Lecture 7.

3. A pleasant detour: Stein's method of normal approximation

3.1. **Basics on Stein's method of normal approximation.** Let us start with the classic *central limit theorem*: suppose $\{X_j, j \geq 1\}$ is a sequence of *i.i.d* random variables with zero mean and unit variance, then

$$\frac{X_1 + \ldots + X_n}{\sqrt{n}} \text{ converges in law to } \mathcal{N}(0,1), \text{ as } n \to +\infty.$$

A standard proof consists of using the Fourier transform so as to establish the convergence of characteristic functions. However, this analytic proof strongly relies on the independence and leaves us no clue about how fast this distributional convergence happens. Moreover, this Fourier-based approach, in general, can not be applied to situations where some dependence arises. The *Stein's method* is a very powerful toolbox of techniques that can be used not only to prove limit theorems, but also to provide rates of convergence in some chosen metrics, for instance the *Berry-Esséen bound* in the *Kolmogorov distance*, see *e.g.* [NP12, Theorem 3.7.1].

The Stein's method is named after Charles Stein, one of the greatest statisticians in the last century. In 1972, Charles Stein published his method concerning normal approximation in [St72], and after its first appearance, this method has been modified and further developed by Stein himself as well as many other mathematicians. We refer interested readers to the treatise [CGS11] and the lecture notes [Ros11].

3.2. **Basics on Stein's method.** We know from Lemma 1.6 that if $Z \sim \mathcal{N}(0,1)$, $f \in C_b^\infty(\mathbb{R})$, then $\mathbb{E}\big[Zf(Z)\big] = \mathbb{E}\big[f'(Z)\big]$. The converse statement is also true: if Z is integrable and $\mathbb{E}\big[f'(Z)\big] = \mathbb{E}\big[Zf(Z)\big]$ for all $f \in C_b^\infty(\mathbb{R})$, then Z must be distributed as the standard Gaussian. Indeed, considering $f(x) = \sin(\lambda x)$ and $f(x) = \cos(\lambda x)$, $\lambda \in \mathbb{R}$, we have $\mathbb{E}\big[Z\sin(\lambda Z)\big] = \lambda\mathbb{E}\big[\cos(\lambda Z)\big]$ and $\mathbb{E}\big[Z\cos(\lambda Z)\big] = -\lambda\mathbb{E}\big[\sin(\lambda Z)\big]$, from which we deduce that $\mathbb{E}\big[Ze^{\mathrm{i}\lambda Z}\big] = \mathrm{i}\lambda\mathbb{E}\big[e^{\mathrm{i}\lambda Z}\big]$ for any $\lambda \in \mathbb{R}$. Recognize that $\varphi_Z(\lambda) = \mathbb{E}\big[e^{\mathrm{i}\lambda Z}\big]$ is the characteristic function of Z and $\mathrm{i}\mathbb{E}\big[Ze^{\mathrm{i}\lambda Z}\big] = \varphi_Z'(\lambda)$, which gives us an ordinary differential equation $\varphi_Z'(\lambda) + \lambda\varphi_Z(\lambda) = 0$ subject to $\varphi_Z(0) = 1$. This ODE has a unique solution $\varphi_Z(\lambda) = \exp(-\lambda^2/2)$, which is the characteristic function of $\mathcal{N}(0,1)$.

The following result, known as *Stein's lemma*, summarizes the above discussion.

Lemma 3.1 (Stein's lemma)**.** *Let Z be an integrable random variable, then $Z \sim \mathcal{N}(0,1)$ if and only if $\mathbb{E}\big[f'(Z)\big] = \mathbb{E}\big[Zf(Z)\big]$ for any continuously differentiable function $f : \mathbb{R} \to \mathbb{R}$ such that $f'(Z), Zf(Z) \in L^1(\mathbb{P})$, and f, f' have at most* polynomial growth at infinity[16].

Heuristically speaking, the above characterization hints that a real random variable Z is *close* to a standard Gaussian, whenever $\mathbb{E}\big[f'(Z) - Zf(Z)\big]$ is close to zero for every $f \in \mathscr{H}$, with $\mathscr{H}$ being some rich enough class of functions.

Then how to quantify the distance between two probability distributions?

