$$
\mathrm{Ent}(X) = \mathbb{KL}(P \,\|\, Q) \tag{56}
$$

7.2 Dual Formulation

- **Lemma 7.1** *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 7.2** ***(Duality Formula of Entropy)*** *[Boucheron et al., 2013]*
Let X *be a non-negative random variable defined on a probability space* $(\Omega, \mathscr{A}, P)$ *such that* $\mathbb{E}[\Phi(X)] < \infty$*. Then we have **the duality formula***

$$\mathit{Ent}(X) = \sup_{U\in\mathcal{U}} \mathbb{E}[U\,X] \tag{57}$$

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

- **Corollary 7.3** ***(Alternative Duality Formula of Entropy)*** *[Boucheron et al., 2013]*

$$\mathit{Ent}(X) = \sup_{T} \mathbb{E}\left[X\left(\log(T) - \log\left(\mathbb{E}[T]\right)\right)\right] \tag{58}$$

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

- **Corollary 7.4** ***(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* $\lambda \in \mathbb{R}$*,*

$$\log \mathbb{E}_{\mathbb{P}}\left[e^{\lambda(X-\mathbb{E}[X])}\right] = \sup_{\mathbb{Q}\ll\mathbb{P}}\left\{\lambda\left(\mathbb{E}_{\mathbb{Q}}[X] - \mathbb{E}_{\mathbb{P}}[X]\right) - \mathbb{KL}\left(\mathbb{Q}\,\|\,\mathbb{P}\right)\right\}, \tag{59}$$

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

- **Corollary 7.5** ***(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*

$$\mathbb{KL}\left(\mathbb{Q}\,\|\,\mathbb{P}\right) = \sup_{X}\left\{\mathbb{E}_{\mathbb{Q}}[X] - \log\mathbb{E}_{\mathbb{P}}\left[e^{X}\right]\right\}, \tag{60}$$

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

- **Definition** (***Bregman Divergence***)
Let $F : \mathcal{X} \to \mathbb{R}$ 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:

$$\mathbb{D}^{F}\left(p\,\|\,q\right) = F(p) - F(q) - \langle \nabla F(q)\,,\,p - q\rangle \tag{61}$$

- **Theorem 7.6** ***(The Expected Value Minimizes Expected Bregman Divergence)*** *[Boucheron et al., 2013]*
  Let $I \subseteq \mathbb{R}$ be an open interval and let $f : I \to \mathbb{R}$ 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*

$$\mathbb{E}\left[f(X) - f(\mathbb{E}\left[X\right])\right] = \inf_{a \in I} \mathbb{E}\left[f(X) - f(a) - f'(a)(X - a)\right] \tag{62}$$

- **Corollary 7.7** ***(Duality Formula of Entropy via Bregman Divergence)*** *[Boucheron et al., 2013]*
  Let X be a non-negative random variable such that $\mathbb{E}\left[\Phi(X)\right] < \infty$. Then

$$\mathit{Ent}(X) = \inf_{u > 0} \mathbb{E}\left[X\left(\log(X) - \log(u)\right) - (X - u)\right] \tag{63}$$

7.3 Tensorization Property

- **Proposition 7.8** ***(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 $Z_1, Z_2, \ldots, Z_n$ be independent random variables taking values in $\mathcal{X}$, and let $f : \mathcal{X}^n \to [0, \infty)$ be a measurable function. Letting $X = f(Z_1, Z_2, \ldots, Z_n)$ such that $\mathbb{E}\left[X \log X\right] < \infty$, we have

$$\mathit{Ent}(X) := \mathbb{E}\left[\Phi(X)\right] - \Phi(\mathbb{E}\left[X\right]) \le \sum_{i=1}^{n} \mathbb{E}\left[\mathbb{E}_{(-i)}\left[\Phi(X)\right] - \Phi(\mathbb{E}_{(-i)}\left[X\right])\right], \tag{64}$$

