Stein's method for Normal Approximations

11.1 Weak convergence and the classical CLT

One of the most celebrated results of probability theory says that if $S_n = \xi_1 + \xi_2 + \ldots + \xi_n$ is a sum of i.i.d. random variables with mean 0 and variance 1, then for large n the distribution of $n^{-1/2}S_n$ can roughly be approximated by a standard Gaussian distribution on the real line. In other words, as $n \to \infty$, the distribution of $n^{-1/2}S_n$ approaches to standard normal distribution distribution with density $(2\pi)^{-1/2}e^{-x^2/2}$ for $x \in \mathbb{R}$.

We have proved this result using Characteristic function approach and Lindeberg technique. However, the proof breaks down for sums of "weakly dependent" random variables. Martingale central limit theorem gives one example where CLT holds for dependent sums. Stein's method, introduced by Charles Stein (1972) and developed by later researchers, is a way of proving CLT for dependent sums that also gives explicit rates of convergence. Recall that,

Definition 11.1. *For P_n, P be two probability measures on $\mathbb{R}$, we say that P_n converges weakly to P, which we write by $P_n \implies P$, if $\int_{\mathbb{R}} f dP_n \to \int_{\mathbb{R}} f dP$ for all continuous bounded functions f.*

For a random variable $X : (\Omega, \mathcal{F}, \mu) \to \mathbb{R}$, by $\mathcal{L}(X)$, we mean the probability measure induced by X on $\mathbb{R}$, *i.e.,* $\mu \circ X^{-1}$. We say that $X \implies Y$ if $\mathcal{L}(X) \implies \mathcal{L}(Y)$.

11.2 Distances between probability measures

The space of probability is metrizable under the weak convergence. One such metric is known as Prohorov metric which is given below.

$$d(P, Q) = \inf\{\varepsilon > 0 : P(A) \leqslant Q(A^\varepsilon) + \varepsilon, Q(A) \leqslant P(A^\varepsilon) + \varepsilon \ \forall A \in \mathcal{B}(\mathbb{R})\},$$

where $A^\varepsilon = \{x : d(x, A) < \varepsilon\}$. As it turns out the above metric is not very useful for calculations. There are several other distances between two probability measures which we can work with and are stronger than the metric of weak convergence. A typical distance between probability measures can be of the following type

$$d_{\mathcal{D}}(P, Q) = \sup\left\{\left|\int f dP - \int f dQ\right| \ : \ f \in \mathcal{D}\right\}, \tag{11.1}$$

where $\mathcal{D}$ is some class of $\mathbb{R}$-valued functions. Note that $d(\cdot, \cdot)$ defined above is always a proper metric as long as $\mathcal{D}$ is a separating class, i.e.

$$\int f dP = \int f dQ \ \ \forall f \in \mathcal{D} \implies P = Q.$$

11.2.1 Total variation distance

Let us take $\mathcal{D}$ to be the class of indicator functions of Borel sets of $\mathbb{R}$ in (11.1). The resulting metric is known as the total variation distance. Thus the total variation distance between two probability measures P and Q on S is given by

$$d_{\mathrm{TV}}(P,Q) = \sup_{A\in\mathcal{B}(\mathbb{R})} |P(A)-Q(A)| = \sup_{h:\ \|h\|_\infty\leqslant 1} \frac{1}{2}\left|\int h dP - \int h dQ\right|.$$

Note that this ranges in $[0,1]$.

11.2.2 Wasserstein distance

This is also known as the Kantorovich-Monge-Rubinstein metric. Here $\mathcal{D}$ = set of all 1-Lipschitz functions on $\mathbb{R}$.

$$d_{\mathrm{W}}(P,Q) := \sup\left\{\left|\int f\,dP - \int f\,dQ\right| : f \text{ is 1-Lipschitz}\right\}.$$

The Wasserstein distance can range in $[0,\infty]$.

11.2.3 Kolmogorov-Smirnov distance

It is only defined for probability measures on $\mathbb{R}$.

$$d_{\mathrm{K}}(P,Q) := \sup_{x\in\mathbb{R}} \left|P\left((-\infty,x]\right) - Q\left((-\infty,x]\right)\right|. \tag{11.2}$$

It also falls under the same general framework of (11.1) when $\mathcal{D} = \{\mathbf{1}_{(-\infty,x]} : x\in\mathbb{R}\}$.

Given any metric d on the set of all probability measures on $\mathbb{R}$ and random variables X, Y, we will abuse notation by writing $d(X,Y)$ for $d(\mathcal{L}(X),\mathcal{L}(Y))$.

Remark 11.2.

- *All the distances defined above are stronger than weak convergence. That is, if any of these distances go to zero as $n\to\infty$, then we have weak convergence. But converse is not true.*
- *Total variation is a very strong notion, often too strong to be useful. Suppose $X_1, X_2, \ldots$ i.i.d. ± 1 with probability $1/2$ each. $S_n = \sum_1^n X_i$. Then*

$$\frac{S_n}{\sqrt{n}} \implies \mathrm{N}(0,1).$$

But $d_{TV}(\frac{S_n}{\sqrt{n}}, Z) = 1$ for all n. But both Wasserstein and Kolmogorov distances go to 0 at rate $n^{-1/2}$.

11.3 Stein's method for normal approximation

Stein's method was introduced by Charles Stein in the early 70's to prove central limit theorem for dependent random variables and more importantly to find explicit estimates of the accuracy of the approximation. Instead of using characteristic functions he used, what is now called Stein's Characterizing operator, to convert the 'global problem' of analyzing the whole random variable into more tractable 'local problems'.

Suppose we want to show that the random variable X has a distribution which is approximately equal to that of another random variable Z. As noted in the previous section, the distributional proximity of two random variables is measured by $\sup_{g\in\mathcal{D}} |\mathbb{E}\, g(X) - \mathbb{E}\, g(Z)|$ for a large family of test functions $\mathcal{D}$. Various choices of family $\mathcal{D}$ lead to different notions of distances between two probability measures. Given $\mathcal{D}$ and the target distribution Z, the main goal of Stein's method is to compute upper bounds of the complicated looking object $\sup_{g\in\mathcal{D}} \mathbb{E}\,|g(X) - g(Z)|$. The basic theme behind the Stein's method will be as follows

1. Identify the "Stein Characterizing operator" for Z, that is, find a suitable operator $\mathcal{A}$ such that for any random variable W, $\mathbb{E}\,\mathcal{A}f(W) = 0$ for all functions f belonging to a suitably large class $\mathcal{D}$ if and only if $W \stackrel{d}{=} Z$. For example, if Z has a standard normal distribution, then

$$(\mathcal{A}f)(x) = f'(x) - xf(x) \text{ for} f \in \mathcal{D}$$

   where $\mathcal{D}$ = set of all locally absolutely continuous functions, is a Stein's characterizing operator.

2. Invert the Stein's operator, that is, for a given function $g \in \mathcal{D}$, find f such that $\mathcal{A}f(x) = g(x) - Eg(Z)$. Establish the smoothness properties of f in terms of g. It is often the case that $\mathcal{D}'$ has more smoothness properties than $\mathcal{D}$. Let $\mathcal{D}'$ be a class of ('smooth') functions so that for each $g \in \mathcal{D}$, there exists $f \in \mathcal{D}'$. For instance, $\mathcal{D} = \{g : g \text{ 1-Lipschitz}\}$ will give rise to Wasserstein metric. We will show that if Z is standard normal, we may take $\mathcal{D}' = \{f : |f|_\infty \leqslant 1, |f'|_\infty \leqslant \sqrt{\frac{2}{\pi}} \text{ and } |f''|_\infty \leqslant 2\}$. By that, we mean that given $g \in \mathcal{D}$ with $\mathbb{E}\, g(Z) = 0$, we can get $f \in \mathcal{D}'$ so that $\mathcal{A}f = g$.

3. $|\mathbb{E}\, g(X) - \mathbb{E}\, g(Z)| = |\mathbb{E}\,\mathcal{A}f(X)|$. So, $\sup_{g\in\mathcal{D}} |\mathbb{E}\, g(X) - \mathbb{E}\, g(Z)| \leqslant \sup_{f\in\mathcal{D}'} |\mathbb{E}\,\mathcal{A}f(X)|$. Since $\mathcal{A}$ is a characterizing operator, $\mathbb{E}\,\mathcal{A}f(Z) = 0$. But if distribution of X is close to that of Z, we should hope that $\mathbb{E}\,\mathcal{A}f(X) \approx 0$. As it turns out in practice that bounding $\sup_{f\in\mathcal{D}'} |\mathbb{E}\,\mathcal{A}f(X)|$ is easier than bounding $\sup_{g\in\mathcal{D}} |\mathbb{E}\, g(X) - \mathbb{E}\, g(Z)|$ itself. In normal example, we have an upper bound in Wasserstein metric as follows

$$d_{\mathrm{W}}(X, Z) \leqslant \sup_{f\in\mathcal{D}'} |\mathbb{E}(f'(X) - Xf(X))|.$$

Let us now analyze the case of normal distribution.

Lemma 11.3 (Stein's Lemma)**.** *A random variable W has a standard normal distribution iff for every piecewise continuously differentiable function $f : \mathbb{R} \to \mathbb{R}$ with $\mathbb{E}\,|f'(Z)| < \infty, Z \sim \mathrm{N}(0, 1)$ we have $\mathbb{E}Wf(W) = \mathbb{E}f'(W)$.*

11.4 Inversion of the Stein operator

If we define the operator $T : \mathcal{F} \to \mathcal{C}$ where $\mathcal{F}$ is the set of all piecewise continuously differentiable function f with $\mathbb{E}\,|f'(Z)| < \infty, Z \sim \mathrm{N}(0,1)$ and $\mathcal{C}$ is the set of all continuous functions, as $Tf(x) = f'(x) - xf(x)$, then the above lemma says that

$$W \sim \mathrm{N}(0,1) \text{ **iff** } \mathbb{E}\,Tf(W) = 0 \text{ for all } f \in \mathcal{F}.$$

We can call T the characterizing operator for standard normal distribution.

