STEIN'S METHOD IN DIVERSE SETTINGS

SHREYAS SRIRAM

Abstract. This paper aims to explore the application of Stein's Method in various formats. Stein's Method allows us to understand the tendency of a collection of random variables towards a normal distribution, even if that collection is not independent or identically-distributed. After highlighting probability-related concepts pertinent to this subject, we will establish Stein's Method through a Central Limit Theorem proof that utilizes Wasserstein distances. We will then examine the implementation of Stein's Method in the realms of dependency graphs and exchangeable pairs.

Contents

1. Foundational Concepts

The main goals of Stein's method are to demonstrate a way to prove the Central Limit Theorem and to understand the tendency of the average of a collection of random variables towards a normal distribution, even if that collection is not independent and identically distributed. This section will highlight fundamental concepts in classical statistics, providing a base for the paper's main subject.

First, we will understand the properties of a random variable.

Definition 1.1. In an experiment, X is a **random variable** which signifies the outcome of an experiment that is based on a random event. For example, if the experiment is the roll of a standard die:

- X is the value of the roll, taking on a value from $\{1, 2, 3, 4, 5, 6\}$, for which the respective probabilities are denoted as $\{p_1, p_2, \ldots, p_6\}$.

In classical statistics, $\{p_1 + p_2 + \cdots + p_6 = 1\}$. $\{p_1, p_2, \ldots, p_6\}$ are fixed, but unknown. To a statistician, $X_1 + X_2 + \cdots + X_n$ are the observations of the experiment– in this example, independent rolls. Using the values of $X_1 + X_2 + \cdots + X_n$, it is their job to predict $\{p_1, p_2, \ldots, p_6\}$.

Definition 1.2. A **probability space** represents three factors of an experiment, $(\Omega, \mathcal{F}, \mathbb{P})$ where [1]:

- Ω is the sample space, which represents all possibilities of the experiment. In the aforementioned die experiment, the sample space, Ω, is $\{1, 2, 3, 4, 5, 6\}$. If the experiment was instead rolling two dice, Ω has 36 distinct items: $\{(1, 1), (1, 2), \ldots, (6, 6)\}$, and if the experiment was the sum of the two rolls, Ω would be $\{2, 3, \ldots, 12\}$.
- $\mathcal{F}$ represents all the subsets of the sample space, including the empty set and the full set. In the case of rolling a die, there are 2^6 many valid subsets of which we can calculate probabilities of.
- $\mathbb{P}$ represents the probability of each subset in $\mathcal{F}$ occurring. $\mathbb{P}(X = 1)$ is the probability of the value of the random variable being 1, which is $\frac{1}{6}$. Here, the subset within $\mathcal{F}$ we are considering is $\{1\}$.

Thus, utilizing Definitions 1.1 and 1.2, a random variable $X : \Omega \to \mathbb{R}$ (a sample space containing only real numbers) represents an outcome that we don't know, on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Two types of random variables that we will discuss in this paper are **discrete**, with X assuming countable values (such as the die example), and **continuous**, where X assumes values on a continuous spectrum (such as experiments where we observe the heights of people or amounts of rain). A discrete random variable X is a random variable that takes countably many values $\{a_1, a_2, \ldots\}$.

Definition 1.3. The **expectation** $\mathbb{E}$ of a random variable is the mean of all the possible outcomes, with those outcomes weighted based on their probabilities of occurring. Generalized formulas exist for calculating the expectation, however those formulas are not necessary for this exposition. For discrete random variables, expectation is calculated by the following formula:

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} a_i p(a_i).$$

For a fair 6-sided die,

$$\mathbb{E}(X) = \sum_{i=1}^{6} i \cdot p_i = \frac{1 + 2 + 3 + 4 + 5 + 6}{6} = 3.5,$$

where $a_i = i$ and $p(a_i) = \frac{1}{6}$. For continuous random variables, expectation is calculated as:

$$\int_{-\infty}^{\infty} x f(x) dx.$$

This means, if an experiment is repeated n times independently, we can attain an approximation of the 'average' value of the random variable.

