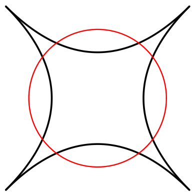
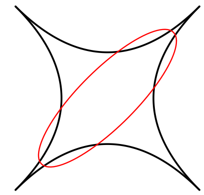


Stein's Method and The Bootstrap in Low and High Dimensions: A Tutorial

Larry Wasserman

February 8 2014



These notes give a brief tutorial on Stein's method and the bootstrap. The notes are divided into two parts. Part I is a review of Stein's method in the low dimensional case. Part II is a review of the results due to Victor Chernozhukov, Denis Chetverikov, Kengo Kato who deal with the high dimensional case. Our ultimate goal, which we consider in Part II, is to discuss the high dimensional bootstrap. The two parts are self-contained and can be read independently of each other.

Part I: Stein's Method in Low Dimensions

The main reference for this part, which I follow very closely, is:
Chen, Goldstein and Shao. (2011). *Normal Approximation by Stein's Method.* Springer.
Other references are given at the end of Part I.

Thanks for Martin Azizyan, Roy Allen and the CMU Statistical Machine Learning Reading Group for helpful comments and for finding typos.

1 Introduction

Let $X_1,\ldots,X_n \in \mathbb{R}$ be iid with mean 0 and variance 1 and let

$$X = \frac{1}{\sqrt{n}}\sum_i X_i$$

and $Y \sim N(0,1)$. We want to bound

$$\Delta_n = \sup_z \Big|\mathbb{P}(X \le z) - \mathbb{P}(Y \le z)\Big| = \sup_z \Big|\mathbb{P}(X \le z) - \Phi(z)\Big| \tag{1}$$

where Φ is the cdf of a standard Normal. "Stein's method" is really a collection of methods for bounding Δ_n (or quantities similar to Δ_n). The key idea is based on the following fact:

$$\mathbb{E}[Y f(Y)] = \mathbb{E}[f'(Y)] \;\text{ for all smooth } f \text{ if and only if } Y \sim N(0,1). \tag{2}$$

This suggests that Y should be almost Normal if $\mathbb{E}[Yf(Y)-f'(Y)]$ is close to 0, More precisely, let h be any function such that $\mathbb{E}|h(Y)| < \infty$. The **Stein function** f_h associated with h is a function satisfying the differential equation

$$f_h'(x) - xf_h(x) = h(x) - \mathbb{E}[h(Y)]. \tag{3}$$

It follows that

$$\mathbb{E}[h(X)] - \mathbb{E}[h(Y)] = \mathbb{E}[f_h'(X) - Xf_h(X)]$$

and showing that X is close to Normal amounts to showing that $\mathbb{E}[f_h'(X) - Xf_h(X)]$ is small. Is there such an f? In fact, the Stein function is

$$f(x) \equiv f_h(x) = e^{x^2/2} \int_{-\infty}^{x} [h(y) - \mu] e^{-y^2/2} dy. \tag{4}$$

where $\mu = \mathbb{E}[h(Y)]$.[1]

Properties of the Stein Function. If h is absolutely continuous, then

$$||f_h||_\infty \le 2||h'||_\infty, \;\; ||f_h'||_\infty \le \sqrt{\frac{2}{\pi}}||h'||_\infty, \;\; ||f_h''||_\infty \le 2||h'||_\infty. \tag{5}$$

If h is bounded then,

$$||f_h||_\infty \le \sqrt{\frac{\pi}{2}}||h - \mu(h)||_\infty, \;\; ||f_h'||_\infty \le 2||h - \mu(h)||_\infty. \tag{6}$$

Example 1 *Choose any $z \in \mathbb{R}$ and let $h(x) = I(x \le z) - \Phi(z)$. Let f_z denote the Stein function for h; thus f_z satisfies*

$$f_z'(x) - xf_z(x) = I(x \le z) - \Phi(z). \tag{7}$$

The unique bounded solution to this equation is

$$f_z(x) = \begin{cases} \sqrt{2\pi} e^{x^2/2} \Phi(x)[1 - \Phi(z)] & x \le z \\ \sqrt{2\pi} e^{x^2/2} \Phi(z)[1 - \Phi(x)] & x > z. \end{cases}$$

f_z is the Stein function associated with $h(x) = I(x \le z) - \Phi(z)$. The function f_z has the following properties:

$$\Big|(x+a)f_z(x+a) - (x+b)f_z(x+b)\Big| \le (|x| + c)(|a| + |b|) \tag{8}$$

where $c = \sqrt{2\pi}/4$. Also,

$$||f_z||_\infty \le \sqrt{\frac{\pi}{2}}, \quad ||f_z'||_\infty \le 2.$$

[1]More precisely, f_h is the unique solution to (3) subject to the side condition $\lim_{x\to\pm\infty} e^{-x^2/2} f(x) = 0$.

Let $\mathscr{F} = \{f_z : z \in \mathbb{R}\}$*. From (7) it follows that*

$$\mathbb{P}(X \le z) - \mathbb{P}(Y \le z) = \mathbb{E}[f'(X) - Xf(X)]$$

and so

$$\Delta_n = \sup_z \Big|\mathbb{P}(X \le z) - \mathbb{P}(Y \le z)\Big| \le \sup_{f \in \mathscr{F}} \Big|\mathbb{E}[f'(X) - Xf(X)]\Big|. \tag{9}$$

We have reduced the problem of bounding $\sup_z \Big|\mathbb{P}(X \le z) - \mathbb{P}(Y \le z)\Big|$ *to the problem of bounding* $\sup_{f \in \mathscr{F}} \Big|\mathbb{E}[f'(X) - Xf(X)]\Big|$.

Notation. Let $X_1, \ldots, X_n$ be iid, mean 0, variance 1. We use the following notation throughout:

$$X = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \qquad \mu_3 = \mathbb{E}[|X_i|^3]$$

$$\xi_i = \frac{1}{\sqrt{n}} X_i \qquad X^i = X - \xi_i.$$

Hence, $X = \sum_i \xi_i$ and also, X^i is independent of ξ_i. Let $Y_1, \ldots, Y_n \sim N(0,1)$ and

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

We also make use of the following simple anti-concentration fact: for any y and a, $|\Phi(y+a) - \Phi(y)| \le a/\sqrt{2\pi}$ since the density of the Normal is bounded above by $1/\sqrt{2\pi}$.

2 A Simple Bound: The Basic Stein Result

Before we bound Δ_n, we first bound a different quantity which is easier to control. We follow Section 1.3 of Chen, Goldstein and Shao (2011).

Let

$$\mathscr{H} = \Big\{h : \mathbb{R} \to \mathbb{R} : |h(y) - h(x)| \le |y - x|\Big\}$$

be the class of Lipschitz functions. We will bound

$$\delta_n \equiv \sup_{h \in \mathscr{H}} \Big|\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]\Big|.$$

Note that δ_n is the Wassertstein distance $W_1(X, Y)$.

Theorem 2 *Suppose that $\mu_3 < \infty$. Then*

$$\delta_n \equiv \sup_{h \in \mathscr{H}} \Big| \mathbb{E}[h(X)] - \mathbb{E}[h(Y)] \Big| \leq \frac{C\mu_3}{\sqrt{n}}$$

Proof. Let f_h be the Stein function associated with h. Hence, $f_h'(x) - x f_h(x) = h(x) - \mathbb{E}[h(Y)]$. Therefore,

$$\mathbb{E}[h(X)] - \mathbb{E}[h(Y)] = \mathbb{E}[f_h'(X) - X f_h(X)].$$

It follows that

$$\delta_n = \sup_{f \in \mathscr{F}} \Big| \mathbb{E}[f'(X) - X f(X)] \Big|$$

where $\mathscr{F} = \{f_h : \ h \in \mathscr{H}\}$. It can be shown that each $f \in \mathscr{F}$ is twice differentiable and that $c = \sup_{f \in \mathscr{F}} \sup_x |f''(x)| < \infty$.

If we can show that $f'(X)$ is close to $X f(X)$ then we are done. Since X^i is independent of ξ_i, $\mathbb{E}[\xi_i f(X^i)] = \mathbb{E}[\xi_i]\mathbb{E}[f(X^i)] = 0$. So,

$$\begin{aligned}
\mathbb{E}[X f(X)] &= \mathbb{E}\big[\sum_i \xi_i f(X)\big] = \mathbb{E}\big[\sum_i \xi_i f(X^i + \xi_i)\big] \\
&= \mathbb{E}\left[\sum_i \xi_i \left(f(X^i) + \xi_i \int_0^1 f'(X^i + u\xi_i) du \right)\right] \\
&= \mathbb{E}\left[\sum_i \xi_i^2 \int_0^1 f'(X^i + u\xi_i) du\right].
\end{aligned}$$

Now

$$\mathbb{E}\Big[\xi_i^2 f'(X^i)\Big] = \mathbb{E}\Big[\xi_i^2\Big]\mathbb{E}\Big[f'(X^i)\Big] = \frac{1}{n}\mathbb{E}\Big[f'(X^i)\Big].$$

Hence,

$$\begin{aligned}
\mathbb{E}[f'(X)] &= \mathbb{E}\Big[\sum_i \frac{1}{n} f'(X)\Big] = \mathbb{E}\Big[\sum_i \frac{1}{n} f'(X^i)\Big] + \mathbb{E}\Big[\sum_i \frac{1}{n}(f'(X) - f'(X^i))\Big] \\
&= \mathbb{E}\Big[\sum_i \xi_i^2 f'(X^i)\Big] + \mathbb{E}\Big[\sum_i \frac{1}{n}(f'(X) - f'(X^i))\Big] \\
&= \mathbb{E}\Big[\sum_i \xi_i^2 \int_0^1 f'(X^i) du\Big] + \mathbb{E}\Big[\sum_i \frac{1}{n}(f'(X) - f'(X^i))\Big].
\end{aligned}$$

Therefore,

$$\mathbb{E}[f'(X) - X f(X)] = \mathbb{E}\left[\sum_i \xi_i^2 \int_0^1 (f'(X^i) - f'(X^i + u\xi_i)) du\right] + \mathbb{E}\Big[\sum_i \frac{1}{n}(f'(X) - f'(X^i))\Big].$$

Now, for all $0 \le u \le 1$,

$$|f'(X^i) - f'(X^i + u\xi_i)| \le c|\xi_i| = \frac{c|X_i|}{\sqrt{n}}$$

where $c = \sup_{F \in \mathscr{F}} \sup_x |f''(x)|$. Similarly,

$$|f'(X) - f'(X^i)| = |f'(X^i + \xi_i) - f'(X^i)| \le c|\xi_i| = \frac{c|X_i|}{\sqrt{n}}.$$

Therefore,

$$|\mathbb{E}[f'(X) - Xf(X)]| \le cn\mathbb{E}[|\xi_1|^3] + c\mathbb{E}[|\xi_1|] \le \frac{C\mu_3}{\sqrt{n}}.$$

□

So much for the Wasserstein distance. The rest of the notes focus on the Kolmogorov-Smirnov distance which is more useful statistically, but also harder to bound.

3 Zero Bias Coupling

Recall that $X \sim N(0, \sigma^2)$ if and only if

$$\sigma^2 \mathbb{E}[f'(X)] = \mathbb{E}[Xf(X)] \tag{10}$$

for all absolutely continuous functions f (for which the expectations exist). Inspired by this, Goldstein and Reinert (1997) introduced the following definition. Let X be any mean 0 random variable with variance σ^2. Say that X^* has the *X-zero bias distribution* if

$$\sigma^2 \mathbb{E}[f'(X^*)] = \mathbb{E}[Xf(X)]. \tag{11}$$

for all absolutely continuous functions f for which $\mathbb{E}|Xf(X)| < \infty$. Zero-biasing is a transform that maps one random variable X into another random variable X^*. (More precisely, it maps the distribution of X into the distribution of X^*.) The Normal is the fixed point of this map. The following result shows that X^* exists and is unique.

Theorem 3 *Let X be any mean 0 random variable with variance σ^2. There exists a unique distribution corresponding to a random variable X^* such that*

$$\sigma^2 \mathbb{E}[f'(X^*)] = \mathbb{E}[Xf(X)]. \tag{12}$$

for all absolutely continuous functions f for which $\mathbb{E}|Xf(X)| < \infty$. The distribution of X^* has density

$$p^*(x) = \frac{\mathbb{E}[XI(X > x)]}{\sigma^2} = -\frac{\mathbb{E}[XI(X \le x)]}{\sigma^2}. \tag{13}$$

Proof. It may be verified that $p^*(x) \geq 0$ and integrates to 1. Let us verify that (12) holds. For simplicity, assume that $\sigma^2 = 1$. Given an absolutely continuous f there is a g such that $f(x) = \int_0^x g$. Then

$$\int_0^\infty f'(u)\, p^*(u) du = \int_0^\infty f'(u)\, \mathbb{E}[XI(X > u)]du = \int_0^\infty g(u)\, \mathbb{E}[XI(X > u)]du$$
$$= \mathbb{E}\left[X \int_0^\infty g(u) I(X > u) du\right] = \mathbb{E}\left[X \int_0^{X \vee 0} g(u) du\right] = \mathbb{E}[Xf(X)I(X \geq 0)].$$

Similarly, $\int_{-\infty}^0 f'(u)\, p^*(u) du = \mathbb{E}[Xf(X)I(X \geq 0)]$. □

Here is a way to construct X^* explicitly when dealing with a sum.

Lemma 4 *Let $\xi_1, \ldots, \xi_n$ be independent, mean 0 random variables and let $\sigma_i^2 = \mathrm{Var}(\xi_i)$. Let $W = \sum_i \xi_i$. Let $\xi_1^*, \ldots, \xi_n^*$ be independent and zero-bias. Define*

$$W^* = W - \xi_J + \xi_J^*$$

where $\mathbb{P}(J = i) = \sigma_i^2$. Then W^* has the W-zero bias distribution. In particular, suppose that $X_1, \ldots, X_n$ have mean 0 common variance and let $W = \frac{1}{\sqrt{n}} \sum_i X_i$. Let J be a random integer from 1 to n. Then $W^* = W - \frac{1}{\sqrt{n}} X_J + \frac{1}{\sqrt{n}} X_J^*$ has the W-zero bias distribution.

Proof. Let W^* be the zero-bias random variable for W. Let $W^i = W - \xi_i$ and note that W^i and ξ_i are independent. Then

$$\mathbb{E}[f'(W^*)] = \mathbb{E}[Wf(W)] = \sum_i \mathbb{E}[\xi_i f(W)] = \sum_i \mathbb{E}[\xi_i f(W^i + \xi_i)]$$
$$= \sum_i \mathbb{E}[\mathbb{E}[\xi_i f(W^i + \xi_i) | W^i]] = \text{check} \sum_i \mathbb{E}[\sigma_i^2 f'(W - \xi_i + \xi_i^*)]$$
$$= \sum_i \sigma_i^2 \mathbb{E}[f'(W - \xi_i + \xi_i^*)] = \sum_i \mathbb{E}[I(J = i)]\, \mathbb{E}[f'(W - \xi_i + \xi_i^*)]$$
$$= \mathbb{E}\left[\sum_i I(J = i) f'(W - \xi_i + \xi_i^*)\right] = \mathbb{E}[f'(W - \xi_J + \xi_J^*)].$$

So $\mathbb{E}[f'(W^*)] = \mathbb{E}[f'(W - \xi_J + \xi_J^*)]$ for all absolutely continuous f which implies that $W - \xi_J + \xi_J^* \stackrel{d}{=} W^*$. □

Now we can prove the following theorem using zero-biasing. We focus on the bounded case for simplicity.