In this section we will study the solution of the differential equation

$$f'(x) - xf(x) = g(x) - \mathbb{E}\,g(Z), \quad Z \sim \mathrm{N}(0,1). \tag{11.3}$$

Lemma 11.4. *Given function $g : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}\,|g(Z)| < \infty$ where $Z \sim \mathrm{N}(0,1)$,*

$$f(x) = e^{x^2/2} \int_{-\infty}^{x} e^{-y^2/2}(g(y) - \mathbb{E}\,g(Z))dy \tag{11.4}$$

is an absolutely continuous solution of (11.3). Moreover, any a.c. solution $\tilde{f}$ of (11.3) is of the form

$$\tilde{f}(x) = f(x) + ce^{x^2/2}, \quad c \in \mathbb{R}.$$

Finally, f is the only solution that satisfies $\lim_{|x|\to\infty} f(x)e^{-x^2/2} = 0$.

Proof. By the method of integrating factors, we have that if f is a solution to (11.3), then

$$\frac{d}{dx}(e^{-x^2/2}f(x)) = e^{-x^2/2}(f'(x) - xf(x)) = e^{-x^2/2}(g(x) - \mathbb{E}\,g(Z)).$$

So, (11.4) is a reasonable candidate as a solution of (11.3). And it is easy to verify directly that (11.4) indeed satisfies (11.3). If $\tilde{f}$ is any other solution of (11.3), then

$$\frac{d}{dx}\left(e^{-x^2/2}(f(x) - \tilde{f}(x))\right) = 0.$$

Hence, $\tilde{f}(x) = f(x) + ce^{x^2/2}$ for some $c \in \mathbb{R}$. Clearly, from definition and DCT, we have $\lim_{x\to-\infty} f(x)e^{-x^2/2} = 0$. Note that since $Z \sim \mathrm{N}(0,1)$, we have $\int_{-\infty}^{\infty} e^{-y^2/2}(g(y) - \mathbb{E}\,g(Z))dy = 0$. So, f can also be written as follows

$$f(x) = -e^{x^2/2} \int_{x}^{\infty} e^{-y^2/2}(g(y) - \mathbb{E}\,g(Z))dy. \tag{11.5}$$

Therefore, by DCT, $\lim_{x\to+\infty} f(x)e^{-x^2/2} = 0$. ∎

When the function g is Lipschitz we can write the solution in a different way, which will help us to find optimal bounds.

Lemma 11.5. *Assume g is Lipschitz. Then*

$$f(x) = -\int_0^1 \frac{1}{2\sqrt{t(1-t)}}\,\mathbb{E}\left[Zg(\sqrt{t}x + \sqrt{1-t}Z)\right]dt, \quad Z \sim \mathrm{N}(0,1) \tag{11.6}$$

is a solution of (11.3). In fact, it is the same as (11.4).

Proof. Let g is C-Lipschitz. Then[1] $|g'|_\infty \leqslant C$. On differentiating f and carrying the derivative inside the integral and expectation which can be justified using DCT, we have

$$f'(x) = -\int_0^1 \frac{1}{2\sqrt{1-t}}\,\mathbb{E}\left[Zg'(\sqrt{t}x + \sqrt{1-t}Z)\right] dt. \tag{11.7}$$

On the other hand, the Stein identity gives us

$$\mathbb{E}\left[Zg(\sqrt{t}x + \sqrt{1-t}Z)\right] = \sqrt{1-t}\,\mathbb{E}\left[g'(\sqrt{t}x + \sqrt{1-t}Z)\right].$$

Thus,

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

■

Now that we know the form of the solution to (11.3) the next obvious question is what can we say about the solution when g has some nice properties like boundedness or differentiability. Our next theorem says that indeed the solution f inherits the 'niceness' of the function g. Recall the notation that $Ng := \mathbb{E}\,g(Z)$.

Lemma 11.6. *If $g : \mathbb{R} \to \mathbb{R}$ is bounded,*

$$\text{I. } |f|_\infty \leqslant \sqrt{\frac{\pi}{2}}\,|g - Ng|_\infty \quad \textit{and}$$
$$\text{II. } \left|f'\right|_\infty \leqslant 2\,|g - Ng|_\infty.$$

and if g is Lipschitz, but not necessarily bounded, then

$$\text{III. } |f|_\infty \leqslant \left|g'\right|_\infty,$$
$$\text{IV. } \left|f'\right|_\infty \leqslant \sqrt{\frac{2}{\pi}}\,\left|g'\right|_\infty \quad \textit{and}$$
$$\text{V. } \left|f''\right|_\infty \leqslant 2\,\left|g'\right|_\infty.$$

Here for a function f we define $|f|_\infty := \sup\{|f(x)| : x \in \mathbb{R}\}$. Moreover, all the above bounds are tight. The bounds (I), (II) and (V) were obtained by Stein.

The proof is given in the Appendix 11.8.

11.4.1 Example: Ordinary CLT in the Wasserstein metric

Suppose $X_1, X_2, \ldots, X_n$ are independent, mean 0, variance 1, $\mathbb{E}\,|X_i|^3 < \infty$. Let $S_n = \sum_1^n X_i$ and $W = n^{-1/2}S_n$. Take any $f \in C^1$ with f' absolutely continuous, and satisfying $|f| \leqslant 1$, $|f'| \leqslant \sqrt{2/\pi}$ and $|f''| \leqslant 2$. First, note that

$$\mathbb{E}\,Wf(W) = n^{-1/2} \sum \mathbb{E}\,(X_i f(W)).$$

[1] Any Lipschitz function g is absolutely continuous. Hence, it is (Lebesgue) almost surely differentiable. Define g' to be derivative of g at the points where it exists and 0 elsewhere.

Now define

$$W_i = W - n^{-1/2}X_i = n^{-1/2}\sum_{j\neq i} X_j, \quad i = 1, 2, \ldots, n.$$

Observe that X_i, W_i are independent. Thus

$$\mathbb{E}\, X_i f(W_i) = \mathbb{E}\, X_i\, \mathbb{E}\, f(W_i) = 0$$

and so, we can write

$$\begin{aligned}\mathbb{E}\,(X_i f(W)) &= \mathbb{E}\,(X_i\,(f(W) - f(W_i))) \\ &= \mathbb{E}\left(X_i\left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right) + \mathbb{E}\left[X_i(W - W_i)f'(W_i)\right].\end{aligned}$$

Note that

$$\left|f(b) - f(a) - (b-a)f'(a)\right| \leqslant \frac{1}{2}(b-a)^2\left|f''\right|_\infty$$

and that $W - W_i = n^{-1/2}X_i$. Thus

$$\begin{aligned}&\left|\mathbb{E}\left[X_i\left(f(W) - f(W_i) - n^{-1/2}X_i f'(W_i)\right)\right]\right| \\ &\leqslant \frac{1}{2}\,\mathbb{E}\left|X_i(n^{-1/2}X_i)^2\right| \cdot \left|f''\right|_\infty \leqslant n^{-1}\,\mathbb{E}\,|X_i|^3\,.\end{aligned}$$

Again,

$$\mathbb{E}\left[X_i\,(W - W_i)\,f'(W_i)\right] = n^{-1/2}\,\mathbb{E}\,X_i^2 f'(W_i) = n^{-1/2}\,\mathbb{E}\,f'(W_i)$$

since $\mathbb{E}\,X_i^2 = 1$ and X_i is independent of W_i.

From (**??**) and the above calculation we see that

$$\left|\mathbb{E}\,W f(W) - n^{-1}\sum_{i=1}^n \mathbb{E}\,f'(W_i)\right| \leqslant n^{-3/2}\sum_{i=1}^n \mathbb{E}\,|X_i|^3\,.$$

Finally, note that

$$\begin{aligned}\left|n^{-1}\sum_{i=1}^n \mathbb{E}\,f'(W_i) - \mathbb{E}\,f'(W)\right| &\leqslant n^{-1}|f''|_\infty \sum_{i=1}^n \mathbb{E}\,|W - W_i| \\ &= n^{-3/2}\left|f''\right|_\infty \sum_{i=1}^n \mathbb{E}\,|X_i| \leqslant 2n^{-3/2}\sum_{i=1}^n \mathbb{E}\,|X_i|\,.\end{aligned}$$

Combining, we have

$$\left|\mathbb{E}\,f(W)W - \mathbb{E}\,f'(W)\right| \leqslant n^{-3/2}\sum_{i=1}^n \mathbb{E}\,|X_i|^3 + 2n^{-3/2}\sum_{i=1}^n \mathbb{E}\,|X_i|\,.$$

Since $\mathbb{E}\,X_i^2 = 1$ we can conclude that $\mathbb{E}\,|X_i|^3 \geqslant 1$ and hence $\mathbb{E}\,|X_i| \leqslant \left(\mathbb{E}\,|X_i|^3\right)^{1/3} \leqslant \mathbb{E}\,|X_i|^3$.

We have now arrived at a 'Berry-Esséen bound' for the Wasserstein metric:

Theorem 11.7. *Suppose $X_1, \ldots, X_n$ are independent with mean 0, variance 1, and finite third moments. Then*

$$d_W(n^{-1/2}\sum_{i=1}^n X_i,\, Z) \leqslant \frac{3}{n^{3/2}}\sum_{i=1}^n \mathbb{E}\,|X_i|^3\,,$$

where $Z \sim \mathrm{N}(0,1)$.

11.5 Dependency graph approach

There are plenty of examples in the literature that suggest that the CLT for the sums of the random variables goes well beyond the regime of independent random variables. As a natural generalization to i.i.d. random variables, we can think of m-dependent sequences for which it is well known that CLT holds as well. In some sense the fact that for a m-dependent sequence the variables that are 'far apart' are independent makes us believe that the CLT should hold and it is indeed the case. In the section, we try to formalize this vague notion of 'local dependence structure' for a general family of random variables and develop procedures to extract CLT out of it, using the example of i.i.d. sum of r.v.s.

Let's now a define what we mean by a dependency graph.

