Stein's Method:

The last gadget under the hood

Elizabeth Meckes

Case Western Reserve University

LDHD Summer School
SAMSI
August, 2013

What is Stein’s method, and what is it good for?

What is Stein's method, and what is it good for?

"Stein's method" refers to a family of techniques for approximating the distribution of a random variable you want to understand by some model distribution that you already understand (normal, Poisson, gamma, semi-circle, etc.)

What is Stein's method and what is it good for?

- The method has no *a priori* requirements for any particular structure of the random variable (e.g., it need not be a sum), or for any independence. This makes it often useful in geometric or topological problems.

What is Stein's method and what is it good for?

- The method has no *a priori* requirements for any particular structure of the random variable (e.g., it need not be a sum), or for any independence. This makes it often useful in geometric or topological problems.

- It is a non-asymptotic method: when used to prove limit theorems, it automatically produces rates of convergence.

What is Stein's method and what is it good for?

- The method has no *a priori* requirements for any particular structure of the random variable (e.g., it need not be a sum), or for any independence. This makes it often useful in geometric or topological problems.
- It is a non-asymptotic method: when used to prove limit theorems, it automatically produces rates of convergence.
- It is quite robust: one can often handle conditions almost being satisfied, but not exactly.

What is Stein's method and what is it good for?

- The method has no *a priori* requirements for any particular structure of the random variable (e.g., it need not be a sum), or for any independence. This makes it often useful in geometric or topological problems.
- It is a non-asymptotic method: when used to prove limit theorems, it automatically produces rates of convergence.
- It is quite robust: one can often handle conditions almost being satisfied, but not exactly.
- It's most useful when you already have a guess as to a good approximating distribution for your random variable, although this is not an absolute requirement.

The Characterizing Operator

The Characterizing Operator

Let X be a random variable. A characterizing operator for X is an operator T_o on some class of functions $\mathcal{A}$, such that, for any random variable Y,

$$\mathbb{E} T_o f(Y) = 0 \quad \forall f \in \mathcal{A} \qquad \text{iff} \qquad Y \stackrel{d}{=} X.$$

The Characterizing Operator

Let X be a random variable. A characterizing operator for X is an operator T_o on some class of functions $\mathcal{A}$, such that, for any random variable Y,

$$\mathbb{E}T_o f(Y) = 0 \quad \forall f \in \mathcal{A} \qquad \text{iff} \qquad Y \stackrel{d}{=} X.$$

Examples:

The Characterizing Operator

Let X be a random variable. A characterizing operator for X is an operator T_o on some class of functions $\mathcal{A}$, such that, for any random variable Y,

$$\mathbb{E}T_o f(Y) = 0 \quad \forall f \in \mathcal{A} \qquad \text{iff} \qquad Y \overset{d}{=} X.$$

Examples:

- Standard Normal: $T_o f(x) = f'(x) - xf(x)$ for $f : \mathbb{R} \to \mathbb{R}$.

The Characterizing Operator

Let X be a random variable. A characterizing operator for X is an operator T_o on some class of functions $\mathcal{A}$, such that, for any random variable Y,

$$\mathbb{E}T_o f(Y) = 0 \quad \forall f \in \mathcal{A} \qquad \text{iff} \qquad Y \stackrel{d}{=} X.$$

Examples:

- Standard Normal: $T_o f(x) = f'(x) - xf(x)$ for $f : \mathbb{R} \to \mathbb{R}$.
- Poisson(λ): $T_o f(j) = \lambda f(j+1) - jf(j)$ for $f : \mathbb{N} \to \mathbb{R}$.

The Characterizing Operator

Let X be a random variable. A characterizing operator for X is an operator T_o on some class of functions $\mathcal{A}$, such that, for any random variable Y,

$$\mathbb{E} T_o f(Y) = 0 \quad \forall f \in \mathcal{A} \qquad \text{iff} \qquad Y \stackrel{d}{=} X.$$

Examples:

- Standard Normal: $T_o f(x) = f'(x) - x f(x)$ for $f : \mathbb{R} \to \mathbb{R}$.
- Poisson(λ): $T_o f(j) = \lambda f(j+1) - j f(j)$ for $f : \mathbb{N} \to \mathbb{R}$.
- Exponential(λ): $T_o f(x) = f'(x) - \lambda f(x)$ for $f : \mathbb{R}^+ \to \mathbb{R}$ with $f(0) = 0$.

Approximation

Approximation

Suppose Y is the random variable you care about, and X is a random variable with characterizing operator T_o which you think is a good approximation of Y.

Approximation

Suppose Y is the random variable you care about, and X is a random variable with characterizing operator T_o which you think is a good approximation of Y.

The Big Idea:

> *Instead of trying to show that $\mathbb{E}T_o f(Y) = 0$ for all $f \in \mathcal{A}$, (which is probably not true), try to show that $\mathbb{E}T_o f(Y)$ is small for all $f \in \mathcal{A}$. This will imply that Y is close to X in some sense.*

Implementing the Big Idea: The Stein Equation

We need to solve the Stein equation: given a function g, find f such that

$$T_o f(x) = g(x) - \mathbb{E}g(X).$$

We use U_o to denote the operator that gives the solution of the Stein equation:

$$f(x) = U_o g(x).$$

Implementing the Big Idea: The Stein Equation

We need to solve the Stein equation: given a function g, find f such that

$$T_o f(x) = g(x) - \mathbb{E}g(X).$$

We use U_o to denote the operator that gives the solution of the Stein equation:

$$f(x) = U_o g(x).$$

If $f = U_o g$, observe that

$$\mathbb{E}T_o f(Y) = \mathbb{E}g(Y) - \mathbb{E}g(X).$$

Implementing the Big Idea: The Stein Equation

We need to solve the Stein equation: given a function g, find f such that

$$T_o f(x) = g(x) - \mathbb{E}g(X).$$

We use U_o to denote the operator that gives the solution of the Stein equation:

$$f(x) = U_o g(x).$$

If $f = U_o g$, observe that

$$\mathbb{E}T_o f(Y) = \mathbb{E}g(Y) - \mathbb{E}g(X).$$

Thus if $\mathbb{E}T_o f(Y)$ is small, then $\mathbb{E}g(Y) - \mathbb{E}g(X)$ is small.

This leads naturally to notions of distance between the random variables X and Y which can be expressed in the form

$$d(X, Y) = \sup_{\mathcal{F}} \left| \mathbb{E}g(X) - \mathbb{E}g(Y) \right|,$$

where the supremum is over some class $\mathcal{F}$ of test functions g.

This leads naturally to notions of distance between the random variables X and Y which can be expressed in the form

$$d(X,Y) = \sup_{\mathcal{F}} \left| \mathbb{E}g(X) - \mathbb{E}g(Y) \right|,$$

where the supremum is over some class $\mathcal{F}$ of test functions g.