Definition 1.4. The **variance** σ^2 of a random variable is how far the possible outcomes are spread out from the expectation, represented by the formula: $\sigma^2 = Var(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \geq 0$. Higher values indicate a more spread out

set from the expectation.

Definition 1.5. For a continuous random variable X, the **probability density function** (PDF) is:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx.$$

Definition 1.6. For a continuous random variable X, the **cumulative density function** (CDF) is:

$$\mathbb{P}(X \leq b) = \int_{-\infty}^b f(x)dx = F(b).$$

There are many classical examples of discrete and continuous random variables. One well-known example is the **normal distribution**, an important mathematical concept due to its observed prevalence among many natural phenomena. Precisely, it is a continuous random variable with the following PDF [4]:

$$\mathbb{P}(a \leq X \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx, \tag{1.7}$$

where variable X has an expectation of 0 and variance of 1. (1.7) was derived since $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} dx = 1$. If X does not have a distribution as mentioned above, where $X \sim N(0,1)$, then we can adjust the function to still yield a normal distribution for $X \sim N(\mu, \sigma^2)$ with

$$\hat{f}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(X-\mu)^2}{2\sigma^2}}.$$

Now, we will understand how random variables interact with each other. From these early definitions, we move onto an important topic: **convergence**. Using the six-sided die example, the more times we roll the die, the more likely it is that the average value of all the rolls tends towards $\mathbb{E}(X) = 3.5$. In other words, the more times the value of random variable X is calculated, the closer the average value of those rolls will approach the expectation.

We say that a sequence of random variables X_n converges in distribution to X when

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x),$$

for all x which are continuity points of the function $F(x) := \mathbb{P}(X \leq x)$. An alternative way to check for convergence in distribution is to compute the **Wasserstein distance** which is defined as

$$Wass(X,Y) = \sup\{|\mathbb{E}[g(X)] - \mathbb{E}[g(Y)]| : g \text{ is 1-Lipschitz}\}, \tag{1.8}$$

where **1-Lipschitz functions** are the set of all functions g satisfying $|g(u) - g(v)| \leq |u - v|$ for all $\{u, v\}$, and the supremum (sup) is the smallest upper bound of a

set. The Wasserstein metric quantifies a distance between two probability distributions, using the notation $Wass(W, Z)$ to denote the quantified closeness in distribution between distribution W and distribution Z. This is significant because it enables us to compare how close a distribution may be to a normal distribution. If $Wass(X_n, X) \to 0$, then a sequence of random variable X converges in distribution to X, in other words: $X_n \xrightarrow{d} X$.

Now, we must understand the concept of **independence**, which is a property among some collections of random variables where the outcome/value of one random variable *does not* affect the outcome/value of another.

Definition 1.9. Given a probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and a collection of random variables $\{X_1, X_2, \ldots, X_n\}$ in that probability space, we say that the collection is independent if

$$\mathbb{P}(X_1 \in A_1, X_2 \in A_2, X_n \in A_n) = \mathbb{P}(X_1 \in A_1) \cdot \mathbb{P}(X_2 \in A_2) \ldots \mathbb{P}(X_n \in A_n),$$

for all sets $A_1, A_2, \cdots A_n$ in $\mathcal{F}$. Furthermore, if all the X_i's have same distribution, they we say they are **IID variables**– independent and identically distributed.

The **Central Limit Theorem** is a specific application of convergence properties discussed above, and states that as a collection of IID variables with the same probability distribution are simulated a very large number of times, the distribution gets closer to a **normal distribution**. Mathematically, if $X_1, X_2, \ldots X_n$ is an IID sequence of random variables with mean μ and variance σ^2 then

$$\frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} \xrightarrow{d} X.$$

For rigorous proofs, refer to *An Intermediate Course in Probability* by Allan Gut[2], for a comprehensive visualization, refer to the *Essence of Probability* series by 3blue1brown[3].

In this section, we have observed how a normal distribution and Central Limit Theorem are results of a convergence in distribution that we can make *if* the collection of the random variables is independent and identically distributed. However, what happens if the random variables that we are concerned with analyzing are not IID when simulated repeatedly? Is there still some way we can identify a meaningful trend towards a normal-like distribution? These are the questions that Stein's method seeks to offer a solution to.

