Chris Junchi Li

Berry-Esseen Bounds and High-Dimensional Statistics

Springer

Contents

Chapter 1
Berry Esseen Theorem

This lecture concentrates on the topic of high dimensional probability: normal approximation. We start with a basic proof of Berry-Esseen Theorem. Roadmap is as follows.

(a) One-dimensional case
(b) Low-dimensional case (Bentkus 2003)
(c) High-dimensional case (CCK 2014)
(d) A Unified Framework
(e) Applications to Bootstrap (Concentration, entropy bound)

1.1 Introduction to Berry-Esseen CLT and Stein's Method

Setting: $X_1, \dots, X_n$ are independent random variables taking values in $\mathbb{R}$ with

$$\mathbb{E}X_i = 0, \quad \mathbb{E}X_i^2 = \sigma_i^2, \qquad \mathbb{E}|X_i|^3 = \rho_i < \infty.$$

Define

$$S_n := \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}.$$

Denote the CDF for $N(0,1)$ by $\Phi(x)$, let $\mathscr{H}$ be a class of test functions, and denote

$$d_{\mathscr{H}}(S_n, Z) = \sup_{h \in \mathscr{H}} \left| \mathbb{E}h(S_n) - \mathbb{E}h(Z) \right|.$$

There are three special examples of test function classes, as follows.

1. $h(x) = \mathbb{1}_{(-\infty,z)}(x) \Rightarrow$ Kolmogorov distance $d_K(\cdot,\cdot)$

$$\sup_x |F(x) - \Phi(x)| = \|F - \Phi\|_\infty.$$

2. $h(x) = \mathbb{1}_{\mathscr{A}}(x)$ where $\mathscr{A}$ is a Borel set $\Rightarrow$ Total variation distance $d_{TV}(\cdot,\cdot)$ **Much more restrictive condition!**
3. $h(x)$ is 1-Lipschitz i.e. $|h(x) - h(y)| \leq |x - y| \Rightarrow$ Wasserstein distance $d_W(\cdot,\cdot)$

Proposition 1.1. *$d_K(X,Y) \leq \sqrt{2Cd_W(X,Y)}$, where C is an uppder bound of the density of Y* ***Q: why C does not depend on the density of X?***

Theorem 1.1 (Berry-Esseen bound: general form).

$$d_K(S_n, Z) \leq C \cdot \left(\sum_{i=1}^n \sigma_i^2\right)^{-3/2} \sum_{i=1}^n \rho_i$$

We focus on the special case where X_i's are i.i.d. with $\mathbb{E}X_i^2 = 1$. Proofs will be presented using 3 methods: (1) Stein's (2) Lindeberg (3) Slepian's.

Main idea for Stein's method: for f, define $\mathscr{A}f$ given by

$$\mathscr{A}f(x) := f'(x) - xf(x).$$

Lemma 1.2 (Stein's lemma). *If $Z \sim N(0,1)$ then $\mathbb{E}\mathscr{A}f(Z) = 0$.*

Proof. Leave as an exercise for readers. Hint: using integration by parts!

For any $h \in \mathscr{H}$, if we find f s.t.

$$\mathscr{A}f(x) = h(x) - \mathbb{E}h(Z), \tag{1.1.1}$$

then $\mathbb{E}\mathscr{A}f(S_n^X) = \mathbb{E}h(S_n) - \mathbb{E}h(Z)$ **We only need to control the LHS!**

Lemma 1.3. *Solution of ODE in Eq.* (1.1.1) *can be represented as either*

$$\begin{aligned} f(x) &= e^{\frac{x^2}{2}} \int_x^{\infty} e^{-\frac{t^2}{2}} \left[\mathbb{E}h(Z) - h(t)\right] \mathrm{d}t \\ &= -e^{\frac{x^2}{2}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \left[\mathbb{E}h(Z) - h(t)\right] \mathrm{d}t, \end{aligned}$$

or Slepian's interpolation

$$f(x) = -\int_0^1 \frac{1}{\sqrt{2t(1-t)}} \mathbb{E}\left[Zh\left(\sqrt{t}x + \sqrt{1-t}Z\right)\right] \mathrm{d}t.$$

Proof. Standard integrating factors via $e^{-x^2/2}f(x)$ immediately gives us the solution. First equation by integrating factors. Second equation is derived by Stein's indentity. **It is straightforward to verify that the solution is unique among bounded solutions.**

Lemma 1.4. *Properties of f are as follows*

1. *If h is bounded i.e. $\|h\|_\infty < \infty$, then f, g and f' are also bounded, where $g(x) = xf(x)$. WLOG assume $h(0) = 0$.* ***we can substitute it to another constant***
2. *If h is L-Lipschitz, $|h(t)| \le L|t|$, then $\|f'\|_\infty < \infty$ and $\|f''\|_\infty < \infty$.*

Proof. **Property 1.** (a) h is bounded $\Rightarrow$ f is bounded. Let $g(t) = h(t) - \mathbb{E}h(Z)$, assume $\|g\|_\infty < C_0$, then from Lemma 1.3

$$\begin{aligned} |f(x)| &\le C_0 e^{\frac{x^2}{2}} \int_{|x|}^{\infty} e^{-\frac{t^2}{2}} \mathrm{d}t \le C_0 e^{\frac{x^2}{2}} \int_{0}^{\infty} e^{-\frac{(t+|x|)^2}{2}} \mathrm{d}t \\ &= C_0 \int_0^\infty e^{-\frac{t^2}{2}} e^{-t|x|} \mathrm{d}t \le C_0 \sqrt{\frac{\pi}{2}}, \end{aligned}$$

where in the second inequality we used a change of variable techniques, and in the last inequality we used the fact that $e^{-t|x|} \le 1$ for $t > 0$, finishing the proof.

(b) To show $g(x) = xf(x)$ is also bounded, we have

$$\begin{aligned} |f(x)| &\le C_0 e^{\frac{x^2}{2}} \int_{|x|}^{\infty} e^{-\frac{t^2}{2}} \mathrm{d}t \\ &\le C_0 e^{\frac{x^2}{2}} \int_{|x|}^{\infty} \frac{t}{|x|} e^{-\frac{t^2}{2}} \mathrm{d}t = C_0 e^{\frac{x^2}{2}} \frac{1}{|x|} e^{-\frac{x^2}{2}} = \frac{C_0}{|x|}. \end{aligned}$$

This leads to $|g(x)| \le |xf(x)| \le C_0$.

(c) To show that f' is also bounded, we have

$$f'(x) = (h(x) - \mathbb{E}h(Z)) + xf(x)$$

where both terms are bounded by C_0, so $|f'(x)| \le 2C_0$.

Property 2. (a) To show that h is L-Lipschitz $\Rightarrow$ f' is bounded, we note the Slepian's interpolation trick in Lemma 1.3 indicates that

$$\begin{aligned} f'(x) &= -\int_0^1 \frac{1}{\sqrt{2t(1-t)}} \mathbb{E}\left[Zh'\left(\sqrt{t}x + \sqrt{1-t}Z\right)\right] \sqrt{t}\mathrm{d}t \\ &= -\int_0^1 \frac{1}{\sqrt{2(1-t)}} \mathbb{E}\left[Zh'\left(\sqrt{t}x + \sqrt{1-t}Z\right)\right] \mathrm{d}t. \end{aligned}$$

Since $\|h'\|_\infty \le L$ we have

$$|f'(x)| \le \int_0^1 \frac{L}{\sqrt{2(1-t)}} \mathbb{E}|Z| \mathrm{d}t \le 2L.$$

(b) Finally we need to show when h is L-Lipschitz then $\|f''\|_\infty \le CL$. First

$$\begin{aligned} &f'(x) - xf(x) = h(x) - \mathbb{E}h(Z) \\ \Rightarrow\ &f''(x) - xf'(x) = h'(x) + f(x) = u(x). \end{aligned}$$

By Stein's Lemma, we have

$$\mathbb{E}\left[f''(Z) - Zf'(Z)\right] = 0 \Rightarrow \mathbb{E}u(Z) = \mathbb{E}\left[f''(Z) - Zf'(Z)\right] = 0,$$

Hence f' satisfies the Stein's ODE wrt $u(x)$, i.e.

$$g'(x) - xg(x) = u(x) - \mathbb{E}u(Z),$$

where $\mathbb{E}u(Z) = 0$. Furthermore, if h is Lipschitz $\Rightarrow$ f is bounded. To see this, we have

$$\begin{aligned} |f(x)| &\le e^{\frac{x^2}{2}} \int_{|x|}^{\infty} e^{-\frac{t^2}{2}} |\mathbb{E}h(Z) - h(t)| \, \mathrm{d}t \\ &\le e^{\frac{x^2}{2}} \int_{|x|}^{\infty} e^{-\frac{t^2}{2}} L\left(\mathbb{E}|Z| + t\right) \\ &\le L\left(\sqrt{\frac{\pi}{2}} + 1\right) \le 3L. \end{aligned}$$

This leads to

$$\|u\|_\infty = \|h'(x) + f(x)\|_\infty \le 4L,$$

since $|h'(x)| \le L$. Therefore

$$f' = f_u \Rightarrow f'' = f_u'.$$

Since u is bounded, f_u' is bounded

$$\Rightarrow \|f''\|_\infty \le 2\|u\|_\infty \le 8L.$$

1.2 Wasserstein distance version

We state the first version of our Berry-Esseen bound.

Theorem 1.5 (Berry-Esseen bound: Wasserstein distance version). *We assume X_i's are i.i.d. with $\mathbb{E}X_1^2 = 1$, then*

$$d_W(S_n, Z) \leq \frac{6\mathbb{E}|X_1|^3}{\sqrt{n}}.$$

Proof. **Goal: bound the absolute value of $\mathbb{E}\left[f'(S_n) - S_n f(S_n)\right]$ for h being 1-Lipschitz. Key Trick: Leave-One-Out then Taylor expansion.**
Define

$$S^{(i)} = S_n - \frac{X_i}{\sqrt{n}}.$$

Taylor expansion with the Lagrange remainder (whose special case is the mean-value theorem) implies there are (possibly random) $t_i, t_i' \in [0,1]$ such that

$$\begin{aligned} f'(S_n) &= f'(S^{(i)}) + \frac{1}{\sqrt{n}} X_i f''\left(S^{(i)} + \frac{t_i X_i}{\sqrt{n}}\right) \\ &= \frac{1}{n}\sum_{i=1}^n f'(S^{(i)}) + \frac{1}{n^{3/2}}\sum_{i=1}^n X_i f''\left(S^{(i)} + \frac{t_i X_i}{\sqrt{n}}\right). \end{aligned}$$

Taking expectation and we have

$$\mathbb{E}\left[f'(S_n)\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[f'(S^{(i)})\right] + \frac{1}{n^{3/2}}\sum_{i=1}^n \mathbb{E}\left[X_i f''\left(S^{(i)} + \frac{t_i X_i}{\sqrt{n}}\right)\right]. \tag{1.2.1}$$

Also, the second term

$$\begin{aligned} S_n f(S_n) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i f\left(S^{(i)} + \frac{X_i}{\sqrt{n}}\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i f(S^{(i)}) + \frac{1}{n}\sum_{i=1}^n X_i^2 f'\left(S^{(i)}\right) + \frac{1}{2n^{3/2}}\sum_{i=1}^n X_i^3 f''\left(S^{(i)} + \frac{t_i' X_i}{\sqrt{n}}\right). \end{aligned}$$

Note $X_i \perp S^{(i)}$ and X_i's are i.i.d. with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}\left[X_1^2\right] = 1$, and hence the expectation

$$\begin{aligned} \mathbb{E}\left[S_n f(S_n)\right] &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbb{E}\left[X_i\right]\mathbb{E}\left[f(S^{(i)})\right] \\ &\quad + \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[X_i^2\right]\mathbb{E}\left[f'\left(S^{(i)}\right)\right] + \frac{1}{2n^{3/2}}\sum_{i=1}^n \mathbb{E}\left[X_i^3 f''\left(S^{(i)} + \frac{t_i' X_i}{\sqrt{n}}\right)\right] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[f'\left(S^{(i)}\right)\right] + \frac{1}{2n^{3/2}}\sum_{i=1}^n \mathbb{E}\left[X_i^3 f''\left(S^{(i)} + \frac{t_i' X_i}{\sqrt{n}}\right)\right]. \end{aligned} \tag{1.2.2}$$

Combining Eqs. (1.2.1) and (1.2.2) we obtain

$$
\begin{aligned}
&\left|\mathbb{E}\left[f'(S_n) - S_n f(S_n)\right]\right| \\
\leq &\left|\frac{1}{n^{3/2}}\sum_{i=1}^{n}\mathbb{E}\left[X_i f''\left(S^{(i)} + \frac{t_i X_i}{\sqrt{n}}\right)\right]\right| + \left|\frac{1}{2n^{3/2}}\sum_{i=1}^{n}\mathbb{E}\left[X_i^3 f''\left(S^{(i)} + \frac{t_i' X_i}{\sqrt{n}}\right)\right]\right| \\
\leq &\|f''\|_\infty \frac{\mathbb{E}|X_i|}{\sqrt{n}} + \|f''\|_\infty \frac{\mathbb{E}|X_i|^3}{2\sqrt{n}},
\end{aligned}
$$

This immediately leads to

$$
\left|\mathbb{E}\left[f'(S_n) - S_n f(S_n)\right]\right| \leq \frac{6\mathbb{E}|X_1|^3}{\sqrt{n}},
$$

by combining the following two displays

1. Hölder's inequality indicates that $1 = \mathbb{E}|X_1|^2 \leq \left(\mathbb{E}|X_1|^3\right)^{2/3}$, and hence

$$
\mathbb{E}|X_1| \leq \left(\mathbb{E}|X_1|^3\right)^{1/3} \leq \mathbb{E}|X_1|^3.
$$

2. Lemma 1.4 implies that f the solution to Stein's ODE satisfies

$$
\|f\|_\infty \leq 2, \quad \|f'\|_\infty \leq \sqrt{\frac{2}{\pi}}, \quad \|f''\|_\infty \leq 4.
$$

Remark 1.1. The key of Stein's method is to leverage the ODE method and transform the bounded, nondifferentiable $h(x)$ to differentiable $f(x)$, or L-Lipschitz $h(x)$ to $f \in C^2$ that has a continuous second-order derivative.

1.3 Kolmogorov distance version

Now we prove the Berry-Esseen theorem for Kolmogorov distance. Recall that

$$d_K(S_n, Z) = \sup_{x\in\mathbb{R}} \left|\mathbb{E}\mathbf{1}_{(-\infty,x)}(S_n) - \mathbb{E}\mathbf{1}_{(-\infty,x)}(Z)\right|.$$

Theorem 1.6 (Berry-Esseen bound: Kolmogorov distance version). *We continue to assume X_i's are i.i.d. with $\mathbb{E}X_1^2 = 1$. Then*

$$d_K(S_n, Z) \le \frac{C}{\sqrt{n}}\mathbb{E}|X_1|^3.$$

To prepare for the proof of Theorem 1.6, we first do the following initial steps:
Step 1: Approximation of $\nVdash_{(-\infty,x)}$

$$h_{x,\varepsilon}(u) = \left\{\left[1 + \frac{x-u}{\varepsilon}\right] \wedge 1\right\} \vee 0.$$

$$h'_{x,\varepsilon}(u) = \frac{1}{\varepsilon}\nVdash_{[x,x+\varepsilon]}(u), \quad a.s.$$

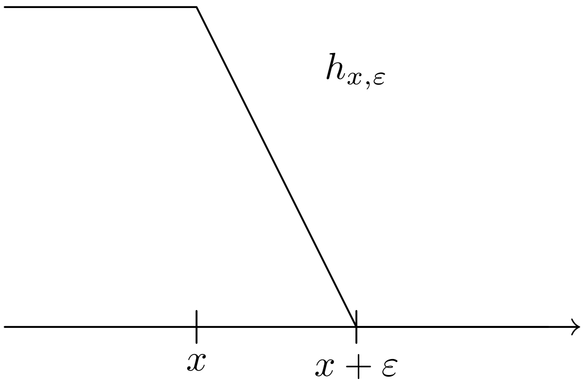

Fig. 1.1 Plot of the function $h_{x,\varepsilon}$

Step 2: Characterize approximation error

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i = \sqrt{\frac{n-1}{n}}S_{n-1} + \frac{1}{\sqrt{n}}X_n.$$

From the definition of the function, it is obvious that

$$h_{x-\varepsilon,\varepsilon} \le \nVdash_{(-\infty,x)} \le h_{x,\varepsilon} \le \nVdash_{(-\infty,x+\varepsilon)}.$$

Therefore

$$\begin{aligned}\mathbb{P}(S_n \le x) - \Phi(x) &\le \mathbb{E}h_{x,\varepsilon}(S_n) - \mathbb{E}h_{x,\varepsilon}(Z) + \Phi(x+\varepsilon) - \Phi(x) \\ &\le |\mathbb{E}h_{x,\varepsilon}(S_n) - \mathbb{E}h_{x,\varepsilon}(Z)| + \frac{\varepsilon}{\sqrt{2\pi}},\end{aligned}$$

where the last inequality we used the fact that the density of standard normal is bounded by $1/\sqrt{2\pi}$. For the other side, we have

$$\Phi(x) - \mathbb{P}(S_n \le x) \le \Phi(x) - \mathbb{E}h_{x-\varepsilon,\varepsilon}(S_n) + |\mathbb{E}h_{x-\varepsilon,\varepsilon}(S_n) - \mathbb{E}h_{x-\varepsilon,\varepsilon}(Z)|.$$

Therefore we have

$$d_K(S_n, Z) \le \sup_{x\in\mathbb{R}} |\mathbb{E}h_{x,\varepsilon}(S_n) - \mathbb{E}h_{x,\varepsilon}(Z)| + \frac{\varepsilon}{\sqrt{2\pi}},$$

where the last term on the right, $\frac{\varepsilon}{\sqrt{2\pi}}$ is called approximation error.

Step 3: Bound Gaussian approximation error of smooth functions

Denote $f_{x,\varepsilon}$ as the solution to the Stein's ODE for $h_{x,\varepsilon}$, i.e.

$$\mathbb{E}h_{x,\varepsilon}(S_n) - \mathbb{E}h_{x,\varepsilon}(Z) = \mathbb{E}f_{x,\varepsilon}(S_n) - \mathbb{E}S_n f_{x,\varepsilon}(S_n).$$

We establish several properties as follows

1. $|h(u) - \mathbb{E}h(Z)| \le 1, \quad \forall u \in \mathbb{R}$.
2. $\|f\|_\infty \le \sqrt{\frac{\pi}{2}}$, $\|f'\|_\infty \le 2$, $\|xf(x)\|_\infty \le 1$, however f'' cannot be bounded by a constant since $h' = O(\varepsilon^{-1})$.
3. Trick: $h' = -\frac{1}{\varepsilon}\nVdash_{(x,x+\varepsilon)}$, so

   $$|h(u+v) - h(u)| = \frac{1}{\varepsilon}\left|\int_u^{u+v} \nVdash_{[x,x+\varepsilon]}(t)\mathrm{d}t\right| = \frac{|v|}{\varepsilon}\int_0^1 \nVdash_{[x,x+\varepsilon]}(u+sv)\mathrm{d}s.$$

   This property characterizes the smoothing error, which later we are going to use.
4. We have

   $$\begin{aligned}&\left|f'_{x,\varepsilon}(u+v) - f'_{x,\varepsilon}(u)\right| \\ &= |(u+v)f_{x,\varepsilon}(u+v) - uf(u) + h(u+v) - h(u)| \\ &\le |u|\,|f_{x,\varepsilon}(u+v) - f(u)| + |v|\,|f(u+v)| + |h(u+v) - h(u)| \\ &\le 2|u||v| + 2|v| + \frac{|v|}{\varepsilon}\int_0^1 \nVdash_{[x,x+\varepsilon]}(u+sv)\mathrm{d}s.\end{aligned}$$

   The last property serves as an alternative for Taylor's expansion with only the first-order derivative.

Proof (Proof of Theorem 1.6). **Goal: bound the absolute value of** $\mathbb{E}\left[f'(S_n) - S_n f(S_n)\right]$ **for** h **being bounded. Key Trick: (1) Independent copies to decouple the expectation; (2) Leave-One-Out then Taylor expansion.**
Define

$$S^{(i)} = S_n - \frac{X_i}{\sqrt{n}}.$$

Then

$$f'(S_n) = f'\left(S^{(i)} + \frac{X_i}{\sqrt{n}}\right) = \frac{1}{n}\sum_{i=1}^{n} f'\left(S^{(i)} + \frac{X_i}{\sqrt{n}}\right).$$

Let X_i' be i.i.d. copies of X_i with $\mathbb{E}\left[X_1^2\right] = 1$, and hence the expectation

$$\mathbb{E}\left[f'(S_n)\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i^2 f'\left(S^{(i)} + \frac{X_i'}{\sqrt{n}}\right)\right]. \tag{1.3.1}$$

For the second term, using the mean-value theorem we conclude there are (possibly random) $t_i \in [0,1]$ such that

$$S_n f(S_n) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i f(S^{(i)}) + \frac{1}{n}\sum_{i=1}^{n} X_i^2 f'\left(S^{(i)} + \frac{t_i X_i}{\sqrt{n}}\right).$$

Since $X_i \perp S^{(i)}$ with $\mathbb{E}X_i = 0$ the expectation

$$\begin{aligned}
\mathbb{E}\left[S_n f(S_n)\right] &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbb{E}\left[X_i\right] \mathbb{E}\left[f(S^{(i)})\right] + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i^2 f'\left(S^{(i)} + \frac{t_i X_i}{\sqrt{n}}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i^2 f'\left(S^{(i)} + \frac{t_i X_i}{\sqrt{n}}\right)\right] \qquad (1.3.2)
\end{aligned}$$

Combining Eqs. (1.3.1) and (1.3.2) we obtain

$$\begin{aligned}
&\left|\mathbb{E}\left[f'(S_n) - S_n f(S_n)\right]\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i^2 f'\left(S^{(i)} + \frac{X_i'}{\sqrt{n}}\right)\right] - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i^2 f'\left(S^{(i)} + \frac{t_i X_i}{\sqrt{n}}\right)\right]\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i^2 \left|f'\left(S^{(i)} + \frac{X_i'}{\sqrt{n}}\right) - f'\left(S^{(i)} + \frac{t_i X_i}{\sqrt{n}}\right)\right|\right].
\end{aligned}$$

Note Property (4) above gives

$$\left|f'(u+v) - f'(u)\right| \leq 2|u||v| + 2|v| + \frac{|v|}{\varepsilon}\int_0^1 \nVdash_{[x,x+\varepsilon]}(u+sv)\mathrm{d}s.$$

Now, we compute the Gaussian approximation error

$$\begin{aligned}
&\left|\mathbb{E}\left[f'(S_n)-S_nf(S_n)\right]\right| \\
\le\ & \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\left|f'\left(S^{(i)}+\frac{X_i'}{\sqrt{n}}\right)-f'\left(S^{(i)}+\frac{t_iX_i}{\sqrt{n}}\right)\right|\right] \\
\le\ & \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\left|f'(S^{(i)})-f'\left(S^{(i)}+\frac{t_iX_i}{\sqrt{n}}\right)\right|\right] \\
&+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i^2\left|f'\left(S^{(i)}+\frac{X_i'}{\sqrt{n}}\right)-f'\left(S^{(i)}\right)\right|\right] \\
\le\ & \frac{1}{n^{3/2}}\sum_{i=1}^{n}\mathbb{E}X_i^2\left[|X_i|+|X_i'|\right]\left(2|S^{(i)}|+2\right) \\
&+\frac{1}{n^{3/2}}\sum_{i=1}^{n}\mathbb{E}\left[|X_i|^3\frac{1}{\varepsilon}\int_0^1 \nVdash_{[x,x+\varepsilon]}\left(S^{(i)}+\frac{st_iX_i}{\sqrt{n}}\right)\mathrm{d}s\right] \\
&+\frac{1}{n^{3/2}}\sum_{i=1}^{n}\mathbb{E}\left[|X_i|^2|X_i'|\frac{1}{\varepsilon}\int_0^1 \nVdash_{[x,x+\varepsilon]}\left(S^{(i)}+\frac{sX_i'}{\sqrt{n}}\right)\mathrm{d}s\right].
\end{aligned}$$

where the third inequality above is due to property (4). Now to analyze the first term we have

$$\text{The first term} \le \frac{\mathbb{E}|X_1|^3+\mathbb{E}|X_1|^2\mathbb{E}|X_1'|}{\sqrt{n}}\cdot\left(2\mathbb{E}|S^{(i)}|+2\right).$$