Let $\mathscr{H}$ be a *separating* class[17] of real bounded measurable functions and we define, for two probability measures μ, ν, that $d_{\mathscr{H}}(\mu,\nu) := \sup_{h\in\mathscr{H}} \big| \int_{\mathbb{R}} h\, d\nu - \int_{\mathbb{R}} h\, d\mu \big|$. It is trivial that $d_{\mathscr{H}}$

[16] We say F has at most *polynomial growth at infinity*, if there are some universal constants $C > 0$ and $d \in N^*$ such that $|F(x)| \leq C + C|x|^d$ for each $x \in \mathbb{R}$. The proof of Lemma 3.1 is omitted here, for it is exactly the same as in previous discussion.

[17] that is, for any two different probability measures μ, ν on $\mathbb{R}$, there exists $f \in \mathscr{H}$ such that $\int_{\mathbb{R}} f\, d\nu \neq \int_{\mathbb{R}} f\, d\mu$.

is a metric on the set of probability distributions on $\mathbb{R}$ and we also write $d_{\mathscr{H}}(X,Y) = d_{\mathscr{H}}(\mu,\nu)$ if $X \sim \mu, Y \sim \nu$. For example, $\mathscr{H}_1 := \big\{\mathbf{1}_A : A \in \mathcal{B}(\mathbb{R})\big\}$, $\mathscr{H}_2 := \big\{\mathbf{1}_{(-\infty,x]} : x \in \mathbb{R}\big\}$ and $\mathscr{H}_3 := \big\{f : \|f\|_\infty + \|f'\|_\infty \leq 1\big\}$ are three separating classes; $d_{\mathscr{H}_1}$, $d_{\mathscr{H}_2}$, $d_{\mathscr{H}_3}$ are known as the *total-variation distance* d_{TV}, the *Kolmogorov distance* d_{Kol} and the *Fortet-Mourier distance* d_{FM} respectively. It is known from [Dud10, Section 11.3] that d_{FM} induces the same topology as the weak convergence of probability measures. It is clear that if μ_n, μ are probability measures on $\mathbb{R}$, then the following implications hold in view of the Portemanteau's theorem:

$$\lim_{n\to+\infty} d_{\mathrm{TV}}(\mu_n,\mu) = 0 \quad\Longrightarrow\quad \lim_{n\to+\infty} d_{\mathrm{Kol}}(\mu_n,\mu) = 0 \quad\Longrightarrow\quad \lim_{n\to+\infty} d_{\mathrm{FM}}(\mu_n,\mu) = 0\,.$$

Another important distance that we will use is the *Wasserstein distance* d_{W}, defined by $d_{\mathrm{W}}(\mu,\nu) := \sup\big\{\int_{\mathbb{R}} h\,d\nu - \int_{\mathbb{R}} h\,d\mu \,:\, \|h'\|_\infty \leq 1\big\}$ for two probability measures μ and ν on $\mathbb{R}$. Note that d_{FM} is sometimes called the "bounded Wasserstein distance".

Ingeniously, Charles Stein introduced the following equation:

$$f'(x) - xf(x) = h(x) - \mathbb{E}\big[h(N)\big]\,, \tag{3.1}$$

where $N \sim \mathcal{N}(0,1)$ and $h : \mathbb{R} \to \mathbb{R}$ is a measurable function satisfying $h(N) \in L^1(\mathbb{P})$. The equation (3.1) is known as the Stein's equation with unknown h. We call f a solution to (3.1), if it is *absolutely continuous* and one version of f' satisfies (3.1) everywhere.

Let us now stop to digest the ingenuity of (3.1): suppose that we can solve (3.1) and get nice properties of f in terms of those of h, then we replace the dummy variable x in (3.1) by the random variable Z, so that after taking expectation on both sides, we get

$$\mathbb{E}\big[f'(Z) - Zf(Z)\big] = \mathbb{E}\big[h(Z) - h(N)\big]\,. \tag{3.2}$$

That is to say, in order to get uniform control of the right-hand side, one can instead try to uniformly control the left-hand side, which fits the heuristic after Stein's lemma. In essence, Stein's method replaces the complex-valued characteristic function by the above real characterizing equation, and many coupling methods have been developed so far to deal with the left-hand side of (3.2); see aforementioned references.