Definition 11.8. *Let $\{X_v\}_{v\in V}$ be a family of random variables. A dependency graph for $\{X_v\}$ is any graph G with vertex set V such that if A and B are two disjoint subsets of V so that there is no edge of G between any vertex in A to any vertex in B, then the families $\{X_v\}_{v\in A}$ and $\{X_v\}_{v\in B}$ are mutually independent.*

For a given family of dependent random variables $\{X_v\}_{v\in V}$, suppose we want to find a CLT for $\sum_{v\in V} X_v$. The very basic idea behind dependency graph is to compactly encode the informations about the dependence between the random variables. If we can find a dependency graph which is sparse, then that ensures that there is a lot of independence in the family $\{X_v\}$ which can be exploited by Stein's method to prove the CLT. Of course, the choice of the dependency graph is not unique, for the complete graph serves as a dependency graph for all families of random variables. But in practical examples, it often turns out that there is a natural choice of dependency graph which is optimal in a sense that it can not be made any sparser by deleting edges.

Given $u \in V$, let $\overline{\mathcal{N}}(u)$ denote the set of all vertices consisting of all the neighbors of u in G and u itself. Let $D := 1 +$ max degree $= \max_{u\in G} |\overline{\mathcal{N}}(u)|$. We have the following theorem.

Theorem 11.9. *Let $\{X_v\}_{v\in V}$ be a family of random variables with dependency graph G. Let $W = \sigma^{-1}\sum_{v\in V} X_v$, where $\sigma^2 = \mathrm{Var}(\sum_{v\in V} X_v)$. Also assume that $\mathbb{E}\, X_v = 0, v \in V$. Then*

$$d_W(W, Z) \leqslant \sigma^{-3} \sum_{v\in V} \sum_{u,w\in\overline{\mathcal{N}}(v)} \Big(3\,\mathbb{E}\,|X_v X_u X_w| + 4\,\mathbb{E}\,|X_v X_u|\,\mathbb{E}\,|X_w|\Big), \tag{11.8}$$

where $Z \sim \mathrm{N}(0, 1)$.

Corollary 11.10. *If, in Theorem 11.9, we additionally assume $E|X_v|^3 < \infty$ for all $v \in V$ then*

$$d_W(W, Z) \leqslant \frac{7D^2}{\sigma^3} \sum_{v\in V} \mathbb{E}\,|X_v|^3. \tag{11.9}$$

Proof of Theorem 11.9. Let $Z_v = \sigma^{-1}\sum_{u\in\overline{\mathcal{N}}(v)} X_u$ and $W_v = \sigma^{-1}\sum_{u\notin\overline{\mathcal{N}}(v)} X_u$ so that $W = W_v + Z_v$ for each $v \in V$. We can further decompose W_v as $W_v = W_{vu} + Y_{vu}, v \in V, u \in \overline{\mathcal{N}}(v)$, where $W_{vu} = \sigma^{-1}\sum_{w\notin\overline{\mathcal{N}}(v)\cup\overline{\mathcal{N}}(u)} X_w$ and $Y_{vu} = \sigma^{-1}\sum_{w\in\overline{\mathcal{N}}(u)\setminus\overline{\mathcal{N}}(v)} X_w$. Note that W_v is independent of X_v and W_{vu} is independent of (X_v, X_u).

Take any function f with $|f''|_\infty \leqslant 2$. We can write

$$\begin{aligned}\mathbb{E}\, Wf(W) - \mathbb{E}\, f'(W) &= \Big\{\mathbb{E}\, Wf(W) - \sigma^{-1}\sum_{v\in V}\mathbb{E}\, X_v Z_v f'(W_v)\Big\}\\ &\quad + \Big\{\sigma^{-1}\sum_{v\in V}\mathbb{E}\, X_v Z_v f'(W_v) - \sigma^{-2}\sum_{v\in V}\sum_{u\in\overline{\mathcal{N}}(v)}\mathbb{E}(X_vX_u)\,\mathbb{E}\, f'(W_{vu})\Big\}\\ &\quad + \sigma^{-2}\sum_{v\in V}\sum_{u\in\overline{\mathcal{N}}(v)}\Big\{\mathbb{E}(X_vX_u)[\mathbb{E}\, f'(W_{vu}) - \mathbb{E}\, f'(W)]\Big\}\end{aligned}$$

where we have used the fact that

$$\sigma^{-2}\sum_{v\in V}\sum_{u\in\overline{\mathcal{N}}(v)}\mathbb{E}(X_vX_u) = \sigma^{-1}\sum_{v\in V}\mathbb{E}\, X_vZ_v = \sigma^{-1}\sum_{v\in V}\mathbb{E}\, X_vW = \mathbb{E}\, W^2 = 1.$$

We are now going to bound each of the three term using Taylor's expansion. For the first term, we have

$$Wf(W) = \sigma^{-1}\sum_{v\in V}X_vf(W) = \sigma^{-1}\sum_{v\in V}\Big\{X_vf(W_v) + X_vZ_vf'(W_v) + \frac{1}{2}X_vZ_v^2f''(W_v^*)\Big\},$$

for some random variable W_v^* between W and W_v. Thus, by using that $\mathbb{E}\, X_vf(W_v) = \mathbb{E}\, X_v\,\mathbb{E}\, f(W_v) = 0$, we obtain

$$\begin{aligned}\Big|\mathbb{E}\, Wf(W) - \sigma^{-1}\sum_{v\in V}\mathbb{E}\, X_vZ_vf'(W_v)\Big| &\leqslant \sigma^{-1}\sum_{v\in V}\mathbb{E}(|X_v|Z_v^2)\\ &\leqslant \sigma^{-3}\sum_{v\in V}\sum_{u,w\in\overline{\mathcal{N}}(v)}\mathbb{E}\,|X_vX_uX_w|.\end{aligned}$$

Again by Taylor's expansion,

$$\begin{aligned}X_vZ_vf'(W_v) &= \sigma^{-1}\sum_{u\in\overline{\mathcal{N}}(v)}X_vX_uf'(W_v)\\ &= \sigma^{-1}\sum_{u\in\overline{\mathcal{N}}(v)}X_vX_u\{f'(W_{vu}) + Y_{vu}f''(W_{vu}^*)\},\end{aligned}$$

where W_{vu}^* is some random variable between W_v and W_{vu}. This gives

$$\begin{aligned}&\Big|\sigma^{-1}\sum_{v\in V}\mathbb{E}\, X_vZ_vf'(W_v) - \sigma^{-2}\sum_{v\in V}\sum_{u\in\overline{\mathcal{N}}(v)}\mathbb{E}(X_vX_u)\,\mathbb{E}\, f'(W_{vu})\Big|\\ &\leqslant 2\sigma^{-2}\sum_{v\in V}\sum_{u\in\overline{\mathcal{N}}(v)}\mathbb{E}\,|X_vX_uY_{vu}| = 2\sigma^{-2}\sum_{u\in V}\sum_{v\in\overline{\mathcal{N}}(u)}\mathbb{E}\,|X_vX_uY_{vu}|\\ &\leqslant 2\sigma^{-3}\sum_{u\in V}\sum_{v\in\overline{\mathcal{N}}(u)}\sum_{w\in\overline{\mathcal{N}}(u)}\mathbb{E}\,|X_vX_uX_w|.\end{aligned}$$

Now since $W_{vu} = W_v - Y_{vu} = W - (Z_v + Y_{vu})$, we have

$$f'(W_{vu}) = f'(W) - (Z_v + Y_{vu})f'(W_{vu}^{**})$$

for some W_{vu}^{**} between W_{vu} and W. Therefore,

$$\begin{aligned}|\,\mathbb{E}\,f'(W_{vu}) - \mathbb{E}\,f'(W)| &\leqslant 2\,\mathbb{E}\,|Z_v + Y_{vu}| \\ &\leqslant 2\sigma^{-1} \sum_{w \in \overline{\mathcal{N}}(v) \cup \overline{\mathcal{N}}(u)} \mathbb{E}\,|X_w|.\end{aligned}$$

Putting the ingredients together, we get the result. ■

Remark 11.11. *Though the above theorem gives a relatively tight error bound, it is hard to believe that it is optimal as it only uses the fact $|f''|_\infty$ is bounded but it never used the boundedness of $|f'|_\infty$.*

11.5.1 Sum of independent random variables

Let's start with the simplest of an example when the family $\{X_i\}_{1 \leqslant i \leqslant n}$ consists of independent random variables with $\mathbb{E}\,X_i = 0$, $\mathrm{Var}(X_i) = 1$ and $\mathbb{E}\,|X_i|^3 < \infty$. Define $W = n^{-1/2} \sum_{i=1}^n X_i$. In this case, the graph G with the vertex set $V = \{1, 2, \ldots, n\}$ and with no edges is a straightforward choice for the dependency graph. For this graph G, $\overline{\mathcal{N}}(i)$ is simply $\{i\}$ itself and $D = 1$. Also $\sigma^2 = \mathrm{Var}(\sum_{i=1}^n X_i) = \sqrt{n}$. To prove CLT, we can now direct apply Corollary 11.10 to obtain the following bound on the Wasserstein distance between W and Z,

$$\mathrm{Wass}\big(W, Z\big) \leqslant \frac{7 \max_i \mathbb{E}\,|X_i|^3}{\sqrt{n}}.$$

11.5.2 m-dependent sequence

The m-dependent sequence is one of the most natural example one can think of where the method dependency graph technique can be applied.

Definition 11.12. *A sequence of a random variables $\{X_i : i \geqslant 0\}$ is called m-dependent if $\{X_i : i \leqslant t\}$ and $\{X_i : i \geqslant s\}$ are independent whenever $s - t > m$.*

Definition 11.13. *A sequence $\{X_i : i \geqslant 0\}$ is said to be weakly stationary if*

1. $\mathbb{E}\,X_i^2 < \infty \quad \forall i \geqslant 0.$
2. $\mathbb{E}\,X_i = \mu \quad \forall i \geqslant 0$ *for some* $\mu \in \mathbb{R}$.
3. $\mathrm{Cov}(X_{r+t}, X_{s+t}) = \mathrm{Cov}(X_r, X_s) \quad \forall r, s, t \geqslant 0.$

For a weakly stationary sequence $\{X_i : i \geqslant 0\}$ let us define the autocovariance function γ as

$$\gamma(h) := \mathrm{Cov}(X_h, X_0) \quad \text{for } h \geqslant 0.$$

Note that the weak-stationarity implies that $\gamma(h) = \mathrm{Cov}(X_{r+h}, X_r)$ for all $r \geqslant 0$.

Next we will illustrate an application of Theorem 11.9 to give an easy prove of the central limit theorem for the sum of weakly stationary m-dependent process.