Further we use Jensen's inequality to obtain $\mathbb{E}|S^{(i)}| \le \sqrt{\mathbb{E}\left[S^{(i)}\right]^2}=\sqrt{\frac{n-1}{n}}\le 1$ and $\mathbb{E}|X_1|^2\mathbb{E}|X_1'|\le\mathbb{E}|X_1|^3$. We conclude that the first term is hence controlled by

$$\text{The first term} \le \frac{8\mathbb{E}|X_1|^3}{\sqrt{n}}.$$

To analyze the second and third terms, by letting $U=\frac{st_iX_i}{\sqrt{n}}$ we have

$$\begin{aligned}
&\mathbb{E}\left[\nVdash_{[x,x+\varepsilon]}\left(S^{(i)}+\frac{st_iX_i}{\sqrt{n}}\right)\,\middle|\,X_i\right] \\
=\ & \mathbb{P}\left[x-U\le S^{(i)}\le x+\varepsilon-U\,\middle|\,X_i\right] \\
=\ & \mathbb{P}\left[\sqrt{\frac{n}{n-1}}(x-U)\le\frac{\sum_{k=1}^{n}X_k-X_i}{\sqrt{n-1}}\le\sqrt{\frac{n}{n-1}}(x+\varepsilon-U)\,\middle|\,X_i\right] \\
=\ & 2d_K(S_{n-1},Z)+\mathbb{P}\left[\sqrt{\frac{n}{n-1}}(x-U)\le Z\le\sqrt{\frac{n}{n-1}}(x+\varepsilon-U)\right] \\
\le\ & 2d_K(S_{n-1},Z)+\sqrt{\frac{n}{n-1}}\frac{\varepsilon}{\sqrt{2\pi}},
\end{aligned}$$

where the second and third equalities above is by $S^{(i)} = S_n - \frac{X_i}{\sqrt{n}}$ and $S_{n-1} =^d \frac{\sum_{k=1}^n X_k - X_i}{\sqrt{n-1}}$. Finally we combine everything

$$\begin{aligned} d_K(S_n, Z) &\le \frac{8\mathbb{E}|X_1|^3}{\sqrt{n}} + \frac{2\mathbb{E}|X_1|^3}{\sqrt{n}\varepsilon}\left(2d_K(S_{n-1}, Z) + \sqrt{\frac{n}{n-1}}\frac{\varepsilon}{\sqrt{2\pi}}\right) \\ &\le \frac{\mathbb{E}|X_1|^3}{\sqrt{n}}\left[8 + 2\sqrt{\frac{n}{n-1}}\frac{1}{\sqrt{2\pi}} + \frac{4d_K(S_{n-1}, Z)}{\varepsilon}\right]. \end{aligned}$$

By choosing $\varepsilon = C_1 \frac{\mathbb{E}|X_1|^3}{\sqrt{n}}$ for some $C_1 > 8\sqrt{2}$, and then choose $C_2 > 2\left(8 + \frac{2}{\sqrt{\pi}}\right)$ sufficiently large such that for all $n' \le n-1$

$$d_K(S_{n'}, Z) \le C_2 \frac{\mathbb{E}|X_1|^3}{\sqrt{n'}},$$

then we have

$$\begin{aligned} d_K(S_n, Z) &\le \frac{\mathbb{E}|X_1|^3}{\sqrt{n}}\left[8 + \frac{2}{\sqrt{\pi}}\right] + \frac{4}{C_1} d_K(S_{n-1}, Z) \\ &\le \frac{\mathbb{E}|X_1|^3}{\sqrt{n}}\left[8 + \frac{2}{\sqrt{\pi}}\right] + \frac{4}{C_1} C_2 \sqrt{\frac{n}{n-1}} \frac{\mathbb{E}|X_1|^3}{\sqrt{n}} \\ &\le \frac{\mathbb{E}|X_1|^3}{\sqrt{n}}\left[8 + \frac{2}{\sqrt{\pi}} + \frac{4\sqrt{2}C_2}{C_1}\right] \le C_2 \frac{\mathbb{E}|X_1|^3}{\sqrt{n}}. \end{aligned}$$

This completes the whole proof.

1.4 Summary

Consider

$$X_1,\ldots,X_n \sim \mathscr{P}, \mathbb{E}X_1 = 0, \mathbb{E}X_1^2 = 1, \mathbb{E}|X_1|^3 \leq \rho$$

$$S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i, \quad Z \sim N(0,1)$$

WTS

$$d_{\mathscr{H}}(S_n, Z) = \sup_{h\in\mathscr{H}} |\mathbb{E}h(S_n) - \mathbb{E}h(Z)| \to 0$$

in the case (1) h is 1-Lip (2) h is bounded

$$\mathscr{A} f = f'(x) - xf(x)$$

$\forall h \; \mathscr{A} f = h(S_n) - \mathbb{E}h(Z)$ so the goal is now to bound for all $f \in \mathscr{H}$

$$\left|\mathbb{E}\mathscr{A} f\right| = \left|\mathbb{E}f'(S_n) - \mathbb{E}S_n f(S_n)\right| = \left|\mathbb{E}f'(S_n) - \frac{1}{n}\sum X_i f'(S_n)\right|$$

1. h is 1-Lip: $\|f\|_\infty \leq 2, \quad \|f'\|_\infty \leq \sqrt{\frac{2}{\pi}}, \quad \|f''\|_\infty \leq 4$. Trick: (1) $\mathbb{E}X_i^2 = 1$, (2) Leave-One-Out.
2. h is bounded: $\|f\|_\infty \leq \sqrt{\frac{\pi}{2}}, \quad \|f'\|_\infty \leq 2, \quad |xf(x)|_\infty < 1$. New Trick (3): Stein's formula to obtain

$$f'(u+v) - f'(u) = (u+v)f(u+v) - uf(u) + h(u+v) - h(u).$$

Remark 1.2. Note one can construct a special example that matches the Berry-Esseen convergence rate $n^{-1/2}$, so the . Consider the case where $\mathbb{P}(X_1 = 1) = \mathbb{P}(X_1 = -1) = 1/2$. Simple random walk theory implies

$$\mathbb{P}(S_{2n} \leq 0) - \Phi(0) = \frac{1}{2} + \frac{1}{2}\mathbb{P}(S_{2n} = 0) - \frac{1}{2} = \frac{1}{2(\pi n)^{1/2}} + o(n^{-1/2}).$$

Chapter 2
Berry Esseen Theorem II

Topic: (1) Lindeberg Replacement Trick / Sleptian Trick (2) Multivariate case d-dim

2.1 Lindeberg's version

Idea: gradually replace the similar-to-Gaussian random variable that you wanna replace by a Gaussian rv.

Lemma 2.1 (Lindeberg's Lemma). *Let $X,Y \in \mathbb{R}$, independent. Suppose random variable W is independent of X,Y, $\mathbb{E}|Y|^3 < \infty$ $\mathbb{E}|W|^3 < \infty$. $\mathbb{E}W = \mathbb{E}Y$, $\mathbb{E}W^2 = \mathbb{E}Y^2$. Then for any $f \in C^3(\mathbb{R})$, $\|f'''\|_\infty \leq Const.$*

$$\mathbb{E}\left|f(X+Y) - \mathbb{E}f(X+W)\right| \leq \frac{1}{6}\|f'''\|_\infty \left(\mathbb{E}|Y|^3 + \mathbb{E}|Z|^3\right).$$

Proof. By Taylor's expansion,

$$f(X+Y) - f(X) = f'(X)\cdot Y + \frac{1}{2}f''(X)\cdot Y^2 + \frac{1}{6}f'''(X+t_1Y)\cdot Y^3, \quad t_1 \in [0,1]$$

and

$$f(X+W) - f(X) = f'(X)\cdot W + \frac{1}{2}f''(X)\cdot W^2 + \frac{1}{6}f'''(X+t_2W)\cdot W^3, \quad t_2 \in [0,1]$$

Subtract the last two displays and taking expectations

$$\begin{aligned}|\mathbb{E}f(X+Y) - \mathbb{E}f(X+W)| = \Big|&\mathbb{E}f'(X)\cdot Y - \mathbb{E}f'(X)\cdot W \\ &+ \mathbb{E}\frac{1}{2}f''(X)\cdot Y^2 - \mathbb{E}\frac{1}{2}f''(X)\cdot W^2 \\ &+ \mathbb{E}\frac{1}{6}f'''(X+t_1Y)\cdot Y^3 - \mathbb{E}\frac{1}{6}f'''(X+t_2W)\cdot W^3\Big|\end{aligned}$$

A weaker version by Lindeberg's Replacement Trick:

$$S_n^X = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i, \qquad S_n^Y = \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i, \quad Y_i \sim N(0,1).$$

$$\sup_{A=(-\infty,t]} \left|\mathbb{P}\left(S_n^X \in \mathscr{A}\right) - \mathbb{P}\left(S_n^Y \in \mathscr{A}\right)\right|$$

1. Find a smooth approximation of $1_{(-\infty,t]}(x)$ denoted by φ_ε (New smooth techniques from Stein's linear interpolation!) so $\|\varphi_\varepsilon^{(k)}\|_\infty \sim \varepsilon^{-k}, k=1,2,3,\ldots$
2. Establish the smoothing error
3. Establish the approximation error for smooth functions (Gaussian approximation)

Step 3: We compute

$$\left|\mathbb{E}\varphi_\varepsilon(S_n^X) - \mathbb{E}\varphi_\varepsilon(S_n^Y)\right|.$$

Denoting

$$S_i = \frac{1}{\sqrt{n}}\left(X_1 + \cdots + X_{i-1} + Y_{i+1} + \cdots + Y_n\right),$$

the difference above can be written in the form of telescoping sum **key trick!**

$$\varphi_\varepsilon(S_n^X) - \varphi_\varepsilon(S_n^Y) = \sum_{i=1}^n \varphi_\varepsilon\left(S_i + \frac{X_i}{\sqrt{n}}\right) - \varphi_\varepsilon\left(S_i + \frac{Y_i}{\sqrt{n}}\right).$$

Since

$$S_{n-1} + \frac{X_n}{\sqrt{n}} = S_n^X$$
$$S_1 + \frac{Y_1}{\sqrt{n}} = S_n^Y$$

and

$$S_{k-1} + \frac{X_{k-1}}{\sqrt{n}} = S_k + \frac{Y_k}{\sqrt{n}},$$

we use Lindeberg's Lemma 2.1 we take expectation on each term

$$\left|\mathbb{E}\varphi_\varepsilon\left(S_i + \frac{X_i}{\sqrt{n}}\right) - \varphi_\varepsilon\left(S_i + \frac{Y_i}{\sqrt{n}}\right)\right| \le \frac{1}{6}\|\varphi_\varepsilon'''\|_\infty \frac{\mathbb{E}|X_i|^3 + \mathbb{E}|Y_i|^3}{n^{3/2}}$$

$$\sim C\cdot\varepsilon^{-3}\cdot\frac{\mathbb{E}|X_1|^3}{n^{3/2}}$$

Summing up gives

$$\varphi_\varepsilon(S_n^X) - \varphi_\varepsilon(S_n^Y) \le \sum_{i=1}^n \frac{1}{6}\|\varphi_\varepsilon'''\|_\infty \frac{\mathbb{E}|X_i|^3 + \mathbb{E}|Y_i|^3}{n^{3/2}} \sim C\cdot\varepsilon^{-3}\cdot\frac{\mathbb{E}|X_1|^3}{\sqrt{n}}$$

Step 2: smoothing error $\le \varepsilon/\sqrt{2\pi}$

$$d_K(S_n^X, S_n^Y) \lesssim \frac{\varepsilon}{\sqrt{2\pi}} + \frac{1}{\varepsilon^3} \cdot \frac{\mathbb{E}|X_1|^3}{\sqrt{n}}$$

Let $\varepsilon \asymp \left(\frac{\mathbb{E}|X_1|^3}{\sqrt{n}}\right)^{1/4}$ We have

$$d_K(S_n^X, S_n^Y) \lesssim \left(\frac{\mathbb{E}|X_1|^3}{\sqrt{n}}\right)^{1/4}.$$

Step 1: Approximate $1_{(-\infty,t]}$

$$\varphi_{t,\varepsilon}(x) = f\left(\frac{|x-t|}{\varepsilon}\right)$$

f is bounded up to third order derivative

$$f = \begin{cases} 1 \; x \in [0,1/2] \\ 0 \quad x > 1 \end{cases}$$

Function $\sigma_\varepsilon(x) = \mathbb{E}|x + \varepsilon Z|$ serves as an approximation to function $|x|$. **Bentkus uses this function to generalize the result to higher dimension**

Step 1': Let $f \in C^3(\mathbb{R})$ with $\max\{\|f'\|_\infty, \|f''\|_\infty, \|f'''\|_\infty\} \le C$ $f : \mathbb{R} \to [0,1]$ where

$$f(x) = \begin{cases} 1 & \text{if } x \le 0 \\ 0 & \text{if } x \ge 1 \\ \in [0,1] & \text{if } x \in [0,1] \end{cases}$$

Picture Define

$$\varphi_{t,\varepsilon}(x) = f\left(\frac{x-t}{\varepsilon}\right)$$

Check

$$\varphi_{t,\varepsilon}(x) = \begin{cases} 1 & \text{if } x \le t \\ 0 & \text{if } x \ge t + \varepsilon \end{cases}$$

and

$$\|\varphi_{\varepsilon,t}^{(k)}\|_\infty = \varepsilon^{-k} \|f^{(k)}\|_\infty$$

Approximate 1_A, where A is convex and closed.

$$\rho_2(x, A) = \inf_{y \in A} \|x - y\|_2$$

$$\varphi_{\varepsilon,A}(x) = f\left(\frac{\rho_2(x,A)}{\varepsilon}\right)$$

$\rho_2(x, (-\infty, t]) = \max\{0, x - t\}$, not differentiable at t

2.2 Slepian's Interpolation

Consider smooth functions. **Dive to Step 3.** φ_ε is the approximation function.

$$\varphi_\varepsilon(S_n^X) - \varphi_\varepsilon(S_n^Y)$$

Beforehand in Lindeberg, $S_i = (X_1 + \cdots + X_{i-1} + Y_{i+1} + \cdots + Y_n)/\sqrt{n}$ is asymmetric and bound

$$\varphi\left(S_i + \frac{X_i}{\sqrt{n}}\right) - \varphi\left(S_i + \frac{Y_i}{\sqrt{n}}\right).$$

Here we want something more symmetric. Define

$$Z_i(t) = \sqrt{t}X_i + \sqrt{1-t}Y_i, \quad t \in [0,1].$$

Then

$$Z_i'(t) = \frac{1}{2}\left(\frac{X_i}{\sqrt{t}} - \frac{Y_i}{\sqrt{1-t}}\right),$$

and hence

$$\begin{aligned} S_n^X &= \sum_{i=1}^n Z_i(1)/\sqrt{n}, \\ S_n^Y &= \sum_{i=1}^n Z_i(0)/\sqrt{n}. \end{aligned}$$

Denote by

$$S(t) = \sum_{i=1}^n \frac{Z_i(t)}{\sqrt{n}},$$

then by Fundamental Theorem of Calculus (Newton-Leibniz Law)

$$\begin{aligned} \varphi_\varepsilon(S_n^Y) - \varphi_\varepsilon(S_n^X) &= \varphi_\varepsilon(S(1)) - \varphi_\varepsilon(S(0)) \\ &= \int_0^1 \varphi_\varepsilon'(S(t)) \cdot \sum_{i=1}^n \frac{Z_i'(t)}{\sqrt{n}} \mathrm{d}t \\ &= \sum_{i=1}^n \int_0^1 \varphi_\varepsilon'(S(t)) \cdot \frac{Z_i'(t)}{\sqrt{n}} \mathrm{d}t. \end{aligned}$$

Define $S_i(t) = S(t) - \frac{Z_i(t)}{\sqrt{n}}$. Expand $\varphi_\varepsilon'(S(t))$ at $S_i(t)$ we have by Taylor expansion with integral forms of remainder

$$\begin{aligned} f(x+a) - f(x) - f'(x) \cdot a &= \int_x^{x+a} (x+a-t) f''(t) \mathrm{d}t \\ &= a^2 \int_0^1 (1-u) f''(x+au) \mathrm{d}u. \end{aligned}$$

so that

$$\varphi_\varepsilon'(S(t)) - \varphi_\varepsilon'(S_i(t)) - \frac{1}{\sqrt{n}} Z_i(t) \cdot \varphi_\varepsilon''(S_i(t))$$
$$= \frac{1}{n} Z_i(t)^2 \cdot \int_0^1 (1-u)\varphi_\varepsilon''' \left(S_i(t) + u \cdot \frac{Z_i(t)}{\sqrt{n}} \right) \mathrm{d}u$$

$$A_i = I_1 + I_2 + I_3$$

where

$$I_1 = \frac{1}{\sqrt{n}} \int_0^1 \varphi_\varepsilon'(S_i(t)) \cdot Z_i'(t) \mathrm{d}t$$
$$I_2 = \frac{1}{n} \int_0^1 \varphi_\varepsilon''(S_i(t)) \cdot Z_i(t) \cdot Z_i'(t) \mathrm{d}t$$
$$I_3 = \frac{1}{n^{3/2}} \int_0^1 \mathrm{d}t \int_0^1 (1-u)\varphi_\varepsilon'' \left(S_i(t) + u \cdot \frac{Z_i(t)}{\sqrt{n}} \right) \cdot Z_i'(t) \cdot Z_i^2(t) \mathrm{d}u$$

Beauty of this method Note

$$S_i(t) \perp Z_i(t)$$

and

$$Z_i(t) = \sqrt{t} X_i + \sqrt{1-t} Y_i$$

therefore

$$Z_i'(t) = \left(\frac{X_i}{\sqrt{t}} - \frac{Y_i}{\sqrt{1-t}} \right) \cdot \frac{1}{2}$$
$$\mathbb{E} Z_i(t) = 0 \Rightarrow \mathbb{E} I_1 = 0$$

Z_i and Z_i' are uncorrelated

$$\mathbb{E} Z_i(t) Z_i'(t) = \frac{1}{2} \mathbb{E} \left(\sqrt{t} X_i + \sqrt{1-t} Y_i \right) \left(\frac{X_i}{\sqrt{t}} - \frac{Y_i}{\sqrt{1-t}} \right)$$

$$\mathbb{E} Z_i'(\alpha) \cdot Z_i(\alpha) = 0.$$

Another way: note $\mathbb{E} Z_i^2(t) = 1$ so taking derivatives (using differentiation under integral)

$$0 = \frac{d}{dt} \mathbb{E} Z_i^2(t) = \mathbb{E} \left[2 \cdot Z_i(t) \cdot Z_i'(t) \right].$$

Chenozhukov's version

More intuitive representation: Let $t = \cos^2 \alpha$

$$\bar{Z}_i(\alpha) = \cos\alpha \cdot X_i + \sin\alpha \cdot Y_i,$$
$$\bar{Z}_i'(\alpha) = -\sin\alpha \cdot X_i + \cos\alpha \cdot Y_i.$$

Then

$$\mathbb{E} \bar{Z}_i'(\alpha) \bar{Z}_i(\alpha) = 0$$

Let

$$\bar{S}(\alpha) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \bar{Z}_i(\alpha).$$

Then

$$S_n^X = \bar{S}(0), \qquad S_n^Y = \bar{S}\left(\frac{\pi}{2}\right).$$

Thus

$$\varphi_\varepsilon'(S(\alpha)) - \varphi_\varepsilon'(S_i(\alpha)) = \int_0^{\frac{\pi}{2}} \varphi_\varepsilon'(\bar{S}(\alpha)) \cdot \sum_{i=1}^{n} \frac{\bar{Z}_i'(\alpha)}{\sqrt{n}} \mathrm{d}\alpha$$

Picture

$$\widetilde{Z}_i(u) = e^{-u} X_i + \sqrt{1 - e^{-2u}} Y_i$$
$$\widetilde{Z}_i'(u) = -e^{-u} X_i + \frac{e^{-2u}}{\sqrt{1 - e^{-2u}}} Y_i$$

Bhattacharya & Holmes (2010). **Relations to OU process? Brownian bridge?**

We have

$$\begin{aligned} \mathbb{E}A_i = \mathbb{E}I_3 &= \frac{1}{n^{3/2}} \mathbb{E} \left| \int_0^1 \mathrm{d}t \int_0^1 (1-u) \varphi_\varepsilon'' \left(S_i(t) + u \cdot \frac{Z_i(t)}{\sqrt{n}} \right) \cdot Z_i'(t) \cdot Z_i^2(t) \mathrm{d}u \right| \\ &\leq \frac{\|\varphi_\varepsilon'''\|_\infty}{n^{3/2}} \int_0^1 \mathrm{d}t \int_0^1 \mathbb{E}\left[|Z_i'(t)| \cdot Z_i^2(t) \right] \mathrm{d}u. \end{aligned}$$

Angle representation: using Hölder's inequality

$$\begin{aligned} \mathbb{E} \int_0^{\frac{\pi}{2}} \bar{Z}_i^2(\alpha) |\bar{Z}_i'(\alpha)| \mathrm{d}\alpha &\leq \int_0^{\frac{\pi}{2}} \left(\mathbb{E}|\bar{Z}_i(\alpha)|^3 \right)^{2/3} \left(\mathbb{E}|\bar{Z}_i'(\alpha)|^3 \right)^{1/3} \mathrm{d}\alpha \\ &\leq \frac{\pi}{2} \left(\mathbb{E}|X_1|^3 + \mathbb{E}|Y_1|^3 \right). \end{aligned}$$

Conclusion: Slepian's interpolation gives the $n^{-1/8}$ rate!

Connection between Stein's method and Slepian's interpolation: Stein's differential equation is as follows: for $x \in \mathbb{R}^d$

$$\Delta f(\mathbf{x}) - \mathbf{x}^\top \nabla f(\mathbf{x}) = h(\mathbf{x}) - \mathbb{E}_{\boldsymbol{Y} \sim N(0, I_d)} h(\boldsymbol{Y}).$$

Solution **Uniqueness?** is

$$f(\mathbf{x}) = -\int_0^1 \frac{1}{2t} \left\{ \mathbb{E}_{\boldsymbol{Y}} h\left(\sqrt{t}\mathbf{x} + \sqrt{1-t}\boldsymbol{Y} \right) - \mu \right\} \mathrm{d}t,$$

where $\mu = \mathbb{E}_{\boldsymbol{Y} \sim N(0, I_d)} h(\boldsymbol{Y})$.

$$\mathbb{E}h(S_n^X)-\mathbb{E}h(S_n^Y)=\mathbb{E}\left[\Delta f(S_n^X)-(S_n^X)^\top\nabla f(S_n^X)\right].$$

By the definition of f

$$\mathbf{x}^\top\nabla f(\mathbf{x})=-\mathbb{E}_{\boldsymbol{Y}}\int_0^1\frac{1}{2}\frac{\boldsymbol{x}^\top}{\sqrt{t}}\nabla h\left(\sqrt{t}\mathbf{x}+\sqrt{1-t}\boldsymbol{Y}\right)\mathrm{d}t.$$

Also using Stein's equation

$$\begin{aligned}\Delta f(\mathbf{x})&=-\mathbb{E}\int_0^1\frac{1}{2}\Delta h\left(\sqrt{t}\mathbf{x}+\sqrt{1-t}\boldsymbol{Y}\right)\mathrm{d}t\\&=-\mathbb{E}\int_0^1-\frac{1}{2}\frac{\boldsymbol{Y}^\top}{\sqrt{1-t}}\cdot\nabla h\left(\sqrt{t}\mathbf{x}+\sqrt{1-t}\boldsymbol{Y}\right)\mathrm{d}t\end{aligned}$$

Let $\mathbf{J}$ be the Jacobian matrix

$$\mathbb{E}\boldsymbol{Y}^\top f(\boldsymbol{Y})=\mathbb{E}\mathbf{J}f(\boldsymbol{Y}).$$

This is just

$$\sum_{j=1}^d\mathbb{E}\boldsymbol{Y}_j f_j(\boldsymbol{Y})=\sum_{j=1}^d\mathbb{E}\frac{\partial f_j}{\partial x_j}(\boldsymbol{Y}).$$

Therefore

$$\begin{aligned}\mathbb{E}h(S_n^X)-\mathbb{E}h(S_n^Y)&=-\mathbb{E}\int_0^1\frac{1}{2}\left(\frac{\boldsymbol{Y}}{\sqrt{1-t}}-\frac{S_n^X}{\sqrt{t}}\right)^\top\cdot\nabla h\left(\sqrt{t}S_n^X+\sqrt{1-t}\boldsymbol{Y}\right)\mathrm{d}t\\&=\mathbb{E}\int_0^1\frac{1}{2}\left(\frac{S_n^X}{\sqrt{t}}-\frac{S_n^Y}{\sqrt{1-t}}\right)^\top\cdot\nabla h\left(\sqrt{t}S_n^X+\sqrt{1-t}S_n^Y\right)\mathrm{d}t\\&=\mathbb{E}\int_0^1\sum_{i=1}^n\left[Z_i'(t)\right]^\top\nabla h(S(t))\mathrm{d}t,\end{aligned}$$

where the last equality is due to $\boldsymbol{Y}=^d S_n^Y$ the standard normal. Take $h=\varphi_{\varepsilon,A}$ works. $\frac{1}{2}\left(\frac{S_n^X}{\sqrt{t}}-\frac{S_n^Y}{\sqrt{1-t}}\right)=\sum\frac{1}{2}\left(\frac{X_i}{\sqrt{t}}-\frac{Y_i}{\sqrt{1-t}}\right)$.