Examples:

This leads naturally to notions of distance between the random variables X and Y which can be expressed in the form

$$d(X, Y) = \sup_{\mathcal{F}} \left| \mathbb{E}g(X) - \mathbb{E}g(Y) \right|,$$

where the supremum is over some class $\mathcal{F}$ of test functions g.

Examples:

- $\mathcal{F} = \{f : \|f\|_\infty \leq 1, continuous\}$ $\longleftrightarrow$ total variation distance.

This leads naturally to notions of distance between the random variables X and Y which can be expressed in the form

$$d(X,Y) = \sup_{\mathcal{F}} \left| \mathbb{E}g(X) - \mathbb{E}g(Y) \right|,$$

where the supremum is over some class $\mathcal{F}$ of test functions g.

Examples:

- $\mathcal{F} = \{f : \|f\|_\infty \leq 1, \textit{continuous}\}$ $\longleftrightarrow$ total variation distance.
- $\mathcal{F} = \{f : \|f'\|_\infty \leq 1\}$ $\longleftrightarrow$ Wasserstein distance.

This leads naturally to notions of distance between the random variables X and Y which can be expressed in the form

$$d(X, Y) = \sup_{\mathcal{F}} \left| \mathbb{E}g(X) - \mathbb{E}g(Y) \right|,$$

where the supremum is over some class $\mathcal{F}$ of test functions g.

Examples:

- $\mathcal{F} = \{f : \|f\|_\infty \leq 1, \mathit{continuous}\}$ $\longleftrightarrow$ total variation distance.
- $\mathcal{F} = \{f : \|f'\|_\infty \leq 1\}$ $\longleftrightarrow$ Wasserstein distance.
- $\mathcal{F} = \{f : \|f\|_\infty + \|f'\|_\infty \leq 1\}$ $\longleftrightarrow$ bounded Lipschitz distance.

So What?

So What?

Instead of trying to estimate the distance between X and Y directly, the problem has been reduced to trying to estimate $\mathbb{E}T_o f(Y)$ for some large class of functions f. Why is this any better?

So What?

Instead of trying to estimate the distance between X and Y directly, the problem has been reduced to trying to estimate $\mathbb{E}T_o f(Y)$ for some large class of functions f. Why is this any better?

Various techniques are in use for trying to estimate $\mathbb{E}T_o f(Y)$. Among them:

So What?

Instead of trying to estimate the distance between X and Y directly, the problem has been reduced to trying to estimate $\mathbb{E}T_o f(Y)$ for some large class of functions f. Why is this any better?

Various techniques are in use for trying to estimate $\mathbb{E}T_o f(Y)$. Among them:

- The method of exchangeable pairs (e.g. Stein's book)

So What?

Instead of trying to estimate the distance between X and Y directly, the problem has been reduced to trying to estimate $\mathbb{E}T_o f(Y)$ for some large class of functions f. Why is this any better?

Various techniques are in use for trying to estimate $\mathbb{E}T_o f(Y)$. Among them:

- The method of exchangeable pairs (e.g. Stein's book)
- The dependency graph method (e.g. Arratia, Goldstein, and Gordon or Barbour, Karoński, and Ruciński)

So What?

Instead of trying to estimate the distance between X and Y directly, the problem has been reduced to trying to estimate $\mathbb{E}T_o f(Y)$ for some large class of functions f. Why is this any better?

Various techniques are in use for trying to estimate $\mathbb{E}T_o f(Y)$. Among them:

- The method of exchangeable pairs (e.g. Stein's book)
- The dependency graph method (e.g. Arratia, Goldstein, and Gordon or Barbour, Karoński, and Ruciński)
- Size-bias coupling (e.g. Goldstein and Rinott)

So What?

Instead of trying to estimate the distance between X and Y directly, the problem has been reduced to trying to estimate $\mathbb{E}T_o f(Y)$ for some large class of functions f. Why is this any better?

Various techniques are in use for trying to estimate $\mathbb{E}T_o f(Y)$. Among them:

- The method of exchangeable pairs (e.g. Stein's book)
- The dependency graph method (e.g. Arratia, Goldstein, and Gordon or Barbour, Karoński, and Ruciński)
- Size-bias coupling (e.g. Goldstein and Rinott)
- Zero-bias coupling (e.g. Goldstein and Reinert)

So What?

Instead of trying to estimate the distance between X and Y directly, the problem has been reduced to trying to estimate $\mathbb{E}T_o f(Y)$ for some large class of functions f. Why is this any better?

Various techniques are in use for trying to estimate $\mathbb{E}T_o f(Y)$. Among them:

- The method of exchangeable pairs (e.g. Stein's book)
- The dependency graph method (e.g. Arratia, Goldstein, and Gordon or Barbour, Karoński, and Ruciński)
- Size-bias coupling (e.g. Goldstein and Rinott)
- Zero-bias coupling (e.g. Goldstein and Reinert)
- The generator method (Barbour)

The idea of the method of exchangeable pairs

The idea of the method of exchangeable pairs

- Suppose you have a random variable W which you conjecture is well-approximated by X. Make a "small random change" to W to get a new random variable W', such that $(W, W') \stackrel{d}{=} (W', W)$.

The idea of the method of exchangeable pairs

- Suppose you have a random variable W which you conjecture is well-approximated by X. Make a "small random change" to W to get a new random variable W', such that $(W, W') \stackrel{d}{=} (W', W)$.

- The goal is to bound $\left|\mathbb{E} T_o f(W)\right|$. Many characterizing operators T_o are defined using derivatives or differences. Use the fact that W and W' are close to express or approximate those derivatives or differences in terms of (W, W').

The idea of the method of exchangeable pairs

- Suppose you have a random variable W which you conjecture is well-approximated by X. Make a "small random change" to W to get a new random variable W', such that $(W, W') \stackrel{d}{=} (W', W)$.

- The goal is to bound $\left|\mathbb{E} T_o f(W)\right|$. Many characterizing operators T_o are defined using derivatives or differences. Use the fact that W and W' are close to express or approximate those derivatives or differences in terms of (W, W').

- Use the fact that W' was constructed explicitly from W together with the nesting property of conditional expectation to help evaluate/estimate the resulting espression.

Exchangeable pairs for normal approximation

Fix h and let $f = U_o h$; in other words,

$$T_o f(x) = h(x) - \mathbb{E}h(Z),$$

where Z is a standard normal random variable.

Suppose (W, W') is exchangeable.

Exchangeable pairs for normal approximation

Fix h and let $f = U_o h$; in other words,

$$T_o f(x) = h(x) - \mathbb{E}h(Z),$$

where Z is a standard normal random variable.