The following lemma gives an explicit form for solutions to Stein's equation (3.1).

Lemma 3.2 (Stein's solution). *Let N, h be given as before, the solutions f to (3.1) are given by*

$$f(x) = ce^{x^2/2} + e^{x^2/2}\int_{-\infty}^x \Big[h(y) - \mathbb{E}\big(h(N)\big)\Big]e^{-y^2/2}\,dy\,, \quad x \in \mathbb{R}\,,$$

where $c \in \mathbb{R}$. In particular, the function

$$f_h(x) := e^{x^2/2}\int_{-\infty}^x \Big[h(y) - \mathbb{E}\big(h(N)\big)\Big]e^{-y^2/2}\,dy\,, \quad x \in \mathbb{R} \tag{3.3}$$

is the unique solution to (3.1) verifying $\lim_{|x|\to+\infty}\exp(-x^2/2)f(x) = 0$. We call f_h the Stein's solution to (3.1).

Proof. Noticing that $e^{x^2/2}\dfrac{d}{dx}\big(f(x)e^{-x^2/2}\big) = f'(x) - xf(x) = h(x) - \mathbb{E}\big[h(N)\big]$, we have[18]

$$f(x)e^{-x^2/2} - f(0) = \int_0^x \Big[h(y) - \mathbb{E}\big[h(N)\big]\Big]e^{-y^2/2}\,dy\,.$$

[18] We follow the convention $\int_0^x = -\int_x^0$ if $x < 0$

By the dominated convergence theorem and

$$\int_{\mathbb{R}} \Big(h(y) - \mathbb{E}\big[h(N)\big]\Big)\, e^{-y^2/2}\, dy = 0\,, \tag{3.4}$$

we have $c := \lim_{x\to-\infty} e^{-x^2/2} f(x) = -\lim_{x\to+\infty} e^{-x^2/2} f(x)$. Thus, for any $x \in \mathbb{R}$,

$$f(x)e^{-x^2/2} - c = \int_{-\infty}^{x} \Big[h(y) - \mathbb{E}\big[h(N)\big]\Big] e^{-y^2/2}\, dy\,,$$

which gives us the desired form for the solutions to (3.1). And the rest is trivial. $\square$

Now starting with the expression in (3.3), we study the properties of the Stein's solution f_h when h is 1-Lipschitz function or h is merely bounded measurable:

(i) Suppose that $h : \mathbb{R} \to [0,1]$ is measurable, then it follows from Stein's equation that

$$|f_h'(x)| \le |h(x) - \mathbb{E}[h(N)]| + |xf_h(x)| \le 1 + |xf_h(x)|$$

and from (3.4) that $\int_{-\infty}^{x} \big(h(y) - \mathbb{E}[h(N)]\big)e^{-y^2/2}dy = -\int_{x}^{\infty} \big(h(y) - \mathbb{E}[h(N)]\big)e^{-y^2/2}dy$. Thus,

$|f_h(x)| \le e^{x^2/2} \int_{|x|}^{\infty} e^{-y^2/2}\, dy \le \sqrt{\pi/2}\,$, and $|xf_h(x)| \le |x| \cdot e^{x^2/2} \cdot \int_{|x|}^{\infty} e^{-y^2/2}\, dy \le 1$. So we can conclude that f_h is 2-Lipschitz and uniformly bounded by $\sqrt{\pi/2}$.

(ii) Suppose that $h : \mathbb{R} \to \mathbb{R}$ is K-Lipschitz for some $K \in (0, +\infty)$, then $\|f_h\|_\infty \le 2K$, $\|f_h'\|_\infty \le \sqrt{2/\pi}K$ and $\|f_h''\|_\infty \le 2K$. The proof for these bounds involves a careful analysis, here we just give a very sketchy one and refer interested readers to [CGS11, Lemma 2.4]: first denote the standard Gaussian density, distributional functions by ϕ, Φ respectively, one starts with $h(x) - \mathbb{E}[h(N)] = \int_{\mathbb{R}}[h(x) - h(u)]\phi(u)\, du$, then it follows from an application of Fubini's theorem that $h(x) - \mathbb{E}\big[h(N)\big] = \int_{-\infty}^{x} h'(t)\Phi(t)\, dt - \int_{x}^{\infty} h'(t)[1 - \Phi(t)]\, dt$, together with which (3.3) implies