Theorem 5 *Suppose that $|X_i| \leq B$. Then*

$$\sup_z \Big| \mathbb{P}(X \leq z) - \mathbb{P}(Y \leq z) \Big| \leq \frac{6B}{\sqrt{n}}.$$

Proof. Let $X_1^*, \ldots, X_n^*$ be zero-bias independent random variables. Let J be chosen randomly from $\{1, \ldots, n\}$ and let

$$X^* = X - \frac{1}{\sqrt{n}}(X_J - X_J^*).$$

Then, by the last lemma, X^* is zero-bias for X and hence

$$\mathbb{E}[Xf(X)] = \mathbb{E}[f'(X^*)] \tag{14}$$

for all absolutely continuous f. Also note that

$$|X^* - X| \le \frac{2B}{\sqrt{n}} \equiv \delta. \tag{15}$$

So,

$$\begin{aligned}
\mathbb{P}(X \le z) - \mathbb{P}(Y \le z) &\le \mathbb{P}(X \le z) - \mathbb{P}(Y \le z + \delta) + \mathbb{P}(Y \le z + \delta) - \mathbb{P}(Y \le z) \\
&\le \mathbb{P}(X \le z) - \mathbb{P}(Y \le z + \delta) + \frac{\delta}{\sqrt{2\pi}} \\
&\le \mathbb{P}(X^* \le z + \delta) - \mathbb{P}(Y \le z + \delta) + \frac{\delta}{\sqrt{2\pi}} \\
&\le \sup_z |\mathbb{P}(X^* \le z + \delta) - \mathbb{P}(Y \le z + \delta)| + \frac{\delta}{\sqrt{2\pi}} \\
&= \sup_z |\mathbb{P}(X^* \le z) - \mathbb{P}(Y \le z)| + \frac{\delta}{\sqrt{2\pi}}.
\end{aligned}$$

By a symmetric argument, we deduce that

$$\sup_z |\mathbb{P}(X \le z) - \mathbb{P}(Y \le z)| \le \sup_z |\mathbb{P}(X^* \le z) - \mathbb{P}(Y \le z)| + \frac{\delta}{\sqrt{2\pi}}.$$

Let $f = f_z$. From (8), (9) and (14),

$$\begin{aligned}
\sup_z |\mathbb{P}(X^* \le z) - \mathbb{P}(Y \le z)| &\le \sup_{f \in \mathscr{F}} \left| \mathbb{E}[f'(X^*) - X^* f(X^*)] \right| \\
&= \sup_{f \in \mathscr{F}} \left| \mathbb{E}[Xf(X) - X^* f(X^*)] \right| \\
&\le \mathbb{E}\left[\left(|X| + \frac{2\pi}{4} \right) |X^* - X| \right] \\
&\le \delta \left(1 + \frac{2\pi}{4} \right).
\end{aligned}$$

Combining these inequalities we have

$$\sup_z |\mathbb{P}(X \le z) - \mathbb{P}(Y \le z)| \le \delta\left(1 + \frac{1}{\sqrt{2\pi}} + \frac{2\pi}{4}\right) \le 3\delta = \frac{6B}{\sqrt{n}}.$$

□

4 The K-function

When X has mean 0, define the K-function

$$K(t) = \mathbb{E}\Big(X\ \Big[I(0 \le t \le X) - I(X \le t < 0)\Big]\Big) = \mathbb{E}[XI(X \ge t)]. \tag{16}$$

In particular, let K_i be the K-function for ξ_i:

$$K_i(t) = \mathbb{E}\Big(\xi_i\ \Big[I(0 \le t \le \xi_i) - I(\xi_i \le t < 0)\Big]\Big) = \mathbb{E}[\xi_i I(\xi_i \ge t)]. \tag{17}$$

K_i has the following properties:

$$K_i(t) \ge 0, \quad \int_{-\infty}^{\infty} K_i(t)dt = \mathbb{E}(\xi_i^2), \quad \sum_i \int_{-\infty}^{\infty} K_i(t)dt = 1, \quad \int_{-\infty}^{\infty} |t|\ K_i(t)dt = \frac{1}{2}\mathbb{E}|\xi_i|^3.$$

Note that the K-function is just the density for the zero-bias distribution.

Theorem 6 *Suppose that $|X_i| \le B$. Then*

$$\sup_z \Big|\mathbb{P}(X \le z) - \mathbb{P}(Y \le z)\Big| \le \frac{3.3B}{\sqrt{n}}.$$

Proof. Before plunging into the details, I'll outline the main idea of the proof which has three steps. Let $f \equiv f_z$. **Step one** is to establish that

$$\sum_i \int_{-\infty}^{\infty} \mathbb{P}(X^i + t \le z)K_i(t)dt - \Phi(z) = \sum_i \int_{-\infty}^{\infty} \mathbb{E}[Xf(X) - (X^i + t)f(X^i + t)]K_i(t)dt. \tag{18}$$

Step two is to use (8) to show that the right hand side is bounded by $C/\sqrt{n}$. **Step three** is to show that

$$\begin{aligned}\sum_i \int_{-\infty}^{\infty} \mathbb{P}(X^i + t \le z) K_i(t)dt - \Phi(z) &\approx \sum_i \int_{-\infty}^{\infty} \mathbb{P}(X \le z) K_i(t)dt - \Phi(z) \\ &= \mathbb{P}(X \le z) \sum_i \int_{-\infty}^{\infty} K_i(t)dt - \Phi(z) \\ &= \mathbb{P}(X \le z) - \Phi(z)\end{aligned}$$

since $\sum_i \int_{-\infty}^{\infty} K_i(t)dt = 1$. Thus, $|\mathbb{P}(X \le z) - \Phi(z)| \le C/\sqrt{n}$ as required. Keep in mind, throughout, that X^i is independent of ξ_i.

Step 1. Recall that $f'(x) - xf(x) = I(x \le z) - \Phi(z)$, so that $f'(x) = xf(x) + I(x \le z) - \Phi(z)$. Since $X = \sum_i \xi_i$, we have

$$\begin{aligned}\mathbb{E}[Xf(X)] &= \sum_i \mathbb{E}[\xi_i f(X)] = \sum_i \mathbb{E}[\xi_i (f(X) - f(X^i) + f(X^i))] \\ &= \sum_i \mathbb{E}[\xi_i (f(X) - f(X^i))] + \sum_i \mathbb{E}[\xi_i f(X^i)] \\ &= \sum_i \mathbb{E}[\xi_i (f(X) - f(X^i))] + \sum_i \mathbb{E}[\xi_i]\, \mathbb{E}[f(X^i)] \\ &= \sum_i \mathbb{E}[\xi_i (f(X) - f(X^i))] = \sum_i \mathbb{E}[\xi_i (f(X^i + \xi_i) - f(X^i))] \\ &= \sum_i \mathbb{E}\left[\xi_i \int_0^{\xi_i} f'(X^i + t)dt\right] \\ &= \sum_i \mathbb{E}\left[\int_{-\infty}^{\infty} f'(X^i + t)\xi_i [I(0 \le t \le \xi_i) - I(\xi_i \le t < 0)\right] dt \\ &= \sum_i \mathbb{E}\left[\int_{-\infty}^{\infty} f'(X^i + t)\right] \mathbb{E}\big[\xi_i [I(0 \le t \le \xi_i) - I(\xi_i \le t < 0)\big] dt \\ &= \sum_i \int_{-\infty}^{\infty} \mathbb{E}[f'(X^i + t)] K_i(t)dt \\ &= \sum_i \int_{-\infty}^{\infty} \mathbb{E}[(X^i + t)f(X^i + t) + I(X^i + t \le z) - \Phi(z)] K_i(t)dt \\ &= \sum_i \int_{-\infty}^{\infty} \mathbb{E}[(X^i + t)f(X^i + t)] K_i(t)dt + \sum_i \int_{-\infty}^{\infty} \mathbb{E}[I(X^i + t \le z) - \Phi(z)] K_i(t)dt \\ &= \sum_i \int_{-\infty}^{\infty} \mathbb{E}[(X^i + t)f(X^i + t)] K_i(t)dt + \sum_i \int_{-\infty}^{\infty} \mathbb{P}(X^i + t \le z) K_i(t)dt - \Phi(z)\end{aligned}$$

Thus we have the key equation:

$$\sum_i \int_{-\infty}^{\infty} \mathbb{P}(X^i + t \le z) K_i(t)dt - \Phi(z) = \sum_i \int_{-\infty}^{\infty} \mathbb{E}[Xf(X) - (X^i+t)f(X^i+t)]K_i(t)dt. \tag{19}$$

To complete the proof we need to show that the right hand side is small and that then left had side is close to $\mathbb{P}(X \le z) - \Phi(z)$.

Step 2. Let RHS denote the right hand side of (19). From (8) and the fact that $X = X^i + \xi_i$, we have

$$\begin{aligned}
|\text{RHS}| &\le \sum_i \mathbb{E} \int_{-\infty}^{\infty} \Big|(X^i + \xi_i) f(X^i + \xi_i) - (X^i + t) f(X^i + t)\Big| K_i(t) dt \\
&\le \sum_i \int_{-\infty}^{\infty} \mathbb{E}(|W^i| + c)\ \mathbb{E}(|\xi_i| + |t|) K_i(t) dt \\
&\le (1+c) \sum_i \int_{-\infty}^{\infty} \mathbb{E}(|\xi_i| + |t|) K_i(t) dt \\
&= (1+c) \sum_i [\mathbb{E}|\xi_i|\ \mathbb{E}(\xi_i^2) + \frac{1}{2}\mathbb{E}|\xi_i|^3] \\
&\le (1+c) \sum_i [\mathbb{E}|\xi_i|^3 + \frac{1}{2}\mathbb{E}|\xi_i|^3] \\
&= \frac{3(1+c)}{2} \sum_i \mathbb{E}|\xi_i|^3 \le 2.44 \sum_i \mathbb{E}|\xi_i|^3 \le 2.44 B_n \sum_i \mathbb{E}|\xi_i|^2 = 2.44 B_n.
\end{aligned}$$

We have now shown that, for each z,

$$\left| \sum_i \int_{-\infty}^{\infty} \mathbb{P}(X^i + t \le z) K_i(t) dt - \Phi(z) \right| \le 2.44 B_n. \tag{20}$$

Step 3. Since $|X_i| \le B$, $|\xi_i| \le B/\sqrt{n} \equiv B_n$. Note that $K_i(t) = 0$ when $|t| > B_n$. When $|\xi_i| \le B_n$ and $|t| \le B_n$ we have

$$\mathbb{P}(X \le z) \le \mathbb{P}(X - \xi_i + t \le z + 2B_n) = \mathbb{P}(X^i + t \le z + 2B_n).$$

Using (20)

$$\begin{aligned}
\mathbb{P}(X \le z) - \Phi(z) &= \sum_i \int \mathbb{P}(X \le z) K_i(t) dt - \Phi(z) \\
&\le \sum_i \int \mathbb{P}(X^i + t \le z + 2B_n) K_i(t) dt - \Phi(z + 2B_n) + \Phi(z + 2B_n) - \Phi(z) \\
&\le 2.44 B_n + \frac{2B_n}{\sqrt{2\pi}} \le \frac{3.3B}{\sqrt{n}}.
\end{aligned}$$

By a similar argument $\Phi(z) - \mathbb{P}(X \le z) \le \frac{3.3B}{\sqrt{n}}$. □

Connection to Zero-Biasing. As I mentioned earlier, $K(t)$ is the zero-bias density. In the K-function proof, we showed that

$$\mathbb{E}[Xf(X)] = \sum_i \int \mathbb{E}[f'(X^i + t)]K_i(t)dt.$$

So, with J randomly chosen from $\{1, \dots, n\}$, we have

$$\mathbb{E}[Xf(X)] = \sum_i \int \mathbb{E}[f'(X^i + t)]K_i(t)dt = \mathbb{E}[f'(X_J + X_J^*)] = \mathbb{E}[f'(X^*)].$$

5 Concentration

We can adapt the K-function approach to deal with the unbounded case, using a trick that Chen, Goldstein and Shao call "concentration." This is different from "concentration of measure" as in things like Bernstein's inequality.

Previously, we used boundedness of X_i in the K-function proof to show that $\mathbb{P}(X^i + t \le z)$ was close to $\mathbb{P}(X \le z) = \mathbb{P}(X^i + \xi_i \le z)$. It turns out that we can show the two quantities are close, without assuming boundedness, if we can show that $\mathbb{P}(a \le X^i \le b) \le (b-a)+$ something small. The same trick will be useful for dealing with nonlinear functions later.

Lemma 7 *For all $a \le b$,*

$$\mathbb{P}(a \le X^i \le b) \le \sqrt{2}(b-a) + \frac{2(\sqrt{2}+1)\mu_3}{\sqrt{n}}. \tag{21}$$

The proof is a bit long so I won't reproduce it here; see Chapter 3 of Chen, Goldstein and Shao (2011).

Theorem 8 *Assuming $\mu_3 < \infty$ we have $\Delta_n \le \frac{9.4\mu_3}{\sqrt{n}}$.*

Proof. We have already shown in (20) that

$$\left| \sum_{i=1}^n \int \mathbb{P}(X^i + t \le z)K_i(t)dt - \Phi(z) \right| \le \frac{2.44\mu_3}{\sqrt{n}}. \tag{22}$$

We use concentration to bound $\mathbb{P}(X^i + t \le z) - \mathbb{P}(X \le z)$. We have

$$
\begin{aligned}
\left|\sum_{i=1}^{n} \int \mathbb{P}(X^i + t \le z) K_i(t)dt - \mathbb{P}(X \le z)\right| &\le \sum_{i=1}^{n} \int |\mathbb{P}(X^i + t \le z) K_i(t)dt - \mathbb{P}(X \le z)| K_i(t)dt \\
&= \sum_{i=1}^{n} \int |\mathbb{P}(X^i + t \le z) K_i(t)dt - \mathbb{P}(X^i + \xi_i \le z)| K_i(t)dt \\
&= \sum_{i=1}^{n} \int \mathbb{P}(z - (t \vee \xi_i) \le X^i \le z - (t \wedge \xi_i)) K_i(t)dt \\
&= \sum_{i=1}^{n} \int \mathbb{E}[\mathbb{P}(z - (t \vee \xi_i) \le X^i \le z - (t \wedge \xi_i) | \xi_i) K_i(t)dt \\
&\le \sum_{i=1}^{n} \int \mathbb{E}[\sqrt{2}(|t| + |\xi_i|) + 2(\sqrt{2}+1)\frac{\mu_3}{\sqrt{n}}] K_i(t)dt \\
&= \sqrt{2} \sum_{i=1}^{n} \left(\frac{\mu_3}{2} + \mathbb{E}|\xi_i|\ \mathbb{E}[\xi_i^2]\right) + \frac{2(\sqrt{2}+1)\mu_3}{\sqrt{n}} \\
&\le \frac{6.95\mu_3}{\sqrt{n}}.
\end{aligned}
$$

The result follows from (22). □

6 Exchangeable Pairs

Another approach to Stein's method is based on finding pairs of variables with a special property. Specifically, we say that (X, X') is an *exchangeable pair* if (X, X') is equal in distribution to (X', X), written $(X, X') \stackrel{d}{=} (X', X)$. We say that (X, X') is a *Stein pair* if they are exchangeable and if there exists $\lambda \in (0, 1)$ such that

$$\mathbb{E}[X' | X] = (1 - \lambda)X. \tag{23}$$

For example, let $\xi_1', \ldots, \xi_n'$ be independent of $\xi_1, \ldots, \xi_n$. Define $X' = X - \xi_J + \xi_J'$ where J is chosen uniformly from $\{1, \ldots, n\}$. Then (X, X') is a Stein pair with $\lambda = 1/n$, so that

$$\mathbb{E}[X' | X] = \left(1 - \frac{1}{n}\right) X. \tag{24}$$

Theorem 9 *Let (X, X') be a Stein pair where W has mean 0 and variance 1. Suppose that $|X' - X| \le \delta$. Let*

$$B = \frac{\sqrt{\mathrm{Var}(\mathbb{E}((X' - X)^2 | X))}}{2\lambda}.$$

Then

$$\sup_z |\mathbb{P}(X \le z) - \Phi(z)| \le 1.1\delta + \frac{\delta^3}{2\lambda} + 2.7B.$$

We will omit the proof but we note that it is similar to the K-function proof.