2. Stein's Method

Stein's Lemma presents the following result: A random variable is normally distributed, such as variable $Z \sim N(0, 1)$. Given such a variable,

$$\mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)].$$

The converse also holds, where if the above equation is true for a large class of functions, then Z is normally distributed. Consequently, supposing we are presented with a random variable $Y_n \to Z$, then we would expect $|\mathbb{E}[Y_n f(Y_n)] -$

$\mathbb{E}[f'(Y_n)]|$ to be very small as $n \to \infty$. Therefore, for a bounded function g, if $\mathbb{E}[g(Y_n) - g(Z)] \to 0$, then $Y_n \xrightarrow{d} Z$. This means that even if Y_n is not an IID variable, we can identify key behavioral patterns to help us conclude a certain similarity level in distribution to Z which *is* normally distributed. In (1.8), we observed the Wasserstein distance formula which connects all these ideas together, and will be used to compute calculations in later sections.

Given a function g such that $\mathbb{E}[g(Z)] < \infty$ and $Z \sim N(0,1)$, Stein's lemma gives us a function f that satisfies the equation

$$f'(x) - xf(x) = g(x) - \mathbb{E}[g(Z)]. \tag{2.1}$$

Through a process of integration, we find the function f that is an absolute, continuous solution of the above equation is

$$f(x) = e^{\frac{x^2}{2}} \int_{-\infty}^{x} e^{\frac{-y^2}{2}} (g(y) - \mathbb{E}g(Z))dy. \tag{2.2}$$

Importantly, from this precise description of f, we can deduce further properties of f: $|f(X)| \leq 1, |f'(X)| \leq \sqrt{2}\pi, |f''(X)| \leq 2$, given g is 1-Lipschitz. By increasing this class we obtain the following inequality:

$$Wass(W, Z) \leq sup\{|\mathbb{E}(f'(W) - Wf(W))| : |f| \leq 1, |f'| \leq \sqrt{2}\pi, |f''| \leq 2\}. \tag{2.3}$$

(2.3) will be the starting point in future sections when we calculate results for our examples. When the collection of random variables is IID, we can obtain a quantitative version of the Central Limit Theorem using the following theorem:

Theorem 2.4 (Stein's method for IID samples [5])**.** *Suppose $X_1, X_2, X_3, \ldots$ are IID random variables with a mean of 0, variance of 1, and a finite 3rd moment ($\mathbb{E}(X^3)$ is a finite number). We have*

$$Wass(\frac{\sum_{i=1}^{n} X_i}{\sqrt{n}}, Z) \leq \frac{3}{n^{\frac{3}{2}}} \sum_{i=1}^{n} \mathbb{E}|X_i|^3, \tag{2.5}$$

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

Proof of Theorem 2.4. Let us expand $\mathbb{E}[Wf(W)]$ as follows:

$$\mathbb{E}[Wf(W)] = \frac{1}{\sqrt{n}}(\mathbb{E}[X_i f(W)].$$

$W_i = W - \frac{X_i}{\sqrt{n}}$, which then makes W_i independent of X_i. Then, we reach

$$\mathbb{E}[X_i f(W_i)] = \mathbb{E}[X_i]\mathbb{E}[f(W_i)] = 0.$$

In the next step, we now see that

$$\mathbb{E}(X_i f(W)) = \mathbb{E}(X_i(f(W) - f(W_i))) =$$
$$\mathbb{E}(X_i(f(W) - f(W_i) - (W - W_i)f'(W_i))) + \mathbb{E}[X_i(W - W_i)f'(W_i)].$$

Here, we note that $W - W_i = \frac{X_i}{\sqrt{n}}$. In light of the above equation, and using Taylor expansion, we get

$$|\mathbb{E}[X_i(f(W) - f(W_i) - f'(W_i)\frac{X_i}{\sqrt{n}})]| \leq \frac{\mathbb{E}|X_i|^3}{n}.$$