Suppose (W, W') is exchangeable. Then

$$0 = \mathbb{E}\left[(W' - W)\big(f(W') + f(W)\big)\right]$$

Exchangeable pairs for normal approximation

Fix h and let $f = U_o h$; in other words,

$$T_o f(x) = h(x) - \mathbb{E}h(Z),$$

where Z is a standard normal random variable.

Suppose (W, W') is exchangeable. Then

$$\begin{aligned} 0 &= \mathbb{E}\left[(W' - W)\big(f(W') + f(W)\big)\right] \\ &= \mathbb{E}\left[(W' - W)\big(f(W') - f(W)\big) + 2(W' - W)f(W)\right] \end{aligned}$$

Exchangeable pairs for normal approximation

Fix h and let $f = U_o h$; in other words,

$$T_o f(x) = h(x) - \mathbb{E}h(Z),$$

where Z is a standard normal random variable.

Suppose (W, W') is exchangeable. Then

$$\begin{aligned}
0 &= \mathbb{E}\left[(W' - W)\big(f(W') + f(W)\big)\right] \\
&= \mathbb{E}\left[(W' - W)\big(f(W') - f(W)\big) + 2(W' - W)f(W)\right] \\
&= \mathbb{E}\Big[(W' - W)^2 f'(W) + 2(W' - W)f(W) + R\Big]
\end{aligned}$$

Exchangeable pairs for normal approximation

Fix h and let $f = U_o h$; in other words,

$$T_o f(x) = h(x) - \mathbb{E}h(Z),$$

where Z is a standard normal random variable.

Suppose (W, W') is exchangeable. Then

$$\begin{aligned}
0 &= \mathbb{E}\left[(W' - W)\big(f(W') + f(W)\big)\right] \\
&= \mathbb{E}\left[(W' - W)\big(f(W') - f(W)\big) + 2(W' - W)f(W)\right] \\
&= \mathbb{E}\left[(W' - W)^2 f'(W) + 2(W' - W)f(W) + R\right] \\
&= \mathbb{E}\left[f'(W)\mathbb{E}\left[(W' - W)^2 \middle| W\right] + 2f(W)\mathbb{E}\left[W' - W \middle| W\right] + R\right].
\end{aligned}$$

$$\mathbb{E}\left[f'(W)\mathbb{E}\left[(W'-W)^2\middle|W\right]+2f(W)\mathbb{E}\left[W'-W\middle|W\right]+R\right]=0$$

$$\mathbb{E}\left[f'(W)\mathbb{E}\left[(W'-W)^2\middle|W\right]+2f(W)\mathbb{E}\left[W'-W\middle|W\right]+R\right]=0$$

Now, suppose that there is a $\lambda \in (0,1)$ such that

$$\mathbb{E}\left[f'(W)\mathbb{E}\left[(W'-W)^2\middle|W\right]+2f(W)\mathbb{E}\left[W'-W\middle|W\right]+R\right]=0$$

Now, suppose that there is a $\lambda \in (0,1)$ such that

- $\mathbb{E}\left[W'-W\middle|W\right] = -\lambda W$

$$\mathbb{E}\left[f'(W)\mathbb{E}\left[(W'-W)^2\middle|W\right]+2f(W)\mathbb{E}\left[W'-W\middle|W\right]+R\right]=0$$

Now, suppose that there is a $\lambda \in (0,1)$ such that

- $\mathbb{E}\left[W'-W\middle|W\right]=-\lambda W$
- $\mathbb{E}\left[(W'-W)^2\middle|W\right]=2\lambda+E.$ (E is a random variable.)

$$\mathbb{E}\left[f'(W)\mathbb{E}\left[(W'-W)^2\middle|W\right]+2f(W)\mathbb{E}\left[W'-W\middle|W\right]+R\right]=0$$

Now, suppose that there is a $\lambda \in (0,1)$ such that

- $\mathbb{E}\left[W'-W\middle|W\right]=-\lambda W$
- $\mathbb{E}\left[(W'-W)^2\middle|W\right]=2\lambda+E.$ (E is a random variable.)

Then

$$2\lambda\mathbb{E}\left[f'(W)-Wf(W)+\frac{f'(W)E+R}{2\lambda}\right]=0.$$

$$\mathbb{E}\left[f'(W)\mathbb{E}\left[(W'-W)^2\middle|W\right]+2f(W)\mathbb{E}\left[W'-W\middle|W\right]+R\right]=0$$

Now, suppose that there is a $\lambda \in (0,1)$ such that

- $\mathbb{E}\left[W'-W\middle|W\right]=-\lambda W$
- $\mathbb{E}\left[(W'-W)^2\middle|W\right]=2\lambda+E.$ (E is a random variable.)

Then

$$2\lambda\mathbb{E}\Bigg[\underbrace{f'(W)-Wf(W)}_{T_o f(W)}+\frac{f'(W)E+R}{2\lambda}\Bigg]=0.$$

$$\mathbb{E}\left[f'(W)\mathbb{E}\left[(W'-W)^2\middle|W\right]+2f(W)\mathbb{E}\left[W'-W\middle|W\right]+R\right]=0$$

Now, suppose that there is a $\lambda \in (0,1)$ such that

- $\mathbb{E}\left[W'-W\middle|W\right]=-\lambda W$
- $\mathbb{E}\left[(W'-W)^2\middle|W\right]=2\lambda+E.$ (E is a random variable.)

Then

$$2\lambda\mathbb{E}\Bigg[\underbrace{f'(W)-Wf(W)}_{T_o f(W)}+\frac{f'(W)E+R}{2\lambda}\Bigg]=0.$$

That is, $\mathbb{E}T_o f(W)=\mathbb{E}h(W)-\mathbb{E}h(Z)=-\frac{1}{2\lambda}\mathbb{E}\left[f'(W)E+R\right]$.

Stein's abstract normal approximation theorem

Stein's abstract normal approximation theorem

Theorem (Stein)

Let (W, W') *be an exchangeable pair of random variables with* $\mathbb{E}W^2 = 1$ *and*

$$\mathbb{E}\left[W' - W \middle| W\right] = -\lambda W$$

for some $\lambda \in (0, 1)$*. Let* $\Delta = W' - W$*,. Then for* Z *a standard normal random variable,*

$$d_{BL}(W, Z) \leq \frac{2}{\lambda}\sqrt{\mathrm{Var}\left(\mathbb{E}\left[\Delta^2 \middle| W\right]\right)} + \frac{1}{2\lambda}\mathbb{E}|\Delta|^3.$$

An infinitesimal version

An infinitesimal version

Theorem (M)

Suppose that (W, W_ϵ) *is a family of exchangeable pairs defined on a common probability space, such that* $\mathbb{E}W = 0$ *and* $\mathbb{E}W^2 = \sigma^2$.