Slepian's seems better

Chapter 3
Berry Esseen Bound in $\mathbb{R}^d$

In this chapter, we focus on multivariate Berry Esseen bound with regard to convex sets. We are curious about the rate's dependence on dimension "d", which will be the dominant factor in high-dimensional regimes. Our main techniques are smooth function approximation, Taylor expansion combined with Stein's method and Slepian's interpolation. Two questions remain open: whether the rate is optimal considering d and if it is possible to replace the L_2 ball by L_∞ ball in the settings to achieve better rate w.r.t. d.

In the following parts, we first define Kolmogorov distance w.r.t. convex sets as our target. Then we introduce the main theorem and prove it in 3 steps.

Notations: We use $\|\|_{op}$ to denote the operator norm, i.e. for a linear operator $\mathscr{P}$ in $\mathbb{R}^d$, we have $\|\mathscr{P}\|_{op} = \sup_{\mathbf{v}\in\mathbb{R}^d, \|\mathbf{v}\|_2=1} |\mathscr{P}(\mathbf{v})|$. Let $\mathscr{D}$ be the differential operator.

Definition 3.1 (Kolomogorov Distance w.r.t. Convex Sets). Let $\boldsymbol{X}$, $\boldsymbol{Y}$ be d dimensional random vectors, $\mathscr{A}$ be the set of convex sets in $\mathbb{R}^d$. We define

$$d_C(\boldsymbol{X},\boldsymbol{Y}) = \sup_{A\in\mathscr{A}} \left|\mathbb{P}\left(\boldsymbol{X}\in A\right) - \mathbb{P}\left(\boldsymbol{Y}\in A\right)\right|.$$

Recall that in univariate case, due to the simplicity of geometry, we only need to consider A to be $(-\infty,t)$. In multivariate scenario, two reasons contribute to why we focus on convex sets: they are well studied due to widespread applications like convex optimization and their properties are favorable in constructing the approximation function and estimating smooth error.

Theorem 3.1. *Let $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ be i.i.d. random vectors in $\mathbb{R}^d$, $d \ll n$. $\mathbb{E}\left[\boldsymbol{X}_1\right] = \mathbf{0}$, $\mathbb{E}\left[\boldsymbol{X}_1\boldsymbol{X}_1^\top\right] = \mathbf{I_d}$ and $\mathbb{E}\left[\|\boldsymbol{X}_1\|_2^3\right] = \rho$. Let $\boldsymbol{Z}\sim N(0,\mathbf{I_d})$, $\boldsymbol{S}_n = \sum_{i=1}^n \boldsymbol{X}_i/\sqrt{n}$, then there exists a constant $C>0$, s.t.*

$$d_C(\boldsymbol{S}_n,\boldsymbol{Z}) \le C\rho d^{\frac{1}{4}} n^{-\frac{1}{2}}. \tag{3.0.1}$$

Compared with univariate case, we have the same rate w.r.t. n and ρ.The proofs follow the same structure, but in multivariate regime we will construct different approximation function, introduce Gaussian surface area to bound smooth error and deal with matrix or even tensor norms. The reason we use $\mathbb{E}\left[\|\boldsymbol{X}_1\|_2^3\right] = \rho$, i.e. L_2 norm here is because we utilize operator (L_2) norm of matrix and inequality $\langle \mathbf{u}, \mathbf{v} \rangle \leq \|\mathbf{u}\|_2 \|\mathbf{v}\|_2$ in the proof. It may be replaced by L_∞ and possibly the rate will improve significantly due to the order of $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_\infty$ w.r.t. d. Notice that ρ is of $\mathscr{O}(d^{3/2})$ and the rate $\mathscr{O}(\rho d^{1/4} n^{-1/2})$ may be larger than 1 and explode if d grows fast with n. So this result does not apply to high-dimensional case.

Proof. Our proof assembles the univariate case and is conducted in three major steps. **Step 1: Approximate $\mathbf{1}_A$ by smooth funtion.** For any convex set A, we hope to find a function f that can approximate the indicator function $\mathbf{1_A}$ well, at least three times differentiable and derivatives up to the third order can be controlled. The first property is natural since we aim to use $\mathbb{E}[\mathbf{f}(\boldsymbol{X})]$ to approximate $\mathbb{E}[\mathbf{1_A}(\boldsymbol{X})]$. The last two requirements make Taylor expansion of f possible, which is the core technique of the proof. So far, there has not been a very suitable closed form of f. We assume that there exists an f satisfying following properties and use it in the proof.

Lemma 3.2 (Approximation Function). *Let $A \in \mathbb{R}^d$ be a convex set, define $A^\varepsilon = \bigcup \left\{\mathscr{B}(y,\varepsilon) \subseteq \mathbb{R}^d | y \in A\right\}$, $A^{-\varepsilon} = \bigcup \left\{y \in \mathbb{R}^d | \mathscr{B}(y,\varepsilon) \subseteq A\right\}$, then $A^{\pm\varepsilon}$ are convex. For any $\varepsilon > 0$, there exists a function $h_{A,\pm\varepsilon}$ that holds following properties*

1. *Smoothness: $h_{A,\pm\varepsilon}$ is three times continuous differentiable;*
2. *Operator norm: operator norms of derivatives of $h_{A,\pm\varepsilon}(x)$ are controlled uniformly w.r.t. $\mathbf{x}$ by*
$$\|\mathscr{D}^{(j)} h_{A,\pm\varepsilon}\|_{op} \leq C\varepsilon^{-j}, \tag{3.0.2}$$
where C is a constant which does not depend on j, d, n, A, ε;
3. *Support: $h_{A,\pm\varepsilon}$ is compactly supported and further*

$$\begin{cases} h_{A,\varepsilon}(\mathbf{x}) = 1, & \mathbf{x} \in A \\ h_{A,\varepsilon}(\mathbf{x}) = 0, & \mathbf{x} \notin A^{2\varepsilon} \\ h_{A,\varepsilon}(\mathbf{x}) \in [0,1], & \mathbf{x} \in A^{2\varepsilon} \setminus A \end{cases}$$

$$\begin{cases} h_{A,-\varepsilon}(\mathbf{x}) = 1, & \mathbf{x} \in A^{-2\varepsilon} \\ h_{A,-\varepsilon}(\mathbf{x}) = 0, & \mathbf{x} \notin A \\ h_{A,-\varepsilon}(\mathbf{x}) \in [0,1], & \mathbf{x} \in A \setminus A^{-2\varepsilon} \end{cases}$$

Step 2: Estimate smooth error. Similar to univariate case, we derive the smooth error by the relationship between $\mathbf{1_A}$ and $h_{A,\pm\varepsilon}$. What is different here is the upper bound of $\sup_{A\in\mathscr{A}} |\mathbb{P}(\boldsymbol{Z} \in A^{2\varepsilon}) - \mathbb{P}(\boldsymbol{Z} \in A)|$. In one-dimension case, we use $2\varepsilon/\sqrt{2\pi}$ directly, but in multivariate regime we need to consider the influence of d carefully. Recall the third property in lemma 3.2, we have

$$\begin{aligned} \mathbb{P}(\boldsymbol{S}_n \in A) - \mathbb{P}(\boldsymbol{Z} \in A) &\leq \mathbb{E}[h_{A,\varepsilon}(\boldsymbol{S}_n)] - \mathbb{E}[h_{A,\varepsilon}(\boldsymbol{Z})] + \mathbb{E}[\mathbf{1}_{\mathbf{A}^{2\varepsilon}}(\boldsymbol{Z})] - \mathbb{E}[\mathbf{1_A}(\boldsymbol{Z})] \\ &\leq \sup_{A\in\mathscr{A}} |\mathbb{E}[h_{A,\varepsilon}(\boldsymbol{S}_n)] - \mathbb{E}[h_{A,\varepsilon}(\boldsymbol{Z})]| + \sup_{A\in\mathscr{A}} \mathbb{P}(\boldsymbol{Z} \in A^{2\varepsilon} \setminus A). \end{aligned}$$

Similarly on the other hand we have

$$\begin{aligned}\mathbb{P}(\boldsymbol{Z}\in A)-\mathbb{P}(\boldsymbol{S}_n\in A)&\le\mathbb{E}\left[\mathbf{1}_{\mathbf{A}}(\boldsymbol{Z})\right]-\mathbb{E}\left[\mathbf{1}_{\mathbf{A}^{-2\varepsilon}}(\boldsymbol{Z})\right]+\mathbb{E}\left[h_{A,-\varepsilon}(\boldsymbol{Z})\right]-\mathbb{E}\left[h_{A,-\varepsilon}(\boldsymbol{S}_n)\right]\\&\le\sup_{A\in\mathscr{A}}\left|\mathbb{E}\left[h_{A,-\varepsilon}(\boldsymbol{S}_n)\right]-\mathbb{E}\left[h_{A,-\varepsilon}(\boldsymbol{Z})\right]\right|+\sup_{A\in\mathscr{A}}\mathbb{P}(\boldsymbol{Z}\in A\setminus A^{-2\varepsilon}).\end{aligned}$$

Combine the above two parts and notice that in our construction $h_{A,-\varepsilon}=h_{A^{-\varepsilon},\varepsilon}$, we obtain

$$\begin{aligned}d_C(\boldsymbol{S}_n,\boldsymbol{Z})&=\sup_{A\in\mathscr{A}}\left|\mathbb{P}(\boldsymbol{S}_n\in A)-\mathbb{P}(\boldsymbol{Z}\in A)\right|\\&\le\sup_{A\in\mathscr{A}}\left|\mathbb{E}\left[h_{A,\varepsilon}(\boldsymbol{S}_n)\right]-\mathbb{E}\left[h_{A,\varepsilon}(\boldsymbol{Z})\right]\right|+\sup_{A\in\mathscr{A}}\max\left\{\mathbb{P}(\boldsymbol{Z}\in A^{2\varepsilon}\setminus A),\,\mathbb{P}(\boldsymbol{Z}\in A\setminus A^{-2\varepsilon})\right\}.\end{aligned}\tag{3.0.3}$$

For the first term in Eq. (3.0.3), we will leave its bound to **Step 3**. For the second term in Eq. (3.0.3), we have the following lemma.

Lemma 3.3.

$$\sup_{A\in\mathscr{A}}\max\left\{\mathbb{P}(\boldsymbol{Z}\in A^{2\varepsilon}\setminus A),\,\mathbb{P}(\boldsymbol{Z}\in A\setminus A^{-2\varepsilon})\right\}\le 8d^{\frac{1}{4}}\varepsilon.\tag{3.0.4}$$

This is a conclusion from convex geometry. Whether the rate of $d^{1/4}$ is sharp or not remains open. We call $\int_{\partial A}\phi(x)\,d\sigma$ the Gaussian surface area of A, where $\phi(x)$ denotes the Gaussian density. We can see that this quantity is similar to what we aim to evaluate. Lemma 3.0.4 is the core to improve the final rate or extend convex sets to other interesting families of sets, e.g. sets in stochastic gradient descent. If we replace convex sets by L_2 balls, the upper bound is of constant level. If we consider polytops, the quantity is of rate $\mathscr{O}(\log(d))$, which may be useful if we consider L_∞ ball. Relevant research topics include iso-parameterization.

Step 3: Bound approximation error of smooth functions. Before we delve into the proof of Theorem 3.1, we first give a straightforward but not sharp upper bound by directly adapting the proof in univariate case using Stein's method.

Recall the multivariate version of Stein's PDE, denote f as the solution, we have

$$\Delta f(\mathbf{x})-\mathbf{x}^\top\nabla f(\mathbf{x})=h(\mathbf{x})-\mathbb{E}_{\boldsymbol{Z}}\left[h(\boldsymbol{Z})\right].\tag{3.0.5}$$

Since we aim to bound the expectation of the right hand side, we use Stein's equation as a bridge and focus on the left hand side. In fact, we have closed form of f, ∇f and $\nabla^2 f$

$$f(\mathbf{x})=-\int_0^{\frac{\pi}{2}}\frac{\cos\theta}{\sin\theta}\left(\mathbb{E}\left[h(\cos(\theta\boldsymbol{Z})+\sin(\theta\mathbf{x}))\right]-\mathbb{E}_{\boldsymbol{Z}}[h(\boldsymbol{Z})]\right)d\theta.\tag{3.0.6}$$

$$\nabla f(\mathbf{x})=-\int_0^{\frac{\pi}{2}}\cos\theta\cdot\mathbb{E}\left[\nabla h(\cos(\theta\boldsymbol{Z})+\sin(\theta\mathbf{x}))\right]d\theta.\tag{3.0.7}$$

$$\nabla^2 f(\mathbf{x})=-\int_0^{\frac{\pi}{2}}\cos\theta\sin\theta\cdot\mathbb{E}\left[\nabla^2 h(\cos(\theta\boldsymbol{Z})+\sin(\theta\mathbf{x}))\right]d\theta.\tag{3.0.8}$$

Define $\boldsymbol{S}^{(i)} = \boldsymbol{S}_n - \boldsymbol{X}_i/\sqrt{n}$, we evaluate two terms on the left hand side of Eq. (3.0.5) respectively and then merge them together.
Let $\left\{\boldsymbol{X}_i^{`}\right\}$ be an independent identical copy of $\{\boldsymbol{X}_i\}$, recall that $\mathbb{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^\top\right] = \mathbf{I_d}$,

$$\mathbb{E}\left[\Delta f(\boldsymbol{S}_{\mathbf{n}})\right] = \mathbb{E}\left[\operatorname{tr}\left(\nabla^2 f(\boldsymbol{S}_{\mathbf{n}})\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\operatorname{tr}\left(\nabla^2 f(\boldsymbol{S}^{(i)} + \frac{\boldsymbol{X}_i^{`}}{\sqrt{n}})\cdot\mathbb{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^\top\right]\right)\right].$$

By the independence of $\boldsymbol{S}^{(i)}$,$\boldsymbol{X}_i$, $\boldsymbol{X}_i^{`}$, we have

$$\mathbb{E}\left[\Delta f(\boldsymbol{S}_{\mathbf{n}})\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\langle\nabla^2 f(\boldsymbol{S}^{(i)} + \frac{\boldsymbol{X}_i^{`}}{\sqrt{n}}), \boldsymbol{X}_i\boldsymbol{X}_i^\top\rangle\right]. \tag{3.0.9}$$

For the other term, we use leave one out trick again and do Taylor expansion

$$\mathbb{E}\left[\boldsymbol{S}_n^\top\nabla f(\boldsymbol{S}_N)\right] = \sum_{i=1}^{n}\mathbb{E}\left[\frac{\boldsymbol{X}_i^\top}{\sqrt{n}}\nabla f(\boldsymbol{S}^{(i)} + \frac{\boldsymbol{X}_i}{\sqrt{n}})\right]$$

$$= \sum_{i=1}^{n}\mathbb{E}\left[\frac{\boldsymbol{X}_i^\top}{\sqrt{n}}\nabla f(\boldsymbol{S}^{(i)}) + \frac{\boldsymbol{X}_i^\top}{\sqrt{n}}\nabla^2 f(\boldsymbol{S}^{(i)} + t_i\frac{\boldsymbol{X}_i}{\sqrt{n}})\frac{\boldsymbol{X}_i}{\sqrt{n}}\right].$$

Since $\mathbb{E}\left[\boldsymbol{X}_i\right] = \mathbf{0}$ and use the independence

$$\mathbb{E}\left[\boldsymbol{S}_n^\top\nabla f(\boldsymbol{S}_N)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\boldsymbol{X}_i^\top\nabla^2 f(\boldsymbol{S}^{(i)} + t_i\frac{\boldsymbol{X}_i}{\sqrt{n}})\boldsymbol{X}_i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\langle\nabla^2 f(\boldsymbol{S}^{(i)} + t_i\frac{\boldsymbol{X}_i}{\sqrt{n}}), \boldsymbol{X}_i\boldsymbol{X}_i^\top\rangle\right]. \tag{3.0.10}$$

Now we merge Eq. (3.0.9) and Eq. (3.0.10) together

$$\left|\mathbb{E}\left[\Delta f(\boldsymbol{S}_{\mathbf{n}})\right] - \mathbb{E}\left[\boldsymbol{S}_n^\top\nabla f(\boldsymbol{S}_N)\right]\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\langle\nabla^2 f(\boldsymbol{S}^{(i)} + \frac{\boldsymbol{X}_i^{`}}{\sqrt{n}}), \boldsymbol{X}_i\boldsymbol{X}_i^\top\rangle - \langle\nabla^2 f(\boldsymbol{S}^{(i)} + t_i\frac{\boldsymbol{X}_i}{\sqrt{n}}), \boldsymbol{X}_i\boldsymbol{X}_i^\top\rangle\right]\right|$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}\left[\langle\nabla^2 f(\boldsymbol{S}^{(i)} + \frac{\boldsymbol{X}_i^{`}}{\sqrt{n}}) - \nabla^2 f(\boldsymbol{S}^{(i)}), \boldsymbol{X}_i\boldsymbol{X}_i^\top\rangle\right]\right| + \frac{1}{n}\sum_{i=1}^{n}\left|\mathbb{E}\left[\langle\nabla^2 f(\boldsymbol{S}^{(i)} + t_i\frac{\boldsymbol{X}_i}{\sqrt{n}}) - \nabla^2 f(\boldsymbol{S}^{(i)}), \boldsymbol{X}_i\boldsymbol{X}_i^\top\rangle\right]\right|$$

$$\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\nabla^2 f(\boldsymbol{S}^{(i)} + \frac{\boldsymbol{X}_i^{`}}{\sqrt{n}}) - \nabla^2 f(\boldsymbol{S}^{(i)})\|_{op}\cdot\|\boldsymbol{X}_i\|_2^2\right]$$

$$+ \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\|\nabla^2 f(\boldsymbol{S}^{(i)} + t_i\frac{\boldsymbol{X}_i}{\sqrt{n}}) - \nabla^2 f(\boldsymbol{S}^{(i)})\|_{op}\cdot\|\boldsymbol{X}_i\|_2^2\right]. \tag{3.0.11}$$

Now we only need to bound $\|\nabla^2 f(\boldsymbol{S}^{(i)} + \boldsymbol{X}_i^{`}/\sqrt{n}) - \nabla^2 f(\boldsymbol{S}^{(i)})\|_{op}$, $\|\nabla^2 f(\boldsymbol{S}^{(i)} + t_i\boldsymbol{X}_i/\sqrt{n}) - \nabla^2 f(\boldsymbol{S}^{(i)})\|_{op}$. Recall the closed form w.r.t. f in Eq. (3.0.6), Eq. (3.0.7) and Eq. (3.0.8), we have for $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^d$

$$\|\nabla^2 f(\mathbf{x}+\mathbf{y}) - \nabla^2 f(\mathbf{x})\|_{op} = \left\| \int_0^{\frac{\pi}{2}} \cos\theta \sin^2\theta \cdot \mathbb{E}\left[\int_0^1 \nabla^3 h(\cos(\theta \boldsymbol{Z}) + \sin(\theta \mathbf{x}) + \tau \sin(\theta \mathbf{y}))\mathbf{y} d\tau \right] d\theta \right\|_{op}$$

Recall Eq. (3.0.2), we obtain

$$\|\nabla^2 f(\mathbf{x}+\mathbf{y}) - f(\mathbf{x})\|_{op} \leq C_0 \cdot \|\nabla^3 h\|_{op} \|\mathbf{y}\|_2 . \leq C_1 \cdot \frac{\|\mathbf{y}\|_2}{\varepsilon^3}.$$

We plug in this general form by what we are considering and get

$$\|\nabla^2 f(\boldsymbol{S}^{(i)} + t_i \frac{\boldsymbol{X}_i}{\sqrt{n}}) - \nabla^2 f(\boldsymbol{S}^{(i)})\|_{op} \leq \frac{C_1}{\sqrt{n}} \cdot \frac{\|\boldsymbol{X}_{\mathbf{i}}\|_2}{\varepsilon^3}$$

$$\|\nabla^2 f(\boldsymbol{S}^{(i)} + \frac{\boldsymbol{X}_i^{`}}{\sqrt{n}}) - \nabla^2 f(\boldsymbol{S}^{(i)})\|_{op} \leq \frac{C_1}{\sqrt{n}} \cdot \frac{\|\boldsymbol{X}_{\mathbf{i}}\|_2}{\varepsilon^3}.$$

Combine each term together in Eq. (3.0.11)

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\nabla^2 f(\boldsymbol{S}^{(i)} + \frac{\boldsymbol{X}_i^{`}}{\sqrt{n}}) - \nabla^2 f(\boldsymbol{S}^{(i)})\|_{op} \cdot \|\boldsymbol{X}_i\|_2^2 \right] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\|\nabla^2 f(\boldsymbol{S}^{(i)} + t_i \frac{\boldsymbol{X}_i}{\sqrt{n}}) - \nabla^2 f(\boldsymbol{S}^{(i)})\|_{op} \cdot \|\boldsymbol{X}_i\|_2^2 \right]$$

$$\leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{C_1}{\varepsilon^3 \sqrt{n}} \|\boldsymbol{X}_i^{`}\|_2 \|\boldsymbol{X}_i\|_2^2 \right] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\frac{C_1}{\varepsilon^3 \sqrt{n}} \|\boldsymbol{X}_i\|_2 \|\boldsymbol{X}_i\|_2^2 \right] \leq \frac{C_2 \rho}{\varepsilon^3 \sqrt{n}}.$$

By adding the smooth error in Lemma 3.0.4 and take $\varepsilon = \rho^{1/4} d^{-1/16} n^{-1/8}$ we obtain

$$d_{\mathscr{C}}(\boldsymbol{S}_n, \boldsymbol{Z}) \leq C \rho^{\frac{1}{4}} d^{\frac{3}{16}} n^{\frac{1}{8}}. \tag{3.0.12}$$

Compared with the rate stated in Theorem 3.1, we see that when $d \geq n^{6/19}$, the rate just obtained is no worse. However, when d grows at such a speed with n, the upper bound will be greater than 1 as n goes to infinity, which is meaningless since d_C is always upper bounded by 1.

To be continued: Theorem 3.1. Where can we improve the estimation and the high-level proof.

Chapter 4
Concentration Inequality

Concentration is a common phenomenon. Let $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ be i.i.d. random variables, then by the law of large numbers, we have

$$\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{X}_{\boldsymbol{k}}-\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\boldsymbol{X}_k\right]\to 0 \quad \text{as} \quad n\to\infty.$$

We can also view this in a more general way: let $f(x_1,...,x_n)=\left(\sum_{k=1}^{n}x_k\right)/n$, and when n is sufficiently large, we have random variable $f(\boldsymbol{X}_{\boldsymbol{1}},...,\boldsymbol{X}_{\boldsymbol{n}})$ "close" to its mean. Actually this is not only valid when f is a linear function. So we first introduce an informal idea of concentration:
Let $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ be independent random variables, then random variable $f(\boldsymbol{X}_{\boldsymbol{1}},...,\boldsymbol{X}_{\boldsymbol{1}})$ is "close" to its mean if $f(x_1,...,x_n)$ is not too "sensitive" to $x_1,\dots,x_n$.
In this chapter, we need to clarify rigorously:

- What is "close"?
- What is "sensitive"?

As for "close", we review central limit theorem: let $\boldsymbol{X}_{\boldsymbol{1}}$,...$\boldsymbol{X}_{\boldsymbol{n}}$ be i.i.d. random variables, $\mathrm{Var}(\boldsymbol{X}_{\boldsymbol{1}})=\sigma^2$ then when n is large enough

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\boldsymbol{X}_{\boldsymbol{k}}-\mathbb{E}\boldsymbol{X}_{\boldsymbol{k}})\approx N(0,\sigma^2).$$

Here $\approx$ means two random variables are similar in distribution. If we treat the left hand side as Gaussian, we get the concentration rate

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\boldsymbol{X}_{\boldsymbol{k}}-\mathbb{E}\boldsymbol{X}_{\boldsymbol{k}})\geq t\right)\leq e^{-t^2/2\sigma^2}.$$

Inspired by CLT, we use sub-Gaussian distribution to characterize "close", which will be explained rigorously later. Note that actually we get a stronger conclusion in CLT than needed here, since we not only state that the random variable concentrates to its mean, but also provide the first order distribution approximation. Though the latter is not to our interest here, we emphasize that above approximation can be viewed as

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\boldsymbol{X_k}-\mathbb{E}\boldsymbol{X_k})\approx\frac{1}{\sqrt{n}}\sum_{k=1}^{n}(\boldsymbol{Z_k}-\mathbb{E}\boldsymbol{Z_k}).$$

where $\boldsymbol{Z_k}$ are i.i.d. Gaussian variables with $\mathrm{Var}(\boldsymbol{Z_1})=\sigma^2$. This demonstration actually shows that in calculating the distribution, we can treat each random variables as Gaussian. This enables us to use a nummber of good properties of Gaussian random variables while other distributions may be hard to compute or bound in high dimensional status.