Using the fact that $\mathbb{E}(X_i)^2 = 1$,

$$\mathbb{E}[X_i(W - W_i)f'(W_i)] = \frac{1}{\sqrt{n}}\mathbb{E}f'(W_i).$$

Using the above two calculations,

$$|\mathbb{E}(W(f(W)) - \frac{1}{n}\sum \mathbb{E}(f'(W_i))| \leq \frac{\mathbb{E}|X_i|^3}{n^{\frac{3}{2}}}.$$

Finally, using the bounds established in (2.3) we also observe that

$$|\frac{1}{n}\sum \mathbb{E}[f'(W_i) - f'(W)]| \leq \frac{|f''|}{n}\sum_{i=1}^{n}\mathbb{E}[W_i - W] \leq \frac{2}{n^{\frac{3}{2}}}\sum_{i=1}^{n}\mathbb{E}|X_i|^3.$$

By combining the above terms, we complete our proof, and thus derive a bound for our Wasserstein distance.

□

In the next two sections, we will apply the concepts discussed in the first section, and the equations derived in this section, to implement Stein's Method across different examples of random variables that model real-world scenarios.

3. Dependency Graphs

W is a large graph with vertices $\{X_1, \ldots, X_n\}$ that may or may not demonstrate a relationship to one another through an edge, where

$$W = \frac{X_1 + \ldots + X_n}{\sqrt{n}}.$$

W_i are the parts of W not influenced by X_i, where $W - W_i$ is thought to be small since in a large field, one vertex X_i may influence others around it, but when considering the total field at large the impact is presumed be small. This can be contextualized to a **population group** where you are attempting to chart out relationships among people. While one individual may be impactful to a group of connections around them, that person has a minimal effect on W, thereby making W_i rather large. Importantly, however, because of this influence that one vertex may have on others in the form of a connected edge, $\{X_1, X_2, \ldots, X_n\}$ are identically distributed but *not* independent.

Theorem 3.1 (Dependency Graph Method [5])**.** *With* $\mathbb{E}(X_i) = 0, \sigma^2 = Var(\sum X_i)$, *$D$ representing the most connected vertex* $(1 + $ *max degree*$)$ *and Z representing a normal distribution:*

$$Wass(W, Z) \leq \frac{4}{\sqrt{\pi}\sigma^2}\sqrt{D^3\sum \mathbb{E}|X_i|^4} + \frac{D^2}{\sigma^3}\sum \mathbb{E}|X_i|^3. \tag{3.2}$$

Proof of Theorem 3.1. Using any function f that is within the bounds established in (2.3),

$$\mathbb{E}[Wf(W)] = \frac{1}{\sigma}\sum_{i=1}^{n}\mathbb{E}[X_i(f(W) - f(W_i))] =$$
$$\frac{1}{\sigma}\sum \mathbb{E}[X_i(f(W) - f(W_i) - (W - W_i)f'(W)]$$
$$+\frac{1}{\sigma}\sum \mathbb{E}[X_i(W - W_i)f'(W)].$$

N_i is considered the neighborhood of dependence, the vertices that are influenced by X_i. If j is not in N_i then X_j is independent of X_i, and $W - W_i$, is equal to $\frac{1}{\sigma}\sum_{j\notin N_i} X_j$. From this, using Lagrangian bounds, we derive

$$\frac{1}{\sigma}\sum \mathbb{E}[X_i(f(W) - f(W_i) - (W - W_i)f'(W)] \leq \frac{1}{\sigma}\sum \frac{1}{2}\mathbb{E}|X_i(W - W_i)^2||f''|$$
$$\leq \frac{1}{\sigma^3}\sum \mathbb{E}|X_i(\sum_{j\in N_i} X_j)^2| \text{ and } \frac{1}{\sigma}\sum \mathbb{E}[X_i(W - W_i)f'(W)] =$$
$$\frac{1}{\sigma}\sum \mathbb{E}X_i(\sum_{j\in N_i} X_j f(W)) = \mathbb{E}(f'(W)[\frac{1}{\sigma^2}\sum X_i(\sum_{j\in N_i} X_j)]).$$