An infinitesimal version

Theorem (M)

Suppose that (W, W_ϵ) *is a family of exchangeable pairs defined on a common probability space, such that* $\mathbb{E}W = 0$ *and* $\mathbb{E}W^2 = \sigma^2$.

Suppose there is a function $\lambda(\epsilon)$ *and random variables* E, E' *such that*

An infinitesimal version

Theorem (M)

Suppose that (W, W_ϵ) *is a family of exchangeable pairs defined on a common probability space, such that* $\mathbb{E}W = 0$ *and* $\mathbb{E}W^2 = \sigma^2$.

Suppose there is a function $\lambda(\epsilon)$ *and random variables* E, E' *such that*

1. $\frac{1}{\lambda(\epsilon)}\mathbb{E}\left[W_\epsilon - W \middle| W\right] \xrightarrow[\epsilon \to 0]{L_1} -W + E'$.

An infinitesimal version

Theorem (M)

Suppose that (W, W_ϵ) *is a family of exchangeable pairs defined on a common probability space, such that* $\mathbb{E}W = 0$ *and* $\mathbb{E}W^2 = \sigma^2$.

Suppose there is a function $\lambda(\epsilon)$ *and random variables* E, E' *such that*

1. $\frac{1}{\lambda(\epsilon)}\mathbb{E}\left[W_\epsilon - W \middle| W\right] \xrightarrow[\epsilon \to 0]{L_1} -W + E'$.
2. $\frac{1}{2\lambda(\epsilon)\sigma^2}\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right] \xrightarrow[\epsilon \to 0]{L_1} 1 + E$.

An infinitesimal version

Theorem (M)

Suppose that (W, W_ϵ) *is a family of exchangeable pairs defined on a common probability space, such that* $\mathbb{E}W = 0$ *and* $\mathbb{E}W^2 = \sigma^2$.

Suppose there is a function $\lambda(\epsilon)$ *and random variables* E, E' *such that*

1. $\frac{1}{\lambda(\epsilon)}\mathbb{E}\left[W_\epsilon - W \middle| W\right] \xrightarrow[\epsilon \to 0]{L_1} -W + E'$.
2. $\frac{1}{2\lambda(\epsilon)\sigma^2}\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right] \xrightarrow[\epsilon \to 0]{L_1} 1 + E$.
3. $\frac{1}{\lambda(\epsilon)}\mathbb{E}|W_\epsilon - W|^3 \xrightarrow{\epsilon \to 0} 0$.

An infinitesimal version

Theorem (M)

Suppose that (W, W_ϵ) *is a family of exchangeable pairs defined on a common probability space, such that* $\mathbb{E}W = 0$ *and* $\mathbb{E}W^2 = \sigma^2$.

Suppose there is a function $\lambda(\epsilon)$ *and random variables* E, E' *such that*

1. $\frac{1}{\lambda(\epsilon)}\mathbb{E}\left[W_\epsilon - W \middle| W\right] \xrightarrow[\epsilon\to 0]{L_1} -W + E'$.
2. $\frac{1}{2\lambda(\epsilon)\sigma^2}\mathbb{E}\left[(W_\epsilon - W)^2 \middle| W\right] \xrightarrow[\epsilon\to 0]{L_1} 1 + E$.
3. $\frac{1}{\lambda(\epsilon)}\mathbb{E}|W_\epsilon - W|^3 \xrightarrow{\epsilon\to 0} 0$.

Then if Z *is a standard normal random variable,*

$$d_{TV}(W, Z) \leq \mathbb{E}|E| + \sqrt{\frac{\pi}{2}}\mathbb{E}|E'|.$$

A now familiar example:

Rank 1 projection of Haar measure on $\mathbb{O}(n)$

A now familiar example:

Rank 1 projection of Haar measure on $\mathbb{O}(n)$

Theorem (M)

Let $M \in \mathbb{O}(n)$ be a random orthogonal matrix.

Let $A \in \mathbb{O}(n)$ be a fixed orthogonal matrix with $\|A\|_{HS} = 1$.

Define the random variable W by

$$W := \mathrm{Tr}(AM).$$

If Z is a standard normal random variable, then

$$d_{TV}(W, Z) \leq \frac{2\sqrt{3}}{n-1}.$$

The exchangeable pair

(first used by Charles Stein)

The exchangeable pair

(first used by Charles Stein)

- Fix $\epsilon > 0$, and let $A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus \mathrm{I}_{n-2}$.

The exchangeable pair

(first used by Charles Stein)

- Fix $\epsilon > 0$, and let $A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus \mathrm{I}_{n-2}$.
- Let U be distributed according to Haar measure on $\mathbb{O}(n)$, and independent of M.

The exchangeable pair

(first used by Charles Stein)

- Fix $\epsilon > 0$, and let $A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus \mathrm{I}_{n-2}$.
- Let U be distributed according to Haar measure on $\mathbb{O}(n)$, and independent of M.

  The matrix $UA_\epsilon U^T$ is a rotation by $\arcsin(\epsilon)$ in a random two-dimensional subspace of $\mathbb{R}^n$.

The exchangeable pair

(first used by Charles Stein)

- Fix $\epsilon > 0$, and let $A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus \mathrm{I}_{n-2}$.
- Let U be distributed according to Haar measure on $\mathbb{O}(n)$, and independent of M.

  The matrix $UA_\epsilon U^T$ is a rotation by $\arcsin(\epsilon)$ in a random two-dimensional subspace of $\mathbb{R}^n$.
- Make an exchangeable pair of random matrices (M, M_ϵ) by randomly rotating M:

$$M_\epsilon := UA_\epsilon U^T M.$$

The exchangeable pair

(first used by Charles Stein)

- Fix $\epsilon > 0$, and let $A_\epsilon = \begin{bmatrix} \sqrt{1-\epsilon^2} & \epsilon \\ -\epsilon & \sqrt{1-\epsilon^2} \end{bmatrix} \oplus \mathrm{I}_{n-2}$.
- Let U be distributed according to Haar measure on $\mathbb{O}(n)$, and independent of M.