$$\begin{aligned} f_h(\mathrm{w}) &= e^{\mathrm{w}^2/2} \int_{-\infty}^{\mathrm{w}} \left(\int_{-\infty}^{x} h'(t)\Phi(t)\, dt - \int_{x}^{\infty} h'(t)[1 - \Phi(t)]\, dt\right) e^{-x^2/2}\, dx \\ &= -\sqrt{2\pi} e^{\mathrm{w}^2/2}\big(1 - \Phi(\mathrm{w})\big) \int_{-\infty}^{\mathrm{w}} h'(t)\Phi(t)\, dt \\ &\qquad - \sqrt{2\pi} e^{\mathrm{w}^2/2}\Phi(\mathrm{w}) \int_{\mathrm{w}}^{\infty} h'(t)[1 - \Phi(t)]\, dt \end{aligned}$$

and

$$\begin{aligned} f_h'(\mathrm{w}) &= \mathrm{w} f_h(\mathrm{w}) + h(\mathrm{w}) - \mathbb{E}[h(N)] \\ &= \big\{1 - \sqrt{2\pi}\mathrm{w} e^{\mathrm{w}^2/2}[1 - \Phi(\mathrm{w})]\big\} \int_{-\infty}^{\mathrm{w}} h'(t)\Phi(t)\, dt \\ &\qquad - \big\{1 + \sqrt{2\pi}\mathrm{w} e^{\mathrm{w}^2/2}\Phi(\mathrm{w})\big\} \int_{\mathrm{w}}^{\infty} h'(t)[1 - \Phi(t)]\, dt. \end{aligned}$$

So the first two bounds follow from some standard Gaussian computations. Similar computations can be done for $\|f_h''\|_\infty$ once we derive $f_h''(\mathrm{w}) = (1 + \mathrm{w}^2)f_h(\mathrm{w}) + \mathrm{w}(h(\mathrm{w}) - \mathbb{E}[h(N)]) + h'(\mathrm{w})$ using (3.1).

(iii) Combining the above two points, we can also assert that if $h \in C_b^\infty(\mathbb{R})$ satisfies $0 \le h \le 1$, then $\|f_h\|_\infty \le \sqrt{\pi/2}$, $\|f_h'\|_\infty \le 2$ and $\|f_h''\|_\infty \le 2\|h'\|_\infty < +\infty$.

Now we give Stein's bounds based on the above discussions.

Proposition 3.3. *Let F be a real integrable random variable and $N \sim \mathcal{N}(0,1)$, then*

(1) $d_{\mathrm{TV}}(F,N) \le \sup \big|\mathbb{E}[\varphi'(F) - F\varphi(F)]\big|$, *where the supremum runs over all* $\varphi \in C^2(\mathbb{R})$ *with* $\|\varphi\|_\infty \le \sqrt{\pi/2}$, $\|\varphi'\|_\infty \le 2$ *and* $\|\varphi''\|_\infty < +\infty$.

(2) $d_{\mathrm{W}}(F,N) \le \sup \big|\mathbb{E}[\varphi'(F) - F\varphi(F)]\big|$, *where the supremum runs over all* $\varphi \in C^2(\mathbb{R})$ *with* $\|\varphi\|_\infty \le 2$, $\|\varphi'\|_\infty \le \sqrt{2/\pi}$ *and* $\|\varphi''\|_\infty \le 2$.

Proof. It follows first from *Lusin's theorem*[19] that

$$\sup_{\substack{h \text{ measurable} \\ 0\le h\le 1}} \big|\mathbb{E}\big(h(N) - h(F)\big)\big| = \sup_{\substack{h\in C(\mathbb{R}) \\ 0\le h\le 1}} \big|\mathbb{E}\big(h(N) - h(F)\big)\big|.$$