Developing our proof further,

$$T = \frac{1}{\sigma^2}\sum X_i(\sum_{j\in N_i} X_j) = \frac{1}{\sigma}\mathbb{E}\sum X_i(W - W_i) = \frac{1}{\sigma}\mathbb{E}\sum X_i W = \mathbb{E}(W^2) = 1.$$

Reincorporating T back into the equation above,

$$|\frac{1}{\sigma}\sum \mathbb{E}[X_i(W - W_i)f'(W)] - f'(W)| = |\mathbb{E}(f'(W)(T - 1)| \leq$$
$$\sqrt{\frac{2}{\pi}}\sqrt{\mathbb{E}(T-1)^2} = \sqrt{\frac{2}{\pi}}\sqrt{Var(T)}.$$

Now, we will expand T to its long form, and substitute it into the Wasserstein bound equation:

$$|\mathbb{E}(Wf(W) - \mathbb{E}f'(W)| \leq \sqrt{\frac{2}{\pi}}\sqrt{Var(\frac{1}{\sigma^2}\sum X_i(\sum_{j\in N_i} X_j))}$$
$$+\frac{1}{\sigma^3}\sum \mathbb{E}|X_i(\sum_{j\in N_i} X_j)^2| \leq \frac{1}{\sigma^3}\sum_i \sum_{j,k\in N_i} \mathbb{E}|X_i X_j X_k|$$
$$\leq \frac{1}{\sigma^3}\sum_i \sum_{j,k\in N_i} \frac{1}{3}(\mathbb{E}|X_i|^3 + \mathbb{E}|X_i|^3 + \mathbb{E}|X_i|^3) \leq \frac{D^2}{\sigma^3}\sum \mathbb{E}|X_i|^3.$$

Thus, we have determined the second term in the right-hand side of (3.2). When substituting in both the respective terms, our proof is completed. We will now calculate the first term through the following operation:

$$Var(\sum_{i,j\in N_i} X_iX_j) \leq 2D^2 \sum_{i\sim j} Var(X_iX_j) \leq 2D^3 \sum \mathbb{E}(X_i^4).$$

□

We can now apply this concept to a relationship triangle. Consider a graph with n nodes $\{1, 2, 3, \ldots, n\}$ where each edge is present with probability p independently. These are called **Erdos-Renyi random graphs**. We would like to prove the Central Limit Theorem for triangles in this random graph as $n \to \infty$, keeping in mind that the probability of having a triangle with $\{1, 2, 3\}$ vertices and with $\{1, 2, 4\}$ vertices are not independent.

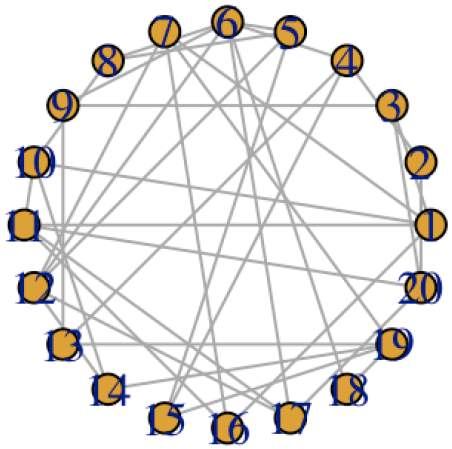

FIGURE 1. Erdos-Renyi random graph [6].

Theorem 3.3 (Triangle Central Limit Theorem)**.** *Let W be the number of triangles in an Erdos-Renyi random graph with probability parameter $p \in (0, 1)$. There exists a constant $C > 0$ depending on p such that*

$$Wass(W, Z) \leq \frac{C}{n}, \tag{3.4}$$

where Z is a standard normal random variable.

Proof of Theorem 3.3. Given edges i and j, $X_{i,j} = 1$ if i and j are connected, and equals 0 if not. Further, $\mathbb{P}(X_{i,j} = 1) = p$ and $\mathbb{P}(X_{i,j} = 0) = 1-p$. The total number of triangles that exist in a given field of edges is thus: $\sum_{1\leq i<j<k\leq n} X_{i,j}X_{j,k}X_{k,i}$. This concept can be used to illustrate a variety of real-world interactions such as friend groups and their potential connections to surrounding people. Because the probability of an edge forming is independent, the expectation $\mathbb{E}[X_{i,j}X_{j,k}X_{k,i}] = p^3$ and the variance is $\mathbb{E}[X_{i,j}X_{j,k}X_{k,i}] - (\mathbb{E}[X_{i,j}X_{j,k}X_{k,i}])^2 = p^3(1 - p^3)$.