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

$$\mathit{Ent}(X) \le \mathbb{E}\left[\sum_{i=1}^{n} \mathit{Ent}_{(-i)}(X)\right]. \tag{65}$$

7.4 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

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

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

$$\frac{\mathrm{Ent}(e^{\lambda Z})}{\mathbb{E}\left[e^{\lambda Z}\right]} = \lambda\, \psi'_{Z - \mathbb{E}[Z]}(\lambda) - \psi_{Z - \mathbb{E}[Z]}(\lambda). \tag{66}$$

  Our strategy is based on using (66) ***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

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

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

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

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

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

and the result follows.

- **Proposition 7.9** ***(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 $\lambda > 0$,

$$
\frac{Ent(e^{\lambda Z})}{\mathbb{E}\left[e^{\lambda Z}\right]} \le \frac{\nu\lambda^2}{2} \tag{67}
$$

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

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

7.5 Connection to Variance Bounds

- **Proposition 7.10** ***(A Modified Logarithmic Sobolev Inequalities for Moment Generating Function)*** *[Boucheron et al., 2013]*
  Consider independent random variables $Z_1, \ldots, Z_n$ taking values in $\mathcal{X}$, a real-valued function $f : \mathcal{X}^n \to \mathbb{R}$ and the random variable $X = f(Z_1, \ldots, Z_n)$. Also denote $Z_{(-i)} = (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n)$ and $X_{(-i)} = f_i(Z_{(-i)})$ where $f_i : \mathcal{X}^{n-1} \to \mathbb{R}$ is an arbitrary function. Let $\phi(x) = e^x - x - 1$. Then for all $\lambda \in \mathbb{R}$,

$$
Ent(e^{\lambda X}) := \mathbb{E}\left[\lambda X e^{\lambda X}\right] - \mathbb{E}\left[e^{\lambda X}\right] \log \mathbb{E}\left[e^{\lambda X}\right] \le \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda X}\phi(-\lambda(X - X_{(-i)}))\right] \tag{68}
$$

- **Proposition 7.11** ***(Symmetrized Modified Logarithmic Sobolev Inequalities)*** *[Boucheron et al., 2013]*
  Consider independent random variables $Z_1, \ldots, Z_n$ taking values in $\mathcal{X}$, a real-valued function $f : \mathcal{X}^n \to \mathbb{R}$ and the random variable $X = f(Z_1, \ldots, Z_n)$. Also denote $\widetilde{X}^{(i)} = f(Z_1, \ldots, Z_{i-1}, Z'_i, Z_{i+1}, \ldots, Z_n)$. Let $\phi(x) = e^x - x - 1$. Then for all $\lambda \in \mathbb{R}$,

$$
\lambda \mathbb{E}\left[X e^{\lambda X}\right] - \mathbb{E}\left[e^{\lambda X}\right] \log \mathbb{E}\left[e^{\lambda X}\right] \le \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda X}\phi(-\lambda(X - \widetilde{X}^{(i)}))\right] \tag{69}
$$

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

$$\lambda \mathbb{E}\left[X e^{\lambda X}\right] - \mathbb{E}\left[e^{\lambda X}\right] \log \mathbb{E}\left[e^{\lambda X}\right] \le \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda X} \tau(-\lambda(X - \widetilde{X}^{(i)})_+)\right],$$
$$\lambda \mathbb{E}\left[X e^{\lambda X}\right] - \mathbb{E}\left[e^{\lambda X}\right] \log \mathbb{E}\left[e^{\lambda X}\right] \le \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda X} \tau(\lambda(\widetilde{X}^{(i)} - X)_-)\right].$$