As for "sensitive", intuitively, if $f(\boldsymbol{X_1},...,\boldsymbol{X_n})$ changes rapidly w.r.t. $\boldsymbol{X_1}$ or the range of function value is rather large, the concentration is unlikely to be good. To be rigorous, we use discrete and continuous gradients to characterize such "sensitiveness" of f to $\boldsymbol{X_i}$. For discrete gradient, the limitations are that we are only able to handle the case where $f(\boldsymbol{X_1},...,\boldsymbol{X_n})$ is $a.s.$ bounded. This is enough for some problems like random matrix. But against our will, Gaussian usually does not lie in this regime, so we need continuous gradient to help handle at least the case of Gaussian. We develop a general tool for continuous gradient and use Markov process ... .

In this chapter, we first define sub-Gaussian, discrete gradient as pre-knowledge.Then we give two upper bounds with sub-Gaussian variance proxies considering discrete gradient. The proofs follow two similar steps: analyse univariate case and bridge through univariate and multivariate regimes. To be specific, we will use Hoeffding lemma and discrete log-Soblev inequality for the first part and introduce martingale and tensorization of entropy for the latter. In terms of continuous gradient, we first introduce the general Markov Semi-group and bounds. Then we focus on Gaussian case. (TBC)

Key Words: Concentration Inequality, Martingale, Entropy, Tensorization

4.1 Sub-Gaussian

Gaussian is a commonly used distribution and compared with other distribution such as exponential and polynomial ones, it has lighter tails which means it concentrates much better. Intuitively, we hope that the "close" should be sharp so we use sub-Gaussian. Another reason why we consider sub-Gaussian is that when $\boldsymbol{X_k}$ follows Gaussian distribution and f is taking means, we have the variable concentrate Guassianly around its expectation. Compared with the Markov inequality, we usually obtain

$$\mathbb{P}(\boldsymbol{X}-\mathbb{E}\boldsymbol{X}\geq t)\leq\frac{\mathrm{Var}\,\boldsymbol{X}}{t^2}.$$

which decays polynomially away from the central, our sub-Gaussian "closeness" is much stronger.

We first state the definition of sub-Gaussian random variable.

Definition 4.1. (Sub-Gaussian) A random variable $\boldsymbol{X}$ is said to be sub-Gaussian with variance proxy σ^2 if it satisfies either of the following inequality for $\forall \lambda \in \mathbb{R}$:

- $\mathbb{P}(\boldsymbol{X} - \mathbb{E}\boldsymbol{X} \geq t) \leq e^{-t^2/2\sigma^2} \quad \text{for } \forall\, t \in \mathbb{R}$
- $\mathbb{E}(e^{\lambda \boldsymbol{X}}) \leq \frac{\lambda^2\sigma^2}{2}$.

Though the above definitions are equivalent with different variance proxies, in practise, the log-moment inequality is easier to prove, considering the properties of the expectation of sum of independent random variables and the fact that the first inequality needs to be checked for $\forall\, t \in \mathbb{R}$. So in the following parts, we focus on how to estimate the centered log-moment.

In the definition, the variance proxy σ^2 is an index of ”close”, and much of our efforts are spared on how to obtain a smaller variance proxy. In the following parts, we use different methods to derive three variance proxies, where the first one is the weakest and the last one the sharpest.

4.2 Discrete Gradient

As we mentioned earlier, we need to define rigorously the concept “sensitive”, and here we first introduce discrete gradient, which characterizes a quantity similar to the range f can fluctuate due to $\boldsymbol{X_k}$. This does not give the sharpest bound, and in some context, we may only need the fluctuation up or down from the expectation. However, in most cases, these discrete gradients are comparable, and for simplicity we analyse the gradient defined as follows through this chapter.

Definition 4.2. (Discrete Gradient) Let $\boldsymbol{X_1}$,... be independent random variables, $f(x_1,...,x_n)$ is a multivariate function taking values in $\mathbb{R}$. Define discrete gradient of f w.r.t. $\boldsymbol{X_k}$

$$D_k f(x_1,...,x_{k-1},x_{k+1},...,x_n) = \sup_{\boldsymbol{X_k}} f(x_1,...,x_{k-1},\boldsymbol{X_k},x_{k+1},...,x_n) - \inf_{t} f(x_1,...,x_{k-1},\boldsymbol{X_k},x_{k+1},...,x_n).$$

4.2.1 Martingale method

To introduce the Martingale method, we first give the road map of the proof. We begin with the univariate case using Hoeffdng’s inequality and then use Martingale method to extend the conclusions to multivariate version. The decomposition of multivariate case to univariate case is striaghtforward, but since we overused the Hoeffding’s inequalities, we only obtained a loose upper bound, expecially when n is large.

Lemma 4.1. ***(Hoeffding)*** *Let random variable $\boldsymbol{X}$ satisfies $a \leq \boldsymbol{X} \leq b$ a.s. for some a, $b \in \mathbb{R}$, then $\boldsymbol{X}$ is $(b-a)^2/4$ sub-Gaussian.*

Proof. Without loss of generality, we assume that $\mathbb{E}\boldsymbol{X} = 0$. We use $\phi(\lambda) = \log\left(\mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right]\right)$, and by calculus we obtain

$$\phi'(\lambda) = \frac{\mathbb{E}\left[\boldsymbol{X}e^{\lambda \boldsymbol{X}}\right]}{e^{\lambda \boldsymbol{X}}}, \quad \phi''(\lambda) = \frac{\mathbb{E}\left[\boldsymbol{X}^2 e^{\lambda \boldsymbol{X}}\right]}{e^{\lambda \boldsymbol{X}}} - \left[\frac{\mathbb{E}\left[\boldsymbol{X}e^{\lambda \boldsymbol{X}}\right]}{e^{\lambda \boldsymbol{X}}}\right]^2.$$

By the form we can introduce a new measure $d\mathbb{Q} = e^{\lambda \boldsymbol{X}}/\mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right] d\mathbb{P}$ and obviously, $\phi'(\lambda)$ can be viewd as the expectation of $\boldsymbol{X}$ under the new measure and $\phi''(\lambda)$ as the new variance of $\boldsymbol{X}$. Since $\boldsymbol{X}$ takes value in $[a,b]$, so the variance, i.e. $\phi''(\lambda) \leq [b-a]^2/4$. Meanwhile, we observe that $\phi(0) = 0$, and $\phi'(0) = 0$ so by Taylor expansion at 0 we get

$$\phi(\lambda) = \phi(0) + \phi'(0)\lambda + \frac{1}{2}\phi''(\theta)\lambda^2 \leq \frac{\lambda^2(b-a)^2}{8},$$

where $\theta \in [\lambda, 0]$ or $[0, \lambda]$ depending on the sign of λ. Above tells us that $\boldsymbol{X}$ is $(b-a)^2/8$ sub-Gaussian.

Next, we hope to extend the univariate case to multivariate regime. For this purpose, we introduce the martingale difference as a bridge.

$$f(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}) - \mathbb{E}\left[f(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})\right] = \sum_{k=1}^{n} \Delta_k,$$

where $\Delta_k = \mathbb{E}\left[f(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})|\boldsymbol{X_1}, \ldots, \boldsymbol{X_k}\right] - \mathbb{E}\left[f(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})|\boldsymbol{X_1}, \ldots, \boldsymbol{X}_{k-1}\right]$. By using exponential of sum can be divided to the multiplication of individual terms, we can have the following lemma.

Lemma 4.2 (Azuma). *Let $\{\Delta_k\}_{1 \leq k \leq n}$ be martingale differences with regard to filtration $\{\mathscr{F}_k\}_{1 \leq k \leq n}$. If $\mathbb{E}\left[e^{\lambda \Delta_k}|\mathscr{F}_{k-1}\right] \leq e^{\lambda^2 \sigma_k^2/2}$ a.s., then $\sum_{k=1}^{n} \Delta_k$ is sub-Gaussian with variance proxy $\left(\sum_{k=1}^{n} \sigma_k^2\right)$.*

Proof. The core of the proof is breaking the expectation of exponential of sum into the multiplication of single exponential where the Hoeffding's lemma can be applied. To be specific,

$$\mathbb{E}\left[e^{\lambda \sum_{i=1}^{k} \Delta_i}\right] = \mathbb{E}\left[e^{\lambda \sum_{i=1}^{k-1} \Delta_i} \mathbb{E}\left[e^{\lambda \Delta_k} \,|\, \mathscr{F}_{k-1}\right]\right] \leq e^{\lambda^2 \sigma_k^2/2}\, \mathbb{E}\left[e^{\lambda \sum_{i=1}^{k-1} \Delta_i}\right].$$

By induction we get $\mathbb{E}\left[e^{\lambda \sum_{i=1}^{k} \Delta_i}\right] \leq e^{\lambda^2 \sum_{i=1}^{n} \sigma_i^2/2}$ and our desired result.

Now we are ready to combine the two lemmas to derive the final theorem.

Theorem 4.3. ***(McDiarmid)*** *Let $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ be independent random variables, then $f(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})$ is $\left(\sum_{k=1}^{n} \|D_k f\|_\infty^2\right)$ sub-Gaussian.*

Proof. We can simply apply the previous lemmas. To be specific, we let

$$A_K = \mathbb{E}\left[\inf_{\boldsymbol{X_k}} f(\boldsymbol{X_1},...,\boldsymbol{X}_k,...,\boldsymbol{X_n}) - f(\boldsymbol{X_1},...,\boldsymbol{X_n})|\boldsymbol{X_1},...,\boldsymbol{X}_{K-1}\right].$$

and

$$B_K = \mathbb{E}\left[\sup_{\boldsymbol{X_k}} f(\boldsymbol{X_1},...,\boldsymbol{X}_k,...,\boldsymbol{X_n}) - f(\boldsymbol{X_1},...,\boldsymbol{X_n})|\boldsymbol{X_1},...,\boldsymbol{X}_{K-1}\right].$$

Since $\boldsymbol{X_i}$ is independent, so we have $|B_k - A_K| \leq \|D_k f\|_\infty$. and by plugging in the A_K, B_K to the Azuma's Lemma we get the expected variance proxy.

Remark 4.1. Viewing that sub-Gaussian variance proxy $\sum_{k=1}^n \|D_k f\|_\infty^2$, we neglect the relationship between the fluctuation introduced by each coordinate. In other words, $\|\sum_{k=1}^n (D_k f)^2\|_\infty$ may be a more reasonable and obviously sharper variance proxy regarding that it reflects the range f can fluctuate as $\{\boldsymbol{X_k}\}$ varies.

4.2.2 Entropy Method

Regarding the limitation of McDiarmid's Theorem, we introduce the Entropy method, which gives us a much satisfying bound. Our road map of the proof is similar to the previous one, i.e. analyse the univariate case using discrete log-Soblev's inequality and use tensorizaion of entropy to extend the conclusions to multivariate version. Since the tensorization is a more significant improvement here, we will explain that first.

First of all, to introduce entropy, recall that in the proof of Hoeffding's lemma, we aim to derive $\phi(\lambda) \leq \lambda^2\sigma^2/2$ and directly Taylor expanded $\phi(\lambda)$ at 0 using the fact that $\phi(0) = \phi'(0) = 0$. The whole problem then is shifted to prove $\phi''(\lambda) \leq \sigma^2/2$. This time we change a little bit and turn to prove $\phi(\lambda)/\lambda \leq \lambda\sigma^2/2$. Though $\phi(0)/0$ is not well defined, we have $\lim_{\lambda\to 0}\phi(\lambda)/\lambda = 0$ by L'Hôpital's law. Thus we do Taylor expansion and what we need now is

$$\left(\frac{\phi(\lambda)}{\lambda}\right)' = \frac{\mathbb{E}\left[\lambda \boldsymbol{X} e^{\lambda \boldsymbol{X}}\right] - \mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right]\log\mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right]}{\lambda^2\mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right]} \leq \frac{\sigma^2}{2}.$$

We separate out the numerator in order to introduce entropy

$$\mathbb{E}\left[\lambda \boldsymbol{X} e^{\lambda \boldsymbol{X}}\right] - \mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right]\log\mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right] \leq \frac{\sigma^2}{2}\lambda^2\mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right].$$

Recall the classic entropy function $f(x) = x\log x$, the numerator is taking expectation outside multiplication minus inside multiplication. Besides, when we calculate the KL differences of two distributions $d\mathbb{P}^\lambda = e^{\lambda \boldsymbol{X}}/\mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right]d\mathbb{P}$ and $d\mathbb{P}$, we have

$$KL(d\mathbb{P}^{\lambda}, d\mathbb{P}) = \int \log\left(\frac{d\mathbb{P}^{\lambda}}{d\mathbb{P}}\right) d\mathbb{P}^{\lambda} = \int \frac{d\mathbb{P}^{\lambda}}{d\mathbb{P}} \cdot \left(\log\frac{d\mathbb{P}^{\lambda}}{d\mathbb{P}}\right) d\mathbb{P}$$

which gives us the entropy form. By further calculation, we have

$$KL(d\mathbb{P}^{\lambda}, d\mathbb{P}) = \frac{\mathbb{E}\left[\lambda \boldsymbol{X} e^{\lambda \boldsymbol{X}}\right] - \mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right] \log \mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right]}{\mathbb{E}\left[e^{\lambda \boldsymbol{X}}\right]},$$

and this may provide the intuition why we call the numerator entropy.

Definition 4.3. (Entropy) Given a random variable $\boldsymbol{X}$, we define its entropy

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

The marvelous point of entropy method is that we have a more natural tensorization, so we first look at it.

Lemma 4.4. ***(Tensorization of Entropy)*** *When $\boldsymbol{X_1}, ..., \boldsymbol{X_n}$ are independent, we have*

$$Ent\left(f(\boldsymbol{X_1}, ..., \boldsymbol{X_n})\right) \le \mathbb{E}\left[\sum_{k=1}^{n} Ent_k\left(f(\boldsymbol{X_1}, ..., \boldsymbol{X_n})\right)\right]$$

where $Ent_k\left(f(\boldsymbol{X_1}, ..., \boldsymbol{X_n})\right)$ only takes expectation w.r.t. $\boldsymbol{X_k}$.

To prove the lemma we need the variational form of entropy, which gives as the counterpart of Cauchy Schwarz inequality here.

Lemma 4.5. ***(Variational Form of Entropy)*** *Given a random variable $\boldsymbol{X}$, we have*

$$Ent(\boldsymbol{X}) = \sup_{\boldsymbol{Z}}\left\{\mathbb{E}[\boldsymbol{X}\cdot\boldsymbol{Z}] \mid \boldsymbol{Z} \textit{ is a random variable, } \mathbb{E}\left[e^{\boldsymbol{Z}}\right] = 1\right\}.$$

Proof.

$$Ent(\boldsymbol{X}) - \mathbb{E}[\boldsymbol{X}\cdot\boldsymbol{Z}] = \mathbb{E}[\boldsymbol{X}\log\boldsymbol{X}] - \mathbb{E}\left[\boldsymbol{X}\log e^{\boldsymbol{Z}}\right] - \mathbb{E}[\boldsymbol{X}]\log(\mathbb{E}[\boldsymbol{X}]),$$

we merge the first and second term on the right hand side together, which gives

$$Ent(\boldsymbol{X}) - \mathbb{E}[\boldsymbol{X}\cdot\boldsymbol{Z}] = \mathbb{E}\left[\boldsymbol{X}\log\left(\boldsymbol{X}e^{-\boldsymbol{Z}}\right)\right] - \mathbb{E}[\boldsymbol{X}]\log(\mathbb{E}[\boldsymbol{X}]).$$

From here, we can plug in an $e^{-\boldsymbol{Z}}$ in each term and change the measure to $d\mathbb{Q} = e^{\boldsymbol{Z}}d\mathbb{P}$, which gives

$$Ent(\boldsymbol{X}) - \mathbb{E}[\boldsymbol{X}\cdot\boldsymbol{Z}] = \mathbb{E}_{\mathbb{Q}}\left[\boldsymbol{X}e^{-\boldsymbol{Z}}\log\left(\boldsymbol{X}e^{-\boldsymbol{Z}}\right)\right] - \mathbb{E}_{\mathbb{Q}}\left[\boldsymbol{X}e^{-\boldsymbol{Z}}\right]\log\left(\mathbb{E}_{\mathbb{Q}}\left[\boldsymbol{X}e^{-\boldsymbol{Z}}\right]\right)$$

and the surprise here is that the left hand side is actually the entropy of $\boldsymbol{X}e^{-\boldsymbol{Z}}$ under the new measure. Due to the non-negativeness of entropy (or use Jensen's inequality), we get $Ent(\boldsymbol{X}) - \mathbb{E}[\boldsymbol{X} \cdot \boldsymbol{Z}] \geq 0$. Here we also need to specify when the equality can be reached. From the condition that entropy of a random variable hits zero, we let $\boldsymbol{Z} = \log(\boldsymbol{X}/\mathbb{E}\boldsymbol{X})$ and achieves the sup. Notice that here we also need to prove that $d\mathbb{Q}$ defined above is a measure. Given the condition that $\mathbb{E}\left[e^{\boldsymbol{Z}}\right] = 1$, we know this is correct.

Proof. Now we dive into the proof of tensorization of entropy. Since we aim to break the multivariate entropy into univariate ones, and we have $Ent(\boldsymbol{X}) = \mathbb{E}[\boldsymbol{X}(\log(\boldsymbol{X}) - \log(\mathbb{E}\boldsymbol{X}))]$ so we split $\log \boldsymbol{X}$ here. Let $\boldsymbol{Y} = f(\boldsymbol{X_1}, ..., \boldsymbol{X_n})$ and,

$$\log(\boldsymbol{Y}) - \log(\mathbb{E}\boldsymbol{Y}) = \sum_{k=1}^{n} \boldsymbol{U_k},$$

where $U_k = \log(\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X_1}, ..., \boldsymbol{X_k}]) - \log(\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{X_1}, ..., \boldsymbol{X}_{k-1}])$. Now we have $Ent(\boldsymbol{Y}) = \mathbb{E}[Y \sum_{k=1}^{n} \boldsymbol{U_k}]$, and further we hope to apply the variational form of entropy $Ent_k(\boldsymbol{Y}) \geq \mathbb{E}\left[\boldsymbol{Y}\boldsymbol{U_k} \mid \mathscr{F}^{-k}\right]$, where $\mathscr{F}^{-k}$ denotes the σ algebra generated by $\boldsymbol{X_1}, ..., \boldsymbol{X}_{k-1}, \boldsymbol{X}_{k+1}, ..., \mathbf{X_n}$. Here we need to check $E\left[e^{\boldsymbol{U_k}} \mid \mathscr{F}^{-k}\right] = 1$ to use the variational form. Since

$$\mathbb{E}\left[e^{\boldsymbol{U_k}} | \mathscr{F}^{-k}\right] = \frac{\mathbb{E}\left[\mathbb{E}\left[e^{\boldsymbol{U_k}} | \mathscr{F}^{k}\right] \mid \mathscr{F}^{-k}\right]}{\mathbb{E}\left[\mathbb{E}\left[e^{\boldsymbol{U_k}} | \mathscr{F}^{k-1}\right] \mid \mathscr{F}^{-k}\right]},$$

where $\mathscr{F}^{k}$ denotes the σ algebra generated by $\boldsymbol{X_1}, ..., \boldsymbol{X}_k$. Due to the independence of $\boldsymbol{X_k}$, we obtain

$$\mathbb{E}\left[e^{\boldsymbol{U_k}} | \mathscr{F}^{-k}\right] = \frac{\mathbb{E}\left[e^{\boldsymbol{U_k}} | \mathscr{F}^{k-1}\right]}{\mathbb{E}\left[e^{\boldsymbol{U_k}} | \mathscr{F}^{k-1}\right]} = 1.$$

Now we plug in the variational form of entropy and get

$$Ent(\boldsymbol{Y}) = \mathbb{E}\left[Y \sum_{k=1}^{n} \boldsymbol{U_k}\right] \leq \mathbb{E}\left[\sum_{k=1}^{n} Ent_k(\boldsymbol{Y})\right].$$

With such a beautiful tensorization of the entropy, we are now confident to deal with the univariate term, i.e. $Ent_k(\boldsymbol{Y})$.

Before we derive the bound rigorously, we do exploratory calculations. If we have an inequality in one dimension taking the form of $Ent(e^g) \leq \mathbb{E}\left[(Dg)^2 \mathbb{E}[e^g]\right]$, then by tensorization, we get $Ent(e^f) \leq \mathbb{E}\left[\|Df\|_2^2 \mathbb{E}\left[e^f\right)\right]$, which is exactly what we want. So now we introduce a special case of log-Soblev inequality, which is sufficient for our proofs here.

Lemma 4.6. ***(Discrete Log-Soblev)*** *Let D_g be the discrete gradient of function g of univariate random variable, we have*

$$Ent(e^g) \leq \mathbb{E}\left[Dg^2 e^g\right].$$

Proof. Using Jensen's inequality, we have $\log(\mathbb{E}[e^g]) \geq \mathbb{E}[g]$, and we can rewrite the entropy in the covariance form,

$$Ent(e^g) \le \mathbb{E}[g \cdot e^g] - \mathbb{E}[g]\mathbb{E}[e^g] = \mathrm{Cov}[g, e^g].$$

Recall that in proving the Hoeffding's lemma, we bounded the variance term by the upper and lower bound of the random variable. Here we do the same to g, but in a more complicated way.

$$\mathrm{Cov}[g, e^g] = \mathbb{E}[(g - \mathbb{E}[g])(e^g - \mathbb{E}[e^g])] \le \mathbb{E}\left[Dg \cdot (e^g - e^{\inf g})\right],$$

by the convexity of e^x, we have

$$\mathrm{Cov}[g, e^g] \le \mathbb{E}[Dg \cdot (g - \inf g) \cdot e^g] \le \mathbb{E}\left[Dg^2 \cdot e^g\right],$$

and we complete the proof. In fact, if we replace Dg by one side fluctuations, i.e. $(\sup g - \mathbb{E}[g])$ or $(\mathbb{E}[g] - \inf g)$ we can adopt stronger conclusions.

With the univariate case and tensorization in hand, we are ready to arrive at our final conclusion.

Theorem 4.7. *(Bounded difference Inequality) Let $\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}$ be independent random variables, then random variable $f(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})$ is $\| \sum_{k=1}^n (D_k f)^2 \|_\infty$ sub-Gaussian.*

Proof. We can simply connect the previous lemmas. Consider applying discrete log-Soblev lemma to Ent_k, we have

$$Ent_k(e^f) \le \mathbb{E}\left[(D_k f)^2 \cdot e^f \mid \mathscr{F}^{-k}\right].$$

By tensorization,

$$Ent(e^f) \le \mathbb{E}\left[\sum_{k=1}^n \mathbb{E}\left[(D_k f)^2 \cdot e^f \mid \mathscr{F}^{-k}\right]\right] \le \mathbb{E}\left[\left\|\sum_{k=1}^n (D_k f)^2\right\|_\infty \cdot e^f\right].$$

Further, replace f by λf, we have $D\lambda f = |\lambda| Df$, thus

$$Ent(e^{\lambda f}) \le \mathbb{E}\left[\sum_{k=1}^n \mathbb{E}\left[(D_k \lambda f)^2 \cdot e^{\lambda f} \mid \mathscr{F}^{-k}\right]\right] = \mathbb{E}\left[\lambda^2 \left\|\sum_{k=1}^n (D_k f)^2\right\|_\infty \cdot e^{\lambda f}\right],$$

which tells us that $f(\boldsymbol{X_1}, \ldots, \boldsymbol{X_n})$ is $\| \sum_{k=1}^n (D_k f)^2 \|_\infty$ sub-Gaussian.

Notice that discrete gradient can't handle unbounded regimes, especially when $\boldsymbol{X_i}$ follows a Gaussian distribution and f is not bounded as a multivariate function. This will be left to the next section talking of continuous gradient.

