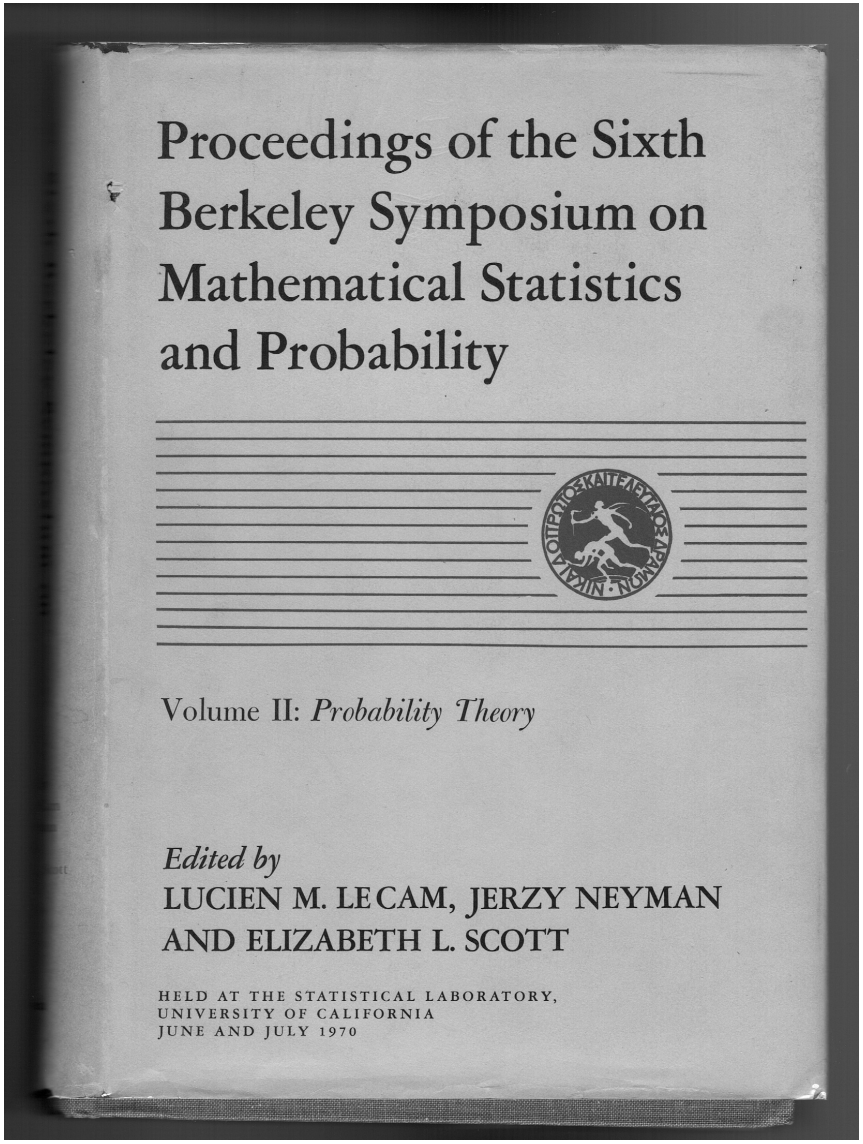
The Many Faces of a Simple Identity

Larry Goldstein
University of Southern California

ICML Workshop, June 15^{th} 2019

Guided Tour

In the Beginning



Itinerary

1. Stein Identity
2. Distributional Approximation
3. Concentration
4. Second order Poincaré Inequalities, and Malliavin Calculus
5. Shrinkage, Unbiased Risk Estimation

Poisson Distribution

(Chen 1975) Non-negative integer valued random variable W is distributed $\mathcal{P}_\lambda$ if and only if

$$E[Wf(W)] = \lambda E[f(W+1)] \quad \text{all } f \in \mathcal{F}.$$

Poisson Distribution

(Chen 1975) Non-negative integer valued random variable W is distributed $\mathcal{P}_\lambda$ if and only if

$$E[Wf(W)] = \lambda E[f(W+1)] \quad \text{all } f \in \mathcal{F}.$$

For any $W \geq 0$ with mean $\lambda \in (0, \infty)$, size bias distribution:

$$E[Wf(W)] = \lambda E[f(W^s)] \quad \text{all } f \in \mathcal{F}.$$

Restatement: $W^s =_d W + 1$ if and only if $W \sim \mathcal{P}(\lambda)$.