  The matrix $UA_\epsilon U^T$ is a rotation by $\arcsin(\epsilon)$ in a random two-dimensional subspace of $\mathbb{R}^n$.
- Make an exchangeable pair of random matrices (M, M_ϵ) by randomly rotating M:

$$M_\epsilon := UA_\epsilon U^T M.$$

- The exchangeable pair descends to W:

$$W_\epsilon := \mathrm{Tr}(AM_\epsilon).$$

Getting our hands (a little) dirty

Getting our hands (a little) dirty

To apply the abstract approximation theorem to this exchangeable pair, we need to evaluate

$$\mathbb{E}\left[W_\epsilon - W \middle| W\right] = \mathbb{E}\left[\mathrm{Tr}\left[A(M_\epsilon - M)\right] \middle| \mathrm{Tr}(AM)\right].$$

Getting our hands (a little) dirty

To apply the abstract approximation theorem to this exchangeable pair, we need to evaluate

$$\mathbb{E}\left[W_\epsilon - W \middle| W\right] = \mathbb{E}\left[\mathrm{Tr}\left[A(M_\epsilon - M)\right] \middle| \mathrm{Tr}(AM)\right].$$

Let K be the $n \times 2$ matrix made of the first two columns of U, let I_2 be the 2×2 identity, and

$$C_2 := \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Getting our hands (a little) dirty

To apply the abstract approximation theorem to this exchangeable pair, we need to evaluate

$$\mathbb{E}\left[W_\epsilon - W \middle| W\right] = \mathbb{E}\left[\mathrm{Tr}\left[A(M_\epsilon - M)\right] \middle| \mathrm{Tr}(AM)\right].$$

Let K be the $n \times 2$ matrix made of the first two columns of U, let I_2 be the 2×2 identity, and

$$C_2 := \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Then

$$M_\epsilon - M = U\big(A_\epsilon - I_n\big)U^T M$$

.

Getting our hands (a little) dirty

To apply the abstract approximation theorem to this exchangeable pair, we need to evaluate

$$\mathbb{E}\left[W_\epsilon - W \middle| W\right] = \mathbb{E}\left[\mathrm{Tr}\left[A(M_\epsilon - M)\right] \middle| \mathrm{Tr}(AM)\right].$$

Let K be the $n \times 2$ matrix made of the first two columns of U, let I_2 be the 2×2 identity, and

$$C_2 := \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Then

$$M_\epsilon - M = U(A_\epsilon - I_n)U^T M = K\left[\left(\sqrt{1-\epsilon^2} - 1\right)I_2 + \epsilon C_2\right]K^T M$$

.

Getting our hands (a little) dirty

To apply the abstract approximation theorem to this exchangeable pair, we need to evaluate

$$\mathbb{E}\left[W_\epsilon - W \middle| W\right] = \mathbb{E}\left[\mathrm{Tr}\left[A(M_\epsilon - M)\right] \middle| \mathrm{Tr}(AM)\right].$$

Let K be the $n \times 2$ matrix made of the first two columns of U, let I_2 be the 2×2 identity, and

$$C_2 := \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

Then

$$\begin{aligned} M_\epsilon - M = U\big(A_\epsilon - I_n\big)U^T M &= K\left[\big(\sqrt{1-\epsilon^2} - 1\big)I_2 + \epsilon C_2\right]K^T M \\ &= K\left[\left(-\frac{\epsilon^2}{2} + O(\epsilon^4)\right)I_2 + \epsilon C_2\right]K^T M. \end{aligned}$$

So:

$$W_\epsilon - W = \left(-\frac{\epsilon^2}{2} + O(\epsilon^4)\right) \mathrm{Tr}(AKK^TM) + \epsilon\, \mathrm{Tr}(AKC_2K^TM).$$

So:

$$W_\epsilon - W = \left(-\frac{\epsilon^2}{2} + O(\epsilon^4)\right) \operatorname{Tr}(AKK^T M) + \epsilon \operatorname{Tr}(AKC_2K^T M).$$

Using symmetry arguments one can easily check that

$$\mathbb{E}\left[KK^T\right] = \frac{2}{n} I_n \qquad\qquad \mathbb{E}\left[KC_2K^T\right] = 0.$$

So:

$$W_\epsilon - W = \left(-\frac{\epsilon^2}{2} + O(\epsilon^4)\right) \mathrm{Tr}(AKK^T M) + \epsilon \, \mathrm{Tr}(AKC_2K^T M).$$

Using symmetry arguments one can easily check that

$$\mathbb{E}\left[KK^T\right] = \frac{2}{n} I_n \qquad\qquad \mathbb{E}\left[KC_2K^T\right] = 0.$$

So out pops:

$$\mathbb{E}\left[W_\epsilon - W \middle| W\right] = \left(-\frac{\epsilon^2}{n} + O(\epsilon^4)\right) W;$$

Condition 1 of the theorem holds with $\lambda(\epsilon) = \frac{\epsilon^2}{n}$.

The error from the theorem is given by

$$\lim_{\epsilon \to 0} \frac{1}{2\lambda(\epsilon)} \mathbb{E}\left| \mathbb{E}\left[|W_\epsilon - W|^2 \middle| W \right] - 1 \right|$$

as long as

$$\lim_{\epsilon \to 0} \frac{1}{\lambda(\epsilon)} \mathbb{E}|W_\epsilon - W|^3 = 0.$$

The error from the theorem is given by

$$\lim_{\epsilon \to 0} \frac{1}{2\lambda(\epsilon)} \mathbb{E}\left| \mathbb{E}\left[|W_\epsilon - W|^2 \middle| W \right] - 1 \right|$$

as long as

$$\lim_{\epsilon \to 0} \frac{1}{\lambda(\epsilon)} \mathbb{E}|W_\epsilon - W|^3 = 0.$$

Using that

$$W_\epsilon - W = \left(-\frac{\epsilon^2}{2} + O(\epsilon^4) \right) \mathrm{Tr}(AKK^T M) + \epsilon \, \mathrm{Tr}(AKC_2 K^T M)$$

and $\lambda(\epsilon) = \frac{\epsilon^2}{n}$,

$$\frac{1}{\lambda(\epsilon)} \mathbb{E}|W_\epsilon - W|^3 = O(\epsilon).$$

Again using

$$W_\epsilon - W = \left(-\frac{\epsilon^2}{2} + O(\epsilon^4)\right) \mathrm{Tr}(AKK^T M) + \epsilon\, \mathrm{Tr}(AKC_2K^T M)$$

and $\lambda(\epsilon) = \frac{\epsilon^2}{n}$,

Again using

$$W_\epsilon - W = \left(-\frac{\epsilon^2}{2} + O(\epsilon^4)\right) \mathrm{Tr}(AKK^T M) + \epsilon\, \mathrm{Tr}(AKC_2K^T M)$$

and $\lambda(\epsilon) = \frac{\epsilon^2}{n}$,

$$\frac{1}{2\lambda(\epsilon)}\mathbb{E}\big[(W_\epsilon - W)2\big|W\big] \sim \frac{n}{2}\mathbb{E}\Big[\big(\mathrm{Tr}(AKCK^T M)\big)^2\Big|W\Big].$$

Again using

$$W_\epsilon - W = \left(-\frac{\epsilon^2}{2} + O(\epsilon^4)\right) \mathrm{Tr}(AKK^T M) + \epsilon \, \mathrm{Tr}(AKC_2 K^T M)$$

and $\lambda(\epsilon) = \frac{\epsilon^2}{n}$,

$$\frac{1}{2\lambda(\epsilon)} \mathbb{E}\big[(W_\epsilon - W)2 \big| W\big] \sim \frac{n}{2} \mathbb{E}\Big[\big(\,\mathrm{Tr}(AKCK^T M)\big)^2 \big| W\Big].$$

The computation thus comes down to some mixed moments of entries of K.