7 Smoothing + Induction

This approach, following section 3.7 of Nourdin and Peccati (2012), combines two ideas: smoothing and induction.

The smoothing idea is based on approximating the indicator function $I(x \le z)$ with a smooth function h and then applying Stein's method directly to h. Specifically, define

$$h_{z,\epsilon}(x) = \begin{cases} 1 & \text{if } x \le z - \epsilon \\ \frac{z+\epsilon-x}{2\epsilon} & \text{if } z-\epsilon < x < z+\epsilon \\ 0 & \text{if } x \ge z+\epsilon. \end{cases}$$

The induction idea is as follows. Assume that $\mathbb{E}[|X_i|^3] < \infty$. Let C_n be the smallest real number such that

$$\sup_z \Big| \mathbb{P}(X \le z) - \mathbb{P}(Z \le z) \Big| \le \frac{C_n \mathbb{E}[|X_i|^3]}{\sqrt{n}}. \tag{25}$$

Such a C_n always exists. But if C_n grows with n then this is a useless bound. The goal is to use an inductive argument to show that there is some fixed $C > 0$ such that $\limsup_{n\to\infty} C_n \le C$.

Theorem 10 *Suppose that* $\mu_3 \equiv \mathbb{E}[|X_i|^3] < \infty$. *Then*

$$\sup_z \Big| \mathbb{P}(X \le z) - \mathbb{P}(Y \le z) \Big| \le \frac{33\,\mu_3}{\sqrt{n}}.$$

Before proving the theorem, we need the following lemma.

Lemma 11 *For all* $a \le b$,

$$\mathbb{P}(a \le X^i \le b) \le \frac{b-a}{\sqrt{2\pi}\sqrt{1-\frac{1}{n}}} + \frac{2C_{n-1}\mu_3}{\sqrt{n-1}} \equiv H_n(a,b). \tag{26}$$

Proof. We can write $X^i = \sqrt{1-\frac{1}{n}}Q_i$ where $Q_i = \frac{1}{\sqrt{n-1}}\sum_{j\neq i} X_j$. Let $Z \sim N(0,1)$. Then

$$\begin{aligned}
\mathbb{P}(a \le X^i \le b) &= \mathbb{P}\left(\frac{a}{\sqrt{1-\frac{1}{n}}} \le Q_i \le \frac{b}{\sqrt{1-\frac{1}{n}}}\right) \\
&= \mathbb{P}\left(\frac{a}{\sqrt{1-\frac{1}{n}}} \le Q_i \le \frac{b}{\sqrt{1-\frac{1}{n}}}\right) - \mathbb{P}\left(\frac{a}{\sqrt{1-\frac{1}{n}}} \le Z \le \frac{b}{\sqrt{1-\frac{1}{n}}}\right) \\
&\quad + \mathbb{P}\left(\frac{a}{\sqrt{1-\frac{1}{n}}} \le Z \le \frac{b}{\sqrt{1-\frac{1}{n}}}\right) \\
&\le \frac{2C_{n-1}\mu_3}{\sqrt{n-1}} + \frac{b-a}{\sqrt{2\pi}\sqrt{1-\frac{1}{n}}}.
\end{aligned}$$

□

Now we prove the theorem.

Proof. We break the proof into two steps.

Part 1. Smoothing. From the definition of $h_{z,\epsilon}$ it follows that

$$\mathbb{E}[h_{z-\epsilon,\epsilon}(X)] \le \mathbb{P}(X \le z) \le \mathbb{E}[h_{z+\epsilon,\epsilon}(X)].$$

Also, if $Z \sim N(0,1)$ then

$$\begin{aligned}
\mathbb{E}[h_{z+\epsilon,\epsilon}(Z)] - \frac{4\epsilon}{\sqrt{2\pi}} &\le \mathbb{E}[h_{z-\epsilon,\epsilon}(Z)] \le \mathbb{P}(Z \le z) \\
&\le \mathbb{E}[h_{z+\epsilon,\epsilon}(Z)] \le \mathbb{E}[h_{z-\epsilon,\epsilon}(Z)] + \frac{4\epsilon}{\sqrt{2\pi}}.
\end{aligned}$$

Therefore,

$$\sup_z \Big|\mathbb{P}(X \le z) - \mathbb{P}(Z \le z)\Big| \le \sup_z \Big|\mathbb{E}[h_{z,\epsilon}(X)] - \mathbb{E}[h_{z,\epsilon}(Z)]\Big| + \frac{4\epsilon}{\sqrt{2\pi}}. \tag{27}$$

Let $f = f_{z,\epsilon}$ be the Stein function for $h_{z,\epsilon}$. It may be verified that

$$||f||_\infty \le \sqrt{\frac{\pi}{2}}, \quad ||f'||_\infty \le 2$$

and that

$$|xf(x) - yf(y)| = |f(x)(x-y) + (f(x) - f(y))\, y| \le \left(\sqrt{\frac{\pi}{2}} + 2|y|\right)|x-y|. \tag{28}$$

Now we again use the leave-one-out trick. Let $X^i = X - n^{-1/2}X_i$. Note that $\mathbb{E}[f(X^i)X_i] = \mathbb{E}[f(X^i)]\mathbb{E}[X_i] = 0$. Hence,

$$\begin{aligned}\mathbb{E}[h(X)] - \mathbb{E}[h(Z)] = \mathbb{E}[f'(X) - Xf(X)] &= \sum_i \mathbb{E}\left[\frac{1}{n}f'(X) - \frac{X_i}{\sqrt{n}}f(X)\right] \\ &= \sum_i \mathbb{E}\left[\frac{1}{n}f'(X) - \frac{X_i}{\sqrt{n}}(f(X) - f(X^i))\right].\end{aligned}$$

Now,

$$\begin{aligned}f(X) - f(X^i) &= f(X^i + X_i/\sqrt{n}) - f(X^i) \\ &= \int_0^{X_i/\sqrt{n}} f'(X^i + t)dt = \frac{X_i}{\sqrt{n}}\int_0^1 f'(X^i + uX_i/\sqrt{n})du \\ &= \frac{X_i}{\sqrt{n}}\mathbb{E}\left[f'(X^i + UX_i/\sqrt{n})\right]\end{aligned}$$

where U is an independent Unif(0,1) random variable and the expectation is over U. Thus,

$$\mathbb{E}[h(X)] - \mathbb{E}[h(Z)] = \sum_i \mathbb{E}\left[\frac{1}{n}f'(X) - \frac{X_i^2}{n}f'(X^i + UX_i/\sqrt{n})\right].$$

The Stein equation $f'(x) - xf(x) = h(x) - \mu$ implies that

$$f'(x) = xf(x) + h(x) - \mu$$

and so

$$\begin{aligned}\frac{1}{n}f'(X) - \frac{X_i^2}{n}f'(X^i + UX_i/\sqrt{n}) = &\frac{1}{n}Xf(X) - \frac{X_i^2}{n}\left(X^i + UX_i/\sqrt{n}\right)f(X^i + UX_i/\sqrt{n}) \\ &+ \frac{1}{n}h(X) - \frac{X_i^2}{n}h(X^i + UX_i/\sqrt{n})\end{aligned}$$

and so

$$\mathbb{E}[h(X)] - \mathbb{E}[h(Z)] = \mathrm{I} - \mathrm{II} + \mathrm{III} - \mathrm{IV}$$

where

$$
\begin{aligned}
\mathrm{I} &= \sum_i \frac{1}{n}\mathbb{E}[Xf(X) - X^i f(X^i)] \\
\mathrm{II} &= \sum_i \mathbb{E}\left[\frac{X_i^2}{n}\left(\left(X^i + UX_i/\sqrt{n}\right)f(X^i + UX_i/\sqrt{n}) - X^i f(X^i)\right)\right] \\
\mathrm{III} &= \sum_i \frac{1}{n}\mathbb{E}[h(X) - h(X^i)] \\
\mathrm{IV} &= \sum_i \mathbb{E}\left[\frac{X_i^2}{n}\left(h(X^i + UX_i/\sqrt{n}) - h(X^i)\right)\right]
\end{aligned}
$$

Now we bound these four terms. We repeatedly use (28) and the fact that X^i is independent of X_i.

Bounding I. We have

$$
\mathrm{I} \le \sum_i \frac{1}{n\sqrt{n}}\left(\sqrt{\frac{\pi}{2}} + 2\mathbb{E}|X^i|\right)\mathbb{E}|X_i| \le \frac{1}{\sqrt{n}}\left(\sqrt{\frac{\pi}{2}} + 2\right)
$$

since $\mathbb{E}[|X^i|] \le 1$ and $\mathbb{E}[|X_i|] \le 1$.

Bounding II. We have

$$
\mathrm{II} \le \sum_i \frac{1}{n\sqrt{n}}\left(\mathbb{E}[U]\mathbb{E}[|X_1|^3]\sqrt{\frac{\pi}{2}} + 2\mathbb{E}[U]\mathbb{E}[|X_1|^3]\mathbb{E}[|X^i|]\right) \le \frac{\mathbb{E}[|X_1|^3]}{\sqrt{n}}\left(\frac{1}{2}\sqrt{\frac{\pi}{2}} + 1\right).
$$

Bounding III. To bound this term we Taylor expand the function h. Note that $h'(x) = -J(x)/(2\epsilon)$ where $J(x)$ is the indicator function for $[z-\epsilon < x < z+\epsilon]$. Hence,

$$
h(y) - h(x) = (y-x)\int_0^1 h'(x + s(y-x))ds = -\frac{y-x}{2\epsilon}\mathbb{E}[J(x + V(y-x))]
$$

where $V \sim U(0,1)$. Hence,

$$
\begin{aligned}
\mathrm{III} &\le \sum_i \frac{1}{2\epsilon n\sqrt{n}}\mathbb{E}\left[|X_i|\ J\left(X^i + V\frac{X_i}{\sqrt{n}}\right)\right] \\
&= \frac{1}{2\epsilon\sqrt{n}}\mathbb{E}\left[|X_i|\ \mathbb{P}\left(z - \frac{VX_i}{\sqrt{n}} - \epsilon \le X^i \le z - \frac{VX_i}{\sqrt{n}} + \epsilon\right)\right] \\
&\le \frac{1}{2\epsilon\sqrt{n}}\sup_{0\le t\le 1}\sup_{y\in\mathbb{R}}\mathbb{P}\left(z - \frac{ty}{\sqrt{n}} - \epsilon \le X^i \le z - \frac{ty}{\sqrt{n}} + \epsilon\right) \\
&\le \frac{1}{\sqrt{2\pi}\sqrt{n-1}} + \frac{C_{n-1}\mu_3}{\sqrt{n}\sqrt{n-1}\epsilon}
\end{aligned}
$$

where we used (26).

Bounding IV. This is bounded using the same argument as III. Let $U,V \sim \text{Unif}(0,1)$ be independent. Then

$$\begin{aligned}
\text{IV} &\leq \frac{1}{2n\sqrt{n}\epsilon}\sum_i \left|\mathbb{E}\left[X_i^3 U J\left(X^i + UV\frac{X_i}{\sqrt{n}}\right)\right]\right| \\
&\leq \frac{\mu_3}{4\sqrt{n}\epsilon}\sup_{0\leq t\leq 1}\sup_{y\in\mathbb{R}}\mathbb{P}\left(z - \frac{ty}{\sqrt{n}} - \epsilon \leq X^i \leq z - \frac{ty}{\sqrt{n}} + \epsilon\right) \\
&\leq \frac{\mu_3}{2\sqrt{2\pi}\sqrt{n-1}} + \frac{C_{n-1}\mu_3^2}{2\sqrt{n}\sqrt{n-1}\epsilon}.
\end{aligned}$$

Combining these bounds, we get

$$\sup_z |\mathbb{E}[h_{z,\epsilon}(X)] - \mathbb{E}[h_{z,\epsilon}(Y)]| \leq \frac{6\mu_3}{\sqrt{n}} + \frac{3C_{n-1}\mu_3^2}{n\epsilon}. \tag{29}$$

From (29) and (27) we have

$$\sup_z |\mathbb{P}(X \leq z) - \mathbb{P}(Z \leq z)| \leq \frac{6\mu_3}{\sqrt{n}} + \frac{3C_{n-1}\mu_3^2}{n\epsilon} + \frac{4\epsilon}{\sqrt{2\pi}}.$$

Setting

$$\epsilon = \mu_3\sqrt{\frac{C_{n-1}}{n}}$$

we get

$$\Delta_n \leq \frac{\mu_3}{\sqrt{n}}\left(6 + \left(3 + \frac{4}{\sqrt{2\pi}}\right)\sqrt{C_{n-1}}\right). \tag{30}$$

Part 2. Induction. First we get a crude bound on C_n. Since $\Delta_n \leq 1$ we must have that $C_n \leq \sqrt{n}/\mu_3$. But $\mu_3 = \mathbb{E}[|Y_1|^3] \geq \mathbb{E}[Y_1^2]^{3/2} = 1$ and so $C_n \leq \sqrt{n}$. Therefore, $C_1 \leq 33$. Equation (30) gives a bound on Δ_n. But, by the definition of C_n, the tightest bound on Δ_n is given by (25). This implies that

$$C_n \leq \left(6 + \left(3 + \frac{4}{\sqrt{2\pi}}\right)\sqrt{C_{n-1}}\right).$$

It follows from induction that $C_n \leq 33$ for all n. □

8 Slepian Interpolation + Smoothing

The approach in this section is due to Chernozhukov, Chetverikov and Kato (2012) and is closely related to the approach in Chatterjee (2008) and Rollin (2013). For one-dimensional problems it leads to sub-optimal rates (due to the smoothing of the indicator functions). In Part II we show how Chernozhukov, Chetverikov and Kato used it successfully for high-dimensional problems. Here, I will focus on one dimension.