Poisson Distribution

(Chen 1975) Non-negative integer valued random variable W is distributed $\mathcal{P}_\lambda$ if and only if

$$E[Wf(W)] = \lambda E[f(W+1)] \quad \text{all } f \in \mathcal{F}.$$

For any $W \geq 0$ with mean $\lambda \in (0, \infty)$, size bias distribution:

$$E[Wf(W)] = \lambda E[f(W^s)] \quad \text{all } f \in \mathcal{F}.$$

Restatement: $W^s =_d W + 1$ if and only if $W \sim \mathcal{P}(\lambda)$.

$$d_{\mathrm{TV}}(W, P_\lambda) \leq (1 - e^{-\lambda}) E|(W^s - 1) - W|.$$

Applications e.g. to matchings in molecular sequence analysis.

$$d_{\mathrm{TV}}(W, P_\lambda) \leq (1 - e^{-\lambda}) E|(W^s - 1) - W|$$

Simple Example: Let

$$W = \sum_{i=1}^{n} X_i \quad \text{with } \lambda = E[W],$$

the sum of independent Bernoullis with $p_i = E[X_i] \in (0, 1)$. Then

$$W^s = W - X_I + 1 \quad \text{where } P(I = i) = p_i/\lambda,\ I \text{ independent.}$$

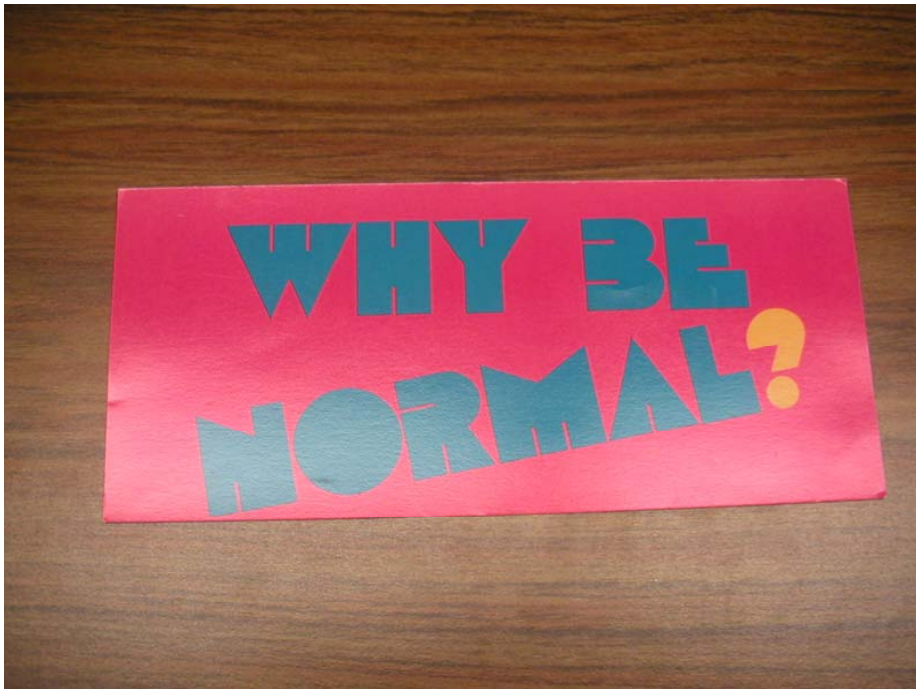
Then

$$d_{\mathrm{TV}}(W, P_\lambda) \leq (1 - e^{-\lambda}) E X_I = \frac{1 - e^{-\lambda}}{\lambda} \sum_{i=1}^{n} p_i^2.$$

If $p_i = \lambda/n$ then the bound specializes to $\lambda(1 - e^{-\lambda})/n$.

The Big Question

The Big Question



Stein Identity for Standard Gaussian

Let Y be normal $\mathcal{N}(\theta, \sigma^2)$ with density $\phi_{\theta,\sigma^2}(t) = e^{-(t-\theta)^2/2\sigma^2}/\sqrt{2\pi\sigma^2}$. Then the law of a random variable W has the same distribution as Y if and only

$$E[(W-\theta)f(W)] = \sigma^2 E[f'(W)] \quad \text{for all } f \in \mathcal{F},$$

where $\mathcal{F}$ is some sufficiently rich class of smooth functions.