Theorem 11.14. *Fix $m \geqslant 0$. Let $\{X_i : i \geqslant 0\}$ be a weakly stationary m-dependent sequence. Assume that $\nu_m := \gamma(0) + 2\sum_{i=1}^{m}\gamma(i) \neq 0$ and $\sup_{i\geqslant 0}\mathbb{E}\,|X_i|^3 < \infty$. Then*

$$d_W\left(\frac{\sum_{i=0}^{n-1} X_i - n\mu}{\sqrt{\mathrm{Var}(\sum_{i=0}^{n-1} X_i)}}, Z\right) \leqslant \frac{7(2m+1)^2 \cdot n \cdot \max_i \mathbb{E}\,|X_i|^3}{\sigma^3},$$

where μ is the common mean of X_i and Z, as usual, denote $\mathrm{N}(0,1)$ random variable and

$$\sigma^2 = n\nu_m - 2\sum_{i=1}^{m} i\gamma(i).$$

Proof. Without loss of generality, we can assume $\mu = 0$. From the definition of m-dependent sequence, it is almost immediate how to construct a dependency graph for $\{X_i : 0 \leqslant i \leqslant n-1\}$. The vertex set is $V = \{0, 1, 2, \ldots, n-1\}$ and vertex i and vertex j is connected by an edge if and only if $1 \leqslant |i-j| \leqslant m$. For $i \in V$, we have

$$\overline{\mathcal{N}}(i) = \{i-m, i-m+1, \ldots, i-1, i, i+1, \ldots, i+m\} \cap V,$$

so that $D = 2m+1$.

From the weak-stationarity of $\{X_i; i \geqslant 0\}$

$$\sigma^2 := \mathrm{Var}\left(\sum_{i=0}^{n-1} X_i\right) = \sum_{i=0}^{n-1}\sum_{j=0}^{n-1}\mathbb{E}\,X_iX_j = \sum_{i=1}^{n-1}\sum_{j\in\overline{\mathcal{N}}(i)}\mathbb{E}\,X_iX_j = n\nu_m - 2\sum_{i=1}^{m} i\cdot\gamma(i). \tag{11.10}$$

Note that as $\nu_m \neq 0$, we have $\sigma = \Omega(n^{1/2})$. Define $W = \sigma^{-1}\sum_{i=0}^{n-1} X_i$. We can again apply Corollary 11.10 to conclude

$$d_{\mathrm{W}}(W, Z) \leqslant \frac{7(2m+1)^2 n \max_i \mathbb{E}\,|X_i|^3}{\sigma^3}.$$

Hence the proof is complete. ∎

11.5.3 Number of triangles in $\mathbb{G}(n,p)$

Let $\mathbb{G}(n,p)$ be the Erdös Rényi random graph on n vertices where each of the $\binom{n}{2}$ possible edges is included in the graph independently with probability p. Let $T = T_n$ denote the number of triangles in $\mathbb{G}(n,p)$. Define $W = \sigma^{-1}(T - \mathbb{E}\,T)$ where $\sigma^2 = \mathrm{Var}(T)$. We will show that the number of triangles in $\mathbb{G}(n,p)$ approximately follows the normal distribution.

Theorem 11.15. *Let W be as above and $p = p(n) \leqslant 1/2$ be such that $np \to \infty$. Then*

$$\mathrm{Wass}(W, Z) \to 0 \quad \text{as } n \to \infty,$$

where $Z \sim \mathrm{N}(0,1)$.

Proof. Let $\mathcal{T}$ be the set of all triangles in the complete graph on $\{1, 2, \ldots, n\}$. We can write T as sum of the indicator random variables

$$T = \sum_{\alpha\in\mathcal{T}} \mathbb{1}(\alpha \text{ is present in } \mathbb{G}(n,p)).$$

Define $X_\alpha := \mathbb{1}(\alpha \text{ is present in } \mathbb{G}(n,p)) - \mathbb{E}\,\mathbb{1}(\alpha \text{ is present in } \mathbb{G}(n,p)), \quad \alpha \in \mathcal{T}$. Then $T - \mathbb{E}\,T = \sum_{\alpha\in\mathcal{T}} X_\alpha$. For $\alpha, \beta \in \mathcal{T}$, we will use the notation $e(\alpha \cap \beta)$ to denote the number of common edges shared by the triangles α and β. Similarly, for three triangles α, β and γ, we define $e(\alpha, \beta, \gamma)$ is the number of edges that are present in each of the triangles.

Define a graph G on the vertex set $\mathcal{T}$ so that two distinct triangles α and β are connected in G if and only if $e(\alpha \cap \beta) > 0$. Clearly, G is a dependency graph for $\{X_\alpha\}_{\alpha\in\mathcal{T}}$. Indeed, X_α and X_β are independent if α and β do not share any common edge. Note that in G,

$$\overline{\mathcal{N}}(\alpha) = \{\beta \in \mathcal{T} : e(\alpha \cap \beta) = 1 \text{ or } 3\},$$

whose size is bounded by $3n$. Note that

$$\mathbb{E}\mathbb{1}(\alpha \text{ is present in } \mathbb{G}(n,p)) = p^3, \qquad \mathbb{E}\,\mathcal{T} = \binom{n}{3}p^3 \text{ and}$$

$$\sigma^2 = \sum_{\alpha\in\mathcal{T},\beta\in\overline{\mathcal{N}}(\alpha)} \mathbb{E}\,X_\alpha X_\beta = \binom{n}{3}p^3(1-p^3) + \binom{n}{3}(n-3)p^5(1-p). \tag{11.11}$$

Now the proof follows by using Theorem 11.9. ■

11.5.4 Volume of a random polytope in unit ball in $\mathbb{R}^d$

Let B be the (closed) unit ball centered at the origin and let $\mathcal{P}_n$ be a Poisson point process with intensity n in $\mathbb{R}^d, d \geqslant 2$. Let $X_1, X_2, \ldots, X_N$ be the random number of points of $\mathcal{P}_n$ that fall within the ball B. Define the random polytope Π_n as the convex hull $[X_1, X_2, \ldots, X_N] = [B \cap \mathcal{P}_n]$ of these random points. Let Volume(Π_n) denote the volume of Π_n. In this section we will illustrate how is to prove central limit theorem for Volume(Π_n) using the dependency graph techniques. We will also establish upper bound for the rate of convergence in Berry-Esséen distance. Let us now state the main theorem.

Theorem 11.16. *With notations as above,*

$$\left|\mathbb{P}\left(\frac{\textit{Volume}(\Pi_n) - \mathbb{E}\ \textit{Volume}(\Pi_n)}{\sqrt{\mathrm{Var}(\ \textit{Volume}(\Pi_n))}} \leqslant x\right) - \phi(x)\right| \leqslant b_1 n^{-\frac{1}{2}+\frac{1}{d+1}} \ln^{2+\frac{2}{d+2}} n,$$

where $\phi(\cdot)$ is the cumulative distribution function of the standard Normal distribution and b_1 is a constant which only depends on d.

The proof of the theorem is long, but the main ingredient is dependency graph structure for the volume.

11.6 Exchangeable pairs approach

In this section we will look at the method of exchangeable pair for Stein's method and apply it to a variety of problems. Method of exchangeable pair is originally due to Charles Stein (1986) himself and it is probably the easiest approach for Stein's method to apply on a concrete problem. The usefulness comes from the fact that in many examples there is a natural way to perturb the random variable by a small amount without changing the distribution. The basic idea for the exchangeable pair approach is the following. Suppose that W is a random variable with mean zero and variance one. Also suppose that it is possible to perturb W by a 'small' amount to get another random variable W' having same marginal as W. Assume that the pair (W, W') is exchangeable, that is (W, W') and (W', W) have the same distribution. Now for any 'nice' function f, by the exchangeability condition and 'smallness' of $|W - W|'$ we have

$$\begin{aligned}
&\mathbb{E}(W' - W)(f(W) + f(W')) = 0 \\
\text{or } &\mathbb{E}(W' - W)(f(W') - f(W)) = -2\,\mathbb{E}(W' - W)f(W) \\
\text{or } &\mathbb{E}(W' - W)^2 \frac{f(W') - f(W)}{W' - W} = -2\,\mathbb{E}(W' - W)f(W) \\
\text{or } &\mathbb{E}\left[\mathbb{E}((W' - W)^2|W) \cdot f'(W)\right] \approx \mathbb{E}\left[-2\,\mathbb{E}(W' - W|W) \cdot f(W)\right]
\end{aligned}$$

Now if we have $\mathbb{E}(W' - W|W) \approx -\lambda W$ for some $\lambda \in (0, 1)$, exchangeability yields that

$$\mathbb{E}(W' - W)^2 = 2\,\mathbb{E}\,W(W - W') \approx 2\lambda\,\mathbb{E}\,W^2 = 2\lambda.$$

So if we further assume that $\frac{1}{2\lambda}\,\mathbb{E}((W'-W)^2|W)$ is concentrated around its mean $\frac{1}{2\lambda}\,\mathbb{E}(W'-W)^2 \approx 1$, then for any 'nice' function f we have

$$\begin{aligned}
\mathbb{E}\left[f'(W)\right] &\approx \frac{1}{2\lambda}\,\mathbb{E}\left[\mathbb{E}((W' - W)^2|W) \cdot f'(W)\right] \\
&\approx \mathbb{E}\left[-\frac{2\,\mathbb{E}(W' - W|W)}{2\lambda} \cdot f(W)\right] \approx \mathbb{E}\left[W f(W)\right] \qquad (11.12)
\end{aligned}$$

that is, W approximately satisfies the characterizing equation for standard normal distribution. Before going to concrete examples let us state and prove the result precisely. Though exchangeability condition was used in the original theorem, in the proof it suffices to consider equidistributed pair.

11.7 Bound on Wasserstein Metric

Theorem 11.17. *Let (W, W') be a pair of random variables defined on the same probability space having same marginal distributions,* ***i.e.*** *$W \stackrel{d}{=} W'$. Suppose $\mathbb{E}\,W = 0, \mathbb{E}\,W^2 = 1$ and*

$$\mathbb{E}(W' - W|W) = -\lambda W \ \textit{a.e.} \qquad (11.13)$$

for some constant $\lambda \in (0, 1)$. Then, we have

$$d_W(W, Z) \leqslant \sqrt{\frac{2}{\pi}\,\mathrm{Var}\left(\mathbb{E}(\frac{1}{2\lambda}(W' - W)^2|W)\right)} + \frac{1}{3\lambda}\,\mathbb{E}\,|W' - W|^3 \qquad (11.14)$$

where $Z \sim N(0, 1)$.