4.3 Continuous Gradient

4.3.1 Markov Process method

Theorem 4.8. ***(Gaussian)***

Definition 4.4. (Markov Semi-group)

Theorem 4.9. ***(Poisson)***

Chapter 5
Concentration Inequality II

5.1 Martingale Method

5.2 Markov Method

In the previous sections, we used discrete gradient to characterize ”sensitive” and utilized entropy method to construct the proof. We now aim to target continuous gradient which is essential for the case $\{\boldsymbol{X}_t\}$ following Gaussian distribution. Recall that the entropy method consists of two parts and the tensorization technique is universal. Therefore what we need is to produce an analogue of discrete log-Sobolev inequality considering continuous gradient and Gaussian $\{\boldsymbol{X}_t\}$. In this section, we introduce the machinery to derive general log-Sobolev inequality and apply the result to Gaussian variables. To achieve this, we will introduce Markov semi-group theory and Dirichlet form and build their connections to entropy. This method can produce the previous theorem and admit many extensions. The core of the proof is deeply related to physics and one of the most important application is the proof of Poincare’s conjecture proposed by Pelerman.

We will first state the goal theorem and then introduce background knowledge of Markov semigroup, generator and Dirichlet form. Next we will introduce general log-Sobolev inequality and finally apply it to the proof of main theorem. Since the theories are rather abstract, we provide concrete examples to assist understanding.

Notations: Let μ be a measure of space $\mathscr{S}$ and f be a function from $\mathscr{S}$ to $\mathbb{R}$. Let $\mu(f)$ stands for $\int_{\mathscr{S}} f d\mu$. We denote the entropy of function f with regard to measure μ by $Ent_\mu(f)$ and it is defined as $Ent_\mu(f) = \mu(f\log(f)) - \mu(f)\log(\mu(f))$. Let $\Phi(x)$ and $\phi(x)$ be the df and pdf of standard normal distribution.

Theorem 5.1 (Gaussian Concentration Inequality). *Let $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ be i.i.d. standard normal variables. Let f be any twice continuous differentiable multivariate function, we have $f(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$ is $\left\| \|\nabla f\|_2^2 \right\|_\infty$-subgaussian.*

Remark 5.1. Theorem 5.1 is stronger then the discrete gradient version in the case that f is not very curvy, and the range of f is rather large. One simple example is the widely used linear function.

A large proportion of the following part is to introduce the background knowledge for the construction of general log-Sobolev inequality. The main proof of Theorem 5.1 will not be hurt if the proofs for background knowledge are skipped.

Definition 5.1 (Stationary Distribution). Let $\{\boldsymbol{X}_t\}$ be a homogeneous Markov process on space Ω. We call a measure μ on Ω the stationary distribution of Markov process $\{\boldsymbol{X}_t\}$ if it satisfies for any univariate function f, any $t > 0$

$$\mu(\mathscr{P}_t f) = \mu(f). \tag{5.2.1}$$

Lemma 5.2 (Markov Semigroup). *Let $\{\boldsymbol{X}_t\}$ be a homogeneous Markov process on space Ω. We define operator $\mathscr{P}_t$ as for any function f, any $t > 0$*

$$\mathscr{P}_t f(x) = \mathbb{E}\left[f(\boldsymbol{X}_t)\middle|\boldsymbol{X}_0 = x\right]. \tag{5.2.2}$$

Then $\{\mathscr{P}_t\}$ forms a semigroup, i.e. we have $\mathscr{P}_{t+s} = \mathscr{P}_t \circ \mathscr{P}_s = \mathscr{P}_s \circ \mathscr{P}_t$.

Proof. By the homogenous Markov property we have for any t, $s > 0$

$$\mathbb{E}\left[f(\boldsymbol{X}_{t+s})\middle|\boldsymbol{X}_t = x\right] = \mathbb{E}\left[f(\boldsymbol{X}_s)\middle|\boldsymbol{X}_0 = x\right],$$

by pluging it in the definition of operator $\mathscr{P}_t$, we obtain

$$\begin{aligned}\mathscr{P}_s \circ \mathscr{P}_t f(x) &= \mathbb{E}\left[\mathbb{E}\left[f(\boldsymbol{X}_{t+s})\middle|\boldsymbol{X}_s\right]\middle|\boldsymbol{X}_0 = x\right] \\ &= \mathbb{E}\left[f(\boldsymbol{X}_{t+s})\middle|\boldsymbol{X}_0 = x\right] = \mathscr{P}_{t+s} f(x).\end{aligned}$$

Similarly we have $\mathscr{P}_{t+s} = \mathscr{P}_t \circ \mathscr{P}_s$.

Definition 5.2 (Generator). Let $\{\boldsymbol{X}_t\}$ be a homogeneous Markov process on space Ω, $\{\mathscr{P}_t\}$ be the Markov semigroup of $\{\boldsymbol{X}_t\}$. We define operator $\mathscr{L}$ as the generator of $\{\mathscr{P}_t\}$ by

$$\mathscr{L} f(x) = \lim_{t \to 0^+} \frac{\mathscr{P}_t f(x) - f(x)}{t} = \lim_{t \to 0^+} \frac{\mathscr{P}_t f(x) - \mathscr{P}_0 f(x)}{t}. \tag{5.2.3}$$

Remark 5.2 (Commutativity of $\mathscr{P}_t$ and $\mathscr{L}$). There are a number of good properties of $\mathscr{L}$, and here we only present the commutativity of $\mathscr{P}_t$ and $\mathscr{L}$. This fact is not used in proof and of independent interest. Consider the definition of $\mathscr{L}$ in Eq. (5.2.3), we have

$$\mathscr{L} \circ \mathscr{P}_t f(x) = \lim_{\delta \to 0} \frac{\mathscr{P}_\delta \circ \mathscr{P}_t f(x) - \mathscr{P}_t f(x)}{\delta} = \lim_{\delta \to 0} \frac{\mathscr{P}_t \circ \mathscr{P}_\delta f(x) - \mathscr{P}_t \circ f(x)}{\delta},$$

we swap the limit and $\mathscr{P}_t$ and get

$$\mathscr{L}\circ\mathscr{P}_t f(x)=\mathscr{P}_t\lim_{\delta\to 0}\frac{\mathscr{P}_\delta f(x)-f(x)}{\delta}=\mathscr{P}_t\circ\mathscr{L}f(x).$$

By considering the relationship of Markov semigroup, generator to PDE, we can give an intuitive reason how $\mathscr{L}$ gets its name. By definition of $\mathscr{P}_t$ and $\mathscr{L}$ in Eq. (5.2.2) and Eq. (5.2.3), we get

$$\begin{cases}\frac{d}{dt}\mathscr{P}_t f=\mathscr{L}\circ\mathscr{P}_t f & x\in\mathscr{X}, t\in\mathbb{R}^+\\ \mathscr{P}_0 f=f & t\in\mathbb{R}^+\end{cases}$$

If we let $u(t,x)=\mathscr{P}_t f(x)$, we can rewrite the above equations as a Cauchy problem of PDE

$$\begin{cases}\frac{\partial}{\partial t}u(t,x)=\mathscr{L}u(t,x)\\ u(0,x)=f(x)\end{cases}$$

So if we start form a homogeneous Markov process $\{\boldsymbol{X}_t\}$ and get $\mathscr{P}_t$, $\mathscr{L}$, we can construct corresponding PDE and $u(t,x)=\mathscr{P}_t$ is exactly the solution. In fact, we can view the procedures inversely. Start with $\mathscr{L}$ and we obtain the PDE problem and solve it to get $\{\mathscr{P}_t\}$. Under some conditions we can construct a homogeneous Markov process from $\{\mathscr{P}_t\}$. This illustrates why we call $\mathscr{L}$ the generator. This relationship can be seen concretely in the Example 5.1.

Definition 5.3 (Dirichlet Form). Let $\{\boldsymbol{X}_t\}$ be a homogeneous Markov process on space Ω with semigroup $\{\mathscr{P}_t\}$, generator $\{\boldsymbol{X}_t\}$ and stationary distribution μ. Let f, g be two arbitrary univariate functions, we define Dirichlet form by

$$\mathscr{E}(f,g)=-\langle f,\mathscr{L}g\rangle_\mu=-\int_{\mathbb{R}}f\cdot\mathscr{L}g d\mu. \tag{5.2.4}$$

The Dirichlet form actually represents a quantity similar to the derivative of entropy with regard to time, or energy. Observe that

$$\begin{aligned}\frac{d}{dt}Ent_\mu(f)&=\frac{d}{dt}\mu\left(\mathscr{P}_t f\cdot\log\left(\mathscr{P}_t f\right)\right)+\frac{d}{dt}\big(\mu\left(\mathscr{P}_t f\right)\cdot\log\left(\mu\left(\mathscr{P}_t f\right)\right)\big)\\ &=\mu\left(\log\left(\mathscr{P}_t f\right)\frac{d}{dt}\mathscr{P}_t f\right)+\mu\left(\mathscr{P}_t f\frac{\frac{d}{dt}\mathscr{P}_t f}{\mathscr{P}_t f}\right)+\frac{d}{dt}\big(\mu\left(\mathscr{P}_t f\right)\log\left(\mu\left(\mathscr{P}_t f\right)\right)\big).\end{aligned}$$

Recall Eq. (5.2.1) and swap differentiation and integration

$$\frac{d}{dt}\big(\mu\left(\mathscr{P}_t f\right)\log\left(\mu\left(\mathscr{P}_t f\right)\right)\big)=0,\quad \mu\left(\mathscr{P}_t f\frac{\frac{d}{dt}\mathscr{P}_t f}{\mathscr{P}_t f}\right)=\frac{d}{dt}\mu(\mathscr{P}_t f)=0,$$

plug in the above terms and we get the desired relationship

$$\frac{d}{dt}Ent_\mu(f)=\mu\left(\log\left(\mathscr{P}_t f\right)\cdot\mathscr{L}f\right)=-\mathscr{E}\big(\log(\mathscr{P}_t f),\mathscr{P}_t f\big). \tag{5.2.5}$$

The negative sign here is because our definition of *Ent* is the negative of that in physics or information theory. This can give a profound explanation of Theorem 5.3 from the angle of physics.

Above is the preknowledge of log-Sobolev inequality. But before we delve into Theorem 5.3, we first present Brownian motion as an example to help understand previous definitions and facts.

Example 5.1 (Brownian Motion). Let $X_t = X_0 + W_t$ be a Brownian motion. We consider stationary distribution μ, Markov semigroup $\mathscr{P}_t$, generator $\mathscr{L}$, corresponding PDE and solution and Dirichlet form $\mathscr{E}(\log(f), f)$.

- (Stationary Distribution).Let μ be the Lesbegue measure on $\mathbb{R}$, for any compactly supported function f, we have for any $t > 0$

$$\mu(\mathscr{P}_t f) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(x + \sqrt{t}y)\phi(y)dydx$$

  by Fubini's law we swap two integration and get

$$\mu(\mathscr{P}_t f) = \int_{\mathbb{R}} \phi(y) \int_{\mathbb{R}} f(x + \sqrt{t}y)dxdy = \int_{\mathbb{R}} \phi(y)\mu(f)dy = \mu(f).$$

  By the definition of stationary distribution Eq. (5.2.1), we claim that Lebesgue measure on $\mathbb{R}$ is the stationary distribution for Brownian motion. Since as $t \to \infty$, we have the variance of $\boldsymbol{X}_t \to \infty$, so we have the pdf of $\boldsymbol{X}_t$ be approximately a flat line when t is sufficiently large, which also explains why the stationary distribution is Lebesgue measure.
- (Markov Semigroup). By definition Eq. (5.2.2), we directly calculate

$$\mathscr{P}_t f(x) = \mathbb{E}\left[f(X_t)\middle|X_0 = x\right] = \int_{\mathbb{R}} f(x + \sqrt{t}y) \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} dy$$

  we do parameterization by letting $z = x + \sqrt{t}y$, then

$$\mathscr{P}_t f(x) = \int_{\mathbb{R}} f(z) \cdot \frac{1}{\sqrt{2\pi t}} \exp\left\{-\frac{(x-z)^2}{2t}\right\} dz = \mathscr{K}_t * f(x),$$

  and we get the operator $\mathscr{P}_t$ is the convolution.
- (Generator). By definition Eq. (5.2.3), let $\xi \sim \mathrm{N}(0,1)$ and we obtain

$$\frac{d}{dt}\mathscr{P}_t f = \frac{d}{dt}\mathbb{E}\left[f(x + \sqrt{t}\xi)\right] = \mathbb{E}\left[\frac{\xi}{2\sqrt{t}} f'(x + \sqrt{t}\xi)\right],$$

  apply Stein's identity to function $y \cdot f'(\sqrt{t}y)$

$$\frac{d}{dt}\mathscr{P}_t f = \mathbb{E}\left[\frac{1}{2} f''(x + \sqrt{t}\xi)\right] = \mathscr{P}_t \frac{\Delta f}{2},$$

  then we get the generator $\mathscr{L} = \Delta/2$.

- (PDE). The PDE generated by $\mathscr{L}$ above is

$$\begin{cases} \frac{\partial}{\partial t}u(t,x) - \frac{1}{2}\Delta u(t,x) = 0 & x \in \mathbb{R},\ t \in \mathbb{R}^+ \\ u(0,x) = f(x) & x \in \mathbb{R} \end{cases}$$

and $u(t,x) = \mathscr{K}_t * f(x)$ is the solution. On the other side, we can directly solve the heat equation through Fourier transformation and obtain the closed form solution which is exactly the same. This can also be interpreted from physics perspective, since in thermodynamics we consider heat from the macroscopic angle and the counterpart in microscopic scale is Brownian motion.
- (Dirichlet Form). By definition Eq. (5.2.4) and use Green's formula

$$\mathscr{E}(f,g) = -\langle f, \frac{1}{2}\Delta g\rangle \mu = \frac{1}{2}\langle \nabla f, \nabla g\rangle_\mu.$$

By setting $g = f$ we get $\mathscr{E}(f,f) = \mathbb{E}_\mu\left[\|\nabla f\|_2^2\right]/2$, which is similar to the form of kinetic energy $mv^2/2$. This helps us to understand the relationship between entropy and Dirichlet form.

Theorem 5.3 (Log-Sobolev Inequality). *Let $\{\boldsymbol{X}_n\}$ be a homogeneous Markov process on space Ω with semigroup $\{\mathscr{P}_t\}$, generator $\{\boldsymbol{X}_n\}$ and stationary distribution μ. For any constant $C > 0$, we have (1) and (2) are equivalent*

1. *(Log-Sobolev Inequality). $Ent_\mu(f) \leq C\mathscr{E}(\log(f), f)$ for any $f \in \mathscr{C}^{(2)}$;*
2. *(Entropic Exponential Ergodicity). $Ent_\mu(\mathscr{P}_t f) \leq e^{-t/C}Ent_\mu(f)$ for any $f \in \mathscr{C}^{(2)}$, $t > 0$;*

If we further assume $Ent_\mu(\mathscr{P}_t f) \to 0$ as $t \to \infty$ (Entropic Ergodicity) then

3. *$\mathscr{E}(\log(\mathscr{P}_t f), \mathscr{P}_t f) \leq e^{-t/C}\mathscr{E}(\log(f), f)$ for any $f \in \mathscr{C}^{(2)}$, $t > 0$ implies (1) and (2).*

Proof. The main tool in proof is calculus and ODE.

- (1) ⇒ (2). Recall Eq. (5.2.5) and assume (1) we have

$$\frac{d}{dt}Ent_\mu(\mathscr{P}_t f) = -\mathscr{E}(\log(\mathscr{P}_t f), \mathscr{P}_t f) \leq -\frac{1}{C}Ent_\mu(\mathscr{P}_t f),$$

this is the standard form of Gronwall's inequality, so we get $Ent_\mu(\mathscr{P}_t f) \leq e^{-t/C}Ent_\mu(f)$.
- (2) ⇒ (1). Again we use Eq. (5.2.5) and assume (2) we obtain

$$\begin{aligned} -\mathscr{E}(\log(f), f) &= \lim_{t\to 0^+} \frac{Ent_\mu(\mathscr{P}_t f) - Ent_\mu(f)}{t} \\ &\leq \lim_{t\to 0^+} \frac{\left(e^{-\frac{t}{C}} - 1\right)Ent_\mu(f)}{t} = -\frac{1}{C}Ent_\mu(f). \end{aligned}$$

- (3) $\Rightarrow$ (1). With assumption $Ent_{\mu}(\mathscr{P}_t f) \to 0$ as $t \to \infty$, by Newton-Leibniz formula

$$Ent_{\mu}(f) = Ent_{\mu}(f) - Ent_{\mu}(\mathscr{P}_{\infty} f) = -\int_0^{\infty} \frac{d}{dt} Ent_{\mu}(\mathscr{P}_t f) dt,$$

use Eq. (5.2.5) and assume (3), we have

$$Ent_{\mu}(f) = \int_0^{\infty} \mathscr{E}(\log(\mathscr{P}_T f), \mathscr{P}_t f) dt \leq \int_0^{\infty} e^{-\frac{t}{C}} \mathscr{E}(\log(f), f) dt = \frac{1}{C} \mathscr{E}(\log(f), f).$$

Remark 5.3 (Discrete Log-Soblev Inequality). By constructing a Poisson process, we can directly get the "Discrete Log-Soblev Inequality".

In order to obtain the desired form in Theorem 5.1, we need a Markov Process that satisfies:

1. Stationary distribution is standard norm;
2. $\mathscr{L}$ contains derivative operator and $\mathscr{E}(\log(f), f)$ has the form $\mu\left((\nabla f)^2, f\right)$
3. Satisfies (2) or the assumption and (3) in Theorem 5.3.

We will use Ornstein–Uhlenbeck process which serves to our needs. Recall that O-U process is defined as $X_t = e^{-t} X_0 + e^{-t} W_{e^{2t}-1}$, homogeneous Markov with standard normal as stationary measure. We are going to compute the Markov semigroup $\mathscr{P}_t$, generator $\mathscr{L}$ and Dirichlet form $\mathscr{E}(\log(f), f)$.

- (Markov Semigroup). By definition Eq. (5.2.2), let $\xi \sim \mathrm{N}(0,1)$ we have

$$\mathscr{P}_t f(x) = \mathbb{E}_{\mu}\left[f(e^{-t}x + \sqrt{1-e^{-2t}}\xi)\right]$$

- (Generator). By definition Eq. (5.2.3), through calculus we obtain

$$\frac{d}{dt}\mathscr{P}_t f = \mathbb{E}_{\mu}\left[f^{`}\left(e^{-t}x + \sqrt{1-e^{-2t}}\xi\right)\left(-e^{-t}x + \frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\xi\right)\right],$$

apply Stein's identity to function $f^{`}(e^{-t}x + \sqrt{1-e^{-2t}}y) \cdot e^{-2t}/\sqrt{1-e^{-2t}}$

$$\frac{d}{dt}\mathscr{P}_t f = \mathbb{E}_{\mu}\left[-e^{-t}xf^{`}\left(e^{-t}x + \sqrt{1-e^{-2t}}y\right) + e^{-2t}f^{``}\left(e^{-t}x + \sqrt{1-e^{-2t}}y\right)\right] \tag{5.2.6}$$

$$= \left(-x\frac{d}{dx} + \frac{d^2}{dx^2}\right)\mathscr{P}_t f = \mathscr{L} \circ \mathscr{P}_t f, \tag{5.2.7}$$

then we have the generator $\mathscr{L} = \Delta - \langle x, \nabla \rangle$, called Stein's operator. In fact, we only need univariate version here. Notice that we have differential operator in $\mathscr{L}$ which is likely to produce Dirichlet Form containing gradient;
- (Dirichlet Form). Let μ be the measure of standard normal, by definition Eq. (5.2.4) and the generator just obtained

$$\mathscr{E}(f,g) = \mathbb{E}_\mu\left[f(\xi)\left(\xi g'(\xi) - g''(\xi)\right)\right],$$

apply Stein's identity to $f(y)g^{\text{`}}(y)$ and we get

$$\mathscr{E}(f,g) = \mathbb{E}_\mu\left[f(\xi)'g'(\xi) + f(\xi)g''(\xi) - f(\xi)g''(\xi)\right] = \mu\left(\nabla f, \nabla g\right).$$

Notice that for $\mathscr{E}(\log(f), f)$, we have

$$\mathscr{E}(\log(f), f) = \mu\left(\frac{\|\nabla f\|_2^2}{f}\right). \tag{5.2.8}$$

Use the concrete form of O-U process and notice that

$$\frac{d}{dx}\mathscr{P}_t f(x) = \mathbb{E}_\mu\left[e^{-t} f^{\text{`}}\left(e^{-t}x + \sqrt{1-e^{-2t}}\xi\right)\right] = e^{-t}\mathscr{P}_t f^{\text{`}}, \tag{5.2.9}$$

then we obtain $\mathscr{E}(\log(f), f) = \mu\left(e^{-2t}(\mathscr{P}_t f)^2/\mathscr{P}_t f\right)$. Next we use Cauchy Schwarz Inequality

$$\left(\mathscr{P}_t f^{\text{`}}\right)^2 = \left(\mathscr{P}_t \frac{f^{\text{`}}\sqrt{f}}{\sqrt{f}}\right)^2 \leq \mathscr{P}_t \frac{(f^{\text{`}})^2}{f} \cdot \mathscr{P}_t f, \tag{5.2.10}$$

Plug Eq. (5.2.9) and Eq. (5.2.10) in Eq. 5.2.8, we get

$$\mathscr{E}(\log(f), f) \leq e^{-2t}\mu\left(\frac{(\mathscr{P}_t f^{\text{`}})^2}{f}\right). \tag{5.2.11}$$

In order to apply Theorem 5.3, notice that Eq. (5.2.11) gives part of the entropic ergodicity. So what remains to be proved is the assumption of $Ent_\mu(\mathscr{P}_t f) \to 0$ as $t \to \infty$. To prove this, we first need to derive $Ent_\mu(\mathscr{P}_t f)$.

$$Ent_\mu(\mathscr{P}_t f) = \mu(\mathscr{P}_t f \log(\mathscr{P}_t f)) - \mu(f)\log(\mu(f)) \Rightarrow \mu(f)\log(\mu(f)) - \mu(f)\log(\mu(f)) = 0$$

Through above analysis, we have proved the conditions of Gaussian log-Sobolev inequality.

Remark 5.4 (Discrete log-Sobolev Inequality). We can obtain discrete log-Sobolev inequality as a special case of Theorem 5.3. Let $\{X_n\}$ be i.i.d. random variables following distribution μ, $\{P_t\}$ be a Poisson process with parameter $\lambda = 1$, we construct homogeneous Markov process $Y_t = \sum_{i=1}^{P_t} X_i$. By computing the entropy and Dirichlet form of $\{Y_n\}$ and using $(1) \Rightarrow (2)$ in Theorem 5.3 we get the discrete log-Sobolev inequality.

Theorem 5.4 (Gaussian Log-Soblev Inequality). *Let μ denote the distribution of standard normal and $f \in \mathscr{C}^{(2)}$, then we have*

$$Ent_\mu(f) \leq \frac{1}{2}\mathbb{E}_\mu\left[\frac{(f')^2}{f}\right]. \tag{5.2.12}$$

Corollary 5.1. *By putting different forms of f, we can get different inequalities.*

- $Ent_\mu e^f \leq \frac{1}{2}\mathbb{E}_\mu\left[(f`)^2 \cdot e^f\right]$.
- $Ent_\mu f^2 \leq 2\mathbb{E}_\mu\left[(f`)^2\right]$.

By Eq. we need $Ent_\mu(e^f)$for Theorem 5.1.

Proof. We divide the proof in two steps and directly plug in the results derived before. Since

$$Ent_\mu(e^f) \leq \frac{1}{2}\mathbb{E}_\mu\left[(f')^2 \cdot e^f\right].$$

and use tensorization of entropy gives us

$$Ent\left(e^{\lambda f(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)}\right) \leq \mathbb{E}\left[\sum_{i=1}^n Ent_i e^{\lambda f}\right] \leq \frac{1}{2}\mathbb{E}\left[\lambda^2\|\nabla f\|_2^2 \cdot e^f\right],$$

which immediately gives us $f(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$ is $\|\nabla f\|_2^2/2$-subgaussian.

5.3 Transportation Method

Recall that in the previous sections we used different gradients to measure the "sensitiveness" of f to x_i. In this section, we propose a new way to characterize "sensitiveness" by using metric and Lipschitz function. Similar to previous techniques, the method is comprised of two parts: the inequality for univariate case and the how to connect the multivairate regime. This general technique gives birth to a variety of inequalities by plugging in different metrics and leads to results which can't be deduced from martingale or entropy method, e.g. the widely used Talagrand concentration inequality.