1. All functions f for which the two sides above exist.
2. All functions in

$$\mathrm{Lip}_1 = \{f : |f(x) - f(y)| \leq |x - y|\}.$$

Proof of Stein Identity; Standard normal case

Direction normality of W implies for all $f \in \mathcal{F}$ equality, some say integration by parts: with $\phi(t) = e^{-t^2/2}/\sqrt{2\pi}$

$$t\phi(t) = -\phi'(t) \quad \text{hence} \quad E[Wf(W)] = E[f'(W)].$$

Requires restricting to finite interval, resulting in boundary terms, on which conditions will be needed for taking limit.

Proof of Stein Identity; Standard normal case

Direction normality of W implies for all $f \in \mathcal{F}$ equality, some say integration by parts: with $\phi(t) = e^{-t^2/2}/\sqrt{2\pi}$

$$t\phi(t) = -\phi'(t) \quad \text{hence} \quad E[Wf(W)] = E[f'(W)].$$

Requires restricting to finite interval, resulting in boundary terms, on which conditions will be needed for taking limit. Use Fubini as Stein did, breaking into positive and negative parts:

$$\int_0^\infty f'(w)\phi(w)dw = -\int_0^\infty f'(w)\int_w^\infty \phi'(t)dtdw$$
$$= \int_0^\infty \int_0^t t\phi(t)f'(w)dwdt = \int_0^\infty t\phi(t)[f(t) - f(0)]dt.$$

Combining with portion on $(-\infty, 0]$, obtain

$$E[f'(W)] = E[W(f(W) - f(0))] = E[Wf(W)].$$

Stein Equation

For a given class of functions $\mathcal{H}$ (e.g. Lip_1), and distributions of random variables X and Y, let (e.g. Wasserstein distance)

$$d_{\mathcal{H}}(X, Y) = \sup_{h \in \mathcal{H}} |Eh(X) - Eh(Y)|.$$

Given a mean zero, variance 1 random variable W, and a test function h in a class $\mathcal{H}$, bound the difference

$$Eh(W) - Eh(Z).$$

Now, reason as follows: since this expectation, and $E[f'(W) - Wf(W)]$ are both zero when W is normal, lets equate them.

Stein Equation (1)

Of course in (2.7), if $W^* > W$, the integral $\int_{W^*}^{W} h(z)\,dz$ is to be interpreted as $-\int_{W}^{W^*} h(z)\,dz$, and similarly with $h(z)$ replaced by $zf(z)$. The conditional expectation signs $E^{\mathscr{F}}$ and $E^{\mathscr{C}}$ in (2.7) could be dropped, but they help suggest the way the lemma will be applied.

PROOF OF LEMMA 2.1. Let f (otherwise arbitrary) be a bounded function, the integral of a bounded measurable function f'. Then

$$\begin{aligned}(2.10)\qquad E[Wf(W)] &= E[(E^{\mathscr{F}}G)f(W)]\\ &= E[Gf(W)]\\ &= E\{G[f(W) - f(W^*)] + (E^{\mathscr{C}}G)f(W^*)\}.\end{aligned}$$

We can rewrite this in the form

$$\begin{aligned}(2.11)\qquad &E[f'(W) - Wf(W)]\\ &= E\{f'(W) - G[f(W) - f(W^*)]\} + E\{G[f(W) - f(W^*)] - Wf(W)\}\\ &= E\{f'(W) - G[f(W) - f(W^*)] - (E^{\mathscr{C}}G)f(W^*)\}.\end{aligned}$$

In order to make conditions for the validity of the normal approximation to the distribution of W more apparent, we express f in terms of an arbitrary bounded measurable function h by (2.9). This function f is the unique bounded solution of the differential equation

$$(2.12)\qquad f'(w) - wf(w) = h(w) - Nh.$$

Then equation (2.11) yields

$$\begin{aligned}(2.13)\qquad Eh(W) &= Nh + E[f'(W) - Wf(W)]\\ &= Nh + E\{f'(W) - G[f(W) - f(W^*)] - (E^{\mathscr{C}}G)f(W^*)\}\\ &= Nh + E\left\{f'(W) - E^{\mathscr{F}}\left[G\int_{W^*}^{W} f'(z)\,dz\right] - (E^{\mathscr{C}}G)f(W^*)\right\}\\ &= Nh + E\left\{\left(h(W) - E^{\mathscr{F}}\left[G\int_{W^*}^{W} h(z)\,dz\right]\right)\right.\\ &\qquad \left. + \left(Wf(W) - E^{\mathscr{F}}\left[G\int_{W^*}^{W} zf(z)\,dz\right]\right) - (E^{\mathscr{C}}G)f(W^*)\right\}.\end{aligned}$$