Corollary 11.18. *Let (W, W') be a pair of random variables defined on the same probability space having same marginal distributions. Suppose $\mathbb{E}\,W = 0, \mathbb{E}\,W^2 = 1$ Define the random variable $R = R(W)$ by*

$$\mathbb{E}(W' - W|W) = -\lambda W + R \text{ a.e.} \tag{11.15}$$

where λ is a number satisfying $0 < \lambda < 1$. Then, we have

$$d_W(W, Z) \leqslant \sqrt{\frac{2}{\pi}\,\mathrm{Var}\left(\mathbb{E}(\frac{1}{2\lambda}(W' - W)^2|W)\right)} + \frac{1}{3\lambda}\,\mathbb{E}\,|W' - W|^3 + 2\frac{\sqrt{\mathbb{E}\,R^2}}{\lambda} \tag{11.16}$$

where $Z \sim N(0, 1)$.

Remark 11.19. *Condition (11.13) is natural in many cases. If (W, W') is close to bivariate normal distribution, then the linearity of conditional expectation as a function of W should hold approximately and one may expect the remainder R to be small. Also, when (W, W') is bivariate normal, it is easy to check that $\frac{1}{2\lambda}(W' - W)^2|W) = \lambda W/2 + 1 - \lambda/2$. This indicates that λ should be small and then, if W is close to normal, one may expect $\mathrm{Var}(\frac{1}{2\lambda}(W' - W)^2|W))$ to be small.*

Proof. (proof of Theorem 11.17) Using the condition $\mathbb{E}(W' - W|W) = -\lambda W$ and $\mathbb{E}\,W^2 = 1$ we have

$$\begin{aligned}\mathbb{E}(W' - W)^2 &= \mathbb{E}[W'^2 + W^2 - 2W'W] = \mathbb{E}[2W(W - W')] \\ &= \mathbb{E}[2W\,\mathbb{E}(W - W'|W)] = \mathbb{E}[2\lambda W^2] = 2\lambda.\end{aligned}$$

Now take any function f with $|f|_\infty \leqslant 1, |f'|_\infty \leqslant \sqrt{2/\pi}, |f''|_\infty \leqslant 2$. Define a new function F by

$$F(x) = \int_0^x f(y)dy. \tag{11.17}$$

Note that for $x < 0$ we define the integral $\int_0^x$ as $-\int_x^0$. Clearly F is three times differentiable and $|F(x)| \leqslant |x|$. Hence $\mathbb{E}\,|F(W)| < \infty$. Now from exchangeabilty we have

$$\begin{aligned}0 &= \mathbb{E}[F(W') - F(W)] \\ &= \mathbb{E}\left[(W' - W)f(W) + \frac{1}{2}(W' - W)^2 f'(W) + \mathbf{Rem}\right]\end{aligned} \tag{11.18}$$

where $|\mathbf{Rem}| \leqslant \frac{1}{6}|W - W'|^3|f''|_\infty \leqslant \frac{1}{3}|W - W'|^3$. In fact using Taylor's formula we can write down the remainder term explicitly as

$$\mathbf{Rem} = \frac{V^3}{2}\int_0^1 (1 - s)^2 f''(W + sV)ds.$$

where $V = W' - W$. This form of the remainder will be used in the Berry-Esseen bound. From equation (11.18), using properties of conditional expectation we have

$$\begin{aligned}\mathbb{E}\left[-2\lambda W f(W) + \mathbb{E}((W' - W)^2|W)f'(W) + 2\mathbf{Rem}\right] = 0 \\ \text{or } \mathbb{E}[W f(W)] = \frac{1}{2\lambda}\,\mathbb{E}\left[\mathbb{E}((W' - W)^2|W)f'(W) + 2\mathbf{Rem}\right].\end{aligned}$$

Hence

$$\begin{aligned}
&|\,\mathbb{E}\, f'(W) - \mathbb{E}\, W f(W)| \\
&\leqslant \left|\mathbb{E}\left[f'(W)\left(\frac{1}{2\lambda}\,\mathbb{E}((W'-W)^2|W) - 1\right)\right]\right| + \frac{1}{\lambda}|\,\mathbb{E}\,\mathbf{Rem}| \\
&\leqslant |f'|_\infty\,\mathbb{E}\left|\frac{1}{2\lambda}\,\mathbb{E}((W'-W)^2|W) - 1\right| + \frac{1}{\lambda}\,\mathbb{E}\,|\mathbf{Rem}| \\
&\leqslant \sqrt{\frac{2}{\pi}\,\mathrm{Var}\left(\frac{1}{2\lambda}\,\mathbb{E}((W'-W)^2|W)\right)} + \frac{1}{3\lambda}\,\mathbb{E}\,|W-W'|^3.
\end{aligned}$$

From the above calculations it follows that

$$\begin{aligned}
&d_{\mathrm{W}}(W,Z) \\
&= \sup\{|\,\mathbb{E}\, g(W) - \mathbb{E}\, g(Z)| : |g'|_\infty \leqslant 1\} \\
&\leqslant \sup\{|\,\mathbb{E}\, f'(W) - \mathbb{E}\, W f(W)| : |f|_\infty \leqslant 1, |f'|_\infty \leqslant \sqrt{2/\pi}, |f''|_\infty \leqslant 2\} \\
&\leqslant \sqrt{\frac{2}{\pi}\,\mathrm{Var}\left(\frac{1}{2\lambda}\,\mathbb{E}((W'-W)^2|W)\right)} + \frac{1}{3\lambda}\,\mathbb{E}\,|W-W'|^3
\end{aligned}$$

where $Z \sim \mathrm{N}(0,1)$. ■

Now let us apply the exchangeable pair method to the simplest possible example, sums of independent random variable.

11.7.1 Example: Sum of i.i.d. random variables

Suppose $X_1, X_2, \ldots$ are i.i.d. random variable with zero mean and unit variance. We want to prove CLT for $W = n^{-1/2}\sum_{i=1}^n X_i$. To construct an exchangeable pair we proceed as follows. Suppose $X_1', X_2', \ldots$ are independent random variables with X_i' having the same distribution as X_1. Choose an index I uniformly at random from $\{1, 2, \ldots, n\}$. Replace X_I in W by X_I'. Let

$$W' = \frac{1}{\sqrt{n}}\sum_{j\neq I} X_j + \frac{X_I'}{\sqrt{n}} = W + \frac{X_I' - X_I}{\sqrt{n}}.$$

Clearly (W, W') is an exchangeable pair. Now we have

$$\begin{aligned}
\mathbb{E}[W'-W|W] = \frac{1}{\sqrt{n}}\,\mathbb{E}[X_I' - X_I|W] &= \frac{1}{\sqrt{n}}\cdot\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i' - X_i|W] \\
&= -\frac{1}{n^{3/2}}\,\mathbb{E}\left[\sum_{i=1}^n X_i|W\right] = -\frac{1}{n}W.
\end{aligned}$$

Hence the condition of Theorem 11.17 is satisfied with $\lambda = n^{-1}$. Note that

$$\begin{aligned}
\frac{1}{3\lambda}\,\mathbb{E}\,|W'-W|^3 &= \frac{n}{3n^{3/2}}\,\mathbb{E}\,|X_I' - X_I|^3 \\
&= \frac{1}{3n^{3/2}}\sum_{i=1}^n \mathbb{E}\,|X_i' - X_i|^3 \leqslant \frac{8}{3\sqrt{n}}\,\mathbb{E}\,|X_1|^3
\end{aligned}$$

and

$$\mathbb{E}\left[\frac{1}{2\lambda}(W'-W)^2|W\right] = \frac{n}{2n}\mathbb{E}((X_I'-X_I)^2|W)$$
$$= \frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}((X_i'-X_i)^2|W) = \frac{1}{2} + \mathbb{E}\left[\frac{1}{2n}\sum_{i=1}^{n}X_i^2\Big|W\right].$$

Hence we have

$$d_{\mathrm{W}}(W,Z) \leqslant \sqrt{\frac{2}{\pi}\operatorname{Var}\left(\mathbb{E}\left[\frac{1}{2n}\sum_{i=1}^{n}X_i^2\Big|W\right]\right)} + \frac{8}{3n^{3/2}}\sum_{i=1}^{n}\mathbb{E}\,|X_i|^3$$
$$\leqslant \sqrt{\frac{1}{2\pi n^2}\operatorname{Var}\left(\sum_{i=1}^{n}X_i^2\right)} + \frac{8}{3\sqrt{n}}\mathbb{E}\,|X_1|^3$$
$$\leqslant \left(\sqrt{\frac{\operatorname{Var}\left(X_1^2\right)}{2\pi}} + \frac{8}{3}\mathbb{E}\,|X_1|^3\right)\cdot\frac{1}{\sqrt{n}}.$$

And this gives another proof of the famous Central Limit theorem. □

11.7.2 Hoeffding's combinatorial central limit theorem

Suppose $A = ((a_{ij}))_{i,j=1}^n$ is an array of real numbers. Let π be a uniform random permutation of $\{1,2,...,n\}$. Define the random variable W by $W = W_A = \sum_{i=1}^n a_{i\pi(i)}$. Define $\mu_A = \mathbb{E}\,W_A$ and $\sigma_A^2 = \operatorname{Var}(W_A)$. This type of statistics arises in nonparametric tests and sampling from finite population. If we look at the normalized random variable $\sigma_A^{-1}(W_A - \mu_A)$, the question then arises is under what conditions on $A = ((a_{ij}))_{i,j=1}^n$ we can approximate the random variable by standard normal distribution. Hoeffding, in 1951, proved that for a sequence of matrices A_n, $\sigma_{A_n}^{-1}(W_{A_n} - \mu_{A_n})$ is approximately the standard Gaussian distribution under certain regularity condition on A_n. His original proof gave conditions for convergence to normality using the method of moments. In this section we will use the exchangeable pair approach to prove the combinatorial central limit theorem. We will bound the Wasserstein distance using theorem (11.17).