- **Remark** In the proof, we have

$$\begin{aligned}
\mathrm{Ent}_{\mu_i}(e^{\lambda X}) &\le \mathbb{E}_{\mu_i}\left[e^{\lambda X}(\log e^{\lambda X} - \log e^{\lambda X_i'}) - (e^{\lambda X} - e^{\lambda X_i'})\right] \\
&= \mathbb{E}_{\mu_i}\left[e^{\lambda X}(\lambda(X - X_i') - (e^{\lambda X} - e^{\lambda X_i'})\right] \\
&\le \mathbb{E}_{\mu_i}\left[(e^{\lambda X} - e^{\lambda X_i'})(\lambda(X - X_i')_+)\right] \\
&\le \lambda^2 \mathbb{E}_{\mu_i}\left[(X - X_i')_+^2\right]
\end{aligned}$$

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

$$\mathrm{Ent}(e^{\lambda X}) \le \lambda^2 \sum_{i=1}^{n} \mathbb{E}\left[\left(X - X_i'\right)_+^2\right].$$

From Efron-Stein inequality, if we can bound

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

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

8 Transportation Method

8.1 Optimal Transport, Wasserstein Distance and its Dual

- **Definition** (***Pushforward Measure***) [Peyr and Cuturi, 2019]
Let $(\mathcal{X}, \mathscr{B}_X)$ and $(\mathcal{Y}, \mathscr{B}_Y)$ 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 $T : \mathcal{X} \to \mathcal{Y}$, the ***push-forward operator*** is defined as $T_{\#} : \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$ that satisfies

$$\forall\, h \in \mathcal{C}(\mathcal{X}), \quad \int_{\mathcal{Y}} h(y)\, d\left(T_{\#}\alpha\right)(y) = \int_{\mathcal{X}} h(T(x))\, d\alpha(x). \tag{70}$$

$$\text{or equivalently,} \quad \left(T_{\#}\alpha\right)(B) := \alpha\left(\{x : T(x) \in B \subset \mathcal{Y}\}\right) = \alpha(T^{-1}(B)) \tag{71}$$

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

$$\rho_\alpha(\boldsymbol{x}) = \left|\det(T'(\boldsymbol{x}))\right| \rho_\beta(T(\boldsymbol{x})) \tag{72}$$

$$\left|\det(T'(\boldsymbol{x}))\right| = \frac{\rho_\alpha(\boldsymbol{x})}{\rho_\beta(T(\boldsymbol{x}))}$$

- **Definition** (***Optimal Transport Problem, Monge Problem***) [Villani, 2009, Santambrogio, 2015, Peyr and Cuturi, 2019]
  Let $(\mathcal{X}, \mathscr{B}_X)$ and $(\mathcal{Y}, \mathscr{B}_Y)$ 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 probabiilty measures on $\mathcal{X}$ and $\mathcal{Y}$. Define a ***cost function*** $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ 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* $T : \mathcal{X} \to \mathcal{Y}$ 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 Relxation***) 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 γ on $\mathcal{X}$. Similarly $\mathbb{Q}$ is the marginal distribution of γ 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 γ with marginal distribution of X and Y being $\mathbb{P}$ and $\mathbb{Q}$. Then the problem is

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

  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:

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

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

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

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

In other words, the dual optimization problem is

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

- **Proposition 8.1** *(**Strong Duality**) [Santambrogio, 2015]*
*Let $\mathcal{X}, \mathcal{Y}$ be **complete separable spaces**, and $c : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$ be **lower semi-continuous and bounded from below**. Then the optimal value of* primal *and* dual problems *are the same*

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

- **Definition** (***Wasserstein Distance***)
Let $((\mathcal{X}, d), \mathscr{B})$ be *a metric measurable space* with *Borel σ-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

$$\mathcal{W}_1(\mathbb{P}, \mathbb{Q}) \equiv \mathcal{W}_d(\mathbb{P}, \mathbb{Q}) := \min_{X\sim\mathbb{P},Y\sim\mathbb{Q}} \mathbb{E}\left[d(X, Y)\right] \tag{73}$$