The key idea is to define the *Slepian Smart Path Interpolation*

$$Z(t) = \sqrt{t}X + \sqrt{1-t}Y = \sum_i \frac{1}{\sqrt{n}}(\sqrt{t}X_i + \sqrt{1-t}Y_i) \equiv \sum_i Z_i(t). \tag{31}$$

Theorem 12 *We have*

$$\sup_t |\mathbb{P}(X \le t) - \mathbb{P}(Y \le t)| \le \frac{C}{n^{1/8}}.$$

Proof. We have

$$Z'(t) \equiv \frac{dZ(t)}{dt} = \left(\frac{X}{\sqrt{t}} - \frac{Y}{\sqrt{1-t}}\right) = \sum_i \frac{1}{\sqrt{n}}\left(\frac{X_i}{\sqrt{t}} - \frac{Y_i}{\sqrt{1-t}}\right) \equiv \sum_i Z_i'(t).$$

Define the leave-one-out quantity $Z^i(t) = Z(t) - Z_i(t)$. Let $g \in C^3$. Then

$$\begin{aligned}\mathbb{E}[g(X)] - \mathbb{E}[g(Y)] &= \mathbb{E}(g(Z(0))) - \mathbb{E}(g(Z(1))) = \int_0^1 \mathbb{E}\left[\frac{dg(Z(t))}{dt}\right]dt = \int_0^1 \mathbb{E}[g'(Z(t))]\ Z'(t)dt \\ &= \frac{1}{2}\sum_i \int_0^1 \mathbb{E}[g'(Z(t))] Z_i'(t).\end{aligned}$$

In general, Taylor's theorem with integral remainder gives

$$\begin{aligned}f(x) &= f(a) + (x-a)f'(a) + \int_a^x (x-t)f''(t)dt \\ &= f(a) + (x-a)f'(a) + (x-a)^2\int_0^1 (1-u)f''(a+u(x-a))du\end{aligned}$$

where we used the transformation $u = \frac{t-a}{x-a}$. Apply this to g' with $x = Z(t)$ and $a = Z^i(t)$ to get

$$g'(Z(t)) = g'(Z^i(t)) + Z_i(t)g''(Z^i(t)) + Z_i^2(t)\int_0^1 (1-\tau)g'''(Z^i(t) + \tau Z_i(t))\,d\tau.$$

Therefore,

$$\mathbb{E}(g(X)) - \mathbb{E}(g(Y)) = \frac{1}{2}(\mathrm{I} + \mathrm{II} + \mathrm{III})$$

where

$$
\begin{aligned}
\mathrm{I} &= \sum_i \int_0^1 \mathbb{E}[g'(Z^i(t))\; Z_i'(t)]dt \\
\mathrm{II} &= \sum_i \int_0^1 \mathbb{E}[g''(Z^i(t))\; Z_i'(t)\; Z_i(t)]dt \\
\mathrm{III} &= \sum_i \int_0^1 \int_0^1 (1-\tau)\mathbb{E}[g'''(Z^i(t)+\tau Z_i(t))\; Z_i'(t)\; Z_i^2(t)]d\tau dt.
\end{aligned}
$$

Note that $g'(Z^i(t))$ and $Z_i'(t)$ are independent and $\mathbb{E}[Z_i'(t)] = 0$. So

$$
\mathrm{I} = \sum_i \int_0^1 \mathbb{E}[g'(Z^i(t))\; Z_i'(t)]dt = \sum_i \int_0^1 \mathbb{E}[g'(Z^i(t))]\; \mathbb{E}[Z_i'(t)]dt = 0.
$$

To bound II, note that $Z^i(t)$ is independent of $Z_i'(t)Z_i(t)$. Let $C = \sup|g''(t)|$. Then,

$$
\begin{aligned}
|\mathrm{II}| &\le \sum_i \int_0^1 \mathbb{E}\left|g''(Z^i(t))\right| \left|\mathbb{E}[Z_i'(t)\; Z_i(t)]\right|dt \\
&\le C\int_0^1 \sum_i \left|\mathbb{E}[Z_i'(t)\; Z_i(t)]\right|dt.
\end{aligned}
$$

Recall that

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

and so

$$
\begin{aligned}
Z_i'(t)\; Z_i(t) &= \frac{1}{\sqrt{n}}\left(\frac{X_i}{\sqrt{t}} - \frac{Y_i}{\sqrt{1-t}}\right)\frac{1}{\sqrt{n}}\left(\sqrt{t}X_i + \sqrt{1-t}Y_i\right) \\
&= \frac{1}{n}\left(X_i^2 - \sqrt{\frac{t}{1-t}}X_iY_i + \sqrt{\frac{1-t}{t}}X_iY_i - Y_i^2\right)
\end{aligned}
$$

and hence $\mathbb{E}[Z_i'(t)\; Z_i(t)] = 0$ and so $\mathrm{II} = 0$.

Next, consider III. The smoothness of g''' implies that

$$
g'''(z) \preceq g'''(z+\tau w) \preceq g'''(z).
$$

So,

$$
\begin{aligned}
|\mathrm{III}| &\le \sum_i \int_0^1 \int_0^1 \mathbb{E}[|g'''(Z^i(t)+\tau Z_i(t))|\ |Z_i'(t)\ Z_i^2(t)|]d\tau dt \\
&\lesssim \sum_i \int_0^1 \mathbb{E}[|g'''(Z^i(t))|\ |Z_i'(t)\ Z_i^2(t)|]dt \\
&= \sum_i \int_0^1 \mathbb{E}[|g'''(Z^i(t))|]\ \mathbb{E}[|Z_i'(t)\ Z_i^2(t)|]dt \\
&\lesssim \sum_i \int_0^1 \mathbb{E}[|g'''(Z(t))|]\ \mathbb{E}[|Z_i'(t)\ Z_i^2(t)|]dt \\
&= \int_0^1 \mathbb{E}[|g'''(Z(t))|]\ n\mathbb{E}[|Z_i'(t)\ Z_i^2(t)|]dt \\
&\le C'n \int_0^1 \mathbb{E}[|Z_i'(t)\ Z_i^2(t)|]dt.
\end{aligned}
$$

Let $\omega(t) = \frac{1}{\sqrt{t}\wedge\sqrt{1-t}}$. Then, using Holder's inequality,

$$
\begin{aligned}
\int_0^1 \mathbb{E}[|Z_i'(t)\ Z_i^2(t)|]dt &= \int_0^1 \omega(t)\mathbb{E}\left[\left|\frac{Z_i'(t)}{\omega(t)}\ Z_i^2(t)\right|\right]dt \\
&\le \int_0^1 \omega(t)\left(\mathbb{E}\left[\left|\frac{Z_i'(t)}{\omega(t)}\right|^3\right]\mathbb{E}|Z_i(t)|^3\mathbb{E}|Z_i(t)|^3\right)^{1/3} dt
\end{aligned}
$$

Now

$$
\left|\frac{Z_i'(t)}{\omega(t)}\right| \le \frac{|X_i|+|Y_i|}{\sqrt{n}}, \quad |Z_i(t)| \le \frac{|X_i|+|Y_i|}{\sqrt{n}}, \quad \int_0^1 \omega(t)dt \lesssim 1.
$$

Also,

$$
\mathbb{E}|Y_i|^3 \lesssim (\mathbb{E}|Y_i|^2)^{3/2} = (\mathbb{E}|X_i|^2)^{3/2} \le \mathbb{E}|X_i|^3.
$$

Hence,

$$
n\int \mathbb{E}|Z_i(t)|^3\mathbb{E}|Z_i(t)|^3 dt \le \frac{1}{\sqrt{n}}\mathbb{E}(|X_i|+|Y_i|)^3 \int_0^1 \omega(t)dt \lesssim \frac{1}{\sqrt{n}}\mathbb{E}|Y_i|^3.
$$

We have shown that

$$
\mathbb{E}|g(X)-g(Y)| \lesssim \frac{1}{\sqrt{n}}\mathbb{E}|Y_i|^3.
$$

We want to bound

$$
\mathbb{P}(X \le t) - \mathbb{P}(Y \le t) = \mathbb{E}(h(X)) - \mathbb{E}(h(Y))
$$

where $h(z) = I(z \le t)$. Let $g_0 : \mathbb{R} \to [0,1]$ be in C^3 be such that $g_0(s) = 1$ for $s \le 0$ and $g_0(s) = 0$ for $s \ge 1$. Define $g(s) = g_0(\psi(s-t))$. Then

$$
\sup_s |g(s)| = 1,\ \sup_s |g'(s)| \lesssim \psi,\ \ \sup_s |g''(s)| \lesssim \psi^2,\ \ \sup_s |g'''(s)| \lesssim \psi^3.
$$

Now

$$\begin{aligned}\mathbb{P}(X \le t) = \mathbb{E}(h(X)) \le \mathbb{E}(g(X)) \le \mathbb{E}(g(Y)) + \frac{\psi + \psi^2 + \psi^3}{\sqrt{n}} \\ \le \mathbb{P}(Y \le t + \psi^{-1}) + \frac{\psi + \psi^2 + \psi^3}{\sqrt{n}} \\ \le \mathbb{P}(Y \le t) + \frac{1}{\psi} + \frac{\psi + \psi^2 + \psi^3}{\sqrt{n}}\end{aligned}$$

Take $\psi = n^{1/8}$. Then

$$\mathbb{P}(X \le t) \le \mathbb{P}(Y \le t) + \frac{1}{n^{1/8}}.$$

A similar proof gives

$$\mathbb{P}(Y \le t) \le \mathbb{P}(X \le t) + \frac{1}{n^{1/8}}.$$

□

Again this is a suboptimal rate for dimension $d = 1$ but leads to the rate $\log d/n^{1/8}$ in the multivariate d-dimensional case as I explain in Part II.

9 Nonlinear Functions

Chapter 10 of Chen, Goldstein and Shao (2011) explains how to get bounds for non-linear functions. In fact, there are two approaches: the direct method and the concentration method. In general, the latter is sharper. However, when the distribution is bounded or sub-Gaussian, they lead to similar results.

Suppose that $T = X + \Delta$ where $X = \sum_{i=1}^{n} \xi_i$ where $\xi_1, \ldots, \xi_n$ are iid, mean 0 and variance 1. Let $Z \sim N(0,1)$. We know how to bound $\mathbb{P}(X \le z) - \mathbb{P}(Z \le z)$. But how do we deal with Δ?

Direct Approach. Let $\mathscr{E} = \{|\Delta| \le \epsilon\}$. Suppose that $\mathbb{E}|\Delta|^p < \infty$. Then

$$\begin{aligned}\mathbb{P}(X + \Delta \le z) - \Phi(z) &= \mathbb{P}(X + \Delta \le z, \mathscr{E}) + \mathbb{P}(X + \Delta \le z, \mathscr{E}^c) - \Phi(z) \\ &\le \mathbb{P}(X \le z + \epsilon) + \mathbb{P}(\mathscr{E}^c) - \Phi(z) \\ &= \mathbb{P}(X \le z + \epsilon) - \Phi(z + \epsilon) + [\Phi(z + \epsilon) - \Phi(z)] + \mathbb{P}(|\Delta| > \epsilon) \\ &\le \mathbb{P}(X \le z + \epsilon) - \Phi(z + \epsilon) + C\epsilon + \frac{\mathbb{E}[|\Delta|^p]}{\epsilon^p}\end{aligned}$$

for some $C > 0$. Optimizing over ϵ we get

$$\mathbb{P}(X+\Delta \le z) - \Phi(z) \le \mathbb{P}(X \le z+\epsilon) - \Phi(z+\epsilon) + C(\mathbb{E}[|\Delta|^p])^{\frac{1}{p+1}}$$

and so

$$\sup_z(\mathbb{P}(X+\Delta \le z) - \Phi(z)) \le \sup_z(\mathbb{P}(X \le z) - \Phi(z)) + C(\mathbb{E}[|\Delta|^p])^{\frac{1}{p+1}}.$$

We get a lower bound similarly and so

$$\sup_z|\mathbb{P}(X+\Delta \le z) - \Phi(z)| \le \sup_z|\mathbb{P}(X \le z) - \Phi(z)| + C(\mathbb{E}[|\Delta|^p])^{\frac{1}{p+1}}.$$

If Δ is bounded, or sub-Gaussian, we can get a tighter bound by using an exponential inequality instead of Markov's inequality. This can give an optimal or near optimal rate. Otherwise, we one should use the following approach.

Concentration Approach. Note that

$$-\mathbb{P}(z-|\Delta| \le X \le z) \le \mathbb{P}(T \le z) - \mathbb{P}(X \le z) \le \mathbb{P}(z \le X \le z+|\Delta|).$$

We discussed bounding quantities of the form $\mathbb{P}(a \le X \le b)$ in Section 5. But in this case, we need to allow a and b to be random. Fortunately, the concentration approach can be extended to this case.

Consider bounding $\mathbb{P}(A \le X \le B)$ where now A and B are random. Assume we can create leave-one-out versions of A and B, denoted A_i and B_i, so that ξ_i is independent of (X^i, A_i, B_i). Then Chen, Goldstein and Shao (2011) prove the following:

Theorem 13 *Let $\delta = C/\sqrt{n}$ where $C > 0$ is a sufficiently large, positive constant. Then,*

$$\mathbb{P}(A \le X \le B) \le 4\delta + \mathbb{E}|X(B-A)| + \sum_i (\mathbb{E}|\xi_i(A-A_i)| + \mathbb{E}|\xi_i(B-B_i)|).$$

In the iid case that we have been focusing on, $\delta = O(1/\sqrt{n})$. This leads to

$$\sup_z|\mathbb{P}(X+\Delta \le z) - \mathbb{P}(X \le z)| \le \frac{1}{\sqrt{n}} + \frac{\mathbb{E}[|\Delta - \Delta_i|]}{\sqrt{n}}.$$

We omit the proof.

10 Multivariate Case

There are a variety of approaches for the multivariate case. Let us first discuss bounding smooth functions. Let $X, Y \in \mathbb{R}^d$ be random vectors where Y is Gaussian. Suppose we want to bound $\left|\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]\right|$ for a smooth function $h : \mathbb{R}^d \to \mathbb{R}$. Rollin (2013) uses the Slepian path approach that we discussed earlier. As before, we create the interpolating path

$$Z_t = \sqrt{t}X + \sqrt{1-t}Y$$

for $0 \le t \le 1$. Suppose that $\mathbb{E}[X] = \mathbb{E}[Y] = (0,\dots,0)^T$ and that $\mathrm{Cov}(X) = \mathrm{Cov}(Y)$. Note that $\mathbb{E}[Z_t] = (0,\dots,0)^T$ and that $\mathrm{Cov}(Z_t) = \mathrm{Cov}(X)$.