If we subtract $Eh(W) - E^{\mathscr{F}}[G\int_{W^*}^{W} h(z)\,dz]$ from both sides of this equation, we obtain (2.7). This completes the proof of Lemma 2.1. We observe that we have

Stein Equation and Couplings

Stein equation for the standard normal:

$$f'(x) - xf(x) = h(x) - Eh(Z).$$

Now to compute the expectation of the right hand side involving h to bound $d_{\mathcal{H}}(W, Z)$, lets solve a differential equation for f and compute the expectation $E[f'(W) - Wf(W)]$ of the left.

Would at first glace appear to make the problem harder. However, there is only one random variable in this expectation, rather than two.

Can handle the left hand side expectation using construction of auxiliary random variables, couplings.

Extend Stein Identity

One direction of the Stein identity, for W with $E[W] = 0$ and $\mathrm{Var}(W) = 1$,

$$E[Wf(W)] = E[f'(W)] \quad \text{for all } f \in \mathcal{F} \tag{1}$$

only if $W \sim \mathcal{N}(0, 1)$. So if W has any other distribution (1) does not hold.

Can we can modify the identity, or make some similar identity, so that it holds for a different W distribution?

Some Options

Feel free to add to the list!

1. Stein's exchangeable pair
2. Stein Kernels
3. Size Bias
4. Zero Bias
5. Score function

Stein Kernels and Zero Bias Coupling

Modify the right hand side of the identity

$$E[Wf(W)] = E[f'(W)] \quad \text{for all } f \in \mathcal{F}$$

in some way to accommodate non-normal distribution.
Stein Kernel (Cacoullos and Papathanasiou '92)

$$E[Wf(W)] = E[Tf'(W)] \quad \text{for all } f \in \mathcal{F}$$

Zero Bias (G. and Reinert '97)

$$E[Wf(W)] = E[f'(W^*)] \quad \text{for all } f \in \mathcal{F}$$

Use of Stein Kernels: $E[Wf(W)] = E[Tf'(W)]$

Given $h \in \mathcal{H}$ let f be the unique bounded solution to

$$f'(x) - xf(x) = h(x) - Eh(Z).$$

Then, using Stein kernels, for $\mathcal{H} = \{f : \mathbb{R} \to [0,1]\}$

$$\begin{aligned} |Eh(W) - Eh(Z)| &= |E[f'(W) - Wf(W)]| = |E[f'(W) - Tf'(W)]| \\ &= |E[(1-T)f'(W)]| \leq \|f'\| E|T-1| \leq 2E|T-1|. \end{aligned}$$

Taking supremum over this choice of $\mathcal{H}$ on the left hand side yields

$$d_{\mathrm{TV}}(W, Z) \leq 2E|T-1|,$$

a bound on the total variation distance.

Use of Zero Bias Coupling: $E[Wf(W)] = E[f'(W^*)]$

Given $h \in \mathcal{H}$ let f be the unique bounded solution to

$$f'(x) - xf(x) = h(x) - Eh(Z).$$

Then, using zero bias, for $\mathcal{H} = \mathrm{Lip}_1$

$$\begin{aligned} |Eh(W) - Eh(Z)| = |E[f'(W) - Wf(W)]| &= |E[f'(W) - f'(W^*)]| \\ &\leq \|f''\| E|W - W^*|. \end{aligned}$$

Taking infimum over all couplings on the right, and then supremum over this choice of $\mathcal{H}$ on the left hand side yields

$$d_1(W, Z) \leq 2d_1(W, W^*),$$

a bound on the Wasserstein distance.

Other Distributions

Classical: Poisson, Gamma, Binomial, Multinomial, Beta, Stable laws, Rayleigh, ...

Not so classical: PRR distribution, Dickman distribution, ...