Again using

$$W_\epsilon - W = \left(-\frac{\epsilon^2}{2} + O(\epsilon^4)\right) \mathrm{Tr}(AKK^T M) + \epsilon\, \mathrm{Tr}(AKC_2 K^T M)$$

and $\lambda(\epsilon) = \frac{\epsilon^2}{n}$,

$$\frac{1}{2\lambda(\epsilon)} \mathbb{E}\left[(W_\epsilon - W)2 \middle| W\right] \sim \frac{n}{2} \mathbb{E}\left[\left(\mathrm{Tr}(AKCK^T M)\right)^2 \middle| W\right].$$

The computation thus comes down to some mixed moments of entries of K. One gets:

$$\frac{n}{2} \mathbb{E}\left[\left(\mathrm{Tr}(AKCK^T M)\right)^2 \middle| W\right] = 1 + \frac{1}{n-1} \left[1 - \mathrm{Tr}\left((AM)^2\right)\right].$$

Again using

$$W_\epsilon - W = \left(-\frac{\epsilon^2}{2} + O(\epsilon^4)\right)\mathrm{Tr}(AKK^TM) + \epsilon\,\mathrm{Tr}(AKC_2K^TM)$$

and $\lambda(\epsilon) = \frac{\epsilon^2}{n}$,

$$\frac{1}{2\lambda(\epsilon)}\mathbb{E}\left[(W_\epsilon - W)2 \middle| W\right] \sim \frac{n}{2}\mathbb{E}\left[\left(\mathrm{Tr}(AKCK^TM)\right)^2 \middle| W\right].$$

The computation thus comes down to some mixed moments of entries of K. One gets:

$$\frac{n}{2}\mathbb{E}\left[\left(\mathrm{Tr}(AKCK^TM)\right)^2 \middle| W\right] = 1 + \frac{1}{n-1}\underbrace{\left[1 - \mathrm{Tr}\left((AM)^2\right)\right]}_{\text{has bounded expectation}}.$$

Dependency Graphs

This is a quite different approach for estimating $\mathbb{E} T_o f(W)$, which is often useful when W is a sum of weakly dependent random variables.

Dependency Graphs

This is a quite different approach for estimating $\mathbb{E}T_o f(W)$, which is often useful when W is a sum of weakly dependent random variables.

Let $\{X_i\}_{i=1}^n$ be a set of random variables. A dependency graph for the X_i is a graph with vertices $\{1, \ldots, n\}$ and edge set E such that, if $K_1, K_2 \subseteq \{1, \ldots, n\}$ are not connected by any edges, then

$$\{X_i\}_{i \in K_1} \text{ and } \{X_i\}_{i \in K_2} \text{ are independent.}$$

Dependency Graphs

This is a quite different approach for estimating $\mathbb{E} T_o f(W)$, which is often useful when W is a sum of weakly dependent random variables.

Let $\{X_i\}_{i=1}^n$ be a set of random variables. A dependency graph for the X_i is a graph with vertices $\{1, \ldots, n\}$ and edge set E such that, if $K_1, K_2 \subseteq \{1, \ldots, n\}$ are not connected by any edges, then

$$\{X_i\}_{i \in K_1} \text{ and } \{X_i\}_{i \in K_2} \text{ are independent.}$$

The idea is to exploit the dependence structure to analyze $\sum_{i=1}^n X_i$.

Poisson approximation via dependency graphs

Theorem (Arratia–Goldstein–Gordon)

Let $\{X_i\}_{i\in V}$ be a finite collection of binary random variables with dependency graph (V, E); let N_i denote the neighborhood of i in V and suppose that

$$\mathbb{P}(X_i = 1) = p_i \qquad \mathbb{P}(X_i = 1, X_j = 1) = p_{ij}.$$

Let $\lambda = \sum p_i$; let $Y \sim Poi(\lambda)$ and $W := \sum X_i$. Then

$$d_{TV}(W, Y) \leq \min(1, \lambda^{-1}) \left[\sum_{i\in I} \sum_{j\in N_i \setminus \{i\}} p_{ij} + \sum_{i\in I} \sum_{j\in N_i} p_i p_j \right].$$

The idea of the proof

Remember that the characterizing operator for Y is

$$T_o f(j) = \lambda f(j+1) - j f(j).$$

The idea of the proof

Remember that the characterizing operator for Y is

$$T_o f(j) = \lambda f(j+1) - j f(j).$$

Let $A \subseteq \mathbb{N}$: if f is such that $T_o f(j) = \mathbb{1}_A(j) - \mathbb{E}\mathbb{1}_A(Y)$, then

$$\mathbb{P}(Y \in A) - \mathbb{P}(W \in A) = \mathbb{E}[W f(W) - \lambda f(W+1)]$$

The idea of the proof

Remember that the characterizing operator for Y is

$$T_o f(j) = \lambda f(j+1) - jf(j).$$

Let $A \subseteq \mathbb{N}$: if f is such that $T_o f(j) = \mathbb{1}_A(j) - \mathbb{E}\mathbb{1}_A(Y)$, then

$$\begin{aligned}\mathbb{P}(Y \in A) - \mathbb{P}(W \in A) &= \mathbb{E}[Wf(W) - \lambda f(W+1)] \\ &= \sum_{i \in V} \mathbb{E}[X_i f(W) - p_i f(W+1)]\end{aligned}$$

The idea of the proof

Remember that the characterizing operator for Y is

$$T_o f(j) = \lambda f(j+1) - jf(j).$$

Let $A \subseteq \mathbb{N}$: if f is such that $T_o f(j) = \mathbb{1}_A(j) - \mathbb{E}\mathbb{1}_A(Y)$, then

$$\begin{aligned}
\mathbb{P}(Y \in A) - \mathbb{P}(W \in A) &= \mathbb{E}[Wf(W) - \lambda f(W+1)] \\
&= \sum_{i \in V} \mathbb{E}[X_i f(W) - p_i f(W+1)] \\
&= \sum_{i \in V} \mathbb{E}\left[X_i f\left(\sum_{j \neq i} X_j + 1\right) - p_i f(W+1)\right].
\end{aligned}$$

$$\mathbb{P}(Y \in A) - \mathbb{P}(W \in A) = \sum_{i \in V} \mathbb{E}\left[X_i f\left(\sum_{j \neq i} X_j + 1\right) - p_i f(W+1)\right].$$

$$\mathbb{P}(Y \in A) - \mathbb{P}(W \in A) = \\ \sum_{i \in V} \mathbb{E}\left[X_i f\left(\sum_{j \neq i} X_j + 1\right) - p_i f(W+1)\right].$$

