Central Limit Theorem via Stein's Method

Kevin Eng

April 19 , 2016

1 Introduction

Stein's method is a collection of tools one can use to determine the distance between two probability distributions with respects to a given metric. In this report we discuss Stein's method with respects to the Wasserstein distance. In particular, we show the bounds generated on the Wasserstein distance via Stein's method provides a direct proof of the Lindeberg-Feller Central Limit Theorem.

2 Wasserstein Distance

This metric goes by several names. Mathematicians call it the Wassertein distance; computer scientist call it the earth mover's distance; statisticians call it Mallow's distance. Here we shall side with the mathematicians. In the context of this report, the Wasserstein distance is useful because convergence in W_1 implies convergence in distribution.

Definition. *The pth Wasserstein distance, W_p, is defined as*

$$W_p(P_1, P_2) = \inf E[d(X, Y)^p].$$

The infimum is over all random variables X and Y with probability distributions P_1 and P_2 respectively.

An important theorem due to Kantorovich and Rubinstein shows that the Wasserstein distance when $p = 1$ is equivalent to

$$W_1(P_1, P_2) = \sup \left\{ \int_\Omega f \, d(P_1 - P_2) : f : \Omega \to \mathbb{R}, Lip(f) \leq 1 \right\}.$$

Here we are assuming Ω is endowed with some metric d and that f is Lipschitz with respects to this metric. Phrased in the framework of probability, the WasserStein distance between the probability distribution of two random variables X and Y is

$$W_1(P_X, P_Y) = \sup_{h \in \mathcal{H}} |E[h(X)] - E[h(Y)]|$$

where $\mathcal{H}$ is the class of all Lipschitz continuous functions with Lipschitz constant one. It is worth noting the class of functions, $\mathcal{H}$, determines the metric. Indeed

- If $\mathcal{H} = \{f(y) = I\{y \leq x\} : x \in \mathbb{R}\}$ then we obtain the Kolmogorov metric.
- If $\mathcal{H} = \{f(y) = I\{y \in A\} : A \in \mathcal{B}(\mathbb{R})\}$ then we obtain the total variation metric.

We can intuitively understand the Wasserstein distance as follows. Suppose we are given two distributions P_1 and P_2. Imagine P_1 is a collection of piles of dirt containing unit mass and P_2 as a collection of holes missing unit mass of dirt. The Wasserstein distance can be thought of as the cost required to move the piles of dirt to fill in the holes (this is why computer scientist like to call it the earth mover's distance). To make this more precise we write

$$P_1 = \{(p_1, w_{p_1}), \ldots, (p_n, w_{p_n})\}$$
$$P_2 = \{(q_1, w_{q_1}), \ldots, (q_n, w_{q_m})\}$$

where $P_1(p_i) = w_{p_i}$ and $P_2(q_i) = w_{q_i}$. Additionally, let $P(X = p_i, Y = q_j) = f_{ij}$. The cost function is then defined as

$$C(P_1, P_2) = \sum_{ij} f_{ij} d_{ij}.$$

Here d_{ij} is some measure of dissimilarity between p_i and q_j and f_{ij} represents the amount of dirt transfered from pile i to hole j. In this example we use $d_{ij} = |p_i - q_j|$. The Wasserstein distance seeks to find the minimum cost over all joint distributions $F \equiv f_{ij}$ of X and Y subject to the constraints

$$\sum_j f_{ij} = w_{p_i}, \quad 1 \leq i \leq n \tag{1}$$
$$\sum_i f_{ij} = w_{q_j}, \quad 1 \leq j \leq m \tag{2}$$

Restriction (1) ensures that all the dirt in each pile is used to fill in the holes. Restriction (2) ensures that no hole receives more dirt than is necessary to fill it.

3 Stein's Method

We first look at the solution to a particular differential equation which will aide us in constructing the center piece of Stein's method.

Lemma 1. *Let $\Phi(x)$ be the c.d.f of the standard normal distribution, then the unique bounded solution f_x of the differential equation*

$$f_x'(w) - w f_x(w) = I\{w \leq x\} - \Phi(x)$$

is given by

$$\begin{aligned} f_x(w) &= e^{w^2/2} \int_w^\infty e^{-t^2/2} (\Phi(x) - I\{t \leq x\})\, dt \\ &= -e^{w^2/2} \int_{-\infty}^w e^{-t^2/2} (\Phi(x) - I\{t \leq x\})\, dt. \end{aligned}$$

The heart of Stein's method relies on the following relationship:

Lemma 2. *Define the functional operator $\mathcal{A}$ by*

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

1 *If $Z \sim N(0,1)$, then $E\mathcal{A}f(Z) = 0$ for all absolutely continuous functions f with $E|f'(Z)| < \infty$.*

2 *If for some random variable W, $E\mathcal{A}f(W) = 0$ for all absolutely continuous functions f with $\|f'\| < \infty$, then $W \sim N(0,1)$. Here $\|\cdot\|$ is the supremum norm.*

Proof. Let $Z \sim N(0,1)$ and let f be absolutely continuous such that $E|f'(Z)| < \infty$. Then

$$\begin{aligned}
Ef'(Z) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-t^2/2} f'(t)\, dt \\
&= \frac{1}{\sqrt{2\pi}} \int_0^\infty f'(t) \left[\int_t^\infty we^{-w^2/2}\, dw \right] dt + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 f'(t) \left[\int_{-\infty}^t we^{-w^2/2}\, dw \right] dt \\
&= \frac{1}{\sqrt{2\pi}} \int_0^\infty we^{-w^2/2} \left[\int_0^w f'(t)\, dt \right] dw + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 we^{-w^2/2} \left[\int_w^0 f'(t)\, dt \right] dw \\
&= E[Zf(Z)].
\end{aligned}$$

On the other hand suppose W is a random variable such that $E[f'(W) - Wf(W)] = 0$ for all absolutely continuous f satisfying $\|f'\| < \infty$. The function f_x from lemma 1 is a function satisfying the previously mentioned conditions. Hence for all $x \in \mathbb{R}$

$$0 = E[f_x'(W) - Wf_x(W)] = P(W \le x) - \Phi(x).$$

Thus $W \sim N(0,1)$. □

Corollary. *If $Z \sim N(0,1)$ and f is Lipschitz continuous then*

$$E[f'(Z) - Zf(Z)] = 0.$$

The idea is that given some random variable X and $Z \sim N(0,1)$, if $E[f'(X) - Zf(Z)]$ is close to zero then X should be similar to a standard normal. To formalize this idea we introduce the Stein equation.

Definition. *Let $h : \mathbb{R} \to \mathbb{R}$ be a Lipschitz continuous function. Then Stein equation of h which we will call f_h is*

$$f_h(x) = e^{-x^2/2} \int_{-\infty}^x [h(t) - Eh(Z)] e^{-t^2/2}\, dt$$

where Z is the standard normal distribution.

Lemma 3 gives us the relationship between lemma 1 and the Stein equation.

Lemma 3. *Let $h : \mathbb{R} \to \mathbb{R}$ be an Lipschitz continuous function. The Stein equation of h satisfies the condition*

$$E[h(X)] - E[h(Z)] = E[f_h'(X) - Xf_h(X)]$$

Proof. From the product rule we have

$$\begin{aligned} f_h'(x) &= e^{x^2/2}\left[(h(x) - Eh(Z))e^{-x^2/2}\right] + xe^{-x^2/2}\int_{-\infty}^{x}[h(x) - Eh(Z)]e^{-t^2/2}\,dt \\ &= [h(x) - Eh(Z)] + xf_h(x) \end{aligned}$$

Hence

$$\begin{aligned} E[f_h'(X) - Xf_h(X)] &= E[h(x) - Eh(Z)] \\ &= Eh(X) - Eh(Z) \end{aligned}$$

□

There are several important facts about the Stein equation that we will need to use in the proof of the central limit theorem. They are presented in lemma 4 without proof.

Lemma 4. *For a given function $h : \mathbb{R} \mapsto \mathbb{R}$, let f_h be the solution to the Stein equation and $Z \sim N(0,1)$. If h is bounded, then*

$$\|f_h\| \leq \sqrt{\pi/2}\|h(.) - Eh(Z)\| \qquad \|f_h'\| \leq 2\|h(.) - Eh(Z)\|.$$

Additionally, if h is absolutely continuous, then

$$\|f_h\| \leq 2\|h'\| \qquad \|f_h'\| \leq \sqrt{2/\pi}\|h'\| \qquad \|f_h''\| \leq 2\|h'\|.$$

Here $\|\cdot\|$ is the supremum norm.

4 CLT via Stein's Method

In this section we provide a direct proof of CLT using Stein's Method. From the previous section we know

$$\begin{aligned} W_1(X,Z) &= \sup_{f\in\mathcal{H}} |Ef(X) - Ef(Z)| \\ &= \sup_{f\in\mathcal{H}} |E[f_h'(X) - Xf_h(X)]|. \end{aligned}$$

It will then follow $X \xrightarrow{d} Z$ since convergence in W_1 implies convergence in distribution.