Other Distributions

Classical: Poisson, Gamma, Binomial, Multinomial, Beta, Stable laws, Rayleigh, ...

Not so classical: PRR distribution, Dickman distribution, ...

Dickman characterizations for $W \geq 0$, independent $U \sim \mathcal{U}[0,1]$,

$$W^s =_d W + U \quad \text{and} \quad W =_d U(W+1)$$

Subgaussian Concentration

Chatterjee 2005: (W, W') exchangeable pair, $F(x, y) = -F(y, x)$

$$E[F(W, W')|W] = f(W)$$

$$v(w) = \frac{1}{2} E[|(f(W) - f(W')F(W, W')|W = w] \leq \sigma^2,$$

then the tail of $f(W)$ decays like a Gaussian with variance σ^2. Recovers Hoeffding's inequality for a sum W of independent, c_i bounded random variables. Taking $F(x, y) = n(x - y)$, $W' = W - X_I + X'_I$, I uniform, yields $f(W) = W$ and

$$v(W) = \frac{1}{2n} \sum_{i=1}^{n} E\left(n(X_i - X'_I)^2|W\right) \leq 2 \sum_{i=1}^{n} c_i^2.$$

Applications to e.g. magnetization in the Curie-Weiss model.

Sub-poisson Concentration

G. Ghosh 2011, Arratia Baxendale 2015, Cook, G. and Johnson 2018. If (W, W^s) is a size biased coupling of a non-negative random variable W with finite, nonzero mean satisfying

$$W^s \leq W + c$$

for some c, then W is sub-Poisson. (Recall $W^s =_d W + 1$ if and only if W is Poisson.)

Sub-poisson Concentration

G. Ghosh 2011, Arratia Baxendale 2015, Cook, G. and Johnson 2018. If (W, W^s) is a size biased coupling of a non-negative random variable W with finite, nonzero mean satisfying

$$W^s \leq W + c$$

for some c, then W is sub-Poisson. (Recall $W^s =_d W + 1$ if and only if W is Poisson.)

Example with dependence, number of fixed point of π, a uniformly chosen random permutation, and

$$W_\pi = \sum_{i=1}^{n} \mathbf{1}(\pi(i) = i).$$

$W_\pi^s \leq W_\pi + c$ for $W_\pi = \sum_{i=1}^n \mathbf{1}(\pi(i) = i)$

For I an independent and uniformly chosen index, with π given by

$$\begin{array}{ccccccc} 1 & \cdots & k & \cdots & I & \cdots & n \\ \pi(1) & \cdots & I & \cdots & \pi(I) & \cdots & \pi(n) \end{array}$$

let π^s be given by

$$\begin{array}{ccccccc} 1 & \cdots & k & \cdots & I & \cdots & n \\ \pi(1) & \cdots & \pi(I) & \cdots & I & \cdots & \pi(n) \end{array}$$

Then W_{π^s} has the W_π size bias distribution and $W_{\pi^s} \leq W_\pi + 2$.

Applications to, e.g. eigenvalues of random regular graphs.

2^{nd} order Poincaré inequality and Malliavin Calculus

Stein Kernel,

$$E[Wf(W)] = E[Tf'(W)]$$

Obtain, for instance, an immediate total variation distance bound of $2E|T-1|$. What's the catch?

2^{nd} order Poincaré inequality and Malliavin Calculus

Stein Kernel,

$$E[Wf(W)] = E[Tf'(W)]$$

Obtain, for instance, an immediate total variation distance bound of $2E|T-1|$. What's the catch?

When W is the sum of independent variables, the Kernel for W is the sum of the kernels of the components. In other situations, determining the kernel may be much more difficult.

2^{nd} order Poincaré inequality and Malliavin Calculus

Stein Kernel,

$$E[Wf(W)] = E[Tf'(W)]$$

Obtain, for instance, an immediate total variation distance bound of $2E|T-1|$. What's the catch?