Given a triangle (edges i, j, k are all connected to each other), the maximum number of triangles that share an edge to that triangle is $3(n - 3)$ which is our D. Consider an example of 4 vertices in which 3 edges connect with each other to

form a triangle. Then, the 1 remaining vertex can at maximum connect with the 3 other vertices, forming 3 guaranteed triangles given we know the other vertices are already connected to each other. Thus, our D value here is $3(4-3)=3$.

Standardizing the distribution of the scenario above gives us

$$W = \frac{\sum_{1\le i<j<k\le n} X_{\tilde{i}\tilde{j}} X_{\tilde{j}\tilde{k}} X_{\tilde{k}\tilde{i}}}{\sigma},$$

where $X_{\tilde{i}\tilde{j}} X_{\tilde{j}\tilde{k}} X_{\tilde{k}\tilde{i}} = X_{i,j} X_{j,k} X_{k,i} - p^3$ in order to center the expectation. The Wasserstein distance between W and Z is thus

(3.5)

$$Wass(W,Z) \le \frac{4}{\sqrt{\pi}} \frac{\sqrt{D^3 \sum_{i,j,k} \mathbb{E}[|X_{\tilde{i}\tilde{j}} X_{\tilde{j}\tilde{k}} X_{\tilde{k}\tilde{i}}|^4]}}{\sigma^2} + \frac{D^2}{\sigma^3} \sum_{1\le i<j<k\le n} \mathbb{E}[|X_{\tilde{i}\tilde{j}} X_{\tilde{j}\tilde{k}} X_{\tilde{k}\tilde{i}}|^3].$$

To simplify the inequality, we can calculate the order of each variable, since that will help us establish the degree of significance that our Wasserstein distance is less than or equal to, giving us a good idea of the distribution's proximity to a normal distribution even if we do not arrive at an exact answer. D is to the order of n; as mentioned earlier $D = 3(n-3)$ while σ is to the order of n^2. The order of expectation $\mathbb{E}$ is derived from the sum of $\binom{n}{3}$ many random variables, which equals $\frac{n(n-1)(n-2)}{6}$ and thus has a power of n^3. Simplifying the above expression, both terms amount to the order of $\frac{C}{n}$ where C is some constant. This completes the proof.

□

The above theorem illustrates that even though the dependency graph question may involve neighborhoods that make the variable not *necessarily* IID, the above method allows to understand how close an observable of interest is to a normal distribution. In fact, with further computation one can even find $Wass(W,Z) \le \frac{C_0}{np^{\frac{9}{2}}}$ where the constant C_0 can be chosen free of p.

4. Exchangeable Pairs

One limitation of the dependency graph is that if a random variable is dependent on a large collection of other random variables (if the D value is large), the method is no longer suitable to use, given that the nature of the problem we considered above is when the effect of one vertex on the total population group is, while noticeable and thus not independent, *minimal*.

There are real-world instances, as we will see below, where each individual vertex is dependent on every other vertex, yet the dependency itself, while there, is minimal. Essentially, while dependency graphs focus on a minimal number of strong relationships, the **exchangeable pairs** method focuses on a large quantity of weaker relationships. Given how weak those relationships may be, we still can prove Central Limit Theorem.

Theorem 4.1 (Exchangeable Pairs Method [5]). *W' is a random variable with the same distribution as W. Suppose*

$$\mathbb{E}[W] = 0,$$
$$\mathbb{E}[W^2] = Var(W) = 1,$$
$$\mathbb{E}[W' - W|W] = -\lambda W,$$

where λ is some small positive number. We then have the following bound on the Wasserstein distance:

$$Wass(W, Z) \leq \sqrt{\frac{2}{\pi} Var(\frac{1}{2\lambda}\mathbb{E}[W' - W|W])} + \frac{1}{3\lambda}\mathbb{E}|W' - W|^3, \tag{4.2}$$

where Z again denotes a standard normal random variable.