Theorem 1 (CLT)**.** *Let $\{X_n\}$ be a sequence of i.i.d random variables such that $E[X_1] = 0$, $E[X_1^2] = 1$ and $E[|X_1|^3] < \infty$. Then*

$$\sqrt{n}\cdot\overline{X}_n \xrightarrow{d} N(0,1)$$

where $\overline{X}_n = (1/n)\sum_{i=1}^n X_i$.

Proof. The proof will use the "leave one out" technique. Specifically, we let $S_n = (X_1 + \ldots + X_n)/\sqrt{n}$ and $S_n' = S_n - X_1/\sqrt{n}$. At first this may seem like a pointless maneuver but its significance will be made clear in the following steps. Recall our goal is now to show $|E[f_h'(S_n) - S_n f_h(S_n)]|$ is small. Since the X_i's are independent we have

$$\begin{aligned} E[S_n f_h(S_n)] &= \frac{1}{\sqrt{n}} \sum_i E[X_i f_h(S_n)] \\ &= E[\sqrt{n} X_1 f_h(S_n)]. \end{aligned}$$

We now need someway to compare $E[\sqrt{n}X_1 f_h'(S_n)]$ to $E[f_h'(S_n)]$. This is where the leave one out technique comes into play. By Taylor's theorem we know for some constant a,

$$f(x+a) = f(x) + f'(x)a + \frac{f''(\xi)}{2}a^2$$

where ξ is some value in $(x, x+a)$. Hence

$$\begin{aligned} E[\sqrt{n}X_1 f_h(S_n)] &= E\left[\sqrt{n}X_1 f_h\left(S_n' + \frac{X_n}{\sqrt{n}}\right)\right] \\ &= E\left[\sqrt{n}X_1\left(f_h(S_n') + \frac{X_1}{\sqrt{n}}f_h'(S_n') + \frac{X_1^2}{2n}f_h''(S_n')\right)\right] \\ &= E[\sqrt{n}X_1 f_h(S_n')] + E[X_1^2 f_h'(S_n')] + E\left[\frac{X_1^3}{2\sqrt{n}}f_h''(S_n')\right]. \end{aligned}$$

Since X_1 and S_n' are independent we have

$$\begin{aligned} E[\sqrt{n}X_1 f_h(S_n')] &= E[X_1]E[\sqrt{n}f_h(S_n')] = 0 \\ E[X_1^2 f_h'(S_n')] &= E[X_1^2]E[f_h'(S_n')] = E[f_h'(S_n')]. \end{aligned}$$

By lemma 4, $|f_h''| \leq 2|h'|$. Now since we are working with the metric W_1 the functions h consist of all Lipschitz functions with Lipschitz constant one and thus $|f_h''| \leq 2|h'| = 2$. Hence

$$\begin{aligned} \left|\frac{X_1^3}{2\sqrt{n}}f_h''(S_n')\right| &= \frac{|X_1|^3}{2\sqrt{n}}|f_h''(\xi)| \\ &\leq \frac{|X_1|^3}{\sqrt{n}}. \end{aligned}$$

From this we have the bound

$$\left|E\left[\frac{X_1^3}{2n}f_h''(S_n')\right]\right| \leq \frac{E|X_1|^3}{\sqrt{n}}.$$

Likewise we can use the first order Taylor approximation on f_h' to deduce

$$f_h'(S_n) = f_h'(S_n') + f_h''(\xi) \cdot \frac{X_1}{\sqrt{n}}.$$

Hence

$$E[X_1^2 f_h'(S_n')] = E[X_1^2 f_h'(S_n)] - E\left[f_h''(\xi) \cdot \frac{X_3}{\sqrt{n}}\right].$$

We can bound the error term like before to get

$$\left|E\left[f_h''(\xi) \cdot \frac{X_3}{\sqrt{n}}\right]\right| \leq \frac{2E|X_1|^3}{\sqrt{n}}.$$

Using the independence condition again we also have $E[X_1^2 f_h'(S_n)] = E[f_h'(S_n)]$. Thus we conclude

$$|E[S_n f_h(S_n)] - E[f_h'(S_n)]| \leq \frac{2E|X_1|^3}{\sqrt{n}} + \frac{|X_1|^3}{\sqrt{n}} = \frac{3E|X_1|^3}{\sqrt{n}}.$$

Since we are given bounded third moment it follows $\lim_n |E[S_n f_h(S_n)] - E[f_h'(S_n)]| \to 0$ and thus $S_n \xrightarrow{d} N(0,1)$. □

5 Zero Bias Coupling

Stein's method can also be used to deal with dependent random variables. Indeed the method of attack for the independent and dependent cases are practically identical. The only caveat being that in the dependent case we need more bells and whistles to get around the dependency condition. We begin by introducing the concept of a zero-bias distribution.