Define

$$a_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n}a_{ij}, \quad a_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n}a_{ij}, \quad a_{\cdot\cdot} = \frac{1}{n^2}\sum_{i,j=1}^{n}a_{ij} \tag{11.19}$$

and

$$\tilde{a}_{ij} = \frac{a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}}{\sigma_A}.$$

Define the matrix $\tilde{A} = ((\tilde{a}_{ij}))_{i,j=1}^n$. Clearly $\mu_A = na_{\cdot\cdot}$ and we have $\tilde{a}_{i\cdot} = \tilde{a}_{\cdot j} = 0$ for all $1 \leqslant i,j \leqslant n$. Define

$$\tilde{W} := W_{\tilde{A}} = \sigma_A^{-1}\sum_{i=1}^{n}\tilde{a}_{i\pi(i)}$$
$$= \sigma_A^{-1}\left(\sum_{i=1}^{n}a_{i\pi(i)} - \sum_{i=1}^{n}a_{i\cdot} - \sum_{i=1}^{n}a_{\cdot\pi(i)} + na_{\cdot\cdot}\right) = \frac{W-\mu_A}{\sigma_A}.$$

Hence $\mathbb{E}\,\tilde{W} = 0$ and $\mathrm{Var}(\tilde{W}) = 1$. Now we have $\mathbb{E}\,\tilde{a}_{i\pi(i)} = n^{-1}\sum_{j=1}^n \tilde{a}_{ij} = 0$ for all $1 \leqslant i \leqslant n$. Hence we can also write the variance as

$$\begin{aligned}\mathrm{Var}(\tilde{W}) &= \sum_{i=1}^n \mathbb{E}(\tilde{a}^2_{i\pi(i)}) + \sum_{i\neq j}\mathbb{E}(\tilde{a}_{i\pi(i)}\tilde{a}_{j\pi(j)}) = \frac{1}{n}\sum_{i,j=1}^n \tilde{a}^2_{ij} + \frac{1}{n(n-1)}\sum_{i\neq j}\sum_{k\neq l}\tilde{a}_{ik}\tilde{a}_{jl}\\ &= \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^n \tilde{a}^2_{ij} + \frac{1}{n(n-1)}\sum_{i,k=1}^n \tilde{a}_{ik}\tilde{a}_{ik} = \frac{1}{n-1}\sum_{i=1}^n\sum_{j=1}^n \tilde{a}^2_{ij}\\ &= \frac{1}{(n-1)\sigma_A^2}\sum_{i=1}^n\sum_{j=1}^n (a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot})^2.\end{aligned}$$

Here we used the fact that for any i, k we have $\sum_{l\neq k}\tilde{a}_{il} = -\tilde{a}_{ik}$ and $\sum_{j\neq i}\tilde{a}_{jk} = -\tilde{a}_{ik}$. This implies that

$$\begin{aligned}\sigma_A^2 &= \frac{1}{(n-1)}\sum_{i=1}^n\sum_{j=1}^n (a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot})^2\\ &= \frac{1}{(n-1)}\left(\sum_{i=1}^n\sum_{j=1}^n a^2_{ij} - n\sum_{i=1}^n a^2_{i\cdot} - n\sum_{j=1}^n a^2_{\cdot j} + n^2 a^2_{\cdot\cdot}\right).\end{aligned} \tag{11.20}$$

Hence without loss of generality we assume the following,

$$\sum_{j=1}^n a_{ij} = 0,\ \sum_{i=1}^n a_{ij} = 0 \text{ and } \frac{1}{n-1}\sum_{i,j=1}^n a^2_{ij} = 1.$$

Under condition (11.21) we have $\mathbb{E}\,W = 0$ and $\mathbb{E}\,W^2 = 1$. Now we will define a random variable W' so that (W, W') satisfies the condition (11.13). Define $\pi' = \pi \circ (I, J)$ where (I, J) is a uniformly chosen random transposition. Clearly (π, π') is an exchangeable pair. Let $W' = \sum_{i=1}^n a_{i\pi'(i)}$. So (W, W') is an exchangeable pair. Note that

$$W' - W = a_{I\pi'(I)} + a_{J\pi'(J)} - a_{I\pi(I)} - a_{J\pi(J)} = a_{I\pi(J)} + a_{J\pi(I)} - a_{I\pi(I)} - a_{J\pi(J)}.$$

So, by summing over the choices for (I, J),

$$\begin{aligned}\mathbb{E}(W' - W|\pi) &= \frac{1}{n(n-1)}\sum_{i,j\neq i}(a_{i\pi(j)} + a_{j\pi(i)} - a_{i\pi(i)} - a_{j\pi(j)})\\ &= \frac{1}{n(n-1)}\left(-\sum_i a_{i\pi(i)} - \sum_i a_{i\pi(i)} - 2(n-1)\sum_i a_{i\pi(i)}\right) = -\frac{2}{n-1}W.\end{aligned}$$

Hence condition (11.13) is satisfied with $\lambda = 2(n-1)^{-1}$. Note that

$$\begin{aligned}\mathbb{E}(|W' - W|^3|\pi) &= \frac{1}{n(n-1)}\sum_{i,j\neq i}|a_{i\pi(i)} + a_{j\pi(j)} - a_{i\pi(j)} - a_{j\pi(i)}|^3\\ &\leqslant \frac{16}{n(n-1)}\sum_{i,j\neq i}\left(|a_{i\pi(i)}|^3 + |a_{j\pi(j)}|^3 + |a_{i\pi(j)}|^3 + |a_{j\pi(i)}|^3\right).\end{aligned}$$

So we have

$$\mathbb{E}(|W'-W|^3) \leqslant \frac{64}{n^2}\sum_{i,j}|a_{ij}|^3.$$

Now we will prove concentration for the conditional variance $\mathrm{Var}((W'-W)^2|\pi)$. We have

$$\begin{aligned}
\mathbb{E}((W'-W)^2 \mid \pi) &= \frac{1}{n(n-1)}\sum_{i,j}(a_{i\pi(i)}+a_{j\pi(j)}-a_{i\pi(j)}-a_{j\pi(i)})^2 \\
&= \frac{1}{n(n-1)}\left(2n\sum_{i=1}^n a_{i\pi(i)}^2 + 2\sum_{i,j}a_{i\pi(j)}^2 + 2(\sum_i a_{i\pi(i)})^2 + 2\sum_{i,j}a_{i\pi(j)}a_{j\pi(i)}\right) \\
&= \frac{2(n+1)}{n(n-1)}\sum_{i=1}^n a_{i\pi(i)}^2 + \frac{2}{n} + \frac{2}{n(n-1)}W^2 + \frac{2}{n(n-1)}\sum_{i,j\neq i}a_{i\pi(j)}a_{j\pi(i)}.
\end{aligned}$$

So it is enough to bound

$$\begin{aligned}
&\frac{1}{2\lambda}\,\mathbb{E}\left|\mathbb{E}((W'-W)^2|\pi) - \mathbb{E}(W'-W)^2)\right| \\
&= \frac{n-1}{4}\left(\frac{2(n+1)}{n(n-1)}\,\mathbb{E}\left|\sum_{i=1}^n a_{i\pi(i)}^2 - \mathbb{E}\sum_{i=1}^n a_{i\pi(i)}^2\right| + \frac{2}{n(n-1)}\,\mathbb{E}\,|W^2-1|\right. \\
&\quad \left. + \frac{2}{n(n-1)}\left|\sum_{i,j\neq i}a_{i\pi(j)}a_{j\pi(i)} - \mathbb{E}\sum_{i,j\neq i}a_{i\pi(j)}a_{j\pi(i)}\right|\right) \\
&\leqslant \sqrt{\mathrm{Var}\left(\sum_{i=1}^n a_{i\pi(i)}^2\right)} + \frac{1}{n} + \frac{1}{2n}\sqrt{\mathrm{Var}\left(\sum_{i,j\neq i}a_{i\pi(j)}a_{j\pi(i)}\right)}.
\end{aligned}$$

Defining $b_{ij} = a_{ij}^2$ and using (11.20) we have

$$\mathrm{Var}\left(\sum_{i=1}^n a_{i\pi(i)}^2\right) = \mathrm{Var}\left(\sum_{i=1}^n b_{i\pi(i)}\right) \leqslant \frac{1}{n-1}\left(\sum_{i,j}b_{ij}^2 + n^2 b_{\cdot\cdot}^2\right) = \frac{1}{n-1}\sum_{i,j}a_{ij}^4 + \frac{n-1}{n^2}.$$

Also we have

$$\begin{aligned}
&\mathrm{Var}\left(\sum_{i,j\neq i}a_{i\pi(j)}a_{j\pi(i)}\right) \leqslant \mathbb{E}\left(\sum_{i,j\neq i}a_{i\pi(j)}a_{j\pi(i)}\right)^2 = \frac{1}{n(n-1)}\sum_{k,l\neq k}\left(\sum_{i,j\neq i}a_{ik}a_{jl}\right)^2 \\
&= \frac{1}{n(n-1)}\sum_{k,l\neq k}\left(\sum_i a_{ik}a_{il}\right)^2 \leqslant \frac{1}{n(n-1)}\sum_{k,l}\left(\sum_i a_{ik}a_{il}\right)^2 \\
&\leqslant \frac{1}{n(n-1)}\sum_{k,l}\left(\sum_i a_{ik}^2\right)\left(\sum_i a_{il}^2\right) = \frac{n-1}{n} \leqslant 1.
\end{aligned}$$

Hence combining everything we have the following theorem.

Theorem 11.20. *Let $(a_{ij})_{i,j=1}^n$ be an array of real numbers and π be a uniform random permutation of $\{1,2,...,n\}$. Let W be the random variable $\sum_{i=1}^n a_{i\pi(i)}$. Define the matrix $((\tilde{a}_{ij}))_{i,j=1}^n$ by*

$$\tilde{a}_{ij} = \frac{a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}}{\sqrt{(n-1)^{-1}\sum_{k=1}^n\sum_{l=1}^n (a_{kl} - a_{k\cdot} - a_{\cdot l} + a_{\cdot\cdot})^2}} \tag{11.21}$$

for all $1 \leqslant i,j \leqslant n$. Then we have

$$d_W\left(\frac{W - \mathbb{E}\,W}{\sqrt{\mathrm{Var}(W)}}, Z\right) \leqslant \sqrt{\frac{2}{\pi}\cdot\frac{1}{n-1}\sum_{i,j}\tilde{a}_{ij}^4} + \frac{2}{\sqrt{n}} + \frac{11}{n}\sum_{i,j}|\tilde{a}_{ij}|^3 \tag{11.22}$$

where Z follows standard Gaussian distribution.