We then have that

$$\begin{aligned}\mathbb{E}[h(X)] - \mathbb{E}[h(Y)] &= \int_0^1 \frac{\partial}{\partial t}\mathbb{E}[h(Y_t)]dt \\ &= \frac{1}{2}\int_0^1 \mathbb{E}\left[\frac{1}{\sqrt{t}}\sum_i X_i h_i(Y_t) - \frac{1}{\sqrt{1-t}}\sum_i Y_i h_i(Z_t)\right]dt\end{aligned}$$

where $h_i = \partial h(x)/\partial x_i$.

In the case where X has independent coordinates, one gets

$$\left|\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]\right| \le \frac{5}{6}\sum_{j=1}^{d} \mathbb{E}|X(j)|^3 \; ||h_{jjj}||_\infty$$

where h_{jjj} denotes the third derivative. Note that for a sum, $\mathbb{E}|X(j)|^3$ will be of order $O(n^{-1/2})$. This suggest that the bound has order $O(d/\sqrt{n})$.

The proof is based on bounding $\frac{\partial}{\partial t}\mathbb{E}[h(Z_t)]$ using a Stein coupling. In this context, a triple (X, X', G) is called a Stein coupling if, for all smooth f,

$$\mathbb{E}[\sum_i X_i f_i(X)] = \mathbb{E}[\sum_i G_i(f_i(X') - f_i(X))].$$

It can be shown that

$$\mathbb{E}[X_i] = 0, \quad \mathbb{E}[G_i D_j + G_j D_i] = 2\mathrm{Cov}(X_i, X_j)$$

where $D = X' - X$. This allows one to bound the derivative. In the independent case, we can construct the coupling as follows. Let I be drawn uniformly from $\{1,\dots,n\}$. Then define $G_i = -d\delta_{iI}X_i$ and $X'(k) = (1-\delta_{kI})X(k)$.

For statistical applications, we need to bound

$$\sup_{A\in\mathscr{A}}\left|\mathbb{P}(X \in A) - \mathbb{E}(Z \in A)\right|$$

for some class of sets $\mathscr{A}$. Let $\mathscr{A}$ be the class of convex sets. The best bound I know of in this case is due to Bentkus (2003). (The fact that this important paper is not in a leading journal suggests it was rejected by one of the main journals. Another example of th failure of our refereeing system.) He uses a complicated induction argument. The result is as follows. Let $X = n^{-1/2}\sum_i X_i$ be a sum of mean 0 random vectors in $\mathbb{R}^d$ with $\text{Cov}(X_i) = I$. Let $Z \sim N(0, I)$. Then

$$\sup_{A\in\mathscr{A}} \Big|\mathbb{P}(X \in A) - \mathbb{E}(Z \in A)\Big| \le \frac{400 d^{1/4}\beta}{\sqrt{n}} \tag{32}$$

where $\beta = \mathbb{E}||X||^3$. We expect that $\beta = O(d^{3/2})$. So the bound has size $O(d^{7/4}/\sqrt{n}) = O(\sqrt{d^{7/2}/n})$.

A key step in Bentkus' proof is to smooth the indicator function for a set A with a function of the form $h(x) = g(d(x, A)/\psi)$ where g is a smooth function, ψ is a real number and $d(x, A) = \inf_{y\in A}||x - y||$. In fact, smoothing can be combined with Stein's method as in Chen and Fang (2011). However, the result is not as tight as the Bentkus result.

A major breakthrough was obtained in Chernozhukov, Chetverikov and Kato (2012). The showed that if we take $\mathscr{A}$ to be the class of rectangles, then we get a rate of the form $\log d/n^{1/8}$. This result is very useful for statistics. Part II of these notes are devoted just to this result.

11 Slepian Versus Stein Versus Lindeberg

There are connections between Stein's method, Slepian's method and Lindeberg's original telescoping sum interpolation. These connections appear, for example, in Chatterjee (2008), Rollin (2013) and Chernozhukov, Chetverikov and Kato (2012). First we discuss the connection between Stein and Slepian, following Appendix E of Chernozhukov, Chetverikov and Kato (2012).

The multivariate Stein equation is

$$\triangle h(x) - x'\nabla h(x) = f(x) - \mu \tag{33}$$

where $Y \sim N(0, I)$, $\mu = \mathbb{E}[f(Y)]$, ∇ is the gradient and $\triangle$ is the Laplacian. A solution is

$$h(x) = -\int_0^1 \frac{1}{2t}\Big[\mathbb{E}[f(\sqrt{t}x + \sqrt{1-t}Y)] - \mu\Big]\,dt.$$

One then needs to bound

$$\mathbb{E}[f(X)] - \mathbb{E}[f(Y)] = \mathbb{E}[\triangle h(X) - X'\nabla h(X)]. \tag{34}$$

Recall that the Slepian path is $Z_t = \sqrt{t}X + \sqrt{1-t}Y$ and one needs to bound

$$\mathbb{E}[f(X)] - \mathbb{E}[f(Y)] = \mathbb{E}\left[\int_0^1 \frac{1}{2}\nabla f(Z_t)^T\left(\frac{X}{\sqrt{t}} - \frac{Y}{\sqrt{1-t}}\right)\right]dt. \tag{35}$$

Using integration by parts, we have that

$$\mathbb{E}\left[\int_0^1 \frac{1}{2}\nabla f(Z_t)^T\left(\frac{X}{\sqrt{t}}\right)dt\right] = -\mathbb{E}[X'\nabla h(X)]$$

and

$$\mathbb{E}\left[\int_0^1 \frac{1}{2}\nabla f(Z_t)^T\left(\frac{Y}{\sqrt{1-t}}\right)dt\right] = -\mathbb{E}[\triangle h(X)].$$

So the right hand side of (34) is the same as the right hand side of (35).

The Slepian path interpolates between X and Y. Lindeberg's original proof also uses an interpolation. Specifically, define

$$Z_i = (X(1),\ldots,X(i),Y(i+1),\ldots,Y(n))^T.$$

Then, Lindeberg's telescopic interpolation is

$$\mathbb{E}[h(X)] - \mathbb{E}[h(Y)] = \sum_{i=1}^n \mathbb{E}[h(Y_i) - h(Y_{i-1})].$$

The right hand side can be bounded by Taylor expanding h; see Chatterjee (2008). The advantage of Slepian-Stein over Lindeberg is that it treats all the coordinates equally. This is important for handling cases with dependence.

All the discussion above refers to smooth functions h. Dealing with indicator functions requires extra work such as smoothing.

References

Bentkus, V. (2003). On the dependence of the Berry-Esseen bound on dimension. *J. Statist. Plann. Inf.*, 113, 385-402.

Chatterjee, S. (2008). A simple invariance theorem. arxiv:math/0508213

Che, L. and Fang, X. (2011). Multivariate Normal approximation by Stein's method: the concentration inequality approach.

Chen, Goldstein and Shao. (2011). *Normal Approximation by Stein's Method.* Springer.

Victor Chernozhukov, Denis Chetverikov, Kengo Kato (2012). Central Limit Theorems and Multiplier Bootstrap when p is much larger than n. http://arxiv.org/abs/1212.6906.

Victor Chernozhukov, Denis Chetverikov, Kengo Kato (2013). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. http://arxiv.org/abs/1301.4807.

Goldstein and Reinert (1997). Stein's method and the zero bias transformation with application to simple random sampling. *The Annals of Applied Probability*, 7, 935-952.

Nourdin and Peccati (2012). *Normal Approximations With Malliavin Calculus.* Cambridge.

Rollin, A. (2013). Stein's method in high dimensions with applications. arXiv:1101.4454.

Ross, Nathan. Fundamentals of Stein's method. *Probab. Surv*, 8, 210–293.

Part II: The High-Dimensional Case and the Bootstrap

12 Introduction

Chernozhukov, Chetverikov and Kato (CCK) have a remarkable set of theorems about central limit theorems and the bootstrap in high dimensions. Part II of these notes is a tutorial on their results. There are two main results from CCK that we are interested in. Let $X_1,\ldots,X_n \in \mathbb{R}^d$ be iid with μ and covariance Σ. The first result is a Berry-Esseen style central limit theorem which says that

$$\sup_z \left| \mathbb{P}\big(\sqrt{n}||\overline{X} - \mu||_\infty \le z\big) - \mathbb{P}\big(||Y||_\infty \le z\big) \right| \preceq \frac{\log d}{n^{1/8}} \tag{36}$$

where $Y \sim N(0,\Sigma)$. The second is a bootstrap theorem that says that

$$\sup_z \left| \mathbb{P}\big(\sqrt{n}||\overline{X}^* - \overline{X}||_\infty \le z \mid X_1,\ldots,X_n\big) - \mathbb{P}\big(\sqrt{n}||\overline{X} - \mu||_\infty \le z\big) \right| = O_P\left(\frac{\log d}{n^{1/8}}\right) \tag{37}$$

where $\overline{X}^* = \frac{1}{n}\sum_i X_i^*$ and $X_1^*,\ldots,X_n^*$ is a sample from the empirical distribution P_n. The proofs make use a variety of tools including: Stein's method, Slepian interpolation, smoothing and a phenomenon called Gaussian anti-concentration.

Main Sources:

(CCK1) Victor Chernozhukov, Denis Chetverikov, Kengo Kato (2012). Central Limit Theorems and Multiplier Bootstrap when p is much larger than n. http://arxiv.org/abs/1212.6906.

(CCK2) Victor Chernozhukov, Denis Chetverikov, Kengo Kato (2013). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. http://arxiv.org/abs/1301.4807.

Other references are given at the end of the document.

13 The Bootstrap and Gaussian Approximations

The bootstrap is strongly tied to Normal approximations. Consider the following simple case. Suppose that $X_1,\dots,X_n$ are random variables with mean μ and variance σ^2. If we knew the distribution of $\sqrt{n}(\overline{X}-\mu)$ then we could construct a confidence interval for μ.

Let P_n be the empirical distribution that puts mass $1/n$ at each X_i. Let $X_1^*,\dots,X_n^* \sim P_n$. The idea is to approximate the distribution of $\sqrt{n}(\overline{X}-\mu)$ with the distribution of $\sqrt{n}(\overline{X}^* - \overline{X})$ conditional on $X_1,\dots,X_n$. The latter we can approximate by simulation. Let

$$F(z) = \mathbb{P}\big(\sqrt{n}(\overline{X}-\mu) \le z\big)$$
$$\widehat{F}(z) = \mathbb{P}\big(\sqrt{n}(\overline{X^*}-\overline{X}) \le z \;\big|\; X_1,\dots,X_n\big).$$

Let Φ_σ denote a Gaussian cdf with mean 0 and variance σ^2 and let $\widehat{\sigma}^2$ be the sample variance. Then,

$$\sup_z |F(z)-\widehat{F}(z)| \le \sup_z |F(z)-\Phi_\sigma(z)| + \sup_z |\Phi_\sigma(z)-\Phi_{\widehat{\sigma}}(z)| + \sup_z |\widehat{F}(z)-\Phi_{\widehat{\sigma}}(z)|.$$

If we can bound how far a distribution is form its Normal approximation, then we can bound the first and third term. Since $\widehat{\sigma}-\sigma = O_P(1/\sqrt{n})$, we can also bound the second term. This leads to

$$\sup_z |F(z)-\widehat{F}(z)| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

It follows that

$$\mathbb{P}(\mu \in C_n) \ge 1-\alpha - O\left(\sqrt{\frac{1}{n}}\right)$$

where $C_n = \left[\overline{X} - \frac{Z_{1-\alpha/2}}{\sqrt{n}}, \overline{X} - \frac{Z_{\alpha/2}}{\sqrt{n}}\right]$ and $Z_\beta = \widehat{F}^{-1}(\beta)$. So we see that bounding the distance to a Normal approximation is a key step in verifying the validity of the bootstrap.

14 Preliminary Result: CCK Theorem in One Dimension

Here we will go through the CCK theorem in the one dimensional case. This proof does not lead to the optimal rate when applied to $d=1$ but it is a good warm-up for the high-dimensional case. In particular, it allows us to introduce the Slepian interpolation and smoothing. We note that the proof is related to the techniques in Chatterjee (2008) and Rollin (2013) although Chatterjee uses a Lindeberg interpolation rather than a Slepian interpolation.

Let $X_1,\dots,X_n \in \mathbb{R}$ be iid with mean 0 and variance σ^2 and let $Y_1,\dots,Y_n \sim N(0,\sigma^2)$. Let

$$X = \frac{1}{\sqrt{n}} = \sum_i X_i, \quad Y = \frac{1}{\sqrt{n}} = \sum_i Y_i$$

and so $Y \sim N(0,\sigma^2)$. We want to bound

$$\Delta_n = \sup_t \Big| \mathbb{P}(|X| \le t) - \mathbb{P}(|Y| \le t) \Big|. \tag{38}$$

Anti-concentration. We will use the following fact: if $Y \sim N(0,\sigma^2)$ then

$$P(Y \le t + \epsilon) \le P(Y \le t) + C\epsilon$$

where $C = (\sigma\sqrt{2\pi})^{-1}$. This follows trivially from the fact that the Normal density is bounded above by C. When we get to the d-dimensional case in Section 15, we will see that ϵ becomes $\epsilon\sqrt{\log d/\varepsilon}$ rather than $d\epsilon$ as one might expect.

Step 1: Smooth Functions. Before bounding Δ_n, we first bound $\mathbb{E}[g(X) - g(Y)]$ where g is a smooth function. In particular, assume that g has three bounded, continuous derivatives and let $C_3 = \sup_z |g'''(z)|$.

Define the *Slepian interpolation*

$$Z(t) = \sqrt{t}X + \sqrt{1-t}\,Y. \tag{39}$$

Also define

$$Z_i(t) = \frac{1}{\sqrt{n}}\left(\sqrt{t}X_i + \sqrt{1-t}\,Y_i\right)$$

and

$$Z^i(t) = Z(t) - Z_i(t).$$

Thus, $Z(t) = \sum_i Z_i(t)$ and note that $Z^i(t)$ is independent of $Z_i(t)$.