And for $h : \mathbb{R} \to [0,1]$ continuous, by standard mollification procedure, one can construct $h_\varepsilon \in C_b^\infty(\mathbb{R})$ such that $0 \le h_\varepsilon \le 1$ and $h_\varepsilon \to h$ pointwise, as $\varepsilon \downarrow 0$. So

$$d_{\mathrm{TV}}(F,N) = \sup_{\substack{h\in C_b^\infty(\mathbb{R}) \\ 0\le h\le 1}} \big|\mathbb{E}\big(h(N) - h(F)\big)\big| \le \sup_{\substack{\|\varphi\|_\infty\le\sqrt{\pi/2} \\ \|\varphi'\|_\infty\le 2 \\ \|\varphi''\|_\infty<+\infty}} \big|\mathbb{E}[\varphi'(F) - F\varphi(F)]\big|,$$

where the last inequality follows from point (iii) on last page and (3.2), this proves (1) while (2) follows from point (ii) on last page and (3.2). ☐

Let us now look at an easy example.

Example 1. Let $\mathbf{Y} = (Y_i, i \in \mathbb{N})$ be i.i.d symmetric Bernoulli random variables, then $F_n := n^{-1/2}(Y_1 + \ldots + Y_n)$ converges in law to the standard Gaussian, according to the classic central limit theorem. Now fix $\varphi \in C^2(\mathbb{R})$ with $\|\varphi\|_\infty \le 2$, $\|\varphi'\|_\infty \le \sqrt{2/\pi}$ and $\|\varphi''\|_\infty \le 2$, then we have

$$\begin{aligned}
\mathbb{E}\big[F_n\varphi(F_n)\big] &= \sqrt{n}\,\mathbb{E}\big[Y_1\varphi(F_n)\big] \quad \text{using symmetry} \\
&= \frac{\sqrt{n}}{2}\mathbb{E}\Big[\varphi\big(\frac{1+Y_2+\ldots+Y_n}{\sqrt{n}}\big) - \varphi\big(\frac{-1+Y_2+\ldots+Y_n}{\sqrt{n}}\big)\Big] \quad \text{using independence} \\
&= \frac{\sqrt{n}}{2}\mathbb{E}\Big[\varphi\big(\frac{1+Y_2+\ldots+Y_n}{\sqrt{n}}\big) - \varphi(F_n) + \varphi(F_n) - \varphi\big(\frac{-1+Y_2+\ldots+Y_n}{\sqrt{n}}\big)\Big] \\
&= \mathbb{E}\big[\varphi'(F_n)\big] + \frac{\sqrt{n}}{2}R_n \quad \text{with } |R_n| \le 2\|\varphi''\|_\infty n^{-1},
\end{aligned} \tag{3.5}$$

where the last line follows from the usual Taylor expansion. Thus, it follows from Proposition 3.3 that[20] $d_{\mathrm{W}}(F_n, \mathcal{N}(0,1)) \le 2n^{-1/2}$. **Exercise**: work out the details for (3.5).

In the above toy-example, independence and symmetry within the structure contribute to the easy proof. Nevertheless, the strategy of combining an application of Taylor expansion with Stein's equation is usually effective in practice; see, e.g., our Proposition 3.6. Stein's ideas have contributed to many excellent results and solutions to numerous problems, such as local dependence, minimal spanning trees, concentration inequalities and many others; see, e.g., Barbour and Chen's Volumes [BC04, BC05] and S. Chatterjee's ICM survey [Cha14a].

$\propto\propto\propto\propto\propto\propto$ **End of Lecture 7 (Sept. 23, 2025)** $\infty\infty\infty\infty\infty\infty\infty\infty$

[19] Let $\mu = (\mathbb{P}\circ F^{-1} + \mathbb{P}\circ N^{-1})/2$, then for h measurable with values in $[0,1]$, by appropriate application of Lusin's Theorem (*e.g.* see the *Red Rudin*) one can find a sequence of continuous functions h_n such that $0 \le h_n \le 1$ and $h_n \to h$ μ-a.s., thus $h_n \to h$ $\mathbb{P}\circ F^{-1}$-a.s. and $\mathbb{P}\circ N^{-1}$−a.s..

[20] By explicit calculation, $\mathbb{E}[F_n^4] - 3 = -2n^{-1}$, so we obtain the fourth moment bound $d_{\mathrm{W}}(F_n, \mathcal{N}(0,1)) \le \sqrt{2(3 - \mathbb{E}[F_n^4])}$.