Well, $W \approx \sum_{j \neq i} X_j$, so

$$\mathbb{P}(Y \in A) - \mathbb{P}(W \in A) \approx \sum_{i \in V} \mathbb{E}\left[(X_i - p_i) f\left(\sum_{j \neq i} X_j + 1\right)\right].$$

$$\mathbb{P}(Y \in A) - \mathbb{P}(W \in A) = \\ \textstyle\sum_{i \in V} \mathbb{E}\left[X_i f\left(\sum_{j \neq i} X_j + 1\right) - p_i f(W+1)\right].$$

Well, $W \approx \sum_{j \neq i} X_j$, so

$$\mathbb{P}(Y \in A) - \mathbb{P}(W \in A) \approx \sum_{i \in V} \mathbb{E}\left[(X_i - p_i) f\left(\sum_{j \neq i} X_j + 1\right)\right].$$

Moreover, X_i and $\sum_{j \notin N_i} X_j$ are independent, so in fact

$$\mathbb{P}(Y \in A) - \mathbb{P}(W \in A) \\ \approx \sum_{i \in V} \mathbb{E}\left[(X_i - p_i)\left(f\left(\sum_{j \neq i} X_j + 1\right) - f\left(\sum_{j \notin N_i} X_j + 1\right)\right)\right].$$

Example: Betti numbers in the "pretty sparse" regime

Recall the set-up: let f be a bounded density on $\mathbb{R}^d$ and choose n points $\{X_1, \ldots, X_n\}$ independently according to f.

Example: Betti numbers in the "pretty sparse" regime

Recall the set-up: let f be a bounded density on $\mathbb{R}^d$ and choose n points $\{X_1, \ldots, X_n\}$ independently according to f.

Construct the random Čech complex $\mathcal{C} = \mathcal{C}(X_1, \ldots, X_n)$ over the points: any subcollection of the points span a face in $\mathcal{C}$ if the collection of balls with those centers and radius r_n intersect nontrivially.

Example: Betti numbers in the "pretty sparse" regime

Recall the set-up: let f be a bounded density on $\mathbb{R}^d$ and choose n points $\{X_1, \ldots, X_n\}$ independently according to f.

Construct the random Čech complex $\mathcal{C} = \mathcal{C}(X_1, \ldots, X_n)$ over the points: any subcollection of the points span a face in $\mathcal{C}$ if the collection of balls with those centers and radius r_n intersect nontrivially.

Theorem (Kahle–M)

If $n^k r_n^{d(k-1)} \to \alpha \in (0, \infty)$ as $n \to \infty$, then

$$d_{TV}(\beta_k(\mathcal{C}(X_1, \ldots, X_n)), Y) \leq c n r_n^d,$$

where Y is a Poisson random variable with $\mathbb{E}[Y] = \mathbb{E}[\beta_k]$ and c is a constant depending only on α, k and f.

Preliminaries

Firstly, we relate β_k to the number of empty $(k+1)$-simplices in $\mathcal{C}(X_1, \ldots, X_n)$

Preliminaries

Firstly, we relate β_k to the number of empty $(k+1)$-simplices in $\mathcal{C}(X_1, \ldots, X_n)$:

$$\widetilde{S}_{n,k+1} \leq \beta_k(\mathcal{C}) \leq S_{n,k+1} + \textit{other stuff},$$

where $S_{n,k+1}$ is the number of empty simplices on $k+2$ vertices in $\mathcal{C}(X_1, \ldots, X_n)$ and $\widetilde{S}_{n,k+1}$ is the number of *isolated* empty simplices on $k+2$ vertices in $\mathcal{C}$.

Preliminaries

Firstly, we relate β_k to the number of empty $(k+1)$-simplices in $\mathcal{C}(X_1, \ldots, X_n)$:

$$\widetilde{S}_{n,k+1} \leq \beta_k(\mathcal{C}) \leq S_{n,k+1} + \textit{other stuff},$$

where $S_{n,k+1}$ is the number of empty simplices on $k+2$ vertices in $\mathcal{C}(X_1, \ldots, X_n)$ and $\widetilde{S}_{n,k+1}$ is the number of *isolated* empty simplices on $k+2$ vertices in $\mathcal{C}$.

Proving that $S_{n,k+1}$ is approximately Poisson in this regime is basically enough; there's no real difference between $S_{n,k+1}$ and $\widetilde{S}_{n,k+1}$ and the other stuff can be estimated away.

The set-up

The set-up

Write

$$S_{n,k} = \sum_{\substack{\mathbf{i}=(i_0,i_1,\ldots,i_k) \\ 1\le i_1<\cdots<i_k\le n}} \xi_{\mathbf{i}},$$

where $\xi_{\mathbf{i}}$ is the indicator that $X_{i_0}, \ldots, X_{i_k}$ form an empty k-simplex; that is, the balls of radius r_n about any k of the X_{i_j} intersect, but the intersection of all $k+1$ balls is empty.

The set-up

Write

$$S_{n,k} = \sum_{\substack{\mathbf{i}=(i_0,i_1,\ldots,i_k)\\ 1\leq i_1<\cdots<i_k\leq n}} \xi_{\mathbf{i}},$$

where $\xi_{\mathbf{i}}$ is the indicator that $X_{i_0}, \ldots, X_{i_k}$ form an empty k-simplex; that is, the balls of radius r_n about any k of the X_{i_j} intersect, but the intersection of all $k+1$ balls is empty.

The dependency graph: If $\mathbf{i} = (i_0, i_1, \ldots, i_k)$ and $\mathbf{j} = (j_0, j_1, \ldots, j_k)$ have no indices in common, then certainly $\xi_{\mathbf{i}}$ and $\xi_{\mathbf{j}}$ are independent – we thus

connect $\mathbf{i}$ and $\mathbf{j}$ if $\mathbf{i} \cap \mathbf{j} \neq \emptyset$.