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

$$\mathcal{W}_p(\mathbb{P}, \mathbb{Q}) \equiv \mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) := \left(\min_{X\sim\mathbb{P},Y\sim\mathbb{Q}} \mathbb{E}\left[(d(X, Y))^p\right]\right)^{1/p}. \tag{74}$$

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

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

- **Remark** (***Wasserstein p-Distance is a Metric in $\mathcal{P}(\mathcal{X})$***)
The ***Wasserstein p-distance*** $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) := \left(\min_{X\sim\mathbb{P},Y\sim\mathbb{Q}} \mathbb{E}\left[(d(X, Y))^p\right]\right)^{1/p}$ 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*): $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) \ge 0$.
    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*): $\mathcal{W}_{p,d}(\mathbb{P}, \mathbb{Q}) \le \mathcal{W}_{p,d}(\mathbb{P}, \mathbb{M}) + \mathcal{W}_{p,d}(\mathbb{M}, \mathbb{Q})$

- **Definition** (***Total Variation / Variational Distance***)
Let P, Q be two probability measures on measurable space $(\Omega, \mathscr{F})$. The ***total variation*** or ***variational distance*** between P and Q is defined by

$$V(P,Q) := \sup_{A \in \mathscr{F}} |P(A) - Q(A)| \tag{75}$$

- **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 μ 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

$$V(P,Q) := P(A^*) - Q(A^*) = \frac{1}{2}\int_{\Omega} |p(x) - q(x)|\, d\mu(x), \tag{76}$$

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

- **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*

$$V(P,Q) = \min P\{X \neq Y\} \tag{77}$$

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 8.2** *(**Pinsker's Inequality**) [Cover and Thomas, 2006, Boucheron et al., 2013]*
Let P, Q be two probability distributions on measurable space $(\Omega, \mathscr{F})$ such that $P \ll Q$. Then

$$V(P,Q)^2 \le \frac{1}{2}\mathbb{KL}\left(P \,\|\, Q\right). \tag{78}$$

- **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* (78) is the special case of *transportation cost inequality* (80).

- **Theorem 8.3** *(**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 σ-algebra induced by metric d, and $\mathbb{P}$ and $\mathbb{Q}$ be probability measures on $\mathcal{X}$. For fixed $p \in [1, \infty)$, let Lip_1 be the space of all 1-**Lipschitz** function with respect to metric d such that*

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

Then

$$\mathcal{W}_d(\mathbb{P},\mathbb{Q}) \equiv \mathcal{W}_{1,d}(\mathbb{P},\mathbb{Q}) = \sup_{f \in Lip_1}\left\{\mathbb{E}_{\mathbb{P}}\left[f(X)\right] - \mathbb{E}_{\mathbb{Q}}\left[f(Y)\right]\right\}. \tag{79}$$

8.2 Concentration via Transportation Cost

- **Lemma 8.4** *(**Transportation Lemma**) [Boucheron et al., 2013]*
*Let X be a real-valued integrable random variable. Let ϕ 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 **the generalized inverse** of ϕ^*. Then the following two statements are equivalent:*

  1. *for every $\lambda \in (0, b)$,*
  $$\psi_{X-\mathbb{E}[X]}(\lambda) \leq \phi(\lambda)$$
  where $\psi_X(\lambda) := \log \mathbb{E}_Q\left[e^{\lambda X}\right]$ is the logarithm of moment generating function;
  2. *for any probability measure P absolutely continuous with respect to Q such that $\mathbb{KL}\left(P \,\|\, Q\right) < \infty$,*
  $$\mathbb{E}_P\left[X\right] - \mathbb{E}_Q\left[X\right] \leq \phi^{*-1}\left(\mathbb{KL}\left(P \,\|\, Q\right)\right). \tag{80}$$

  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 $\mathbb{KL}\left(P \,\|\, Q\right) < \infty$,*
  $$\mathbb{E}_P\left[X\right] - \mathbb{E}_Q\left[X\right] \leq \sqrt{2\nu\mathbb{KL}\left(P \,\|\, Q\right)}. \tag{81}$$