Proof of Theorem 4.1. Using our above definitions of W and W',

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

Instead of working with f, we can work with $F(x) = \int_0^x f(y)dy$ so that $F' = f$. Then,

$$0 = \mathbb{E}[(W' - W)f'(W)] + \frac{1}{2}\mathbb{E}[(W -' W)^2 f''(W)] + \mathbb{E}[remainder].$$

Utilizing Taylor Series properties given the bounds established in (2.3),

$$|remainder| \leq \frac{1}{6}|W - W'|^3|f''| \leq \frac{1}{3}|W - W'|^3.$$

The calculation of the first-term of (4.2) is,

$$-\lambda\mathbb{E}[Wf(W)] = \mathbb{E}[(W' - W)f(W)] = -\mathbb{E}[\frac{1}{2}(W - W')^2 f'(W) + remainder]$$
$$= -\mathbb{E}[\frac{1}{2}\mathbb{E}[(W - W')^2|W]f'(W)] + \mathbb{E}[remainder].$$

In the final step, we combine the results to derive an upper bound for the Wasserstein distance, calculated as

$$\mathbb{E}|[Wf(W) - f'(W)]| = |\mathbb{E}[f'(W)(\mathbb{E}(\frac{(W - W')^2|W}{2\lambda} - 1))]| + \frac{1}{3\lambda}\mathbb{E}|W - W'|^3 \leq$$
$$\sqrt{\frac{2}{\pi} Var(\mathbb{E}[|\frac{(W - W')^2|W}{2\lambda} - 1|])} + \frac{1}{3\lambda}\mathbb{E}[|W - W'|^3].$$

□

We will now try to understand how to apply this method to an **Ising Model**. Let us envision a scenario where we have a magnetic charge consisting of n spins

$X_1, \ldots, X_n$ that are either a $+1$ or -1. We consider the following probability distribution on the spins:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \propto \exp(\sum_{i=1}^{n-1} x_i x_{i+1}).$$

The above probability distribution is quite natural in the sense that similar charges attract one another, meaning that there is a higher probability that $x_i x_{i+1} = 1$ since both $+1$ and -1 multiplied by itself is 1. This is why the variable we are concerned with is not IID; the positive or negative tilt of a magnetic charge affects the next one and so on so forth. However, Stein's method is still applicable for the average spin under this setup.

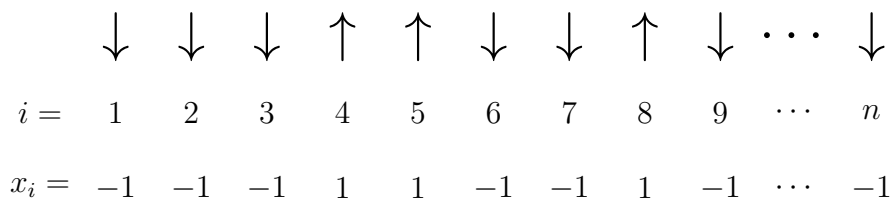

FIGURE 2. 1 dimension Ising Model where the influence of x_i on x_{i+1} is determined by some probability p [7].

Theorem 4.3 (1D Ising model)**.** *Consider the above Ising model. We then have*

$$Wass(\frac{X_1 + X_2 + \cdots + X_n}{\sigma}, Z) \leq \frac{8}{3\sqrt{n}},$$

where $\sigma^2 = Var(X_1 + X_2 + \cdots + X_n)$ *and* Z *is a standard normal random variable.*

Proof of Theorem 4.3. Given the properties of W and W' discussed earlier, $W = \frac{1}{\sigma}\sum_{i=1}^{n} X_i$ and $\sigma^2 = Var(\sum_{i=1}^{n} X_i)$. $W' - W = \frac{X_I' - X_I}{\sigma}$ where I is uniformly picked from $1, \ldots, n$. Taking the expectation with respect to I,

$$\begin{aligned}&\frac{X_1' - X_1}{\sigma} * \frac{1}{n} + \frac{X_2' - X_2}{\sigma} * \frac{1}{n} + \ldots + \frac{X_n' - X_n}{\sigma} * \frac{1}{n} \\ &= \frac{1}{n\sigma}(\sum_{i=1}^{n} X_i' - \sum_{i=1}^{n} X_i) = \frac{1}{n\sigma}\sum_{i=1}^{n} X_i' - \frac{W}{n}.\end{aligned}$$