Now

$$\mathbb{E}[g(X) - g(Y)] = \mathbb{E}[g(Z(1)) - g(Z(0))] = \int_0^1 \frac{dg(Z(t)}{dt} dt = \int_0^1 g'(Z(t))\, Z'(t) dt \tag{40}$$

where

$$Z'(t) = \frac{dZ(t)}{dt} = \sum_i \frac{dZ_i(t)}{dt} = \frac{1}{2}\sum_i \frac{1}{\sqrt{n}}\left(\frac{X_i}{\sqrt{t}} - \frac{Y_i}{\sqrt{1-t}}\right) \equiv \sum_i Z_i'(t).$$

To bound $g'(Z(t))$ we use an expansion. In general, Taylor's theorem with integral remainder gives

$$
\begin{aligned}
f(x) &= f(a) + (x-a)f'(a) + \int_a^x (x-t)f''(t)dt \\
&= f(a) + (x-a)f'(a) + (x-a)^2 \int_0^1 (1-u)f''\big(a+u(x-a)\big)\,du
\end{aligned}
$$

where we used the transformation $u = \frac{t-a}{x-a}$. Apply this expansion to the function $g(Z(t))$ with $x = Z(t)$ and $a = Z^i(t)$, and noting that $x - a = Z_i(t)$, we get

$$
g'(Z(t)) = g'(Z^i(t)) + Z_i(t)g''(Z^i(t)) + Z_i^2(t)\int_0^1 (1-u)g'''\big(Z^i(t) + uZ_i(t)\big)\,du. \tag{41}
$$

Inserting (41) into (40) we get

$$
\mathbb{E}[g(X) - g(Y)] = \mathrm{I} + \mathrm{II} + \mathrm{III}
$$

where

$$
\begin{aligned}
\mathrm{I} &= \sum_i \int_0^1 \mathbb{E}[g'(Z^i(t))\ Z_i'(t)]dt \\
\mathrm{II} &= \sum_i \int_0^1 \mathbb{E}[g''(Z^i(t))\ Z_i'(t)\ Z_i(t)]dt \\
\mathrm{III} &= \sum_i \int_0^1 \int_0^1 (1-u)\mathbb{E}[g'''(Z^i(t) + uZ_i(t))\ Z_i'(t)\ Z_i^2(t)]dudt.
\end{aligned}
$$

Note that $g'(Z^i(t))$ and $Z_i'(t)$ are independent and $\mathbb{E}[Z_i'(t)] = 0$. So

$$
\mathrm{I} = \sum_i \int_0^1 \mathbb{E}[g'(Z^i(t))\ Z_i'(t)]dt = \sum_i \int_0^1 \mathbb{E}[g'(Z^i(t))]\ \ \mathbb{E}[Z_i'(t)]dt = 0.
$$

To bound II, note that $Z^i(t)$ is independent of $Z_i'(t)Z_i(t)$. So

$$
\mathbb{E}[g''(Z^i(t))\ Z_i'(t)\ Z_i(t)] = \mathbb{E}[g''(Z^i(t))]\ \mathbb{E}[Z_i'(t)\ Z_i(t)].
$$

Recall that

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

and so

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

and hence $\mathbb{E}[Z_i'(t)\, Z_i(t)] = (1/2n)(\sigma^2 - 0 + 0 - \sigma^2) = 0$ and so $\mathrm{II} = 0$.

Next, consider III. Recall that $\sup_z |g'''(z)| \le C_3$. So,

$$|\mathrm{III}| \le \sum_i \int_0^1 \int_0^1 \mathbb{E}\left[|g'''(Z^i(t) + uZ_i(t))|\ |Z_i'(t)\, Z_i^2(t)|\right] du dt \le C_3 n \int_0^1 \mathbb{E}[|Z_i'(t)\, Z_i^2(t)|] dt.$$

Let

$$\omega(t) = \frac{1}{\sqrt{t} \wedge \sqrt{1-t}}.$$

Then, using Holder's inequality,

$$\int_0^1 \mathbb{E}[|Z_i'(t)\, Z_i^2(t)|] dt = \int_0^1 \omega(t) \mathbb{E}\left[\left|\frac{Z_i'(t)}{\omega(t)}\, Z_i^2(t)\right|\right] dt$$
$$\le \int_0^1 \omega(t) \left(\mathbb{E}\left[\left|\frac{Z_i'(t)}{\omega(t)}\right|^3\right] \mathbb{E}|Z_i(t)|^3 \mathbb{E}|Z_i(t)|^3\right)^{1/3} dt$$

Now

$$\left|\frac{Z_i'(t)}{\omega(t)}\right| \le \frac{|X_i| + |Y_i|}{2\sqrt{n}}, \quad |Z_i(t)| \le \frac{|X_i| + |Y_i|}{\sqrt{n}}, \quad \int_0^1 \omega(t) dt \le 1.$$

Also,

$$\mathbb{E}|Y_i|^3 \le (\mathbb{E}|Y_i|^2)^{3/2} = (\mathbb{E}|X_i|^2)^{3/2} \le \mathbb{E}|X_i|^3.$$

Hence,

$$n \int \mathbb{E}|Z_i(t)|^3 \mathbb{E}|Z_i(t)|^3 dt \le \frac{1}{\sqrt{n}} \mathbb{E}(|X_i| + |Y_i|)^3 \int_0^1 \omega(t) dt \preceq \frac{\mu_3}{\sqrt{n}}$$

where $\mu_3 = \mathbb{E}(|X_i|^3)$. So we have shown that

$$\mathbb{E}|g(X) - g(Y)| \preceq \frac{\mu_3 C_3}{\sqrt{n}}. \tag{42}$$

Step 2: Back to Indicator Functions. We want to bound

$$\mathbb{P}[X \le t] - \mathbb{P}[Y \le t] = \mathbb{E}[h(X)] - \mathbb{E}[h(Y)]$$

where $h(z) = I(z \le t)$. We will find a smooth function g to approximate the indicator function h.

Let $g_0 : \mathbb{R} \to [0,1]$ be in C^3 be such that $g_0(s) = 1$ for $s \le 0$ and $g_0(s) = 0$ for $s \ge 1$. Let ψ be a real number and define $g(s) = g_0(\psi(s-t))$. Then

$$\sup_s |g(s)| = 1,\ \sup_s |g'(s)| \le \psi,\ \ \sup_s |g''(s)| \le \psi^2,\ \ C_3 = \sup_s |g'''(s)| \le \psi^3 and$$

$$I(z \le t) \le g(t) \le I\left(z \le t + \frac{1}{\psi}\right). \tag{43}$$

From (42) and (43),

$$\begin{aligned}\mathbb{P}(X \le t) = \mathbb{E}[h(X)] \le \mathbb{E}[g(X)] \le \mathbb{E}[g(Y)] + \frac{\psi^3 \mu_3}{\sqrt{n}} \\ \le \mathbb{P}(Y \le t + \psi^{-1}) + \frac{\psi^3 \mu_3}{\sqrt{n}} \le \mathbb{P}(Y \le t) + \frac{1}{\psi} + \frac{\psi^3 \mu_3}{\sqrt{n}}.\end{aligned}$$

In this last step, we used anti-concentration. To balance the last two terms take $\psi = n^{1/8}$. Then

$$\mathbb{P}(X \le t) \le \mathbb{P}(Y \le t) + \frac{\mu_3}{n^{1/8}}.$$

A similar proof gives

$$\mathbb{P}(Y \le t) \le \mathbb{P}(X \le t) + \frac{\mu_3}{n^{1/8}}.$$

We conclude that

$$\sup_t |\mathbb{P}(|X| \le t) - \mathbb{P}(|Y| \le t)| \le \frac{\mu_3}{n^{1/8}}. \tag{44}$$

This completes the one-dimensional proof. Note that the smoothing step changed the $n^{-1/2}$ rate to $n^{-1/8}$. Clearly this is too slow for $d = 1$ but will give us the desired result in high dimensions. In fact, the high-dimensional proof is almost the same except for two crucial differences: we need to approximate the max function in addition to the indicator function and we need to use Gaussian anti-concentration. These changes only add a term that is logarithmic in dimension.

There are other smoothing techniques; see, for example Chapter 5 of Chen, Goldstein and Shao (2011) and Section 3.7 of Nourdin and Peccati (2012). It is not clear, however, if they could be used with the present proof to get a better rate.

15 High Dimensional CLT

The main theorem is in Section 2 of CCK. The proof is quite involved because of some subtle truncation arguments. Appendix F gives a much shorter simpler proof using stronger moment conditions. We will focus on this simpler version.

A crucial part of the proof is the CCK anti-Gaussian concentration result that allows us to go from smooth functions to indicator functions. We will use the Gaussian anti-concentration result here but we defer an explanation of the result until Section 16.

Let $X_1,\ldots,X_n \in \mathbb{R}^d$ be iid with mean 0 and covariance Σ. Let X_{ij} denote the j^{th} element of the vector X_i. Hence, $X_i = (X_{i1},\ldots,X_{id})^T$. Let

$$X = \frac{1}{\sqrt{n}}\sum_i X_i, \quad Y = \frac{1}{\sqrt{n}}\sum_i Y_i$$

where $Y_i \sim N(0,\Sigma)$. Of course, $Y \sim N(0,\Sigma)$. We want to bound

$$\Delta_n = \sup_t \Big|\mathbb{P}(||X||_\infty \le t) - \mathbb{P}(||Y||_\infty \le t)\Big|. \tag{45}$$

Theorem 14 (Chernozhukov, Chetverikov and Kato 2012) *Suppose that* $0 < c \le \mathbb{E}[X_{ij}^2] \le C$ *for all* j *where* c *and* C *and finite, positive constants. Then*

$$\Delta_n \equiv \sup_t \Big|\mathbb{P}(||X||_\infty \le t) - \mathbb{P}(||Z||_\infty \le t)\Big| \le M_d^{1/4}\left(\frac{(\log(dn))^7}{n}\right)^{1/8} \tag{46}$$

where $M_d = \mathbb{E}\left[\left(\max_j\left(|X_{ij}|+|Y_{ij}|\right)\right)^3\right]$.

The main Theorem in CCK does not have the term M_d. This is the price we pay for focusing on the simpler version of the theorem. In fact, we will assume form now on that $M_d = O(\sqrt{\log d})$. (This is true for sub-Gaussian distributions wit $n \le d$ for example.) In this case, the bound becomes

$$\Delta_n \equiv \sup_t \Big|\mathbb{P}(||X||_\infty \le t) - \mathbb{P}(||Z||_\infty \le t)\Big| \preceq \frac{\log d}{n^{1/8}}. \tag{47}$$

Proof. We divide the proof into two steps: the first uses smooth functions and the second derives a bound for indicator functions.

Step I: Smooth Functions. Define the *Slepian interpolation*

$$Z(t) = \sqrt{t}X + \sqrt{1-t}Y = \sum_i Z_i(t) \tag{48}$$

where

$$Z_i(t) = \frac{1}{\sqrt{n}}\left(\sqrt{t}X_i + \sqrt{1-t}Y_i\right).$$

Let

$$Z^i(t) = Z(t) - Z_i(t).$$

and note that $Z^i(t)$ is independent of Z_i.

Now we define a smooth approximation g to the indicator function and another function F_β which is a smooth approximation to the max function. Specifically, let $m = g \circ F_\beta$ where

$$g(s) = g_0\left(\psi\left(s - t - \frac{\log d}{\beta}\right)\right)$$

and $g_0 \in C^3$ with $g_0(s) = 1$ for $s \le 0$ and $g_0(s) = 0$ for $s \ge 1$. Also let

$$F_\beta(z) = \frac{1}{\beta}\log\left(\sum_{j=1}^{d} e^{\beta z_j}\right).$$

Let $G_j = \sup_z |g^{(j)}(z)|$. Then

$$G_0 = 1,\ \ G_1 \le \psi,\ \ G_2 \le \psi^2,\ \ G_3 \le \psi^3.$$

Define $\Psi(t) = \mathbb{E}[m(Z(t))]$. Then

$$\Psi'(t) = \sum_{j=1}^{d}\sum_{i=1}^{n}\mathbb{E}\Big[\partial_j m(Z(t))\ Z'_{ij}(t)\Big]$$

where

$$Z'_{ij}(t) = \frac{d}{dt}Z_{ij}(t) = \frac{1}{2\sqrt{n}}\left(\frac{X_{ij}}{\sqrt{t}} - \frac{Y_{ij}}{\sqrt{1-t}}\right).$$

We will use the following version of Taylor's theorem with integral remainder. Let $f : \mathbb{R}^d \to \mathbb{R}$ and denote its first and second derivatives by $f_k = \partial f/\partial x_k$ and $f_{k\ell} = \partial^2 f/\partial x_k \partial x_\ell$. Then

$$f(x) = f(a) + \sum_k (x_k - a_k) f_k(a) + \sum_{k,\ell}(x_k - a_k)(x_\ell - a_\ell)\int_0^1 (1-u) f_{k\ell}(a + u(x-a))du.$$

Take $x = Z(t)$, $a = Z^i(t)$ and $f(x) = \partial_j m(Z(t))$. Note that $x - a = Z_i(t)$. Hence,

$$\begin{aligned}\partial_j m(Z(t)) = \partial_j m(Z^i(t)) + Z_i(t)\sum_k \partial_k \partial_j m(Z^i(t))\\ + \sum_k\sum_\ell Z_{ik}(t) Z_{i\ell}(t)\int_0^1 (1-u)\partial_j\partial_k\partial_\ell m\big(Z^i(t) + uZ_i(t)\big)\,du.\end{aligned}$$

Recalling that $\Psi'(t) = \mathbb{E}\Big[\partial_j m(Z(t))\ Z'_{ij}(t)\Big]$, we therefore have,

$$\begin{aligned}\mathbb{E}[m(X) - m(Y)] = \Psi(1) - \Psi(0) &= \int_0^1 \Psi'(t)dt \\ &= \sum_{j=1}^{d}\sum_{i=1}^{n}\int_0^1 \mathbb{E}[\partial_j m(Z(t))\ Z'_{ij}(t)]dt \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}\end{aligned}$$

where

$$\begin{aligned}\mathrm{I} &= \sum_j \sum_i \int_0^1 \mathbb{E}[\partial_j m(Z^i(t)) Z'_{ij}(t)]\,dt \\ \mathrm{II} &= \sum_{j,k}\sum_i \int_0^1 \mathbb{E}[\partial_j \partial_k m(Z^i(t)) Z'_{ij}(t) Z_{ik}]\,dt \\ \mathrm{III} &= \sum_{j,k,\ell}\sum_i \int_0^1\int_0^1 \mathbb{E}\Big[\partial_j\partial_k\partial_\ell m\Big(Z^i(t) + uZ_i(t)\Big)\ Z'_{ij}(t) Z_{ik}(t) Z_{i\ell}(t)\Big]\,du\,dt.\end{aligned}$$

Now $Z^i(t)$ is independent of $(Z'_{ij}(t), Z_{ij}(t))$ and

$$\mathbb{E}[Z'_{ij}(t)] = \frac{1}{2\sqrt{n}}\left(\frac{\mathbb{E}[X_{ij}]}{\sqrt{t}} - \frac{\mathbb{E}[Y_{ij}]}{\sqrt{1-t}}\right) = 0$$

and so $\mathrm{I} = 0$.

For II we have

$$\mathbb{E}\big[\partial_j\partial_k m(Z^i(t))\ Z'_{ij}(t) Z_{ik}(t))\big] = \mathbb{E}[\partial_j\partial_k m(Z^i(t))]\ \ \mathbb{E}[Z'_{ij}(t) Z_{ik}(t)].$$

But

$$\begin{aligned}Z'_{ij}(t) Z_{ik}(t) &= \frac{1}{2n}\left(\frac{X_{ij}}{\sqrt{t}} - \frac{Y_{ij}}{\sqrt{1-t}}\right)\left(\sqrt{t}X_{ik} + \sqrt{1-t}\,Y_{ik}\right) \\ &= \frac{1}{2n}\left(X_{ij}X_{ik} + \sqrt{\frac{1-t}{t}}X_{ij}Y_{ij} - \sqrt{\frac{t}{1-t}}Y_{ij}X_{ik} - Y_{ij}Y_{ik}\right).\end{aligned}$$

Recall that X and Y are independent so the middle two terms have mean 0. Thus

$$\mathbb{E}[Z'_{ij}(t) Z_{ik}(t)] = \mathbb{E}[X_{ij}X_{ik}] - \mathbb{E}[Y_{ij}Y_{ik}] = \Sigma_{jk} - \Sigma_{jk} = 0.$$

Thus $\mathrm{II} = 0$.