Note that if a_{ij}'s are of constant order, then $\tilde{a}_{ij}$'s are of order $n^{-1/2}$. And hence the above theorem gives a rate of convergence of $n^{-1/2}$.

11.7.3 CLT for magnetization in Ising model

Consider a graph $G = (V, E)$ with vertex set V of size n and edge set E. The ferromagnetic Ising model on G is defined as follows. At each vertex v of G there is a particle with spin σ_i taking values in the set $\{+1, -1\}$. Denote by $\boldsymbol{\sigma} = (\sigma_v)_{v\in V}$ the spin configuration of the whole system. The spins σ_i interact in pairs. In particular the negative Hamiltonian[2] of the system with configuration $\boldsymbol{\sigma}$ is defined by

$$H(\boldsymbol{\sigma}) = \sum_{u\sim v}\sigma_u\sigma_v \tag{11.23}$$

where $u \sim v$ means u, v are neighbors in G. At inverse temperature β the probability of a configuration $\boldsymbol{\sigma}$ is proportional to $e^{\beta H(\boldsymbol{\sigma})}$. In particular, the probability of a configuration $\boldsymbol{\sigma}$ is

$$\mathbb{P}(\boldsymbol{\sigma}) = Z_\beta^{-1} e^{\beta H(\boldsymbol{\sigma})}$$

where $Z_\beta = \sum_{\boldsymbol{\sigma}} \exp(\beta H(\boldsymbol{\sigma}))$ is the partition function. The magnetization corresponding to the configuration $\boldsymbol{\sigma}$ is defined as the average spin

$$m(\boldsymbol{\sigma}) = \frac{1}{n}\sum_{v\in V}\sigma_v. \tag{11.24}$$

Also for a given vertex $v \in V$ we define

$$N_v(\boldsymbol{\sigma}) = \sum_{u\sim v}\sigma_u \tag{11.25}$$

as the sum of spins of the neighbors of v. Note that from the fact that $\mathbb{P}(\boldsymbol{\sigma}) = -\,\mathbb{P}(\boldsymbol{\sigma})$ for all configurations $\boldsymbol{\sigma}$ we have $\mathbb{E}\,\sigma_v = 0$ for all $v \in V$. Also note that conditional on the neighbors of a vertex v the spin σ_v is independent of the spins of remaining vertices. In particular we have

$$\mathbb{P}(\sigma_v = \varepsilon | \sigma_u : u \neq v) = \mathbb{P}(\sigma_v = \varepsilon \mid \sigma_u : u \sim v) = \frac{\exp(\varepsilon\beta N_v(\boldsymbol{\sigma}))}{\exp(\beta N_v(\boldsymbol{\sigma})) + \exp(-\beta N_v(\boldsymbol{\sigma}))}$$

[2]Physicists use the formula $H(\boldsymbol{\sigma}) = -\sum_{u\sim v}\sigma_u\sigma_v$ for defining Hamiltonian and they define the probability of a configuration $\boldsymbol{\sigma}$ as $\mathbb{P}(\boldsymbol{\sigma}) \propto e^{-\beta H(\boldsymbol{\sigma})}$. For simplicity, instead of taking two negative sign, we take the definition of the Hamiltonian having positive sign.

for $\varepsilon \in \{+1, -1\}$ and

$$\mathbb{E}(\sigma_v | \sigma_u : u \neq v) = \frac{\exp(\beta N_v(\boldsymbol{\sigma})) - \exp(-\beta N_v(\boldsymbol{\sigma}))}{\exp(\beta N_v(\boldsymbol{\sigma})) + \exp(-\beta N_v(\boldsymbol{\sigma}))} = \tanh(\beta N_v(\boldsymbol{\sigma})). \tag{11.26}$$

Define $\nu^2 = \mathrm{Var}(\sum_{v \in V} \sigma_v)$. Define the random variable $W := W_n = n\sigma^{-1} m(\boldsymbol{\sigma}) = \nu^{-1} \sum_{v \in V} \sigma_v$ where n is the size of the vertex set V. Clearly $\mathbb{E}\, W = 0$ and $\mathrm{Var}(W) = 1$. In this section we will prove that for certain sequences of graphs and values of the inverse temperature β, W_n is approximately standard Gaussian. In particular we will prove CLT for magnetization in Ising model in n-cycles and complete graphs under appropriate conditions. We will use the exchangeable pair theorem.To define the exchangeable pair we will use the Glauber dynamics, which is defined as follows.

Given $\boldsymbol{\sigma}$ from the Gibbs distribution $\mathbb{P}$, construct $\boldsymbol{\sigma}'$ as follows. Choose an vertex I uniformly at random from V. Replace σ_I by σ'_I drawn from the conditional distribution of σ_I given $\{\sigma_u : u \neq I\}$. Keep all other $\sigma'_u, u \neq I$ same as $\sigma_u, u \neq I$. If we define $W' = \nu^{-1} \sum_{v \in V} \sigma'_v$ we have $W' \stackrel{\mathrm{d}}{=} W$ and

$$W' - W = \nu^{-1}(\sigma'_I - \sigma_I).$$

Hence $|W' - W| \leqslant 2\nu^{-1}$. Also note that

$$\mathbb{E}(W' - W | \boldsymbol{\sigma}) = \frac{1}{\nu} \mathbb{E}(\sigma'_I - \sigma_I | \boldsymbol{\sigma}) = \frac{1}{n\nu} \sum_{v \in V} (\tanh(\beta N_v(\boldsymbol{\sigma})) - \sigma_v).$$

Now we will go into the details.

11.7.3.1 Ising model on complete graph: Curie-Weiss model

Here we consider the complete graph $G = G_n$ with vertex set $V = [n]$ and egde set $E = \{(i, j) : 1 \leqslant i < j \leqslant n\}$. Note that here every vertex has degree $n - 1$. So instead of β we generally take β/n as the inverse temperature parameter. This model is also known as Curie-Weiss model. So here the probability of a configuration $\boldsymbol{\sigma} \in \{+1, -1\}^n$ is $\mathbb{P}(\sigma) = Z_\beta^{-1} \exp(\beta H(\sigma)/n)$ where $H(\sigma) = \sum_{i<j} \sigma_i \sigma_j$ is the negative Hamiltonian and $Z_\beta = \sum_{\sigma \in \{+1,-1\}^n} \exp(\beta H(\sigma)/n)$ is the partition function. We will prove CLT for the magnetization $m(\sigma) = n^{-1} \sum_{i=1}^n \sigma_i$ when $\beta < 1$.

Again let $\boldsymbol{\sigma}, \boldsymbol{\sigma}', W, W', \nu$ be as defined before. Note that here for $i \in [n]$ we have $N_i(\boldsymbol{\sigma}) = \sum_{j \neq i} \sigma_j$. Hence using (11.26) we have

$$\mathbb{E}(\sigma_i | \sigma_j, j \neq i) = \tanh(\beta m_i(\boldsymbol{\sigma})) \tag{11.27}$$

where $m_i(\boldsymbol{\sigma}) = n^{-1} N_i(\boldsymbol{\sigma}) = n^{-1} \sum_{j \neq i} \sigma_j$. Now using the conditional expectation formula we can write

$$\begin{aligned}
\mathbb{E}(W' - W | \boldsymbol{\sigma}) = &= -\frac{1}{n} \left(W - \frac{1}{\nu} \sum_{i=1}^n \tanh(\beta m_i(\boldsymbol{\sigma})) \right) \\
&= -\frac{1 - \beta}{n} W + \frac{1}{n\nu} \left(\sum_{i=1}^n \tanh(\beta m_i(\boldsymbol{\sigma})) - n\beta m(\boldsymbol{\sigma}) \right) \\
&= -\frac{1 - \beta}{n} \left(W - \nu^{-1} D \right)
\end{aligned}$$

where $D := D(\boldsymbol{\sigma}) = (1-\beta)^{-1}(\sum_{i=1}^n \tanh(\beta m_i(\boldsymbol{\sigma})) - n\beta m(\boldsymbol{\sigma}))$. Hence the conditions in corollary (11.18) is satisfied with

$$\lambda = \frac{1-\beta}{n} \text{ and } \frac{R}{\lambda} = \frac{D}{\nu}.$$

Note that here we have

$$\begin{aligned}(1-\beta)\nu \cdot \frac{|R|}{\lambda} &\leqslant \sum_{i=1}^n |\tanh \beta m_i(\sigma) - \beta m(\sigma)| \\ &\leqslant \sum_{i=1}^n |\tanh \beta m_i(\sigma) - \tanh \beta m(\sigma)| + n\,|\tanh \beta m(\sigma) - \beta m(\sigma)| \\ &\leqslant \beta + \frac{n\beta^3\,|m(\sigma)|^3}{3}.\end{aligned}$$

Here we use the fact that

$$|\tanh(x) - \tanh(y)| \leqslant |x-y| \tag{11.28}$$

and $|m(\boldsymbol{\sigma}) - m_i(\boldsymbol{\sigma})| \leqslant n^{-1}$. We will prove that ν^2 is asymptotically $n/(1-\beta)$ and $m(\sigma)$ is concentrated around its mean 0. Note that $\mathbb{E}(m(\boldsymbol{\sigma}))^2 = \nu^2/n^2 \approx (n(1-\beta))^{-1}$. Hence indeed $\lambda^{-1}\,\mathbb{E}\,|R|$ is small compared to W. Expanding the squares and using the conditional expectation formula again we have

$$\begin{aligned}\mathbb{E}((W'-W)^2|\boldsymbol{\sigma}) &= \frac{1}{n\nu^2}\sum_{i=1}^n \mathbb{E}((\sigma_i'-\sigma_i)^2|\boldsymbol{\sigma}) \\ &= \frac{2}{n\nu^2}\left(n - \sum_{i=1}^n \sigma_i \tanh(\beta m_i(\boldsymbol{\sigma}))\right) = \frac{2}{n\nu^2}(n-Y)\end{aligned}$$

where $Y := Y(\boldsymbol{\sigma}) = \sum_{i=1}^n \sigma_i \tanh(\beta m_i(\boldsymbol{\sigma}))$. So

$$\frac{1}{2\lambda}\,\mathbb{E}((W'-W)^2|\boldsymbol{\sigma}) = \frac{n-Y}{(1-\beta)\nu^2} \approx 1 - \frac{Y}{n}. \tag{11.29}$$