- **Remark** (***Transportation Method***)
Let $\mathbb{P} = \otimes_{i=1}^n \mathbb{P}_i$ be the product measure for $Z := (Z_1, \ldots, Z_n)$ on $\mathcal{X}^n$ and $f : \mathcal{X}^n \to \mathbb{R}$ 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 $X = f(Z_1, \ldots, Z_n)$ 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
  $$\mathbb{E}_{\mathbb{Q}}\left[f(Y)\right] - \mathbb{E}_{\mathbb{P}}\left[f(Z)\right] \leq \min_{\gamma \in \Pi(\mathbb{Q},\mathbb{P})} \mathbb{E}_\gamma\left[d(Y, Z)\right] := \mathcal{W}_{1,d}(\mathbb{Q}, \mathbb{P})$$
  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 $(\mathcal{X}, \mathscr{B})$ be a *measurable space*, where $\mathscr{B}$ is *the Borel σ-algebra* induced by metric d, ***the probability measure*** $\mathbb{P}$ is said to satisfy a ***d-transportation cost inequality*** with parameter $\nu > 0$ if
  $$\mathcal{W}_{1,d}(\mathbb{Q}, \mathbb{P}) \leq \sqrt{2\nu\mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)} \tag{82}$$
  for all probability measure $\mathbb{Q} \ll \mathbb{P}$ on $\mathscr{B}$.

- **Theorem 8.5** ***(Isoperimetric Inequality via Transportation Cost)****[Wainwright, 2019]*
  Consider a metric measure space $(\mathcal{X}, \mathscr{B}, \mathbb{P})$ with metric d, and suppose that $\mathbb{P}$ satisfies the ***d-transportation cost inequality*** *in (82) Then its* ***concentration function*** *satisfies the bound*
$$
\alpha_{\mathbb{P},(\mathcal{X},d)}(t) \le \exp\left(-\frac{(t-t_0)_+^2}{2\nu}\right), \text{ for } t \ge t_0 \tag{83}
$$
  where $t_0 := \sqrt{2\nu \log 2}$. Moreover, for any $Z \sim \mathbb{P}$ and any L-Lipschitz function $f : \mathcal{X} \to \mathbb{R}$, we have the ***concentration inequality***
$$
\mathbb{P}\left\{|f(Z) - \mathbb{E}\left[f(Z)\right]| \ge t\right\} \le 2\exp\left(-\frac{t^2}{2\nu L^2}\right). \tag{84}
$$

8.3 Tensorization for Transportation Cost

- **Proposition 8.6** ***(Tensorization for Transportation Cost)*** *[Boucheron et al., 2013]*
  Suppose that, for each $k = 1, 2, \ldots, n$, the univariate distribution $\mathbb{P}_k$ satisfies a ***d_k-transportation cost inequality*** *with parameter ν_k. Then* ***the product distribution*** *$\mathbb{P} = \otimes_{k=1}^n \mathbb{P}_k$ satisfies the transportation cost inequality*
$$
\mathcal{W}_{1,d}(\mathbb{Q}, \mathbb{P}) = \sqrt{2\left(\sum_{k=1}^n \nu_k\right) \mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)}, \quad \text{for all distributions } \mathbb{Q} \ll \mathbb{P} \tag{85}
$$
  where the Wasserstein metric is defined using the distance $d(x, y) := \sum_{k=1}^n d_k(x_k, y_k)$.