Now we bound III. Because m is smooth, it can be shown that

$$\sum_{j,k,\ell}\left|\partial_j\partial_k\partial_\ell m\big(Z^i(t)+uZ_i(t)\big)Z'_{ij}(t)Z_{ik}(t)Z_{i\ell}(t)\right|$$

is bounded above (up to a constant) by

$$(G_1\beta^2+G_2\beta+G_3)\max_{j,k,\ell}\left|Z'_{ij}(t)Z_{ik}(t)Z_{i\ell}(t)\right|.$$

where we recall that $G_j=\sup_z|g^{(j)}(z)|$. One might expect the triple sum above to introduce a term of order $O(d^3)$. The reason this does not happen is because of the properties of the function F_β. For example, $\partial_j F_\beta(z)=e^{\beta z_j}/\sum_{m=1}^d e^{\beta z_m}$. The sum of this term over j is $O(1)$ rather than $O(d)$.

Let

$$\omega(t)=\frac{1}{\sqrt{t}\wedge\sqrt{1-t}}.$$

Note that

$$\left|\frac{Z'_{ij}(t)}{\omega(t)}\right|\le\frac{|X_{ij}|+|Y_{ij}|}{2\sqrt{n}},\qquad |Z_{ij}(t)|\le\frac{|X_{ij}|+|Y_{ij}|}{\sqrt{n}}.$$

Then, using Holder's inequality,

$$\begin{aligned}\int_0^1\mathbb{E}[\max_{j,k,\ell}|Z'_{ij}(t)\,Z_{ik}(t)\,Z_{i\ell}(t)|]dt&=\int_0^1\omega(t)\mathbb{E}\left[\max_{j,k,\ell}\left|\frac{Z'_{ij}(t)}{\omega(t)}\,Z_{ik}(t)Z_{i\ell}(t)\right|\right]dt\\&\le\int_0^1\omega(t)\left(\mathbb{E}\left[\max_j\left|\frac{Z'_{ij}(t)}{\omega(t)}\right|^3\right]\mathbb{E}|\max_j Z_{ij}(t)|^3\;\mathbb{E}|\max_j Z_{ij}(t)|^3\right)^{1/3}dt\\&\le\frac{1}{n^{3/2}}\int_0^1\omega(t)dtM_d\preceq\frac{M_d}{n^{3/2}}.\end{aligned}$$

Thus,

$$\text{III}\preceq\frac{G_1\beta^2+G_2\beta+G_3}{\sqrt{n}}M_d.$$

To summarize, so far we have shown that:

$$\left|\mathbb{E}[m(X)]-\mathbb{E}[m(Y)]\right|\preceq\frac{G_1\beta^2+G_2\beta+G_3}{\sqrt{n}}M_d.\tag{49}$$

Now, the function F_β satisfies

$$0\le F_\beta(z)-\max_j z_j\le\frac{\log d}{\beta}.$$

Thus

$$g(F_\beta(z)) \le g\left(\max_j z_j + \frac{\log d}{\beta}\right) \le g(\max_j z_j) + \frac{G_1 \log d}{\beta}$$

and we conclude that

$$\left|\mathbb{E}[g(\max_j X_j) - g(\max_j Y_j)]\right| \le \frac{G_1\beta^2 + G_2\beta + G_3}{\sqrt{n}} M_d + \frac{G_1 \log d}{\beta}. \tag{50}$$

Step 2: Back to Indicator Functions. The last step is to replace the smooth function g with the indicator function. Recall that

$$0 \le F_\beta(z) - \max_j z_j \le e_\beta$$

where $e_\beta = \frac{\log d}{\beta}$. Also,

$$I(z \le t) \le g(t) \le I\left(z \le t + \frac{1}{\psi}\right).$$

We have

$$\begin{aligned}
\mathbb{P}(\max_j X_j \le t) &\le \mathbb{P}(F_\beta(X) \le t + e_\beta) \le \mathbb{E}[g(F_\beta(X))] \\
&\le \mathbb{E}[g(F_\beta(Y))] + \frac{G_1\beta^2 + G_2\beta + G_3}{\sqrt{n}} M_d + \frac{G_1 \log d}{\beta} \\
&\le \mathbb{P}\left(F_\beta(Y) \le t + \frac{1}{\psi}\right) + \frac{G_1\beta^2 + G_2\beta + G_3}{\sqrt{n}} M_d + \frac{G_1 \log d}{\beta} \\
&\le \mathbb{P}\left(\max_j Y_j \le t + e_\beta + \frac{1}{\psi}\right) + \frac{G_1\beta^2 + G_2\beta + G_3}{\sqrt{n}} M_d + \frac{G_1 \log d}{\beta}.
\end{aligned}$$

By the Gaussian Anti-concentration result in the next section,

$$\mathbb{P}\left(\max_j Y_j \le t + e_\beta + \frac{1}{\psi}\right) \le \mathbb{P}\left(\max_j Y_j \le t\right) + \left(e_\beta + \frac{1}{\psi}\right)\sqrt{\log(d\psi)}.$$

So

$$\mathbb{P}(\max_j X_j \le t) \le \mathbb{P}\left(\max_j Y_j \le t\right) + \left(e_\beta + \frac{1}{\psi}\right)\sqrt{\log(d\psi)} + \frac{G_1\beta^2 + G_2\beta + G_3}{\sqrt{n}} M_d + \frac{G_1 \log d}{\beta}.$$

To minimize the last three terms we take

$$\beta = \psi \log d, \qquad \psi = \frac{n^{1/8}}{(\log d)^{3/8} M_d^{1/4}}.$$

Thus we get

$$\mathbb{P}(\max_j X_j \le t) \le \mathbb{P}(\max_j Y_j \le t) + \frac{(\log(nd))^{7/8} M_d^{1/4}}{n^{1/8}}.$$

A similar argument provides the bound in the other direction. □

16 Gaussian Anti-Concentration

Now we give the CCK Gaussian anti-concentration result. Let ϕ and Φ denote the density and cdf of a standard Gaussian.

Theorem 15 *Let $X \sim N_d(0,\Sigma)$. Let $\sigma_j^2 = \Sigma_{jj}$ and define $\underline{\sigma} = \min_j \sigma_j$ and $\overline{\sigma} = \max_j \sigma_j$. Then*

$$\sup_{x\in\mathbb{R}} \mathbb{P}(|\max_j X_j - x| \le \varepsilon) \le C\varepsilon\sqrt{1 \vee \log(d/\varepsilon)} \tag{51}$$

where C depends only on $\underline{\sigma}$ and $\overline{\sigma}$.

Proof Outline. We will fix some $x \ge 0$. The proof for negative x is analogous. The first step is to standardize the variables. Let

$$W_j = \frac{X_j - x}{\sigma_j} + \frac{x}{\underline{\sigma}}.$$

Then

$$\mu_j \equiv \mathbb{E}[W_j] = \frac{x}{\underline{\sigma}} - \frac{x}{\sigma_j} \ge 0$$

and $\mathrm{Var}(X_j) = 1$. Define $Z = \max_j W_j$. So

$$\begin{aligned}
\mathbb{P}(|\max_j X_j - x| \le \varepsilon) &\le \mathbb{P}\left(\left|\max_j \frac{X_j - x}{\sigma_j}\right| \le \frac{\varepsilon}{\underline{\sigma}}\right) \\
&\le \sup_y \mathbb{P}\left(\left|\max_j \frac{X_j - x}{\sigma_j} + \frac{x}{\underline{\sigma}} - y\right| \le \frac{\varepsilon}{\underline{\sigma}}\right) \\
&= \sup_y \mathbb{P}\left(|Z - y| \le \frac{\varepsilon}{\underline{\sigma}}\right).
\end{aligned}$$

So we can now assume that the variables have non-negative mean and variance 1.

The next step is to find the density $f(z)$ for z. This part of the proof is long and technical; I'll just state the result, namely,

$$f(z) = \phi(z)G(z) \tag{52}$$

where G is a non-decreasing function. Now we derive an upper bound on G as follows. Let

$$\overline{z} = x\left(\frac{1}{\underline{\sigma}} - \frac{1}{\overline{\sigma}}\right), \quad \overline{Z} = \max_j (W_j - \mu_j)$$

and let

$$a \equiv \mathbb{E}[\overline{Z}] = \mathbb{E}\left[\max_j \frac{X_j}{\sigma_j}\right] \le \sqrt{2\log d}.$$

Then, using the Gaussian tail inequality,

$$\begin{aligned} G(z)(1-\Phi(z)) &= G(z)\int_z^\infty \phi(u)du \le \int_z^\infty \phi(u)G(u)du \\ &= \mathbb{P}(Z > z) \le \mathbb{P}(\overline{Z} > z - \overline{z}) \le \exp\left(-\frac{(z-\overline{z}-a)_+^2}{2}\right). \end{aligned}$$

Hence,

$$G(z) \le \frac{1}{1-\Phi(z)} \exp\left(-\frac{(z-\overline{z}-a)_+^2}{2}\right)$$

and so

$$f(z) = G(z)\phi(z) \le \frac{\phi(z)}{1-\Phi(z)} \exp\left(-\frac{(z-\overline{z}-a)_+^2}{2}\right) \le 2(z \vee 1)\exp\left(-\frac{(z-\overline{z}-a)_+^2}{2}\right)$$

where we used the fact that

$$\frac{\phi(z)}{1-\Phi(z)} \le 2(z \vee 1).$$

For any $y \in \mathbb{R}$ and $t > 0$, we thus have that

$$\mathbb{P}(|Z-y| \le t) = \int_{y-t}^{y+t} f(z)dz \le 2t \max_{y-t\le z\le y+t} f(z) \le 4t(\overline{z}+a+1)$$

and hence,

$$\mathbb{P}(|\max_j X_j - x| \le \varepsilon) \le \frac{4\varepsilon}{\underline{\sigma}}\left(|x|\left(\frac{1}{\underline{\sigma}} - \frac{1}{\overline{\sigma}}\right) + a + 1\right). \tag{53}$$

We only need to show that the right hand side has the required logarithmic bound.

In what follows, recall that $a \le \sqrt{2\log d}$. If $\underline{\sigma} = \overline{\sigma} = \sigma$ then we have from (53) that

$$\mathbb{P}(|\max_j X_j - x| \le \varepsilon) \le \frac{4\varepsilon(a+1)}{\sigma}$$

and we are done.

Now suppose that $\underline{\sigma} < \overline{\sigma}$. Fist suppose that $0 < \varepsilon < \underline{\sigma}$. We consider two sub-cases. First suppose that

$$|x| \le \varepsilon + \overline{\sigma}\left(a + \sqrt{2\log(\underline{\sigma}/\varepsilon)}\right).$$

Inserting this into (53) and using $\varepsilon \le \underline{\sigma}$ we have

$$\mathbb{P}(|\max_j X_j - x| \le \varepsilon) \le \frac{4\varepsilon}{\underline{\sigma}}\left((\overline{\sigma}/\underline{\sigma})a + \left(\frac{\overline{\sigma}}{\underline{\sigma}} - 1\right)\sqrt{2\log(\underline{\sigma}\varepsilon)} + 2 - \underline{\sigma}/\overline{\sigma}\right)$$

which has the required form. The second sub-case is

$$|x| > \varepsilon + \overline{\sigma}\left(a + \sqrt{2\log(\underline{\sigma}/\varepsilon)}\right).$$

Now

$$\mathbb{E}[\max_j X_j] = \mathbb{E}\left[\max_j \frac{X_j}{\sigma_j}\sigma_j\right] \le a\overline{\sigma}.$$

So,

$$\begin{aligned}\mathbb{P}(|\max_j X_j - x| \le \varepsilon) &\le \mathbb{P}(\max_j X_j \ge |x| - \varepsilon) \le \mathbb{P}\left(\max_j X_j \ge \overline{\sigma}a + \overline{\sigma}\sqrt{2\log(\underline{\sigma}/\varepsilon))}\right)\\ &\le \mathbb{P}\left(\max_j X_j \ge \mathbb{E}[\max_j X_j] + \overline{\sigma}\sqrt{2\log(\underline{\sigma}/\varepsilon)}\right) \le \frac{\varepsilon}{\underline{\sigma}}\end{aligned}$$

where we used the fact (concentration of measure for maxima of Gaussians) that

$$\mathbb{P}\left(\max_j X_j \ge \mathbb{E}[\max_j X_j] + r\right) \le \exp\left(-\frac{r^2}{2\overline{\sigma}^2}\right).$$

Similar bounds hold, with different constants, when $\varepsilon > \underline{\sigma}$. Combining these cases completes the proof. □.

An immediate consequence of the result is that the perimeter of a rectangle gets small probability.

Corollary 16 *Assume the conditions of Theorem 15. Then*

$$\sup_t \left|\mathbb{P}(\max_j X_j \le t + \epsilon) - \mathbb{P}(\max_j X_j \le t)\right| \le C\varepsilon\sqrt{1 \vee \log(d/\varepsilon)}$$

and

$$\sup_t \left|\mathbb{P}(\max_j |X_j| \le t + \epsilon) - \mathbb{P}(\max_j |X_j| \le t)\right| \le C\varepsilon\sqrt{1 \vee \log(d/\varepsilon)}.$$

Some Geometric Intuition. Figure 1 gives some geometric intuition about Gaussian anti-concentration. The left plot shows a rectangle and a contour of a Gaussian with no correlation. The rectangle is distorted to show that, in high dimensions, the corners are far from

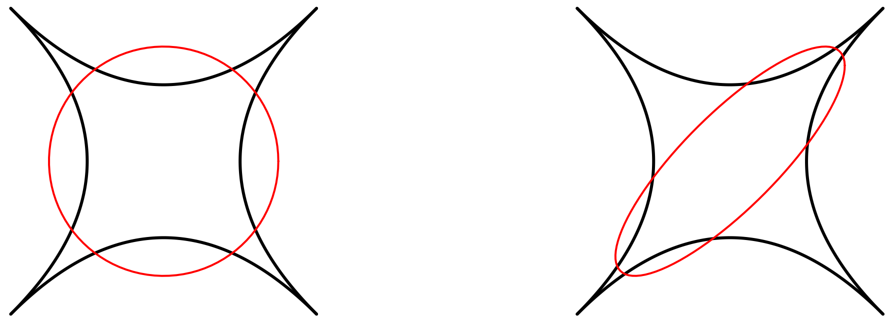


Figure 1: A geometric view of Gaussian anti-concentration. Left: A rectangle and a contour of a Gaussian with no correlation. The rectangle is distorted to show that, in high dimensions, the corners are far from the origin. The Gaussian cuts through the rectangle in a very small region. Right: Gaussian with high correlation. In this case, we are in the tail of the Gaussian.

the origin. The Gaussian cuts through the rectangle in a very small region which is why we get anti-concentration. The right plot shows a Gaussian with high correlation. In this case, the Gaussian cuts through the rectangle in a more substantial way. However, in this case, we are in the tail of the Gaussian so again the probability near the rectangle is small.