When W is the sum of independent variables, the Kernel for W is the sum of the kernels of the components. In other situations, determining the kernel may be much more difficult.
Note

$$\mathrm{Var}(W) = E[T].$$

2^{nd} order Poincaré inequality

Chatterjee 09: For a sufficently smooth $H : \mathbb{R}^d \to \mathbb{R}$, the Stein Kernel T for $H(\mathbf{g})$, where $\mathbf{g} \sim \mathcal{N}(0, I_d)$, is given by

$$T = \int_0^\infty e^{-t} \langle \nabla H(\mathbf{g}), \widehat{E}(\nabla H(\widehat{\mathbf{g}}_t)) \rangle dt.$$

where for $t \geq 0$, $\widehat{\mathbf{g}}_t = e^{-t}\mathbf{g} + \sqrt{1 - e^{-2t}}\widehat{\mathbf{g}}$, where $\widehat{\mathbf{g}}$ is an independent copy of $\mathbf{g}$, and $\widehat{E}$ indicates expectation with respect to $\widehat{\mathbf{g}}$. (Recovers the Poincaré inequality via Cauchy-Schwarz)

Applications include results on the behavior of eigenvalues of random matrices with independent Gaussian entries.

The Malliavin Calculus connection

Nourdin and Peccati 2009 (see their Cambridge Univerity text 2012). Specializing their work to the Hilbert space of functions of Brownian motion $B(t)$ with inner product $\langle F, G \rangle = E[FG]$, for some F we have

$$T = \langle DF, -DL^{-1}F \rangle$$

where L is the Ornstein-Uhlenbeck generator, and D is the Malliavin derivative, which extends

$$DF = \sum_{i=1}^{n} \partial_i g(I(\psi_1), \ldots, I(\psi_n))\psi_i$$

for $F = g(I(\psi_1), \ldots, I(\psi_n))$ and $I(\psi) = \int \psi dB$. Applications: Functions of stochastic integrals.

The Malliavin Calculus connection

Nourdin and Peccati 2009 (see their Cambridge Univerity text 2012). Specializing their work to the Hilbert space of functions of Brownian motion $B(t)$ with inner product $\langle F, G \rangle = E[FG]$, for some F we have

$$T = \langle DF, -DL^{-1}F \rangle$$

where L is the Ornstein-Uhlenbeck generator, and D is the Malliavin derivative, which extends

$$DF = \sum_{i=1}^{n} \partial_i g(I(\psi_1), \ldots, I(\psi_n))\psi_i$$

for $F = g(I(\psi_1), \ldots, I(\psi_n))$ and $I(\psi) = \int \psi dB$. Applications: Functions of stochastic integrals.

Similar results for functions of Poisson processes, applications include to Voronoi tessellations. (Need to start with structure)

Stein Shrinkage Estimation

To estimate an unknown $\boldsymbol{\theta} \in \mathbb{R}^d$ based on an observation $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\theta}, I_d)$, it seems natural, and even optimal, to use $\mathbf{X}$, which has mean squared error $E\|\mathbf{X} - \boldsymbol{\theta}\|^2 = d$.

Stein Shrinkage Estimation

To estimate an unknown $\boldsymbol{\theta} \in \mathbb{R}^d$ based on an observation $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\theta}, I_d)$, it seems natural, and even optimal, to use $\mathbf{X}$, which has mean squared error $E\|\mathbf{X} - \boldsymbol{\theta}\|^2 = d$.

Surprisingly, for $d \geq 3$, we can do better using (Stein '56, James-Stein '61)

$$T(\mathbf{X}) = \mathbf{X}\left(1 - \frac{d-2}{||\mathbf{X}||^2}\right).$$

Expanding, we see that the mean squared error of $T(\mathbf{X})$ is

$$E_{\boldsymbol{\theta}}\left[||(\mathbf{X} - \boldsymbol{\theta})||^2 - \frac{2(d-2)(\mathbf{X} - \boldsymbol{\theta})'\mathbf{X}}{||\mathbf{X}||^2} + \frac{(d-2)^2}{||\mathbf{X}||^2}\right].$$

We improve on $\mathbf{X}$ if the remaining two terms are negative.

Stein Shrinkage Estimation

Mean squared error of James-Stein

$$E_{\boldsymbol{\theta}}\left[||(\mathbf{X}-\boldsymbol{\theta})||^2 - \frac{2(d-2)(\mathbf{X}-\boldsymbol{\theta})'\mathbf{X}}{||\mathbf{X}||^2} + \frac{(d-2)^2}{||\mathbf{X}||^2}\right]$$

Improvement results when

$$2E_{\boldsymbol{\theta}}\left[\frac{(\mathbf{X}-\boldsymbol{\theta})'\mathbf{X}}{||\mathbf{X}||^2}\right] > E_{\boldsymbol{\theta}}\left[\frac{d-2}{||\mathbf{X}||^2}\right].$$

Apply Stein identity on the left, coordinate-wise, to the function $f(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|^2$.