We will first introduce the preknowledge and then state the Bobkov-Götze theorem, which is an analogue to log-Sobolev inequaltiy. Next we delve to transportation and derive Marton's theorem for tensorization. Finally we will combine two ingredients to prove Talagrand concentration inequality.

Definition 5.4 (Lipschitz Function). Let $(\mathscr{S}, \mathsf{d})$ be a metric space. We call a function $f : \mathscr{S} \to \mathbb{R}$ to be L-Lipschitz w.r.t. metric d if for any $x, y \in \mathscr{S}$, $|f(x), f(y)| \leq L \cdot \mathsf{d}(x,y)$. We denote the set of 1-Lipschitz functions as $Lip(\mathscr{S}, \mathsf{d})$.

Now we have three ways to quatify sensitiveness: discrete gradient, continuous gradient and Lipschitz, which requires the function to be bounded, the first order of derivative to be bounded and nearly linear. The Lipschitz properties are equivalent to previous two considering different metrics. We propose Example 5.2 and Example 5.3. From this point of view, we can see the Lipschitz constraint is more general. However, notice that both three quantities will blow up if we consider f to be quadratic, which unables us to produce any meaningful subgaussian estimation. This is reasonable since if we consider the simple case of i.i.d. standard normal $\boldsymbol{X}_i$ and let $f = \sum_{i=1}^n x_i^2$, then $f(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$ follows χ_2^n distribution and not subgaussian.

Example 5.2 (McDiarmid Inequality). Let $f : \mathbb{R}^d \to \mathbb{R}$, then $\|D_i f\|_\infty \le c_i < \infty, i = 1,\dots,d$ if and only if $f \in Lip((\mathbb{R}^d), \mathsf{d}_c)$ *w.r.t.* d_c, where d_c denotes the weighted hamming distance $\mathsf{d}_c(\mathbf{x},\mathbf{y}) = \sum_{i=1}^n c_i \mathbf{1}_{\{x_i \neq y_i\}}$. This can be proved by considering triangle inequality

$$|f(\mathbf{x}) - f(\mathbf{y})| \le \sum_{i=1}^n |f(x_1,\dots,x_i,y_{i+1},\dots,y_n) - f(x_1,\dots,x_{i-1},y_i,\dots,y_n)| \le \sum_{i=1}^n c_i \mathbf{1}_{\{x_i \neq y_i\}}.$$

The other side can be derived by considering $\mathbf{x}$ and $\mathbf{y}$ with only one different coordinate. We just found the counterpart of discrete gradient in the form of Lipschitz constraint and we are going to state the McDiarmid inequality in the new language. For random variables $\boldsymbol{X}_1,\dots,\boldsymbol{X}_d$ taking values in $\mathbb{R}$, $f \in Lip(\mathbb{R}^d, \mathsf{d}_c)$, then we have by McDiarmid inequality that $f(\boldsymbol{X}_1,\dots,\boldsymbol{X}_d)$ is $\sum_{i=1}^d c_i^2/4$-subgaussian.

Example 5.3 (Gaussian Concentration Inequality). If we take $(\mathbb{R}^d, \mathsf{d}_E)$ with d_E the Euclidean distance, $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable to the first order, then $\left\| \|\Delta f\|_2^2 \right\|_\infty \le L^2$ if and only if f is L-Lipschitz. Having obtained the continuous gradient considering Lipschitz properties, we state that according to Gaussian concentration inequality, for random variables $\boldsymbol{X}_1,\dots,\boldsymbol{X}_d \sim$ i.i.d. $\mathrm{N}(0,1)$, $f \in Lip(\mathbb{R}^d, \mathsf{d}_E)$, then $f(\boldsymbol{X}_1,\dots,\boldsymbol{X}_d)$ is $1/4$-subgaussian.

Considering the similar forms of inequalities in Example 5.2 and Example 5.3, we are inspired to propose the following question:*For which distribution μ on the metric space $(\mathscr{S}, \mathsf{d})$ do we have for any $f \in Lip(\mathscr{S}, \mathsf{d})$ we have $f(\boldsymbol{X})$ is σ^2-subgaussian, where random vector $\boldsymbol{X} \sim \mu$.*
We answer this question by Theorem 5.5. Before we formally introduce the theorem, we need to define two ways to measure how far two probability measures are.

Definition 5.5 (Wasserstein Distance). For metric space $(\mathscr{S}, \mathsf{d})$, two probability measures μ, ν that satisfies $\int d(x,\cdot)\rho(dx) < \infty$, $\rho = \mu, \nu$, define the Wasserstein distance of μ, ν by

$$W_1(\mu,\nu) = \sup_{f \in Lip(\mathscr{S}, \mathsf{d})} \left| \int_{\mathscr{S}} f d\mu - \int_{\mathscr{S}} f d\nu \right|.$$

Definition 5.6 (KL Divergence). For metric space $(\mathscr{S}, \mathsf{d})$ and two probability measures μ, ν, define the KL divergence of μ, ν by

$$D(\nu \| \mu) = \begin{cases} \int \log(d\nu/d\mu) d\nu = Ent_\mu(d\nu/d\mu) & \nu \ll \mu \\ \infty & \text{otherwise} \end{cases}$$

Theorem 5.5 (Bobkov-Götze Theorem). *For metric space $(\mathscr{S}, \mathsf{d})$ and probability measure μ, random vectors $\boldsymbol{X} \sim \mu$, then the following statements are equivalent*

1. *$f(\boldsymbol{X})$ is σ^2-subgaussian for any $f \in Lip(\mathscr{S}, \mathsf{d})$;*
2. *For any probability measure ν, $W_1(\nu,\mu) \le \sqrt{2\sigma^2 D(\nu\|\mu)}$.*

Since Wasserstein distance and KL divergence are not comparable, (2) in theorem 5.5 characterizes a special property of probability measure μ w.r.t. the metric d. Unfortunately (2) is not easy to prove generally since we need to justify the inequality for any probability measure ν.

Before we enter the proof, we first provide an example to show the significance of Theorem 5.5.

Example 5.4 (Pinsker's Inequality). For metric space $(\mathscr{S},\mathsf{d})$ with $\mathsf{d}=\mathbf{1}_{\{x\neq y\}}$, by Hoeffding's inequality we get for any $f\in Lip(\mathscr{S},\mathsf{d})$, probability measure μ, random variable/vector $\boldsymbol{X}\sim\mu$, $f(\boldsymbol{X})$ is $1/4$-subgaussian. Then apply $(1)\to(2)$ in Theorem 5.5 we get for any probability measure ν

$$W_1(\nu,\mu)=\sup_{0\leq f\leq 1}\left|\int f d\nu-\int f d\mu\right|=\|\nu-\mu\|_{TV}\leq\sqrt{D(\nu\|\mu)/2}.$$

This gives us the important relationship between KL divergence and total variance. In fact, we can do inversely by proving Pinsker's inequality independently and apply $(2)\to(1)$ in Theorem 5.5 to obtain Hoeffding's inequality.

To prove Theorem 5.5, we need the following variational formula.

Lemma 5.6 (Gibbs Variational Principle). *For a probability measure μ and function f, we have*

$$\log\left(\mathbb{E}_\mu\left[e^f\right]\right)=\sup_\nu\{\mathbb{E}_\nu[f]-D(\nu\|\mu)\}\tag{5.3.1}$$

Proof. Notice that we have

$$\mathbb{E}_\nu[f]-\log\left(\mathbb{E}_\mu\left[e^f\right]\right)=\int\log\left(\frac{e^f}{\mathbb{E}_\mu\left[e^f\right]}\right)d\nu,$$

then we can define a new measure $d\widehat{\mu}=(e^f d\mu)/(\mathbb{E}_\mu\left[e^f\right])$,

$$\mathbb{E}_\nu[f]-\log\left(\mathbb{E}_\mu\left[e^f\right]\right)-D(\nu\|\mu)=\int\log\frac{d\widehat{\mu}}{d\mu}d\nu-\int\log\frac{d\nu}{d\mu}d\nu=-D(\nu\|\widehat{\mu})\leq 0,$$

and the equality can be achieved when $\nu=\widehat{\mu}$

Recall that we have encountered another variational formula in entropy method, which can derive Lemma 5.6. In fact, the following three variational principles all imply each other:

1. $\log\left(\mathbb{E}_\mu\left[e^f\right]\right)=\sup_\nu\{\mathbb{E}_\nu[f]-D(\nu\|\mu)\}$
2. $Ent_\mu(d\nu/d\mu)=\sup_f\left\{\mathbb{E}_\nu[f]\,\middle|\,\mathbb{E}_\mu\left[e^f\right]=1\right\}$
3. $D(\nu\|\mu)=\sup_f\left\{\mathbb{E}_\nu[f]-\log\left(\mathbb{E}_\mu\left[e^f\right]\right)\right\}$

Proof. Now we prove Theorem 5.5. By definition of subgaussian, we have (2) in Theorem 5.5 is equivalent to for all λ, $f\in Lip(\mathscr{S},\mathsf{d})$,

$$\log\left(\mathbb{E}_{\mu}\left[e^{\lambda(f-\mathbb{E}_{\mu}[f])}\right]\right) \leq \frac{\lambda^2\sigma^2}{2}.$$

Recall lemma 5.6 and we have the equivalent form

$$\sup_{\lambda\in\mathbb{R}}\sup_{f\in Lip(\mathscr{S},\mathsf{d})}\sup_{\nu}\left\{\lambda\left(\mathbb{E}_{\nu}f-\mathbb{E}_{\mu}f\right)-D(\nu\|\mu)-\frac{\lambda^2\sigma^2}{2}\right\}\leq 0.$$

Notice that changing the sequence of taking suprema over f and ν brings Wasserstein distance,

$$\sup_{\lambda\in\mathbb{R}}\sup_{\nu}\left\{\lambda W_1(\nu,\mu)-D(\nu\|\mu)-\frac{\lambda^2\sigma^2}{2}\right\}\leq 0.$$

Observe that left hand side is a quadratic form of λ, so we can easily optimize over λ and get

$$\sup_{\nu}\left\{\frac{W_1(\nu,\mu)}{2\sigma^2}-D(\nu\|\mu)\right\}\leq 0.$$

this is exactly (2) and we have completed the proof.

Since we have obtained the univariate case Theorem5.5 and in order to consider multivariate case we need tensorization of Wasserstein distance or KL divergence. We introduce optimal transportation to help us evaluate the problem, which gives us an equivalent form of Wasserstein distance. Combining the the equivalent form and chain rule of KL divergence, we can prove the tensorization. This is inspiring for the proof of Theorem 5.10, which is powerful in many practical cases.

Definition 5.7 (Coupling). For two probability measures μ, ν on Ω, let $\mathscr{C}(\mu,\nu) = \{\text{joint distribution: } \Omega\times\Omega \text{ with marginal distribution } \mu,\nu\}$. Then any $M\in\mathscr{C}(\mu,\nu)$ is called a coupling of μ and ν and let $\boldsymbol{M}$ denote the corresponding measure.

In this way, when considering Wasserstein distance between two probability measures we can utilize the Lipschitz condition of f and get for any $M\in\mathscr{C}(\mu,\nu)$, $f\in Lips(\mathscr{S},\mathsf{d})$:

$$\int f d\mu - \int f d\nu = \int f(\boldsymbol{X})-f(\boldsymbol{Y})\, d\boldsymbol{M} \leq \inf_{M\in\mathscr{C}(\mu,\nu)}\int d(\boldsymbol{X},\boldsymbol{Y})\, d\boldsymbol{M}. \qquad (5.3.2)$$

In fact, we can prove the following duality theorem showing that the above inequality is actually an equation, which gives an equivalent form of Wasserstein distance. To prove it, we are considering the *optimal transportation problem*. Given a pile of bricks and a pit to fill, μ and ν denote the density of bricks and depth of pit respectively and $d(x,y)$ represent the cost of moving a unit of brick x to the place y. In order to minimize the total cost, we are actually minimizing $\int d(x,y)d\boldsymbol{M}$ with $M\in\mathscr{C}(\mu,\nu)$. This problem is surprisingly connected to linear programing.

Theorem 5.7 (Monge-Kantorovich Duality). *Let $(\mathscr{S},\mathsf{d})$ be a separable metric place. μ, ν be two probability measures, then*

$$W_1(\mu,\nu) = \sup_{f\in Lips(\mathscr{S},\mathrm{d})} \left|\mathbb{E}_\mu[f] - \mathbb{E}_\nu[f]\right| = \inf_{M\in\mathscr{C}(\mu,\nu)} \mathbb{E}\left[d(\boldsymbol{X},\boldsymbol{Y})\right].$$

Proof. We will focus on the discrete setting and then introduce how to generalize to continuous measure. The core to prove Theorem 5.7 is the strong duality theorem in linear programming, thus we restate the problem in the form of linear programming.

Let μ, ν be probability measures of finite space $\mathscr{S} = (1,2,\ldots,p)$, then the optimal transportation problem can be written as

$$\min_M \sum_{i,j=1}^{p} d(i,j)M(i,j),$$

$$s.t. \begin{cases} \sum_{i=1}^{p} M(i,j) = \nu(j), \\ \sum_{j=1}^{p} M(i,j) = \mu(i), \\ M(i,j) \geq 0. \end{cases}$$

By the law to write the dual problem, we have the following dual form

$$\max_{p,q} \sum_{i}^{p} \mu(i)p(i) + \nu(j)q(j),$$

$$s.t. \begin{cases} p(i) + q(j) \leq d(i,j), \\ p(i), q(j) \text{ free.} \end{cases}$$

since we know that $M(i,j) \in [0,1]$ and $\sum_{i,j=1}^{p} d(i,j)M(i,j)$ is bounded and the primal problem is feasible. So by strong duality theorem in linear programming, we have $\inf_{M\in\mathscr{C}(\mu,\nu)} \mathbb{E}\left[d(\boldsymbol{X},\boldsymbol{Y})\right]$ equals the solution of the dual problem. Now it suffices to show that the solution to the dual problem is exactly the Wasserstein distance. However, this is not very straight and we need to utilize the fact that d is a metric. Recall that for any f, g that satisfies the constraints in dual problem, we have for all x

$$f(x) \leq \widetilde{f}(x) := \inf_z \{d(x,z) - g(z)\} \leq -g(x).$$

So the objective function of dual problem satisfies

$$\mathbb{E}_\mu[f] + \mathbb{E}_\nu[g] \leq \mathbb{E}_\mu[\widetilde{f}] - \mathbb{E}_\nu[\widetilde{f}].$$

Now we prove that $\widetilde{f}$ is Lipschitz. In fact, for any $\varepsilon > 0$, let z_ε satisfies $d(y,z_\varepsilon) - g(z_\varepsilon) \leq \inf_z\{d(y,z) - g(z)\} + \varepsilon$, then

$$\widetilde{f}(x) - \widetilde{f}(y) \leq d(x,z_\varepsilon) - d(y,z_\varepsilon) + \varepsilon \leq d(x,y) + \varepsilon,$$

since ε is arbitrary, we have $\widetilde{f} \in Lips(\mathscr{S},\mathrm{d})$. Therefore, the maximal value of dual problem is no bigger than $W_1(\mu,\nu)$, and we know that there exists a $M \in \mathscr{C}(\mu,\nu)$ satisfying $\mathbb{E}_{\boldsymbol{M}}[d(\boldsymbol{X},\boldsymbol{Y})] \leq W_1(\mu,\nu)$. Recall Eq. (5.3.2), we have $W_1(\mu,\nu) = \inf_{M\in\mathscr{C}(\mu,\nu)} \mathbb{E}_{\boldsymbol{M}}[d(\boldsymbol{X},\boldsymbol{Y})]$. As for continuous measure, we use discrete distance to approximate and evaluate the limit.

Notice that we assumed the separability of metric space $\mathscr{S}$ (where to use?), but when we consider a space $\mathscr{S}$ endowed with trivial distance $\mathbf{1}_{x\neq y}$ is not separable when $\mathscr{S}$ is not countable. However, similar conclusion in Theorem 5.7 still holds and we show it as an example.

Example 5.5 (Total Variation). Let $\mathrm{d} = \mathbf{1}_{x\neq y}$ be the trivial metric, by previous example we know that $W_1(\mu,\nu) = \|\mu-\nu\|_{TV}$. So the Monte Kantorovich duality takes the form

$$\|\mu-\nu\|_{TV} = \inf_{M\in\mathscr{C}(\mu,\nu)} \boldsymbol{M}(\boldsymbol{X}\neq\boldsymbol{Y}).$$

To obtain the conclusion, we only need to show that the equality holds. Unlike the previous proof in Theorem 5.7, we construct a joint measure that attains equality. This optimal coupling is also used in the proof of the second Marton Theorem. For simplicity, we assume that $d\mu = fdx$ and $d\nu = gdx$. In order to minimize $\int d(x,y)M(x,y)dxdy$ and recall the trivial metric which is 0 if and only if $x=y$, we hope to give more weight to points $\{(x,x)\}$. We write $M(x,y) = h(x)dx\cdot\delta_x(dy) + N(x,y)$ and since $\int M(x,y)dy = \mu(x) \geq h(x)$ and $\int M(x,y)dx = \nu(y) \geq h(y)$, so we set $h(x) = \min\{f(x),g(x)\}$. For $x\neq y$, there is no difference in setting the value of $M(x,y)$. Let $d\widetilde{\mu} = \{f-f\wedge g\}dx$, $d\widetilde{\nu} = \{g-f\wedge g\}dx$ and $d\widetilde{\eta} = \{f\wedge g\}dx$, we construct joint distribution $\boldsymbol{M}$

$$\boldsymbol{M}(dx,dy) = \eta(dx)\delta_x(dy) + \frac{\widetilde{\mu}(dx)\widetilde{\nu}(dy)}{1-\eta(\mathscr{S})} \tag{5.3.3}$$

where $\eta(\mathscr{S})$ denotes the integral over the whole space $\mathscr{S}$ w.r.t. measure $d\eta$.

Now we prove that above $M(x,y)$ satisfies the conditions. Since

$$\boldsymbol{M}(\boldsymbol{X}\neq\boldsymbol{Y}) = 1-\eta(\mathscr{S}) = \int (f-f\wedge g)dx,$$

and notice that

$$\int (f-f\wedge g)dx = \int (f-g)_+dx = \sup_{0\leq h\leq 1}\int h(f-g)dx = \|\mu-\nu\|_{TV}.$$

Thus the obtained coupling attains the infinity and equals the Wasserstein distance.

Recall that we aim to use Theorem 5.5 to prove subgaussian results, so we hope in multivariate case we have

$$W_1(\nu,\mu_1\otimes\cdots\otimes\mu_n) \leq \sqrt{2\sigma^2 D(\nu\|\mu_1\otimes\cdots\otimes\mu_n)} \text{ for all } \nu.$$

Notice that to obtain tensorization, we first need to specify the distance. The following theorem gives us a general form of tensorization which allows us to plug in various metrics.

Theorem 5.8 (Marton's Theorem). *Let $\phi:\mathbb{R}\to\mathbb{R}$ be a convex function, $w_i:\mathscr{S}_i\times\mathscr{S}_i\to R_+$ be positive weight function. Suppose that for all i we have*

$$\inf_{M\in\mathscr{C}(\mu_i,\nu)} \phi(\mathbb{E}_M[w_i(\boldsymbol{X},\boldsymbol{Y})]) \le 2\sigma^2 D(\nu\|\mu_i) \textit{ for all } \nu.$$

Then we have in multivariate case

$$\inf_{M\in\mathscr{C}(\mu_1\otimes\cdots\otimes\mu_n,\nu)} \sum_{i=1}^n \phi(\mathbb{E}_M[w_i(\boldsymbol{X}_i,\boldsymbol{Y}_i)]) \le 2\sigma^2 D(\nu\|\mu_1\otimes\cdots\otimes\mu_n) \textit{ for all } \nu.$$

Before we delve into the proof, we first provide several examples to show the flexibility and power of Theorem 5.8.

Corollary 5.2. *Let $w(x,y) = d(x,y)$ and $\phi(x) = x^2$, we have*

$$\inf_{M\in\mathscr{C}(\mu_1\otimes\cdots\otimes\mu_n,\nu)} \sum_{i=1}^n [\mathbb{E}_M[w_i(\boldsymbol{X}_i,\boldsymbol{Y}_i)]]^2 \le 2\sigma^2 D(\nu\|\mu_1\otimes\cdots\otimes\mu_n) \textit{ for all } \nu. \quad (5.3.4)$$

Notice that the left hand side is not exactly the form of Wasserstein distance. However, define $d_c(x,y) = \sum_{i=1}^n c_i d_i(x_i,y_i)$ where $\sum_{i=1}^n c_i^2 = 1$, then by Cauchy Schwarz inequality and Theorem 5.7 we have

$$W_1(\mu,\nu) = \inf_{M\in\mathscr{C}(\mu,\nu)} \sum_{i=1}^n c_i \mathbb{E}_M[d_i(\boldsymbol{X}_i,\boldsymbol{Y}_i)] \le \left[\inf_{M\in\mathscr{C}(\mu,\nu)} \sum_{i=1}^n [\mathbb{E}_M[d_i(\boldsymbol{X}_i,\boldsymbol{Y}_i)]]^2\right]^{\frac{1}{2}}. \quad (5.3.5)$$

We obtain the desired result by combining Eq. (5.3.4) *and Eq.* (5.3.5).

Example 5.6 (McDiarmid's Inequality). Consider the trivial metric d_i and joint metric $d_c(x,y) = \sum_{i=1}^n c_i d_i(x_i,y_i)/(\sum_{i=1}^n c_i^2)^{0.5}$ where $c_i = \|D_i\|_\infty$. By Pinsker's Inequality 5.4, we have for univariate case $\|\mu_i,\nu\|_{TV} \le \sqrt{1/2\cdot D(\nu\|\mu_i)}$, so by Theorem 5.8 we have $W_1(\mu_1\otimes\cdots\otimes\mu_n,\nu) \le \sqrt{1/2\cdot D(\nu\|\mu_1\otimes\cdots\otimes\mu_n)}$ for all ν. and we obtain $f(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$ is $\sum_{i=1}^n \|D_i\|_\infty^2$ subgaussian.

Now we trun to prove Theorem 5.8. Since by Therorem 5.7 we have an equivalent form of Wasserstein distance. In order to derive the tensorization w .r.t. Wasserstein distance and KL divergence, we need the chain rule of KL divergence.

Lemma 5.9 (Chain rule for KL divergence). *Let $\boldsymbol{M}$ and $\boldsymbol{N}$ be two probability measures corresponding to joint distributions of random variables $\boldsymbol{X}$, $\boldsymbol{Y}$, then*

$$D(\boldsymbol{M}(\boldsymbol{X},\boldsymbol{Y}\in\cdot)\|\boldsymbol{N}(\boldsymbol{X},\boldsymbol{Y}\in\cdot)) = D(\boldsymbol{M}(\boldsymbol{X}\in\cdot)\|\boldsymbol{N}(\boldsymbol{X}\in\cdot)) + \mathbb{E}_{\mathscr{M}}[D(\boldsymbol{M}(\boldsymbol{Y}\in\cdot|\boldsymbol{X})\|\boldsymbol{N}(\boldsymbol{Y}\in\cdot|\boldsymbol{X}))].$$

We see that the KL divergence of two measures consists of the KL divergence w.r.t. marginal measures and the expectation of KL divergence of measures conditional on previous marginal variables. This allows us to split the KL divergence of high dimension into low dimensional combinations and thus use induction.