17 Gaussian Comparison

Before getting to the bootstrap, we need one more results from CCK. This result compares the distribution of two different Gaussian distributions.

Theorem 17 *Let* $X = (X_1, \ldots, X_d)^T \sim N(0, \Sigma^X)$ *and* $Y = (Y_1, \ldots, Y_d)^T \sim N(0, \Sigma^Y)$*. Define*

$$\Delta = \max_{1 \leq j,k, \leq d} |\Sigma^X_{jk} - \Sigma^Y_{jk}|.$$

Suppose that $\min_j \Sigma^Y_{jj} > 0$*. Then*

$$\sup_x \left| \mathbb{P}(\max_j X_j \leq x) - \mathbb{P}(\max_j Y_j \leq x) \right| \leq C \Delta^{1/3} (1 \vee \log(d/\Delta))^{2/3} \tag{54}$$

where C depends on $\min_j \Sigma^Y_{jj}$ *and* $\max_j \Sigma^Y_{jj}$*.*

Proof Outline. The proof follows the same strategy as the proof of Theorem 14. First we approximate the indicator function and the max function by a smooth function m. Then we use the Slepian interpolation to bound $\mathbb{E}[m(X)] - \mathbb{E}[m(Y)]$. Bounding this difference is easier since both distributions are Gaussian. Thus m' can be evaluated explicitly. Applying Stein's identity leads to a simple expression for m' that can be bounded in terms of the variances. The smoothing in this case only requires a function that has two continuous derivatives. This is what gives rise to the power 1/3. The error introduced by the smoothing is bounded by Gaussian anti-concentration. □

18 The Bootstrap

The validity of the bootstrap is proved in Section 3 of CCK1 and in Appendix H of the same paper. I'll present a simpler, albeit less rigorous, result here.

Let $X_1,\ldots,X_n \sim P$. Let P_n be the empirical distribution and let $X_1^*,\ldots,X_n^* \sim P_n$.

Theorem 18 *Under appropriate conditions we have,*

$$\sup_z \left| \mathbb{P}\big(\sqrt{n}||\overline{X}^* - \overline{X}||_\infty \le z \mid X_1,\ldots,X_n\big) - \mathbb{P}\big(\sqrt{n}||\overline{X} - \mu||_\infty \le z\big)\right| = O_P\left(\frac{\log d}{n^{1/8}}\right) \tag{55}$$

where $\overline{X}^* = \frac{1}{n}\sum_i X_i^*$.

Proof. Let $\Sigma = \mathrm{Var}(X_i)$ and let $\widehat{\Sigma}$ denote the sample covariance. For simplicity, assume that $\mu = (0,\ldots,0)^T$. Let

$$F(z) = \mathbb{P}\big(\sqrt{n}||\overline{X} - \mu||_\infty \le z\big)$$
$$\widehat{F}(z) = \mathbb{P}\big(\sqrt{n}||\overline{X}^* - \overline{X}||_\infty \le z \mid X_1,\ldots,X_n\big).$$

Let $Y \sim N(0,\Sigma)$ and $\widetilde{Y} \sim N(0,\widehat{\Sigma})$ where we treat $\widehat{\Sigma}$ as fixed. Then

$$\sup_z |F(z) - \widehat{F}(z)| \le \mathrm{I} + \mathrm{II} + \mathrm{III}$$

where

$$\mathrm{I} = \sup_z |F(z) - \mathbb{P}(||Y||_\infty \le z)|$$
$$\mathrm{II} = \sup_z |\mathbb{P}(||Y||_\infty \le z) - \mathbb{P}(||\widetilde{Y}||_\infty \le z)|$$
$$\mathrm{III} = \sup_z |\widehat{F}(z) - \mathbb{P}(||\widetilde{Y}||_\infty \le z)|.$$

By Theorem 14,

$$\mathrm{I} \le M_d^{1/4}\left(\frac{(\log(dn))^7}{n}\right)^{1/8}.$$

Another application of Theorem 14 yields

$$\mathrm{III} \leq \widehat{M}_d^{1/4} \left(\frac{(\log(dn))^7}{n} \right)^{1/8}$$

where $\widehat{M}_d$ refers to the empirical version of M_d. Assuming suitable moments conditions, we have $\widehat{M}_d \leq M_d(1 + o_P(1))$ and so

$$\mathrm{III} = O_P\left(M_d^{1/4} \left(\frac{(\log(dn))^7}{n} \right)^{1/8} \right).$$

To bound II, apply Theorem 17. Then

$$\mathrm{II} \leq C\Delta^{1/3}(1 \vee \log(d/\Delta))^{2/3}$$

where

$$\Delta = \max_{1\leq j,k,\leq d} |\widehat{\Sigma}_{jk} - \Sigma_{jk}|.$$

Again, assuming suitable moments conditions, we can apply concentration of measure and the union bound to conclude that

$$\Delta = O_P\left(\sqrt{\frac{\log d}{n}} \right).$$

Hence,

$$\mathrm{II} \leq C\left(\frac{\log d}{n} \right)^{1/6} \preceq \frac{\log d}{n^{1/8}}$$

and the result follows. □

Since the limiting distribution is continuous we get the follow corollary.

Corollary 19 *Let*

$$Z_\alpha = \inf\left\{ z : \ \mathbb{P}\left(\sqrt{n}||\overline{X}^* - \overline{X}||_\infty > z \ \Big| \ X_1,\ldots,X_n \right) \leq \alpha \right\}.$$

Let

$$R_n = \left\{ \mu : \ ||\overline{X} - \mu||_\infty \leq \frac{Z_\alpha}{\sqrt{n}} \right\}.$$

Then

$$\mathbb{P}(\mu \in R_n) = 1 - \alpha - O\left(\frac{\log d}{n^{1/8}} \right).$$

19 Functional Version

CCK4 has an extension to empirical processes. Let

$$\mathbb{G}_n f = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - \mathbb{E}[f(X_i)]), \quad f \in \mathscr{F}_n$$

be an empirical process over the class of functions $\mathscr{F}_n$. Intuitively, we can get similar results by approximating $\mathscr{F}$ with a finite cover and the adapting the previous results.

Let $Z_n = \sup_{f\in\mathscr{F}_n} \mathbb{G}_n f$ and let $Z = \sup_{f\in\mathscr{F}_n} \mathbb{G} f$ where $\mathbb{G}$ is a centered Gaussian process with covariance $\mathrm{Cov}(f(X_i), g(X_i))$. Let F be a measurable envelope for $\mathscr{F}$ and assume that $||F||_{P,q} = \int |F(x)|^q dP(x) < \infty$. Let $\kappa = (\mathbb{E}[|f(X)|^3])^{1/3}$, let $N(\mathscr{F}, \epsilon)$ be the covering number under the metric $\sqrt{\int |f-g|^2 dP}$ and let $H_n(\epsilon) = \log(N(\mathscr{F}, \epsilon||F||_{P,2}) \vee n)$.

Theorem 20 (CCK4) *There are constants K and C such that, for all* $0 < \epsilon \leq 1$ *and* $0 < \gamma < 1$*, Then*

$$\mathbb{P}\Big(|Z_n - Z| > K\Delta_n(\epsilon,\gamma)\Big) \leq \gamma(1 + \delta_n(\epsilon,\gamma) + \frac{C\log}{n}$$

where

$$\Delta_n(\epsilon,\gamma) = \phi(\epsilon) + \gamma^{-1/q}\epsilon||F||_{P,2} + n^{-1/2}\gamma^{-1/q}||M||_q + n^{-1/2}\gamma^{-2/q}||M||_2$$
$$+ n^{-1/4}\gamma^{-1/2}\sqrt{\mathbb{E}[||\mathbb{G}_n||_{\mathscr{F}\cdot\mathscr{F}}]H_n(\epsilon)} + n^{-1/6}\gamma^{-1/3}\kappa H_n^{2/3}(\epsilon)$$

and

$$\delta_n(\epsilon,\gamma) = \frac{1}{4}\mathbb{E}[(F/\kappa)^3 I(F/\kappa > c\gamma^{-1/3}n^{1/3}H_n(\epsilon)^{-1/3})].$$

This is not easy to parse. Let us look at an example. Let $\widehat{p}_h$ be the kernel density estimator with bandwidth h and let p_h be the mean of $\widehat{p}_h$. Let

$$Z_n = \sup_x \sqrt{nh^d}(\widehat{p}_h(x) - p_h(x))$$

and let $\widetilde{Z}_n$ be the corresponding supremum for the Gaussian approximation. The theorem then gives

$$|Z_n - \widetilde{Z}_n| = O_P\left(\frac{\log n}{(nh^d)^{1/6}}\right).$$

More on density estimation in the next section.

20 The Bootstrap For Density Estimators

Let $\mathscr{F}_n$ be a class of functions and consider the empirical process $\{\mathbb{G}_n(f): f \in \mathscr{F}_n\}$ where

$$\mathbb{G}_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - \mathbb{E}[f(X_i)]).$$

The results can be extended to such an empirical process by forming a finite covering of $\mathscr{F}_n$ and applying the previous results. The details are non-trivial; see CCK3 and CCK4. An important application is for constructing confidence bands in density estimation.

Let $X_1, \ldots, X_n \sim P$ where P has density p. Consider the usual kernel estimator

$$\widehat{p}_h(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{||x - X_i||}{h}\right).$$

Let $p_h(x) = \mathbb{E}[\widehat{p}_h(x)]$ and define

$$[\ell(x), u(x)] = \left[\widehat{p}_n(x) - \frac{Z_\alpha}{\sqrt{nh^d}},\ \widehat{p}_n(x) + \frac{Z_\alpha}{\sqrt{nh^d}}\right]$$

where Z_α is the bootstrap quantile defined by

$$\mathbb{P}\left(\sqrt{nh^d}||\widehat{p}_h^*(x) - \widehat{p}_h||_\infty > Z_\alpha \,\middle|\, X_1, \ldots, X_n\right) = \alpha.$$

Theorem 21 *We have*

$$\mathbb{P}\Big(\ell(x) \le p_h(x) \le (x) \ \text{ for all } x\Big) \ge 1 - \alpha - O\left(\frac{\log n}{(nh^d)^{1/8}}\right).$$

Actually, the result as I have stated it does not appear explicitly in CCK3. They use a studentized version of the process and they use a multiplier bootstrap instead of the usual bootstrap. Nonetheless, the version I have stated above appears to follow essentially from Theorem 3.1 of CCK3.

Direct Coupling. There is another way to get a bound for the bootstrap for density estimation: one creates a coupling between the data and the bootstrap sample. Neumann (1998) does this as follows. First, he couples the data $(X_1, \ldots, X_n)$ to the output $(\widetilde{X}_1, \ldots, \widetilde{X}_n)$ of a smoothed bootstrap. Then he couples $(\widetilde{X}_1, \ldots, \widetilde{X}_n)$ to the bootstrap output $(X_1^*, \ldots, X_n^*)$.

Draw $\widetilde{X}_1, \ldots, \widetilde{X}_n$ from a kernel density estimate $\widehat{p}_g$ with bandwidth g. Let $\pi = \int (p(x) \wedge \widehat{p}_g(x))dx$. Draw $B \sim$ Bernoulli(π). If $B = 1$, draw X_i from the density

$$\frac{p(x) \wedge \widehat{p}_g(x)}{\pi}$$

and set $\widetilde{X}_i = X_i$. If $B = 0$, draw X_i from the density

$$\frac{p(x) - (p(x) \wedge \widehat{p}_g(x))}{1-\pi}$$

and independently draw $\widetilde{X}_i$ from the density

$$\frac{\widehat{p}_g(x) - (p(x) \wedge \widehat{p}_g(x))}{1-\pi}.$$

Then $X_1, \ldots, X_n$ are iid from p and $\widetilde{X}_1, \ldots, \widetilde{X}_n$ are iid from $\widehat{p}_g$. Furthermore, $\mathbb{P}(X_i = \widetilde{X}_i) = \pi$. Now construct $\widehat{p}_h^*$ from $(\widetilde{X}_1, \ldots, \widetilde{X}_n)$. By dividing the space into small cubes and bounding the difference over the cubes, Neumann proves that,

$$\sup_x \left| (\widehat{p}_h(x) - \mathbb{E}[\widehat{p}_h(x)]) - (\widehat{p}_h^*(x) - \mathbb{E}[\widehat{p}_h^*(x)]) \right| = O_P\left(\sqrt{\frac{\log n}{nh^d}\left(g^2 + \sqrt{\frac{\log n}{ng^d}} \right)} \right). \tag{56}$$

For the regular bootstrap, we do not directly couple X_i^* to X_i; this won't work since the distribution of X_i^* is discrete and has no density. Instead we couple X_i^* to $\widetilde{X}_i$. Indeed, we can think of $\widetilde{X}_i$ as X_i^* plus noise. Hence, we have the coupling $||\widetilde{X}_i - X_i^*|| \le \sqrt{d}g$. This yields, with $\widehat{p}_h^*$ now denoting the estimator based on the usual bootstrap,

$$\sup_x \left| (\widehat{p}_h(x) - \mathbb{E}[\widehat{p}_h(x)]) - (\widehat{p}_h^*(x) - \mathbb{E}[\widehat{p}_h^*(x)]) \right| = O_P\left(\sqrt{\frac{\log n}{nh^d}\left(\frac{g}{h} + \sqrt{\frac{\log n}{ng^d}} \right)} \right). \tag{57}$$

Choosing $g = (h^2/n)^{1/(2+d)}$, it appears that

$$\sup_x \left| (\widehat{p}_h(x) - \mathbb{E}[\widehat{p}_h(x)]) - (\widehat{p}_h^*(x) - \mathbb{E}[\widehat{p}_h^*(x)]) \right| = O_P\left(\frac{\log n}{nh^d} \right)^{\frac{4+d}{2(2+d)}}. \tag{58}$$

References

The main sources are:

(CCK1) Victor Chernozhukov, Denis Chetverikov, Kengo Kato (2012). Central Limit Theorems and Multiplier Bootstrap when p is much larger than n. http://arxiv.org/abs/1212.6906.

(CCK2) Victor Chernozhukov, Denis Chetverikov, Kengo Kato (2013). Comparison and anti-concentration bounds for maxima of Gaussian random vectors. http://arxiv.org/abs/1301.4807.

(CCK3) Victor Chernozhukov, Denis Chetverikov, Kengo Kato (2013). Anti-Concentration and Honest Adaptive Confidence Bands. http://arxiv.org/abs/1303.7152.

(CCK4) Victor Chernozhukov, Denis Chetverikov, Kengo Kato (2013). Gaussian approximation of suprema of empirical processes. http://arxiv.org/abs/1212.6885.

Other useful references are:

Chatterjee, S. (2008). A simple invariance theorem. arxiv:math/0508213

Chen, Goldstein and Shao. (2011). *Normal Approximation by Stein's Method.* Springer.

Neumann, M. (1998). Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations. *The Annals of Statistics*, 26, 2014-2048.

Nourdin and Peccati (2012). *Normal Approximations With Malliavin Calculus.* Cambridge.

Rollin, A. (2013). Stein's method in high dimensions with applications. arXiv:1101.4454.