Estimates

Recall: the theorem says that

$$d_{TV}(S_{n,k}, Y) \leq \min(1, \lambda^{-1}) \left[\sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}} \setminus \{\mathbf{i}\}} p_{\mathbf{ij}} + \sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}}} p_{\mathbf{i}} p_{\mathbf{j}} \right]$$

Estimates

Recall: the theorem says that

$$d_{TV}(S_{n,k}, Y) \leq \min(1, \lambda^{-1}) \left[\sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}} \setminus \{\mathbf{i}\}} p_{\mathbf{ij}} + \sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}}} p_{\mathbf{i}} p_{\mathbf{j}} \right]$$

Recall also from the last lecture: for $0 \leq k \leq d-1$, there is a constant μ depending only on f and k such that

$$\frac{\mathbb{E}\left[\beta_k(\mathcal{C})\right]}{n^k r_n^{d(k-1)}} \longrightarrow \frac{\mu}{(k+1)!} \qquad \text{as} \qquad n \to \infty.$$

Estimates

Recall: the theorem says that

$$d_{TV}(S_{n,k}, Y) \leq \min(1, \lambda^{-1}) \left[\sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}} \setminus \{\mathbf{i}\}} p_{\mathbf{ij}} + \sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}}} p_{\mathbf{i}} p_{\mathbf{j}} \right]$$

Recall also from the last lecture: for $0 \leq k \leq d-1$, there is a constant μ depending only on f and k such that

$$\frac{\mathbb{E}\left[\beta_k(\mathcal{C})\right]}{n^k r_n^{d(k-1)}} \longrightarrow \frac{\mu}{(k+1)!} \qquad \text{as} \qquad n \to \infty.$$

This actually comes from getting the corresponding asymptotics for $\widetilde{S}_{n,k}$ and $S_{n,k}$; in particular,

$$\lambda = \left(\frac{\mu}{k!}\right) n^{k+1} r_n^{dk}.$$

$$d_{TV}(S_{n,k}, Y) \leq \left(\frac{k!}{\mu}\right) n^{-(k+1)} r_n^{-dk} \left[\sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}} \setminus \{\mathbf{i}\}} p_{\mathbf{ij}} + \sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}}} p_{\mathbf{i}} p_{\mathbf{j}}\right]$$

$$d_{TV}(S_{n,k}, Y) \le \left(\frac{k!}{\mu}\right) n^{-(k+1)} r_n^{-dk} \left[\sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}} \setminus \{\mathbf{i}\}} p_{\mathbf{ij}} + \sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}}} p_{\mathbf{i}} p_{\mathbf{j}}\right]$$

Now, for $k+1$ i.i.d. points to form a simplex, the first k all have to be within $2r_n$ of the last:

$$p_{\mathbf{i}} = \mathbb{E}\xi_{\mathbf{i}} \le \left[(2r_n)^d \theta_d \|f\|_\infty\right]^k,$$

where θ_d is the volume of the unit sphere in $\mathbb{R}^d$.

$$d_{TV}(S_{n,k}, Y) \leq \left(\frac{k!}{\mu}\right) n^{-(k+1)} r_n^{-dk} \left[\sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}} \setminus \{\mathbf{i}\}} p_{\mathbf{ij}} + \sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}}} p_{\mathbf{i}} p_{\mathbf{j}}\right]$$

Now, for $k+1$ i.i.d. points to form a simplex, the first k all have to be within $2r_n$ of the last:

$$p_{\mathbf{i}} = \mathbb{E}\xi_{\mathbf{i}} \leq \left[(2r_n)^d \theta_d \|f\|_\infty\right]^k,$$

where θ_d is the volume of the unit sphere in $\mathbb{R}^d$.

Given $\mathbf{i} \in I$, the number of $\mathbf{j} \in I$ with $\mathbf{i} \sim \mathbf{j}$ is

$$\binom{n}{k+1} - \binom{n-k-1}{k+1} = \frac{(k+1)^2 n^k}{(k+1)!} + O\left(n^{k-1}\right).$$

$$d_{TV}(S_{n,k}, Y) \le \left(\frac{k!}{\mu}\right) n^{-(k+1)} r_n^{-dk} \left[\sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}} \setminus \{\mathbf{i}\}} p_{\mathbf{ij}} + \sum_{\mathbf{i}} \sum_{\mathbf{j} \in N_{\mathbf{i}}} p_{\mathbf{i}} p_{\mathbf{j}}\right]$$

Now, for $k+1$ i.i.d. points to form a simplex, the first k all have to be within $2r_n$ of the last:

$$p_{\mathbf{i}} = \mathbb{E}\xi_{\mathbf{i}} \le \left[(2r_n)^d \theta_d \|f\|_\infty\right]^k,$$

where θ_d is the volume of the unit sphere in $\mathbb{R}^d$.

Given $\mathbf{i} \in I$, the number of $\mathbf{j} \in I$ with $\mathbf{i} \sim \mathbf{j}$ is

$$\binom{n}{k+1} - \binom{n-k-1}{k+1} = \frac{(k+1)^2 n^k}{(k+1)!} + O\left(n^{k-1}\right).$$

$\Longrightarrow$ The $p_{\mathbf{i}} p_{\mathbf{j}}$ term above is, to top order,

$$\frac{(2\theta_d \|f\|_\infty)^{2k}}{k!\mu} (n r_n^d)^k.$$

Similarly, if $\left|\mathbf{i} \cap \mathbf{j}\right| = \ell$, then

$$p_{\mathbf{ij}} = \mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{j}}\right] \leq \left[(2r_n)^d \theta_d \|f\|_\infty\right]^{2k-\ell+1}.$$

Similarly, if $\left|\mathbf{i} \cap \mathbf{j}\right| = \ell$, then

$$p_{\mathbf{ij}} = \mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{j}}\right] \leq \left[(2r_n)^d \theta_d \|f\|_\infty\right]^{2k-\ell+1}.$$

Given $\mathbf{i}$, the number of $\mathbf{j}$ with $\left|\mathbf{i} \cap \mathbf{j}\right| = \ell$ is

$$\binom{k+1}{\ell}\binom{n-k-1}{k+1-\ell}.$$

Similarly, if $\left|\mathbf{i} \cap \mathbf{j}\right| = \ell$, then

$$p_{\mathbf{ij}} = \mathbb{E}\left[\xi_{\mathbf{i}}\xi_{\mathbf{j}}\right] \leq \left[(2r_n)^d \theta_d \|f\|_\infty\right]^{2k-\ell+1}.$$

Given $\mathbf{i}$, the number of $\mathbf{j}$ with $\left|\mathbf{i} \cap \mathbf{j}\right| = \ell$ is

$$\binom{k+1}{\ell}\binom{n-k-1}{k+1-\ell}.$$

$\Longrightarrow$ the $p_{\mathbf{ij}}$ term above is,

$$\frac{1}{\lambda}\binom{n}{k+1}\sum_{\ell=1}^{k}\binom{k+1}{\ell}\binom{n-k-1}{k+1-\ell}\left[(2r_n)^d \theta_d \|f\|_\infty\right]^{2k-\ell+1} \lesssim n r_n^d.$$