Proof. The core of the proof is the Bayes rule. Recall that

$$d\boldsymbol{M}(\boldsymbol{X},\boldsymbol{Y}\in\cdot) = d\boldsymbol{M}(\boldsymbol{Y}\in\cdot|\boldsymbol{X})\cdot d\boldsymbol{M}(\boldsymbol{X}\in\cdot)$$

so we have
$$\frac{d\boldsymbol{M}(\boldsymbol{X},\boldsymbol{Y}\in\cdot)}{d\boldsymbol{N}(\boldsymbol{X},\boldsymbol{Y}\in\cdot)}=\frac{d\boldsymbol{M}(\boldsymbol{Y}\in\cdot|\boldsymbol{X})}{d\boldsymbol{N}(\boldsymbol{Y}\in\cdot|\boldsymbol{X})}\cdot\frac{d\boldsymbol{M}(\boldsymbol{X}\in\cdot)}{d\boldsymbol{N}(\boldsymbol{X}\in\cdot)}.$$

We take log on both sides and use the definition of KL divergence we obtain

$$\begin{aligned}D(\boldsymbol{M}(\boldsymbol{X},\boldsymbol{Y}\in\cdot)\|\boldsymbol{N}(\boldsymbol{X},\boldsymbol{Y}\in\cdot)) &= D(\boldsymbol{M}(\boldsymbol{X}\in\cdot)\|\boldsymbol{N}(\boldsymbol{X}\in\cdot))\\ &+\mathbb{E}_{\mathscr{M}}\left[\mathbb{E}_{\mathscr{M}}\left[D(\boldsymbol{M}(\boldsymbol{Y}\in\cdot|\boldsymbol{X})\|\boldsymbol{N}(\boldsymbol{Y}\in\cdot|\boldsymbol{X}))\big|\boldsymbol{X}\right]\right]. \qquad (5.3.6)\end{aligned}$$

Proof. Now we prove Theorem 5.8 by induction. When $n=1$, it is trivial by assumptions. Suppose we have obtained the result for $n=k$, we consider $n=k+1$. By Eq. (5.3.6) we have

$$D(\nu\|\mu_1\otimes\cdots\mu_{k+1})=D(\nu^{(k)}\|\mu_1\otimes\cdots\mu_k)+\mathbb{E}_{\mathscr{M}}\left[D(\nu_{y_1,\ldots,y_k}\|\mu_{k+1})\right],$$

where $\nu^{(k)}$ denotes the marginal distribution w.r.t. $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_k$ and $\nu^{y_1,\ldots,y_k}$ be the distribution conditional on $\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_k$. By induction assumption we have

$$2\sigma^2D(\nu^{(k)}\|\mu_1\otimes\cdots\mu_k)\geq\inf_{M\in\mathscr{C}(\mu_1\otimes\cdots\mu_k,\nu^{(k)})}\sum_{i=1}^{k}\phi(\mathbb{E}_M[w_i(\boldsymbol{X}_i,\boldsymbol{Y}_i)]).$$

By the convexity of ϕ and the assumption we get $M\in\mathscr{C}(\mu_1\otimes\cdots\otimes\mu_{k+1})$ and $M_1\in\mathscr{C}(\boldsymbol{X}_1,\boldsymbol{X}_1)$

$$\begin{aligned}\mathbb{E}_M\left[2\sigma^2D(\nu_{y_1,\ldots,y_k}\|\mu_{k+1})\right]&\geq\mathbb{E}_M\left[\phi(\mathbb{E}_{M_{y_1,\ldots,y_k}}[w_{k+1}(\boldsymbol{X}_{k+1},\boldsymbol{Y}_{k+1})])\right]\\&\geq\phi(\mathbb{E}_M\left[\mathbb{E}_{M_{y_1,\ldots,y_k}}[w_{k+1}(\boldsymbol{X}_{k+1},\boldsymbol{Y}_{k+1})]\right]).\end{aligned}$$

We take M_{k+1} and $M^{(k)}$ be two ε minimizer (M_{k+1} depends on $y_1,\ldots,y_k$), and we construct a joint measure by M_{k+1} and $M^{(k)}$

$$M_\varepsilon(x_1,\ldots,x_{k+1},y_1,\ldots,y_{k+1})=M_{k+1}(x_{k+1},y_{k+1}\big|x_1,\ldots,x_k,y_1,\ldots,y_k)\cdot M^{(k)}(x_1,\ldots,x_k,y_1,\ldots,y_k).$$

So we have by combining previous displays

$$\begin{aligned}2\sigma^2D(\nu\|\mu_1\otimes\cdots\otimes\mu_{k+1})&\geq\sum_{i=1}^{k+1}\phi(\mathbb{E}_{M_\varepsilon}[w_i(\boldsymbol{X}_i,\boldsymbol{Y}_i)])-2\varepsilon\\&\geq\inf_{M\in\mathscr{C}(\mu_1\otimes\cdots\otimes\mu_{k+1}),\nu)}\sum_{i=1}^{k+1}\phi(\mathbb{E}_M[w_i(\boldsymbol{X}_i,\boldsymbol{Y}_i)])-2\varepsilon.\end{aligned}$$

Since ε is arbitrary, we can get the desired tensorization result.

Upon proving such a useful tensorization, we are ready to apply them and introduce the major theorem - Talagrand's inequality.

Theorem 5.10 (Talagrand's Inequality). *Let random variables $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ be independent, and assume that*

$$f(x)-f(y)\le\sum_{i=1}^n c_i(x)\cdot\mathbf{1}_{x_i\neq y_i} \textit{ for all } x,y,$$

then $f(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$ is $\|\sum_{i=1}^n c_i^2\|_\infty$-subgaussian.

Notice that compared with Lipschitz condition, here we only assume one-sided bound and $c_i(x)$ is dependent on x. If we consider convex function with bounded random variables, Talagrand's inequality always applies. Thus, this theorem is of great use.

As for the proof, we mostly follow the previous analysis. But unlike the previous settings, here we don't have a mature metric or a Wasserstein distance. However, recall Theorem 5.5, we only need $|\mathbb{E}_\mu f-\mathbb{E}_\nu|\le\sqrt{2\sigma^2 D(\nu\|\mu)}$, so we hope to find an intermediate counterpart of $\inf_{M\in\mathscr{C}(\mu,\nu)}\mathbb{E}_M[d(x,y)]$. Since

$$\mathbb{E}_\mu f-\mathbb{E}_\nu f\le\inf_{M\in\mathscr{C}(\mu,\nu)}\mathbb{E}_M\left[\sum_{i=1}^n c_i(x)\mathbf{1}_{x_i\neq y_i}\right].$$

In order to compare with KL divergence and notice that we don't have a normalization condition $\sum_{i=1}^n c_i^2=1$ as before, we introduce the following "distance"

$$d_2(\mu,\nu)=\inf_{M\in\mathscr{C}(\mu,\nu)}\sup_{\mathbb{E}_M[\sum_{i=1}^n c_i^2(\boldsymbol{X})]\le 1}\mathbb{E}_M\left[\sum_{i=1}^n c_i(\boldsymbol{X})\mathbf{1}_{\boldsymbol{X}_i\neq\boldsymbol{Y}_i}\right].$$

and we have the following theorem which gives the inequality.

Theorem 5.11 (Marton's Theorem).

$$d_2(\nu,\mu_1\otimes\cdots\otimes\mu_n)\le\sqrt{2D(\nu\|\mu_1\otimes\cdots\otimes\mu_n)} \textit{ for all } \nu,$$

$$d_2(\mu_1\otimes\cdots\otimes\mu_n,\nu)\le\sqrt{2D(\nu\|\mu_1\otimes\cdots\otimes\mu_n)} \textit{ for all } \nu.$$

Proof. Using Theorem 5.11, we are ready to prove Theorem 5.10. Since

$$\mathbb{E}_\mu f-\mathbb{E}_\nu f=\mathbb{E}_{M\in\mathscr{C}(\mu,\nu)}\left[f(\boldsymbol{X})-f(\boldsymbol{Y})\right]\le[\mathbb{E}_\mu[\sum_{i=1}^n c_i^2]]^{\frac{1}{2}}\cdot d_2(\mu,\nu),$$

$$\mathbb{E}_\nu f-\mathbb{E}_\mu f=\mathbb{E}_{M\in\mathscr{C}(\nu,\mu)}\left[f(\boldsymbol{X})-f(\boldsymbol{Y})\right]\le[\mathbb{E}_\nu[\sum_{i=1}^n c_i^2]]^{\frac{1}{2}}\cdot d_2(\nu,\mu).$$

Combine the above two displays and by Theorem 5.11 we obtain

$$\left|\mathbb{E}_\mu f-\mathbb{E}_\nu f\right|\le\|\sum_{i=1}^n c_i^2\|_\infty^{\frac{1}{2}}\cdot\sqrt{2D(\nu\|\mu)},$$

and use Theorem 5.5 we get $f(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$ is $\|\sum_{i=1}^n c_i^2\|_\infty$-subgaussian.

Now we focus on the proof of Theorem 5.11. And we divide the proof into two parts: univariate case and tensorization. We first state the result of tensorization, where the proof is almost the same with previous ones.

Proposition 5.1. *Let $\{\mu_i\}$ be probability measures, if we have for all i,*

$$\inf_{M\in\mathscr{C}(\mu_i,\nu)} \mathbb{E}_M[M[\boldsymbol{X}\neq\boldsymbol{Y}\big|\boldsymbol{X}]^2] \leq 2D(\nu\|\mu_i) \textit{ for all } \nu,$$

then we have

$$\inf_{M\in\mathscr{C}(\mu_1\otimes\cdots\otimes\mu_n,\nu)} \sum_{i=1}^n \mathbb{E}_M[M[\boldsymbol{X}_i\neq\boldsymbol{Y}_i\big|\boldsymbol{X}]^2] \leq 2D(\nu\|\mu_1\otimes\cdots\otimes\mu_n) \textit{ for all } \nu.$$

Similar results hold if we replace $M\in\mathscr{C}(\mu_i,\nu)$ by $M\in\mathscr{C}(\nu,\mu_i)$ and $M\in\mathscr{C}(\mu_1\otimes\cdots\otimes\mu_n,\nu)$ by $M\in\mathscr{C}(\nu,\mu_1\otimes\cdots\otimes\mu_n)$. This is necessary due to the asymmetry.

Proof. The proof follows the idea of that in Theorem 5.8 and the fact that $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ are independent, which gives us for $i\leq k$ and $M\in\mathscr{C}(\nu,\mu_1\otimes\cdots\otimes\mu_{k+1})$

$$\mathbb{E}_M[M[\boldsymbol{X}_i\neq\boldsymbol{Y}_i\big|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k]^2] = \mathbb{E}_M[M[\boldsymbol{X}_i\neq\boldsymbol{Y}_i\big|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{k+1}]^2].$$

We also needs Jensen to get

$$\mathbb{E}_M[M[\boldsymbol{X}_{k+1}\neq\boldsymbol{Y}_{k+1}\big|\boldsymbol{Y}_1,\ldots,\boldsymbol{Y}_k,\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{k+1}]^2] \geq \mathbb{E}_M[M[\boldsymbol{X}_{k+1}\neq\boldsymbol{Y}_{k+1}\big|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{k+1}]^2].$$

the reverse case?

It remains to show that $n=1$, $\inf_{M\in\mathscr{C}(\mu,\nu)}\mathbb{E}_M[M[\boldsymbol{X}\neq\boldsymbol{Y}\big|\boldsymbol{X}]^2]\leq 2D(\nu\|\mu)$ for all ν. We first try to build up the connection between $\inf_{M\in\mathscr{C}(\mu,\nu)}\mathbb{E}_M[M[\boldsymbol{X}\neq\boldsymbol{Y}\big|\boldsymbol{X}]^2]$ with $\mathbb{E}_\mu f-\mathbb{E}_\nu f$ and derive the inequality by pure calculation.

Proof. By Cauchy Schwarz inequality, we have

$$\sup_{f\geq 0,\mu(f^2)\leq 1}\{\mathbb{E}_\mu f-\mathbb{E}_\nu f\} = \sup_{f\geq 0,\mu(f^2)\leq 1}\left\{\mathbb{E}_\mu\left[f\left(1-\frac{d\nu}{d\mu}\right)\right]\right\} = \left[\mathbb{E}_\mu\left[\left(1-\frac{d\nu}{d\mu}\right)_+^2\right]\right]^{\frac{1}{2}}.$$

We have obtained the term related to KL divergence on the right hand side and we aim to connect joint distribution with $\sup_{f\geq 0,\mu(f^2)\leq 1}\{\mathbb{E}_\mu f-\mathbb{E}_\nu f\}$. Since for all $f\geq 0,\mu(f^2)\leq 1$ by the non-negativity of f and Cauchy Schwarz formula

$$\mathbb{E}_\mu f-\mathbb{E}_\nu f = \mathbb{E}_{M\in\mathscr{C}(\mu,\nu)}[f(\boldsymbol{X})-f(\boldsymbol{Y})] \leq \mathbb{E}_{M\in\mathscr{C}(\mu,\nu)}[f(\boldsymbol{X})\mathbf{1}_{\boldsymbol{X}\neq\boldsymbol{Y}}]$$

$$= \mathbb{E}_{M\in\mathscr{C}(\mu,\nu)}[f(\boldsymbol{X})\boldsymbol{M}[\boldsymbol{X}\neq\boldsymbol{Y}\big|\boldsymbol{X}]] \leq \left[\mathbb{E}_{M\in\mathscr{C}(\mu,\nu)}[\boldsymbol{M}[\boldsymbol{X}\neq\boldsymbol{Y}\big|\boldsymbol{X}]^2]\right]^{\frac{1}{2}}.$$

Combine the above two displays, we have approached our final conclusion one more step by obtaining

$$\sup\left\{\mathbb{E}_{M\in\mathscr{C}(\mu,\nu)}[\boldsymbol{M}[\boldsymbol{X}\neq\boldsymbol{Y}\big|\boldsymbol{X}]^2]\right\}\leq\mathbb{E}_\mu\left[\left(1-\frac{d\nu}{d\mu}\right)_+^2\right].$$

We will show that the above inequality can be obtained. Where we conl need to show that there exists a joint distribution M satisfies that $\boldsymbol{M}[\boldsymbol{X}\neq\boldsymbol{Y}\big|\boldsymbol{X}]=(1-d\nu/d\mu)_+$ and by Eq. (5.3.3) we have constructed the suitable joint distribution.

The final step we turn our attention back to $d_2(\mu,\nu)$. Since

$$d_2(\mu,\nu)^2+d_2(\nu,\mu)^2=\int\left(1-\frac{d\nu}{d\mu}\right)_+^2+\int\left(1-\frac{d\mu}{d\nu}\right)_+^2\frac{d\nu}{d\mu}d\mu$$

$$\leq 2\int\left(\frac{d\nu}{d\mu}\log\left(\frac{d\nu}{d\mu}\right)-\frac{d\nu}{d\mu}+1\right)d\mu=2D(\nu\|\mu),$$

where the inequality comes from the fact for all $x>0$

$$x\log(x)-x+1\geq\frac{(1-x)_+^2+(1-x^{-1})_+^2x}{2}$$

and plug in $x=d\nu/d\mu$. By now we have derived the univariate inequality and we can arrive at Theorem 5.10 by further using tensorization.

Chapter 6
High-Dimensional Statistics for GLM

In this chapter, we aim to analyse a family of M-estimators in high-dimensional regime and bound $\|\widehat{\beta}-\beta\|_q$ $(q \geq 1)$.
We first define the models and M-estimators for analysis. Then we give the outline of proving error bound in low-dimensional regime and point out the difficulties in its extension to high-dimension. To overcome the obstacles, we replace strong convexity by a restricted one (RSC) with regard to a certain subspace. In order to enforce the estimators fall into the subspace, we add decomposable regularizers which penalize heavily on the perpendicular subspace. Assuming these two conditions we give the high-level proof assuming the above two conditions, then analyse concrete examples (LASSO, GLM, Gaussian graphical model).

It will be greatly beneficial for you to bear the LASSO example in mind through this chapter.

Key Contents: M-Estimators Decomposable regularizers Restricted strong convexity

For notations, let $\| \ \|_q$ $(q > 1)$ denote the L_q norm. For simplicity, we use $\| \ \|$ for $\| \ \|_2$.
Let $\rho(\beta)$ be a norm in $\mathbb{R}^d$, and $\rho^*(\beta)$ the dual norm.
Assume $\mathscr{M}$ is a linear subspace (or convex set) of $\mathbb{R}^d$, let $\beta_{\mathscr{M}}$ denote the projection of β into $\mathscr{M}$.
We call β is s-sparse, if the number of non-zero components of β is less than s.

6.1 M-Estimators for GLM

We first introduce the models and M-estimators of interest.

Definition 6.1. Let $X \in \mathbb{R}^d$, $Y \in \mathscr{Y} \subseteq \mathbb{R}$, (X,Y) follows GLM if

$$\mathbb{P}_\beta(Y=y|X=x) \propto \exp\left\{\frac{y\beta^T x - A(\beta^T x)}{\tau}\right\}$$

where β is the oracle parameter and A is a known link function.

Example 6.1. **Linear Regression**
We can take $\mathscr{Y}=\mathbb{R}, A(t)=t^2/2$ and $\tau=\sigma^2$.

Example 6.2. **Logistic Regression**
We can take $\mathscr{Y}=\{0,1\}, A(t)=\log(1+e^t)$ and $\tau=1$.

Basically, M-Estimators are to minimize the empirical loss. However, in high-dimensional statistics, classical estimators fail and we need regularizers considering sparsity assumption of β. Therefore, we focus on the following family of estimators,

$$\widehat{\beta}^\lambda \in \underset{\beta\in\mathbb{R}^d}{\operatorname{argmin}}\left\{\frac{1}{n}\sum_{i=1}^n \left(y_i\beta^T x_i - A(\beta^T x_i)\right\} + \lambda\rho(\beta)\right\},$$

We see that the former part corresponds to the empirical loss (denoted by L_n) and the latter is a regularizer.

6.2 Low-Dimenional M-Estimators

We present the big picture of how to bound the error in low-dimensional scenario, which is rather inspiring.

Assume that L_n is differentiable and convex. Let $\{\Lambda_i(\beta)\}$ be the eigen-values of $\triangle L_n(\beta)$ in decreasing order. Assume that $\Lambda_{min}(\beta)$ is larger than $\Lambda_0>0$ in a sufficiently large neighbourhood of β. Define $\widehat{\Delta}=\widehat{\beta}^\lambda-\beta$, then by the definition of M-Estimator

$$0 \geq L_n(\beta+\widehat{\Delta}) - L_n(\beta) \geq \langle \bigtriangledown L_n(\beta), \widehat{\Delta}\rangle + \frac{1}{2}\Lambda_0\|\widehat{\Delta}\|^2$$

we use dual norm here and get

$$\|\bigtriangledown L_n(\beta)\|\|\widehat{\Delta}\| \geq \frac{1}{2}\Lambda_0\|\widehat{\Delta}\|^2,$$

and the rate

$$\|\widehat{\Delta}\| = \mathscr{O}(\sqrt{d}).$$

We can also decompose the inner product using other norms which will introduce a $\sqrt{\log d}$ and not tight.

Remark 6.1. **SC**
According to the graphs, if we require the error to be small, then the Ł$_n$ should not be too flat, which implies strong convexity.

Remark 6.2. **Difficulties in high-dimension**
In high-dimensional regime, especially $d > n$, the strong convexity of L_n is invalid almost everywhere and the rate $\mathscr{O}(\sqrt{d})$ is too loose and not adaptive to sparsity.

6.3 High-Dimenional M-Estimators

To overcome the difficulties brought by high-dimension, we observe that if we can guarantee $\widehat{\beta}^{\lambda}$ falls into a certain space $\mathscr{M}$ and L_n is strong convex in a neighbourhood of β intersecting $\mathscr{M}$, then we can approximately follow the above proofs.

By the definition of norm we have

$$\rho(\theta + r) \leq \rho(\theta) + \rho(r),$$

and if we can obtain equality here when $\theta \in \mathscr{M}$ and $r \in \mathscr{M}_{\perp}$, then intuitively projection of $\widehat{\beta}_{\lambda}$ into $\mathscr{M}^{\perp}$ is small and we can restrict our estimators to $\mathscr{M}$.

6.3.1 Decomposability of Regularizer

To make the above analysis more rigorous, we introduce the definition of decomposable regularizer.

Definition 6.2. We say a regularizer ρ is decomposable with regard to subspace $\mathscr{M} \subseteq \mathbb{R}^d$, if $\forall \theta \in \mathscr{M}$, $r \in \mathscr{M}^{\perp}$, we have

$$\rho(\theta + r) = \rho(\theta) + \rho(r).$$

Example 6.3. **LASSO**
L_1 regularizer is decomposable with regard to s-sparse subspace $\mathscr{M}_s$.

Now we prove that the decomposable regularizer can achieve the desired restriction.

Lemma 6.1. *Assume that L_n is convex, differentiable, ρ is a norm and decomposable w.r.t. subspace $\mathscr{M}$ and $\lambda \geq 2\rho^*(\triangledown L_n(\beta))$. Let $\mathscr{C}(\mathscr{M}, \beta) = \left\{\Delta \in \mathbb{R}^d \mid \rho(\Delta_{\mathscr{M}^{\perp}}) \leq 3\rho(\Delta_{\mathscr{M}}) + 4\rho(\beta_{\mathscr{M}^{\perp}})\right\}$, we have the error of M-Estimator*

$$\widehat{\Delta} \in \mathscr{C}(\mathscr{M}, \beta).$$

Example 6.4. **LASSO**
Before dive into the proof, we first give a intuitive picture of how decomposability works. Consider $d = 3$, *support*$\beta = \{3\}$, then if $\beta \in \mathscr{M}_s$,

$$|\rho(\Delta_1)| + |\rho(\Delta_2)| \leq |\rho(\Delta_3)|.$$

If $\beta \notin \mathscr{M}_s$,

$$|\rho(\Delta_1)| + |\rho(\Delta_2)| \leq |\rho(\Delta_3)| + |\beta_3|.$$

Plot the graph, it gives us a star-shape. By observation, we see that the regularizer does keep $\widehat{\beta}_\lambda$ away from $\mathscr{M}_\perp$.

Proof. By optimality condition, we have

$$L_n(\widehat{\beta}_\lambda) - L_n(\beta) \leq \lambda(\rho(\beta) - \rho(\widehat{\beta}_\lambda)),$$

by convexity and the choice of λ, we analyse the left hand side

$$L_n(\widehat{\beta}_\lambda) - L_n(\beta) \geq \langle \bigtriangledown L_n(\beta), \widehat{\Delta} \rangle \geq -\frac{\lambda}{2}(\rho(\widehat{\Delta}_{\mathscr{M}}) + \rho(\widehat{\Delta}_{\mathscr{M}_\perp})).$$

By decomposablity of ρ w.r.t. $\mathscr{M}$ and triangle inequality of norm, we have the right hand side

$$\rho(\beta) - \rho(\widehat{\beta}_\lambda) \leq (\rho(\beta_{\mathscr{M}}) + \rho(\beta_{\mathscr{M}_\perp})) - \left(\rho(\beta_{\mathscr{M}}) + \rho(\widehat{\Delta}_{\mathscr{M}_\perp}) - \rho(\beta_{\mathscr{M}_\perp}) - \rho(\widehat{\Delta}_{\mathscr{M}})\right),$$

combine two sides, we obtain

$$\rho(\widehat{\Delta}_{\mathscr{M}^\perp}) \leq 3\rho(\widehat{\Delta}_{\mathscr{M}}) + 4\rho(\widehat{\beta}_{\mathscr{M}^\perp}),$$

that is $\widehat{\Delta} \in \mathscr{C}(\mathscr{M}, \beta)$.

Remark 6.3. **Choice of λ**
In LASSO, we can view the oder of λ from different perspectives.
(1) Optimization: due to the necessary condition implied by minimality, we have

$$\frac{\partial}{\partial t}\left(L_n(\beta + t\widehat{\Delta}) + \lambda \|\beta + t\Delta\|_1\right)(1) = \langle \bigtriangledown L_n(\beta + \widehat{\Delta}), \widehat{\Delta} \rangle + \lambda \langle sgn(\beta + \widehat{\Delta}), \widehat{\Delta} \rangle = 0,$$

then we get

$$\lambda \approx \| \bigtriangledown L_n(\beta + \widehat{\Delta})\|_\infty.$$

However, this is dependent on samples. To eliminate the randomness, we expect

$$\lambda \approx \| \bigtriangledown L_n(\beta)\|_\infty,$$

and this matches the order.
(2) Statistics: it is obvious that the loss function consists of two parts and intuitively we hope the two terms balance

$$\left| L_n(\beta + \widehat{\Delta}) - L_n(\beta) \right| \approx \lambda \left| \|\beta + \widehat{\Delta}\|_1 - \|\beta\|_1 \right|,$$

$$\| \bigtriangledown L_n(\beta)\|_\infty \|\widehat{\Delta}\|_1 \approx \lambda \left| \|\beta + \widehat{\Delta}\|_1 - \|\beta\|_1 \right|,$$

and we get the order.

In fact, we see that such order of λ is sufficient for our proofs.

6.3.2 RSC

Now we move to restricted strong convexity.

Definition 6.3. We call that L_n satisfies RSC condition w.r.t. subspace $\mathscr{M}$ and β if $\forall \Delta \in \mathscr{C}(\mathscr{M}, \beta)$, $\exists\ \kappa_L > 0$, s.t.

$$L_n(\beta + \Delta) - L_n(\beta) \geq \langle \bigtriangledown L_n(\beta), \Delta \rangle + \kappa_L \|\Delta\|^2$$

and we call κ_L the curvature parameter.

6.3.3 Error Bounds for M-Estimators

With RSC, we are ready to bound the error.

Lemma 6.2. *Define* $F(\Delta) = L_n(\beta + \Delta) + \lambda\rho(\beta + \Delta) - L_n(\beta) - \lambda\rho(\beta)$*. If for* $\forall \Delta \in \mathscr{C}(\mathscr{M}, \beta) \cap \{\|\Delta\|_q = \delta\}$ *s.t.* $F(\Delta) > 0$*, then we have*

$$\|\widehat{\Delta}\|_q < \delta$$

.