Definition. *Let W be a random variable such that $E[W] = 0$ and $Var(W) = \sigma^2 < \infty$. We say the random variable W^z has the zero-bias distribution with respects to W if for all absolutely continuous f such that $E[Wf(W)] < \infty$ we have*

$$E[Wf(W)] = \sigma^2 E[f'(W^z)].$$

The next theorem gives us a way to bound W_1.

Theorem 2. *Let W be a random variable such that $E[W] = 0$ and $Var(W) < \infty$. Suppose W^z and W are defined on the same space. If $Z \sim N(0,1)$, then*

$$W_1(W, Z) \leq 2E|W^z - W|$$

Proof. Using $\mathcal{H}$ as defined before we have

$$\begin{aligned} W_1(W, Z) &\leq \sup_{f \in \mathcal{H}} |E[f'(W) - Wf(W)]| \\ &= \sup_{f \in \mathcal{H}} |E[f'(W) - f'(W^z)]| \\ &\leq \sup_{f \in \mathcal{H}} \|f''\| E|W - W^z|. \end{aligned}$$

□

Unfortunately there are some difficulties with bounding W_1 using W^z. Although we may know W^z exists it is of little use in Stein's method if we do not know what it is. Indeed, we are trying to construct an explicit bound. As it turns out in the case W is a sum of independent random variables we can construct W^z explicitly.

Lemma 5. *Let $\{X_i\}_{i=1}^n$ be a sequence of random variables such that $E[X_i] = 0$ and $\sum Var(X_i) = 1$. If $W = \sum X_i$ then W^z is constructed as follows.*

1. *For each $i \in [n]$ let X_i^z have the zero-bias distribution of X_i independent of $(X_j)j \neq i$ and $(X_j^z)_{j \neq i}$.*
2. *Choose a random summand X_I, where the index I satisfies $P(I = i) = \sigma_i^2$ and is independent of all else.*
3. *Define $W^z = \sum_{j \neq I} X_j + X_I^z$*

The follow lemma gives two important properties of zero-bias distributions.

Lemma 6. *Let W be a random variable such that $E[W] = 0$ and $Var(W) = \sigma^2 < \infty$.*

1. *There is a unique probability distribution for a random variable W^z satisfying*
$$E[Wf(W)] = \sigma^2 E[f'(W^z)] \tag{3}$$
for all absolutely continuous functions f such that $E[Wf(W)] < \infty$.
2. *The distribution of W^z in (3) is absolutely continuous with respect to the Lebesgue measure with density*
$$p^z(w) = \sigma^{-2} E[WI\{W > w\}] = -\sigma^{-2} E[WI\{W \leq w\}]$$

The main take away is that W^z exists and that there is a unique distribution associated with W^z. This is essentially extending lemma 2 to deal with the zero-bias case.

Theorem 3. *Let $(X_{i,n})$, $n \geq 1$, $1 \leq i \leq n$ be a triangular array and suppose I_n is a random variable independent of the $X_{i,n}$ with $P(I_n = i) = \sigma_{i,n}^2$. For each $1 \leq i \leq n$, let $X_{i,n}^z$ have the zero-bias distribution of $X_{i,n}$ independent of the others. Then the Lindeberg-Feller condition holds if and only if*
$$X_{I_n,n}^z \xrightarrow{P} 0$$

Proof. For some fixed $\epsilon > 0$, define function f such that $f'(x) = I\{|x| \geq \epsilon\}$. This way $xf(x) = (x^2 - \epsilon|x|)I\{|x| \geq \epsilon\}$. Then by definition of a zero-bias transformation

$$\begin{aligned}
P(|X_{I_n,n}^z| \geq \epsilon) &= \sum_{i=1}^n P(I_n = i) P(|X_{i,n}^z| \geq \epsilon) \\
&= \sum_{i=1}^n \sigma_{i,n}^2 P(|X_{i,n}^z| \geq \epsilon) \\
&= \sum_{i=1}^n \sigma_{i,n}^2 E[f'(X_{i,n}^z)] \\
&= \sum_{i=1}^n E[(X_{i,n}^2 - \epsilon|X_{i,n}|)I\{|X_{i,n}| \geq \epsilon\}.
\end{aligned}$$

We have the relation

$$\frac{x^2}{2} I\{|x| \geq 2\epsilon\} \leq (x^2 - \epsilon|x|) I\{|x| \geq \epsilon\} \leq x^2 I[|x| \geq \epsilon].$$