We now can arrive to a precise definition for λ by calculating that

$$\mathbb{E}[W' - W|W] = \frac{1}{n\sigma}\sum_{i=1}^{n} \mathbb{E}[X_i] - \frac{W}{n} = -\frac{W}{n}.$$

Thus, using the earlier equation $\mathbb{E}[W' - W|W] = -\lambda W$, $\lambda = \frac{1}{n}$. Calculating σ^2:

$$\sigma^2 = Var(\sum X_i) = \mathbb{E}((\sum^{n} X_i)^2) = \mathbb{E}[\sum_{i=1}^{n}\sum_{j=1}^{n} X_i X_j] =$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[X_i X_j] = \sum_{i=1}^{n}\mathbb{E}[X_i^2] + \sum\sum_{i\neq j}\mathbb{E}[X_i X_j] \geq n.$$

Now that we have reached a value for λ and for σ^2 that only involve n which designates the number of trials we are running of said random variable, we can plug in those values into the original Wasserstein equation.

$$\frac{1}{3\lambda}\mathbb{E}[|W - W'|^3] = \frac{n}{3} * \frac{\mathbb{E}|X_I - X_I'|^3}{\sigma^3} \leq \frac{8n}{3n^{\frac{3}{2}}} \leq \frac{8}{3\sqrt{n}}.$$

Now we will figure out the first term of (4.2) by calculating the value of the expectation, where

$$\mathbb{E}[(W' - W)^2|W] = \frac{1}{\sigma^2}\mathbb{E}[(X_I' - X_I)^2|W] = \frac{1}{n\sigma^2}\mathbb{E}[\sum_{i=1}^{n} X_i' - X_i)^2|W] =$$

$$\frac{1}{n\sigma}\sum_{i=1}^{n}\mathbb{E}[X_i^2 + X_i^2 - 2X_i' X_i|W] = 2 - 2\mathbb{E}[X_i' X_i|W] = \frac{2}{\sigma}.$$

Thus, $E[\frac{1}{2\lambda}(W'-W)^2|W] = \frac{n}{\sigma}$. When plugging this in to the original Wasserstein equation, the variance of $\frac{n}{\sigma}$ is 0, and the full term that comes under the square root gets eliminated. Therefore, $Wass(W, Z) \leq \frac{8}{3\sqrt{n}}$, which decreases in value as the number of experiments n increases. Therefore, as more trials of this random variable are conducted, the random variable comes closer and closer to a normal distribution, thereby validating initial beliefs of W being close to normal despite not being an independent, identically distributed variable.

□

Acknowledgments

I would like to sincerely thank Professor Sayan Das for his help in introducing me to this paper's topic and instructing me in detail, including teaching me about certain related topics that venture beyond the scope of this paper. I also thank Professor Peter May for organizing this REU, through which I have gained an appreciation for the processes involved in conducting mathematical research.

References

[1] Probability Space. https://en.wikipedia.org/wiki/Probability_space
[2] Allan Gut. An Intermediate Course in Probability, 2009
[3] 3blue1brown. Essence of Probability. https://www.3blue1brown.com/topics/probability
[4] Grant Sanderson. Gaussian Integral, 2023. https://www.3blue1brown.com/lessons/gaussian-integral
[5] Sourav Chatterjee. Stein's method and its applications, 2007. https://souravchatterjee.su.domains/AllLectures.pdf
[6] Dai Shizuka. Random Graphs, 2019. https://dshizuka.github.io/networkanalysis/09_randomnets.html
[7] Kathleen McNamara. Ising Model of a Ferromagnet, 2014. https://www.slideserve.com/lula/ising-model-of-a-ferromagnet

[8] Mathematical Expressions in LaTeX. https://www.overleaf.com/learn/latex/Mathematical_expressions