Proof. We observe that $F(\Delta)$ satisfies:
(1) $F(0) = 0$; (2) $F(\Delta)$ is convex; (3) $F(\widehat{\Delta}) < 0$. And we use the convexity of $F(\Delta)$, properties of the star-shape of $\mathscr{C}(\mathscr{M}, \beta)$ and prove by contradiction.
If $\|\widehat{\Delta}\|_q \geq \delta$, since $\mathscr{C}(\mathscr{M}, \beta)$ is star-shape w.r.t 0, there exists $\alpha \in (0, 1]$ s,t,

$$\|\alpha\widehat{\Delta}\|_q = \delta.$$

Since $F(\Delta)$ is convex,

$$0 < F(\alpha\widehat{\Delta})) \leq \alpha F(\widehat{\Delta})) + (1-\alpha)F(0) \leq 0,$$

contradiction! So

$$\|\widehat{\Delta}\|_q < \delta.$$

What we need now is a valid δ, which is actually of constant level.
First we define subspace compatibility constant, which serves as a bridge connecting different norms.

Definition 6.4. κ_q is called the subspace compatibility constant w.r.t. subspace $\mathscr{M}$ and norms ρ, L_q norms,

$$\kappa_q(\mathscr{M}) = \sup_{u \in \mathscr{M}\setminus\{0\}} \frac{\rho(v)}{\|v\|_q}.$$

Example 6.5. **LASSO**
Taking $\mathscr{M} = M_s$ and $\rho = \|\ \|_1$, $q = 2$, we have $\kappa_2 = \sqrt{s}$, which is much smaller than $\sqrt{d}$ considering $\mathbb{R}^d$. In the following theorem, we can see that the "restrictiveness" gives a much sharper upper bound in high-dimensional settings.

Theorem 6.3. *Assume that L_n is convex, differentiable, and satisfies RSC condition w.r.t. $\mathscr{C}(\mathscr{M},\beta) \cap \{\|\Delta\|_q\}$ with curvature parameter κ_L. ρ is decomposable w.r.t. $\mathscr{M}$ and $\lambda \geq 2\rho^*(\bigtriangledown L_n(\beta))$. We have*

$$\|\widehat{\beta}^\lambda - \beta\|^2 \leq 9\frac{\lambda^2}{\kappa_L^2}\kappa_2^2(\mathscr{M}) + 4\frac{\lambda}{\kappa_L}\rho(\beta_{\mathscr{M}^\perp})$$

We see that if $\beta \in \mathscr{M}$, then we have

$$\|\widehat{\beta}^\lambda - \beta\| \leq 3\frac{\lambda}{\kappa_L}\kappa_2(\mathscr{M}).$$

The upper bound decreases as curvature parameter κ_L increases or λ, $\kappa_2(\mathscr{M})$ and $\rho(\beta_{\mathscr{M}^\perp)}$ decreases. By introducing "restrictiveness" we can significantly reduce $\kappa_2(\mathscr{M})$ and increase κ_L, thus obtain a tighter bound.

Proof. Using RSC w.r.t. $\mathscr{C}(\mathscr{M},\beta) \cap \{\|\Delta\|_q\}$,

$$L_n(\beta + \widehat{\Delta}) - L_n(\beta) \geq \langle \bigtriangledown L_n(\beta), \widehat{\Delta}\rangle + \kappa_L\|\widehat{\Delta}\|^2,$$

and by previous proof,

$$\rho(\beta + \Delta) - \rho(\beta) \geq \rho(\widehat{\Delta}_{\mathscr{M}^\perp}) - \rho(\widehat{\Delta}_{\mathscr{M}}) - 2\rho(\beta_{\mathscr{M}^\perp}),$$

we combine two parts and get

$$F(\widehat{\Delta}) \geq \langle \bigtriangledown L_n(\beta), \widehat{\Delta}\rangle + \kappa_L\|\widehat{\Delta}\|^2 + \lambda\left(\rho(\widehat{\Delta}_{\mathscr{M}^\perp}) - \rho(\widehat{\Delta}_{\mathscr{M}}) - 2\rho(\beta_{\mathscr{M}^\perp})\right).$$

By the assumption of λ we can rearrange the right hand side into two parts

$$F(\widehat{\Delta}) \geq \kappa_L\|\widehat{\Delta}\|^2 - \frac{\lambda}{2}\left(3\rho(\widehat{\Delta}_{\mathscr{M}}) + 4\rho(\beta_{\mathscr{M}^\perp})\right).$$

We have two norms here, and we need κ_2 here to bridge over

$$\rho(\widehat{\Delta}_{\mathscr{M}}) \leq \kappa_2(\mathscr{M})\|\widehat{\Delta}_{\mathscr{M}}\| \leq \kappa_2(\mathscr{M})\|\widehat{\Delta}\|,$$

thus we get a quadratic function of $\|\widehat{\Delta}\|$

$$F(\widehat{\Delta}) \geq \kappa_L \|\widehat{\Delta}\|^2 - \frac{\lambda}{2}\left(3\kappa_2(\mathscr{M})\|\widehat{\Delta}\| + 4\rho(\beta_{\mathscr{M}^\perp})\right).$$

By taking

$$\delta^2 = \|\widehat{\Delta}\|^2 = \frac{9\lambda^2}{\kappa_L^2}\kappa_2^2(\mathscr{M}) + \frac{4\lambda}{\kappa_L}\rho(\beta_{\mathscr{M}^\perp}),$$

we get

$$F(\widehat{\Delta}) > 0 \quad \forall\, \|\widehat{\Delta}\| = \delta \text{ and } \widehat{\Delta} \in \mathscr{C}(\mathscr{M}, \beta).$$

By previous lemma, we have

$$\|\widehat{\beta}^\lambda - \beta\|^2 \leq 9\frac{\lambda^2}{\kappa_L^2}\kappa_2(\mathscr{M})^2 + 4\frac{\lambda}{\kappa_L}\rho(\beta_{\mathscr{M}^\perp}).$$

Remark 6.4. In fact, we can follow the proof and get the bound of $\|\widehat{\beta}^\lambda - \beta\|_q^2$ by redefining RSC w.r.t. L_q norm

$$L_n(\beta + \widehat{\Delta}) - L_n(\beta) \geq \langle \bigtriangledown L_n(\beta), \widehat{\Delta}\rangle + \kappa_{L,q}\|\widehat{\Delta}\|_q^2$$

and use compatibility parameter $\kappa_q(\mathscr{M})$. The result is exactly the same. Another way is to directly derive the bound of $\|\widehat{\beta}^\lambda - \beta\|_q^2$ from $\|\widehat{\beta}^\lambda - \beta\|^2$, since all norms are equivalent. However, this may not be as sharp as the former due to $\kappa_q(\mathscr{M})$ is restricted to space $\mathscr{M}$.

6.3.4 Applications

6.3.4.1 LASSO

We use random matrix theories to prove that w.h.p. RSC condition w.r.t. $\mathscr{C}(\mathscr{M}_s, \beta) \cap \{\|\Delta\| = \delta\}$ and $\lambda \geq 2\rho^*(\bigtriangledown L_n(\beta))$ hold.

6.3.4.2 GLM

6.3.4.3 Gaussian Graphical Model

Chapter 7
Persistency and Model Selection Consistency

In evaluation of an estimator, we are curious about whether it is consistent. Three among the most concerned consistencies are *persistency*, *model selection consistency* and *parameter consistency*. These three properties correspond to how well the estimator is for prediction, how well it can recover the true support and further the signs of true parameters and how close it is to the underlying truth coefficients under certain measure. It seems that the above three consistencies's hardness increase with order, however, it is not always true. This will be explained more explicitly in following sections, but here we first take a brief look. In fact, we can view these three properties in a unified framework considering risk functions. For example, in LASSO, when we take the loss function

$$\mathscr{L}_{pre} = \frac{1}{n}\|\mathbf{Y}-\mathbf{X}\widehat{\beta}\|_2^2,$$

we are concerned about prediction performance. If we consider loss function

$$\mathscr{L}_{\text{supp}} = \begin{cases} 0 & \text{if } \operatorname{supp}(\widehat{\beta}) = \operatorname{supp}(\beta) \\ 1 & \text{if } \operatorname{supp}(\widehat{\beta}) \neq \operatorname{supp}(\beta) \end{cases}$$

We are considering model selection consistency. Besides, if we use loss function

$$\mathscr{L}_{para} = \|\widehat{\beta}-\beta\|_2^2,$$

the goal is to evaluate the distance of the estimator to true value. Since neither of three loss functions is definitely larger than another, we can't say which consistency is stronger, or can imply another.

In this chapter, we will focus on three consistencies respectively, analysing the performances of estimators in both non-asymptotic and asymptotic regimes.

7.1 Persisteny

We consider persistency first because classic M-estimators usually minimize the prediction error, intuitively persistency is easier to analyse. We begin by introducing the general framework of persistency which doesn't rely on well-specified models, but depends on risk functions. Then we apply the high-level theories to two types of LASSO and derive non-asymptotic results.

Definition 7.1 (Persistency). Given a loss function $\mathscr{L}$, its correspondent risk function $\mathscr{R}$, and a sequence of parameter spaces $\{\mathscr{B}_n\}$, we define oracle estimators w.r.t. $\{\mathscr{B}_n\}$ as $\beta_{ora,n} \in \operatorname{argmin}_{\beta \in \mathscr{B}_n} \mathscr{R}(\beta)$. For any estimator sequence $\left\{\widehat{\beta}_n \in \mathscr{B}_n\right\}$, if $\mathscr{R}(\widehat{\beta}_n) - \mathscr{R}(\beta_{ora,n}) \to 0$ in probability as $n \to \infty$, we call $\left\{\widehat{\beta}_n\right\}$ persistent w.r.t. $\{\mathscr{B}_n\}$.

Here $\mathscr{B}_n$, $\beta_{ora,n}$ and $\widehat{\beta}_n$ are dependent on sample size n, but for simplicity, we use symbol $\mathscr{B}$ for $\mathscr{B}_n$, β_{ora} for $\beta_{ora,n}$ and $\widehat{\beta}$ for $\widehat{\beta}_n$. In the following analysis, we focus on M-estimators. In order to bound $\mathscr{R}(\widehat{\beta}) - \mathscr{R}(\beta_{ora})$ we first see how the gap comes into being. The classical form of M-estimator is to minimize the empirical loss, thus the difference comes mainly from the gap of $\mathscr{R}$ and $\widehat{\mathscr{R}}$. To be more rigorous, we split the excessive risk $\mathscr{R}(\widehat{\beta}) - \mathscr{R}(\beta_{ora})$ in the following manner

$$\mathscr{R}(\widehat{\beta}) - \mathscr{R}(\beta_{ora}) = \left(\mathscr{R}(\widehat{\beta}) - \widehat{\mathscr{R}}(\widehat{\beta})\right) + \left(\widehat{\mathscr{R}}(\widehat{\beta}) - \widehat{\mathscr{R}}(\beta_{ora})\right) + \left(\widehat{\mathscr{R}}(\beta_{ora})) - \mathscr{R}(\beta_{ora})\right).$$

The left hand side is non-negative by the optimality of β_{ora} and the second term on right hand side is non-positive since $\widehat{\beta}$ is an M-estimator. For the left two terms on the right, we have the inequality

$$\left(\mathscr{R}(\widehat{\beta}) - \widehat{\mathscr{R}}(\widehat{\beta})\right), \left(\widehat{\mathscr{R}}(\beta_{ora})) - \mathscr{R}(\beta_{ora})\right) \le \sup_{\beta \in \mathscr{B}} \left|\mathscr{R}(\beta) - \widehat{\mathscr{R}}(\beta)\right|,$$

which immediately gives us a upper bound of excessive risk

$$\mathscr{R}(\widehat{\beta}) - \mathscr{R}(\beta_{ora}) \le 2 \cdot \sup_{\beta \in \mathscr{B}} \left|\mathscr{R}(\beta) - \widehat{\mathscr{R}}(\beta)\right|. \tag{7.1.1}$$

This guides us to focus on bounding the infinity norm of real risk and empirical risk function.

Above analysis is general and adaptive, however, it can't offer more fruitful and concrete results. Next as examples, we derive the persistency of two LASSO estimators, i.e. naive LASSO and LASSO with cross validation.

7.1.1 Naive LASSO

Consider linear regression model: $\boldsymbol{y}$ is a random variable, $\boldsymbol{X}$ is a d dimension random vector, $\boldsymbol{\varepsilon}$ is a random noise, β is a d dimension coefficient. The linear model can be expressed by

$$\boldsymbol{y} = \beta^\top \boldsymbol{X} + \boldsymbol{\varepsilon}.$$

Definition 7.2 (Naive LASSO Estimator). For the above linear regression model, we have two equivalent ways to define naive LASSO estimator. Let $\mathbf{y}$, $\mathbf{X}$ be the observed samples, tune parameter $\lambda > 0$, constraint $L > 0$, then we have the Lagrangian form estimator

$$\widehat{\beta} \in \operatorname*{argmin}_{\beta} \left\{ \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right\},$$

and constrained form estimator

$$\widehat{\beta} \in \operatorname*{argmin}_{\|\beta\|_1 \le L} \|\mathbf{y} - \mathbf{X}\beta\|_2^2.$$

In the following analysis for persistency, we mainly focus on the second version of LASSO estimator which is simpler in proof. As for the first version, we need to specify λ first and then investigate the bound.

Theorem 7.1 (Persistency of Naive LASSO). *Assume that $|\boldsymbol{y}| \le B$ and $\|\boldsymbol{X}\|_\infty \le B$, we have*

$$\mathbb{P}\left(\mathscr{R}(\widehat{\beta}) - \mathscr{R}(\beta_{ora}) \le 2(1+L)^2 \sqrt{\frac{2B^4 \log\left(2(d+1)^2/\delta\right)}{n}} \right) \ge 1 - \delta.$$

Before we delve into the proof, we first present a straightforward corollary.

Corollary 7.1 (Persistency of Naive LASSO). *If $d = \mathscr{O}(n^k)$ for arbitrary k, $L_n = o(n/\log n)^{1/4}$, $\mathscr{B}_n = \{\|\beta\|_1 \le L_n\}$, then naive LASSO estimator is persistent w.r.t. $\mathscr{B}_n$.*

As can be seen, as d increases polynomially with n and L_n increases at a rate slower than $(n/\log n)^{1/4}$, we have the persistency result. This seems to be enough in practise if we assume sparsity on β. However, one thing that can't be neglected is the condition of bound, which is not always valid or may dominate the upper bound.

Proof. The proof contains two parts: the first part has already been done in the high-level analysis, and the latter half is a utilization of Hoeffding's inequality, where the condition of bound comes into play.

Recall Eq. (7.1.1), we only need to bound $\sup_{\beta\in\mathscr{B}} \left| \mathscr{R}(\beta) - \widehat{\mathscr{R}}(\beta) \right|$. We do a parametrization to simplify the form. Let

$$\boldsymbol{Z} = (\boldsymbol{y}, \boldsymbol{X^T})^\top, \quad r = (-1, \beta)^\top,$$

then by inequality of matrix norm, we have

$$\left(\mathscr{R}(\beta)-\widehat{\mathscr{R}}(\beta)\right)=r^T\left(\mathbb{E}\left[\boldsymbol{ZZ^T}\right]-\frac{1}{n}\sum_{k=1}^{n}\mathbf{Z_k Z_k^T}\right)r\le\|r\|_1^2\max_{1\le i,j\le(d+1)}\left|\mathbb{E}\left[\boldsymbol{ZZ^T}\right]_{ij}-\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{Z_k Z_k^T}\right)_{ij}\right|.$$

Since $|\boldsymbol{y}|\le B$ and $\|\boldsymbol{X}\|_\infty\le B$, so $\max_{ij}|\boldsymbol{ZZ^T}|_{ij}\le B^2$. By Hoeffding's inequality and union bound we get

$$\mathbb{P}\left(\max_{1\le i,j\le(d+1)}\left|\mathbb{E}\left[\boldsymbol{ZZ^T}\right]_{ij}-\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{Z_k Z_k^T}\right)_{ij}\right|\ge t\right)\le 2(d+1)^2\exp\left\{\frac{-2nt^2}{(2B^2)^2}\right\}.$$

Combine two parts and let $t=(1+L)^2\sqrt{2B^4\log\left(2(d+1)^2/\delta\right)/n}$ we have

$$\mathbb{P}\left(\mathscr{R}(\widehat{\beta})-\mathscr{R}(\beta_{ora})\le 2t\right)\ge 1-\delta.$$

7.1.2 LASSO with Cross-validation

Definition 7.3 (LASSO with Cross-validation).

Theorem 7.2 (Persistency of LASSO with Cross-validation).

Proof.

7.2 Model Selection Consistency

In high dimensional linear regression, we usually assume β to be sparse, which makes selecting the variables with nonzero coefficients of great importance. Model selection consistency is to some extent harder to achieve than parameter consistency, since we aim to get $\widehat{\beta}_{\mathscr{S}^c}$ to be exact zero rather than very small. It is highly possible that $\|\widehat{\beta}-\beta\|_2^2$ is small while $\mathrm{supp}(\widehat{\beta})$ and $\mathrm{supp}(\beta)$ are significantly different. Similarly, small prediction error can't guarantee the true recovery of parameter support, neither. While persistency and parameter consistency focus on the loss viewing β as a vector, model selection emphasizes on element-wise results. Though for $\widehat{\beta}_{\mathscr{S}}$, model selection problem has relatively weak requirements, but if the support is sparse, it is usually easy to get an estimator with small prediction error and close to the true parameter by performing standard linear regression w.r.t. the true support.

In the following subsections, we concentrate on LASSO. First we define rigorously what we need in model selection and then give necessary and sufficient conditions of the settings that LASSO can recovery the support. Further we will explore the primal-dual witness (PDW) technique and use it to prove the model selection consistency theorem.

Definition 7.4. (Model Selection Consistency) In LASSO problem, let $\mathscr{S}$ denote the support of β, we are concerned about

1. Uniqueness or uniform properties of solutions;
2. Whether $\text{supp}(\widehat{\beta}) \subseteq \text{supp}(\beta)$;
3. Whether $\text{sign}(\widehat{\beta}_{\mathscr{S}}) = \text{sign}(\beta_{\mathscr{S}})$.

Obviously, difficulty increases with the order. Before further analysis, we first list a few straightforward observations of the solutions:

1. The solution of LASSO is not always unique;
2. $\mathbf{X}\widehat{\beta}$ is unique;
3. Let $\widehat{z} \in \partial\|\widehat{\beta}\|_1$, then for any k,

$$\widehat{z}_k\widehat{\beta}_k = \widehat{\beta}_k \tag{7.2.1}$$

The second term is due to the triangle inequality of l_1 norm and strong convexity of the empirical loss w.r.t. $\mathbf{X}\widehat{\beta}$ as a whole.

Besides above facts, recall that for element-wise properties and optimality of the estimator, we can use KKT condition. Here we write it in block matrix form: there exists $\widehat{z} \in \partial\|\widehat{\beta}\|_1$ s.t.

$$\frac{1}{n}\begin{pmatrix} \mathbf{X}_{\mathscr{S}}^{\mathbf{T}}\mathbf{X}_{\mathscr{S}} & \mathbf{X}_{\mathscr{S}}^{\mathbf{T}}\mathbf{X}_{\mathscr{S}^{\mathbf{c}}} \\ \mathbf{X}_{\mathscr{S}^{\mathbf{c}}}^{\mathbf{T}}\mathbf{X}_{\mathscr{S}} & \mathbf{X}_{\mathscr{S}^{\mathbf{c}}}^{\mathbf{T}}\mathbf{X}_{\mathscr{S}^{\mathbf{c}}} \end{pmatrix}\begin{pmatrix} \widehat{\beta}_{\mathscr{S}} - \beta_{\mathscr{S}} \\ \widehat{\beta}_{\mathscr{S}^c} - \beta_{\mathscr{S}^c} \end{pmatrix} - \frac{1}{n}\mathbf{X}\mathbf{e} + \lambda\begin{pmatrix} \widehat{z}_{\mathscr{S}} \\ \widehat{z}_{\mathscr{S}^c} \end{pmatrix} = \mathbf{0}. \tag{7.2.2}$$

Since $\mathbf{X}_{\mathscr{S}}^{\mathbf{T}}\mathbf{X}_{\mathscr{S}}$ is the empirical estimator of $\text{Cov}(\boldsymbol{X}_{\mathscr{S}}, \boldsymbol{X}_{\mathscr{S}})$, so if $\beta_{\mathscr{S}}$ has no zero coordinate, the corresponding variables should not be highly linear dependent, thus $\mathbf{X}_{\mathscr{S}}^{\mathbf{T}}\mathbf{X}_{\mathscr{S}}$ is likely to be invertible. If that is true then the object function of LASSO w.r.t. $\beta_{\mathscr{S}}$ is stongly convex thus the optimal solution of restricted LASSO exists, is unique and this can actually gives us $\widehat{\beta}_{\mathscr{S}}$ and $\widehat{z}_{\mathscr{S}}$. Now if we assume $\widehat{\beta}_{\mathscr{S}^c}$ to be zero (suppose we know the support), then we can turn to the bottom row of Eq. (7.2.2) and obtain $\widehat{z}_{\mathscr{S}^c}$. If in this setting there exists a LASSO solution that can truly recover the support of β, then the pair $(\widehat{\beta}, \widehat{z})$ derived above should be the only model-selection-consistent primal-dual pair. However, we may not be able to arrive at this exact solution in practise, so we are also interested in other solutions of LASSO, and hopefully, under certain constraints the solution is unique and equals the desired one above. Notice that, to achieve our goal, we only need all solutions $\widetilde{\beta}$ satisfy $\widetilde{\beta}_{\mathscr{S}} = \mathbf{0}$. Recall Eq. (7.2.1) and uniqueness of $\widehat{z}$, if $\|\widehat{z}_{\mathscr{S}}\|_\infty < 1$, then $\widetilde{\beta}_{\mathscr{S}}$ is forced to be zero. To summarize the above analysis, we get a way to construct a primal-dual pair, a sufficient and a necessary condition that LASSO can recover the true support. We write them in the following lemma.

Lemma 7.3 (Primal-dual Witness). *Assume that $\mathbf{X}_{\mathscr{S}}^{\mathbf{T}}\mathbf{X}_{\mathscr{S}}$ is invertible and we know the information of true support, the primal-dual witness technique consists of following three steps*

1. *Get $\widehat{\beta}_{\mathscr{S}}$ by solving a LASSO problem w.r.t. to the variables with nonzero coefficient:*
$$\beta_{\mathscr{S}} \in \operatorname{argmin}\left\{\frac{1}{2n}\|\mathbf{y}-\mathbf{X}_{\mathscr{S}}\beta_{\mathscr{S}}\|_2^2+\lambda\|\beta_{\mathscr{S}}\|_1\right\},$$
and let $\widehat{\beta}_{\mathscr{S}^c}=0$, we get the solution to the primal;
2. *To derive the dual solution, we obtain $\widehat{z}$ through KKT condition*
$$\frac{1}{n}\mathbf{X}^\top\mathbf{X}\widehat{\beta}-\frac{1}{n}\mathbf{y}^\top\mathbf{X}+\lambda\widehat{z}=\mathbf{0}.$$
3. *Check whether $\|\widehat{z}_{\mathscr{S}^c}\|_\infty<1$.*

If $\|\widehat{z}_{\mathscr{S}^c}\|_\infty<1$, then LASSO solution is unique and supp$(\widehat{\beta})\subseteq$ supp(β); If $\|\widehat{z}_{\mathscr{S}^c}\|_\infty>1$, then there is no LASSO solution that satisfies supp$(\widehat{\beta})\subseteq$ supp(β).

Proof. The proof can be obtained directly from above analysis. Since $\|\widehat{z}_{\mathscr{S}^c}\|_\infty<1$, so $\operatorname{supp}(\widehat{\beta})\subseteq\operatorname{supp}(\beta)$. By the KKT condition and shared $\widehat{z}$, we know any LASSO solution $\widetilde{\beta}$ satisfies $\operatorname{supp}(\widetilde{\beta})\subseteq\operatorname{supp}(\beta)$. In this way, we can solve the first row of Eq. (7.2.2) by plugging in $\widehat{z}_{\mathscr{S}}$ and get $\widetilde{\beta}_{\mathscr{S}}=\widehat{\beta}_{\mathscr{S}}$, which implies the uniqueness of solution.

Remark 7.1. Notice that PDW is only a technique and can't be applied in real practise since we don't know the true support. However, PDW can help us prove the following theorem regarding model selection consistency of LASSO.

Theorem 7.4 (Model Selection Consistency).

Proof.

7.3 Consistency