Hence we have the bounds

$$\frac{1}{2} \sum_{i=1}^{n} E[X_{i,n}^2 I\{|X_{i,n}| \geq 2\epsilon\} \leq P(|X_{I_n,n}^z| \geq \epsilon) \leq \sum_{i=1}^{n} E[X_{i,n}^2 I\{|X_{i,n} \geq \epsilon\}].$$

Looking and the lower and upper bound we can see the Lindeberg-Feller condition holds if and only if $X_{I_n,n}^z \xrightarrow{P} 0$. □

Theorem 4 (Lindeberg-Feller CLT)**.** *Suppose an array of random variables* $(X_{i,n})$*,* $n \geq 1$*,* $1 \leq i \leq n$ *satisfies the Lindeberg condition: for all* $\epsilon > 0$*.*

$$\sum_{i=1}^{n} E[X_{i,n}^2 I\{|X_{i,n} > \epsilon\}] \to 0.$$

If $X_{I_n,n}^z \xrightarrow{P} 0$*, then* $W_n \xrightarrow{d} N(0,1)$*.*

Proof. We have

$$W_1(W_n, Z) \leq \sup_{f \in \mathcal{H}} |E[f'(W) - f'(W^z)]|$$

Hence if we can show $|E[f'(W) - f'(W^z)]| \to 0$ we are done. This is the analogous setup we had in the independent case. Using theorem 2

$$\begin{aligned}
|E[f'(W) - f'(W^z)]| &\leq E|f'(W) - f'(W^z)| \\
&= \int_0^\infty P(|f'(W_n) - f'(W_n^2)| \geq t)\, dt \\
&= \int_0^{2\|f'\|} P(|f'(W_n) - f'(W_n^2)| \geq t)\, dt \\
&\leq \int_0^{2\|f'\|} P(\|f''\||W_n - W_n^z| \geq t)\, dt
\end{aligned}$$

If we can show $P(\|f''\||W_n - W_n^z| \geq t) \xrightarrow{P} 0$ then by DCT $|E[f'(W) - f'(W^z)]| \to 0$. From theorem 3, $X_{I_n,n}^z \xrightarrow{P} 0$ and by construction $|W_n^z - W_n| = |X_{I_n^z,n} - X_{I_n,n}|$. Hence $|W_n^z - W_n| = o_p(1)$ if $X_{I_n,n} = o_p(1)$. Let $M_n = \max_{i \in [n]} \sigma_{i,n}^2$. Then by Markov's inequality

$$\begin{aligned}
P(|X_{I_n}, n| \geq \epsilon) &\leq \frac{Var(X_{I_n,n})}{\epsilon^2} \\
&= \frac{1}{\epsilon^2} \sum_{i=1}^{n} \sigma_{i,n}^4 \\
&\leq \frac{M_n}{\epsilon^2} \sum_{i=1}^{n} \sigma_{i,n}^2 \\
&= \frac{M_n}{\epsilon^2}
\end{aligned}$$

All that is left to show is $M_n \to 0$. Fix some $\delta > 0$ and decompose the variance

$$\begin{aligned}\sigma_{i,n}^2 &= E[X_{i,n}^2 I\{|X_{i,n}| \leq \delta\}] + E[X_{i,n}^2 I\{|X_{i,n}| > \delta\}] \\ &\leq \delta^2 + E[X_{i,n}^2 I\{|X_{i,n}| > \delta\}].\end{aligned}$$

Now using the bound

$$\frac{1}{2}\sum_{i=1}^{n} E[X_{i,n}^2 I\{|X_{i,n}| \geq 2\delta\} \leq P(|X_{I_n,n}^z| \geq \delta)$$

and the fact $X_{I_n,n}^z = o_p(1)$, it follows $E[X_{i,n}^2 I\{|X_{i,n}| > \delta\}] \to 0$. Hence $\limsup M_n \leq \delta$. Since $\delta > 0$ was arbitrary it follows $M_n \to 0$. □

References

[1] Louis H.Y Chen, Larry Goldstein, and Qi-Man Shao. *Normal Approximation by Stein Method*. Springer, 2011.

[2] Nathan Ross. "Fundamentals of Stein's method". In: *Probability Surveys* 8 (2011).

[3] Gregory Valiant. *Stein's Method*. 2014. URL: `http://theory.stanford.edu/~valiant/teaching/CS362/cs362_steins.pdf`.

[4] *Wasserstein Metric*. URL: `https://en.wikipedia.org/wiki/Wasserstein_metric`.