Stein Identity with $f(\mathbf{x}) = \mathbf{x}/||\mathbf{x}||^2$

Yields:

$$2E_{\boldsymbol{\theta}}\left[(\mathbf{X}-\boldsymbol{\theta})'\frac{\mathbf{X}}{||\mathbf{X}||^2}\right] = 2E_{\boldsymbol{\theta}}\sum_{j=1}^{d}\frac{\partial f_i(\mathbf{X})}{\partial x_i}$$

$$= 2E_{\boldsymbol{\theta}}\sum_{j=1}^{d}\left(\frac{||\mathbf{X}||^2 - 2X_i^2}{||\mathbf{X}||^4}\right) = 2E_{\boldsymbol{\theta}}\left(\frac{d}{||\mathbf{X}||^2} - \frac{2||\mathbf{X}||^2}{||\mathbf{X}||^4}\right)$$

$$= 2E_{\boldsymbol{\theta}}\left(\frac{d-2}{||\mathbf{X}||^2}\right) > E_{\boldsymbol{\theta}}\left(\frac{d-2}{||\mathbf{X}||^2}\right).$$

We have shown that

$$E_{\boldsymbol{\theta}}||T(\mathbf{X}) - \boldsymbol{\theta}||^2 < d = E_{\boldsymbol{\theta}}||\mathbf{X} - \boldsymbol{\theta}||^2 \quad \text{for all } \boldsymbol{\theta} \in \mathbf{R}^d.$$

Stein's Unbiased Risk Estimator

Observe $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\theta}, \mathbf{I}_d)$ with $\boldsymbol{\theta}$ unknown. We want to compute an unbiased estimate of the MSE of an estimator the form $S(\mathbf{X}) = \mathbf{X} + h(\mathbf{X})$, that is, of the expectation of

$$\begin{aligned} \|S(\mathbf{X}) - \boldsymbol{\theta}\|^2 &= \|\mathbf{X} - \boldsymbol{\theta} + h(\mathbf{X})\|^2 \\ &= \|\mathbf{X} - \boldsymbol{\theta}\|^2 + \|h(\mathbf{X})\|^2 + 2\langle h(\mathbf{X}), \mathbf{X} - \boldsymbol{\theta}\rangle. \end{aligned}$$

The expectation of the first term is d, and $\|h(\mathbf{X})\|^2$ is an unbiased estimator of its own expectation.

Stein's Unbiased Risk Estimator

Observe $\mathbf{X} \sim \mathcal{N}_d(\boldsymbol{\theta}, \mathbf{I}_d)$ with $\boldsymbol{\theta}$ unknown. We want to compute an unbiased estimate of the MSE of an estimator the form $S(\mathbf{X}) = \mathbf{X} + h(\mathbf{X})$, that is, of the expectation of

$$\begin{aligned}\|S(\mathbf{X}) - \boldsymbol{\theta}\|^2 &= \|\mathbf{X} - \boldsymbol{\theta} + h(\mathbf{X})\|^2 \\ &= \|\mathbf{X} - \boldsymbol{\theta}\|^2 + \|h(\mathbf{X})\|^2 + 2\langle h(\mathbf{X}), \mathbf{X} - \boldsymbol{\theta}\rangle.\end{aligned}$$

The expectation of the first term is d, and $\|h(\mathbf{X})\|^2$ is an unbiased estimator of its own expectation.
Applying the Stein identity coordinate-wise on the last term eliminates the unknown $\boldsymbol{\theta}$,

$$E[\langle \mathbf{X} - \boldsymbol{\theta}, h(\mathbf{X})\rangle] = E\left[\sum_{i=1}^{n} \frac{\partial h_i(\mathbf{X})}{\partial x_i}\right].$$

Hence

$$\mathrm{SURE}(h, \mathbf{X}) := d\sigma^2 + \|h(\mathbf{X})\|^2 + 2\nabla \cdot h(\mathbf{X})$$

is unbiased for the MSE, and computable from the data.

End of Tour

Thanks!