8.4 Induction Lemma

8.5 Marton's Transportation Inequality

- **Theorem 8.7** ***(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 $X = (X_1, \ldots, X_n), Y = (Y_1, \ldots, Y_n)$ in $\mathcal{X}^n$ with distribution $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then
$$
\begin{aligned}
\mathcal{W}_{2,d_H}(\mathbb{Q}, \mathbb{P}) := \sqrt{\min_{\gamma \in \Pi(\mathbb{Q},\mathbb{P})} \sum_{i=1}^n \gamma^2\left\{X_i \neq Y_i\right\}} &\le \sqrt{\frac{1}{2}\mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)} \\
\Leftrightarrow \min_{\gamma \in \Pi(\mathbb{Q},\mathbb{P})} \sum_{i=1}^n \gamma^2\left\{X_i \neq Y_i\right\} &\le \frac{1}{2}\mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)
\end{aligned} \tag{86}
$$

- **Theorem 8.8** ***(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 $X = (X_1, \ldots, X_n), Y = (Y_1, \ldots, Y_n)$ in $\mathcal{X}^n$ with distribution $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then
$$
\min_{\gamma \in \Pi(\mathbb{Q},\mathbb{P})} \mathbb{E}_\gamma\left[\sum_{i=1}^n \left(\gamma^2\left\{X_i \neq Y_i | X_i\right\} + \gamma^2\left\{X_i \neq Y_i | Y_i\right\}\right)\right] \le 2\mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right) \tag{87}
$$

- **Proposition 8.9** ***(Concentration of Lipschitz Function with Function Weighted Hamming Distance)*** *[Boucheron et al., 2013]*
  Let $f : \mathcal{X}^n \to \mathbb{R}$ be a measurable function and let $Z_1, \ldots, Z_n$ be independent random variables taking their values in $\mathcal{X}$. Define $X = f(Z_1, \ldots, Z_n)$. Assume that there exist ***measurable functions*** $c_i : \mathcal{X}_n \to [0, \infty)$ *such that for all $x, y \in \mathcal{X}^n$,*

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

  Setting

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

  for all $\lambda > 0$, we have

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

  In particular, for all $t > 0$,

$$\begin{aligned} \mathbb{P}\{X \ge \mathbb{E}[X] + t\} &\le \exp\left(-\frac{t^2}{2\nu}\right) \\ \mathbb{P}\{X \le \mathbb{E}[X] - t\} &\le \exp\left(-\frac{t^2}{2\nu_\infty}\right). \end{aligned} \tag{88}$$

- **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* $(z_{i,1}, \ldots, z_{i,m})$ satisfying property Π (so that $f(z) = m$). Let $c_i(z)$ denote *the indicator that z_i belongs to the sub-sequence* $(z_{i,1}, \ldots, z_{i,m})$. 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 $z, y \in \mathcal{X}^n$,

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

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

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

8.6 Talagrand's Gaussian Transportation Inequality

- **Theorem 8.10** ***(Talagrand's Gaussian Transportation Inequality)*** *[Boucheron et al., 2013]*
  Let $\mathbb{P}$ be 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 $X = (X_1, \ldots, X_n), Y = (Y_1, \ldots, Y_n)$ in $\mathcal{X}^n$ with distribution $\mathbb{P}$ and $\mathbb{Q}$ respectively. Then

$$
\begin{aligned}
\mathcal{W}_{2,d}(\mathbb{Q}, \mathbb{P}) := \sqrt{\min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^{n} \mathbb{E}_{\gamma}\left[(X_i - Y_i)^2\right]} &\le \sqrt{2\mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)} \qquad (89)\\
\Leftrightarrow \min_{\gamma \in \Pi(\mathbb{Q}, \mathbb{P})} \sum_{i=1}^{n} \mathbb{E}_{\gamma}\left[(X_i - Y_i)^2\right] &\le 2\mathbb{KL}\left(\mathbb{Q} \,\|\, \mathbb{P}\right)
\end{aligned}
$$

9 Proofs of Bounded Difference Inequality

9.1 Martingale Method

9.2 Entropy Method

9.3 Isoperimetric Inequality on Binary Hypercube

9.4 Transportation Method

9.5 Comparison of Different Proofs

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.