Note that

$$Y \approx nm(\boldsymbol{\sigma})\tanh(\beta m(\boldsymbol{\sigma})) \approx n\beta(m(\boldsymbol{\sigma}))^2 \approx \beta\nu^2/n \approx \beta/(1-\beta).$$

So indeed the conditional expectation $\frac{1}{2\lambda}\,\mathbb{E}((W'-W)^2|\boldsymbol{\sigma})$ is concentrated around its mean 1. Hence from the result of corollary (11.18), we have

$$d_{\mathrm{W}}(W,Z) \leqslant \sqrt{\frac{2}{\pi}}\,\mathbb{E}\left|\frac{n-Y}{(1-\beta)\nu^2} - 1\right| + \frac{8n}{3(1-\beta)\nu^3} + 2\frac{\mathbb{E}\,|R|}{\lambda}.$$

One can prove that $\left|\frac{n}{(1-\beta)\nu^2} - 1\right| \leqslant \frac{const}{\sqrt{n}}, \nu^{-2}\,\mathbb{E}\,|Y| \leqslant \frac{const}{\sqrt{n}}$ and $\lambda^{-1}\,\mathbb{E}\,|R| \leqslant \frac{const}{\sqrt{n}}$, thus we have

$$d_{\mathrm{W}}(W,Z) \leqslant \frac{\text{const.}}{\sqrt{n}}. \tag{11.30}$$

11.8 Appendix: Proof of Lemma 11.6

Proof. **Proof of bound (I) :** Take f as in (11.4). Suppose $x > 0$. Using the representation in (11.5), we have

$$|f(x)| \leqslant |g - Ng|_\infty \left(e^{x^2/2} \int_x^\infty e^{-y^2/2} dy \right).$$

Recall the Mill's ratio inequality on $\Phi(x)$ for $x > 0$:

$$\frac{xe^{x^2/2}}{\sqrt{2\pi}(1+x^2)} \leqslant 1 - \Phi(x) \leqslant \frac{e^{x^2/2}}{x\sqrt{2\pi}}. \tag{11.31}$$

Now, $\frac{d}{dx} e^{x^2/2} \int_x^\infty e^{-y^2/2} dy = -1 + xe^{x^2/2} \int_x^\infty e^{-y^2/2} dy \leqslant 0 \; \forall \; x > 0$ by Mill's ratio inequality (11.31). So, $e^{x^2/2} \int_x^\infty e^{-y^2/2} dy$ is maximized at $x = 0$ on $[0, \infty)$ where its value is $\sqrt{\frac{\pi}{2}}$. Hence,

$$|f(x)| \leqslant \sqrt{\frac{\pi}{2}} |g - Ng|_\infty \quad \forall x > 0.$$

For $x < 0$, use the form (11.4) and proceed in the similar manner.

Proof of bound (II) : Again, we will only consider $x > 0$ case. The other case will be similar. Note that

$$f'(x) = g(x) - Ng + xf(x) = g(x) - Ng - xe^{x^2/2} \int_x^\infty e^{-y^2/2} (g(y) - Ng) dy.$$

Therefore,

$$\begin{aligned} |f'(x)| &\leqslant |g - Ng|_\infty \left(1 + xe^{x^2/2} \int_x^\infty e^{-y^2/2} dy \right) \\ &\leqslant 2|g - Ng|_\infty \quad \text{by Mill's ratio inequality (11.31).} \end{aligned}$$

Proof of bound (III) : Applying Stein's identity on (11.6), we have

$$f(x) = -\int_0^1 \frac{1}{2\sqrt{t}} \, \mathbb{E}\left[g'(\sqrt{t}x + \sqrt{1-t}Z) \right] dt. \tag{11.32}$$

Hence,

$$|f|_\infty \leqslant |g'|_\infty \int_0^1 \frac{1}{2\sqrt{t}} = |g'|_\infty.$$

Proof of bound (IV) : From (11.7), it follows that

$$|f|_\infty \leqslant (\mathbb{E}\,|Z|) \, |g'|_\infty \int_0^1 \frac{1}{2\sqrt{1-t}} = \sqrt{\frac{2}{\pi}} |g'|_\infty.$$

Proof of bound (V) : On differentiating (11.3) and rearranging

$$\begin{aligned} f''(x) &= g'(x) + f(x) + xf'(x) \\ &= g'(x) + f(x) + x(g(x) - Ng + xf(x)) \\ &= g'(x) + x(g(x) - Ng) + (1 + x^2) f(x). \end{aligned} \tag{11.33}$$

We can write $g(x) - Ng$ in terms of g' as follows,

$$\begin{aligned}
g(x) - Ng &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2}(g(x) - g(y))dy \\
&= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{x} \int_{y}^{x} g'(z) e^{-y^2/2} dz dy - \int_{x}^{\infty} \int_{x}^{y} g'(z) e^{-y^2/2} dz dy \right] \\
&= \frac{1}{\sqrt{2\pi}} \left[\int_{-\infty}^{x} g'(z) \int_{-\infty}^{z} e^{-y^2/2} dy dz - \int_{x}^{\infty} g'(z) \int_{z}^{\infty} e^{-y^2/2} dy dz \right] \\
&= \int_{-\infty}^{x} g'(z)\Phi(z)dz - \int_{x}^{\infty} g'(z)\overline{\Phi}(z)dz
\end{aligned}$$

where $\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$ is the distribution function for standard normal and $\overline{\Phi}(z) = 1 - \Phi(z)$. Similarly,

$$\begin{aligned}
f(x) &= e^{x^2/2} \int_{-\infty}^{x} e^{-y^2/2}(g(y) - \mathbb{E}\, g(Z))dy \\
&= e^{x^2/2} \int_{-\infty}^{x} e^{-y^2/2} \left(\int_{-\infty}^{y} g'(z)\Phi(z)dz - \int_{y}^{\infty} g'(z)\overline{\Phi}(z) \right) dz dy \\
&= e^{x^2/2} \left(\int_{-\infty}^{x} g'(z)\Phi(z) \int_{z}^{x} e^{-y^2/2} dy dz - \int_{-\infty}^{\infty} g'(z)\overline{\Phi}(z) \int_{-\infty}^{z\wedge x} e^{-y^2/2} dy dz \right) \\
&= \sqrt{2\pi} e^{x^2/2} \left(\int_{-\infty}^{x} g'(z)\Phi(z)(\overline{\Phi}(z) - \overline{\Phi}(x))dz \right. \\
&\qquad \left. - \int_{-\infty}^{x} g'(z)\overline{\Phi}(z)\Phi(z)dz - \int_{x}^{\infty} g'(z)\overline{\Phi}(z)\Phi(x)dz \right) \\
&= -\sqrt{2\pi} e^{x^2/2} \left[\overline{\Phi}(x) \int_{-\infty}^{x} g'(z)\Phi(z)dz + \Phi(x) \int_{x}^{\infty} g'(z)\overline{\Phi}(z)dz \right].
\end{aligned}$$

Substituting the above expressions for $g - Ng$ and f in (11.33), we get

$$\begin{aligned}
f''(x) = g'(x) &+ \left(x - \sqrt{2\pi}(1 + x^2) e^{x^2/2} \overline{\Phi}(x) \right) \int_{-\infty}^{x} g'(z)\Phi(z)dz \\
&+ \left(-x - \sqrt{2\pi}(1 + x^2) e^{x^2/2} \Phi(x) \right) \int_{x}^{\infty} g'(z)\overline{\Phi}(z)dz.
\end{aligned}$$

This gives

$$\begin{aligned}
|f''(x)|_\infty \leqslant |g'|_\infty \Big[1 &+ \left| x - \sqrt{2\pi}(1 + x^2) e^{x^2/2} (1 - \Phi(x)) \right| \int_{-\infty}^{x} \Phi(z)dz \\
&+ \left| -x - \sqrt{2\pi}(1 + x^2) e^{x^2/2} \Phi(x) \right| \int_{x}^{\infty} (1 - \Phi(z))dz \Big].
\end{aligned} \tag{11.34}$$

There is a similar bound for $x \leqslant 0$. To proceed, we wish to remove the absolute values in equation (11.34), by determining the sign of the expressions within the absolute value. From the Mill's ratio (11.31) we have

$$x + \sqrt{2\pi}(1 + x^2) e^{x^2/2} \Phi(x) > 0 \tag{11.35}$$

and

$$-x + \sqrt{2\pi}(1+x^2)e^{x^2/2}(1-\Phi(x)) > 0. \tag{11.36}$$

One can check (11.36) by noting that for $x < 0$ the inequality is obvious, and for $x > 0$ use the lower Mill's ratio inequality; (11.35) follows similarly. Hence both expressions within the absolute values in equation (11.34) are negative.

To finish the simplification, observe that integration by parts gives

$$\int_{-\infty}^{x} \Phi(z)dz = x\Phi(x) + \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

and

$$\int_{x}^{\infty} (1-\Phi(z))dz = -x(1-\Phi(x)) + \frac{e^{-x^2/2}}{\sqrt{2\pi}}.$$

Combining, we get

$$\begin{aligned} |f''(x)| &\leqslant |g'|_\infty \Bigg[1 + \left(-x + \sqrt{2\pi}(1+x^2)e^{x^2/2}(1-\Phi(x))\right)\left(x\Phi(x) + \frac{e^{-x^2/2}}{\sqrt{2\pi}}\right) \\ &\quad + \left(x + \sqrt{2\pi}(1+x^2)e^{x^2/2}\Phi(x)\right)\left(-x(1-\Phi(x)) + \frac{e^{-x^2/2}}{\sqrt{2\pi}}\right)\Bigg] \\ &= 2|g'|_\infty. \end{aligned}$$

This proves the desired bound. ■