Comparing distributions via their canonical Stein operators: a new view of Stein's method

Gesine Reinert
Department of Statistics
University of Oxford

International Colloquium on Stein's Method, Concentration Inequalities, and Malliavin Calculus
June 30, 2014

Joint work with Christophe Ley (Brussels) and Yvik Swan (Liége)

Outline

1. Stein's method
2. A canonical Stein operator
3. Examples
4. Distances between expectations
5. Distance between posteriors
6. Last words

Stein's method in a nutshell

For μ a target distribution, with support $\mathcal{I}$:

1. Find a suitable operator $\mathcal{A}$ (called Stein operator) and a wide class of functions $\mathcal{F}(\mathcal{A})$ (called Stein class) such that $X \sim \mu$ if and only if for all functions $f \in \mathcal{F}(\mathcal{A})$,
$$\mathbb{E}\mathcal{A}f(X) = 0.$$
2. Let $\mathcal{H}(\mathcal{I})$ be a measure-determining class on $\mathcal{I}$. For each $h \in \mathcal{H}$ find a solution $f = f_h \in \mathcal{F}(\mathcal{A})$ of the
$$h(x) - \mathbb{E}h(X) = \mathcal{A}f(x),$$
where $X \sim \mu$. If the solution exists and if it is unique in $\mathcal{F}(\mathcal{A})$ then we can write
$$f(x) = \mathcal{A}^{-1}(h(x) - \mathbb{E}h(X)).$$
We call $\mathcal{A}^{-1}$ the *inverse Stein operator* (for μ).

Comparison of distributions

Let X and Y have distributions μ_X and μ_Y with Stein operators $\mathcal{A}_X$ and $\mathcal{A}_Y$, so that $\mathcal{F}(\mathcal{A}_X) \cap \mathcal{F}(\mathcal{A}_Y) \neq \emptyset$ and choose $\mathcal{H}(\mathcal{I})$ such that all solutions f of the Stein equation belong to this intersection. Then

$$\mathbb{E}h(X) - \mathbb{E}h(Y) = \mathbb{E}\mathcal{A}_Y f(X) = \mathbb{E}\mathcal{A}_Y f(X) - \mathbb{E}\mathcal{A}_X f(X)$$

and

$$\sup_{h \in \mathcal{H}(\mathcal{I})} |\mathbb{E}h(X) - \mathbb{E}h(Y)| \leq \sup_{f \in \mathcal{F}(\mathcal{A}_X) \cap \mathcal{F}(\mathcal{A}_Y)} |\mathbb{E}\mathcal{A}_X f(X) - \mathbb{E}\mathcal{A}_Y f(X)|.$$

If $\mathcal{H}(\mathcal{I})$ is the set of all Lipschitz-1-functions then the resulting distance is $d_{\mathcal{W}}$, the Wasserstein distance. For examples see for example *Holmes (2004), Eichelsbacher and R. (2008), Döbler (2012)*.

Outline

1. Stein's method
2. A canonical Stein operator
3. Examples
4. Distances between expectations
5. Distance between posteriors
6. Last words

Our set-up

Let $(\mathcal{X}, \mathcal{B}, \mu)$ be a measure space, with $\mathcal{X} \subset \mathbb{R}$.

Let $\mathcal{X}^\star$ be the set of real-valued functions on $\mathcal{X}$.

Let $\mathcal{D} : dom(\mathcal{D}) \subset \mathcal{X}^\star \to im(\mathcal{D})$ be a linear operator and $dom(\mathcal{D}) \setminus \{0\} \neq \emptyset$.

Let $\mathcal{D}^{-1} : im(\mathcal{D}) \to dom(\mathcal{D})$ be the linear operator which sends any $h = \mathcal{D}f$ onto f.

Then

$$\mathcal{D}\left(\mathcal{D}^{-1}h\right) = h$$

for all $h \in im(\mathcal{D})$ whereas, for $f \in dom(\mathcal{D})$,

$$\mathcal{D}^{-1}(\mathcal{D}f)$$

is only defined up to addition with an element of $ker(\mathcal{D})$.

Assumption

There exists a linear operator $\mathcal{D}^\star : dom(\mathcal{D}^\star) \subset \mathcal{X}^\star \to im(\mathcal{D}^\star)$ and a constant $l := l_{\mathcal{X},\mathcal{D}}$ such that

$$\mathcal{D}(f(x)g(x+l)) = g(x)\mathcal{D}f(x) + f(x)\mathcal{D}^\star g(x)$$

for all $(f,g) \in dom(\mathcal{D}) \times dom(\mathcal{D}^\star)$.

Under this assumption, $\mathcal{D}$ and $\mathcal{D}^\star$ are skew-adjoint in the sense that

$$\int_{\mathcal{X}} g\mathcal{D}fd\mu = -\int_{\mathcal{X}} f\mathcal{D}^\star gd\mu$$

for all $(f,g) \in dom(\mathcal{D}) \times dom(\mathcal{D}^\star)$ such that $g\mathcal{D}f \in L^1(\mu)$ or $f\mathcal{D}^\star g \in L^1(\mu)$ and $\int_{\mathcal{X}} \mathcal{D}(f(\cdot)g(\cdot+l))d\mu = 0$.

Example 1

Let μ be the Lebesgue measure on $\mathcal{X} = \mathbb{R}$ and take $\mathcal{D}$ the usual strong derivative. Then

$$\mathcal{D}^{-1}f(x) = \int_{\bullet}^{x} f(u)du,$$

the usual antiderivative. Our assumption

$$\mathcal{D}(f(x)g(x+l)) = g(x)\mathcal{D}f(x) + f(x)\mathcal{D}^{\star}g(x)$$

is satisfied with $\mathcal{D}^{\star} = \mathcal{D}$ and $l = 0$.

Example 2

Let μ be the counting measure on $\mathcal{X} = \mathbb{Z}$ and take $\mathcal{D} = \Delta^+$, the forward difference operator. Then

$$\mathcal{D}^{-1} f(x) = \sum_{k=\bullet}^{x-1} f(k).$$

Also we have the discrete product rule

$$\Delta^+(f(x)g(x-1)) = g(x)\Delta^+ f(x) + f(x)\Delta^- g(x)$$

for all $f, g \in \mathbb{Z}^\star$ and all $x \in \mathbb{Z}$. Hence our assumption

$$\mathcal{D}(f(x)g(x+l)) = g(x)\mathcal{D}f(x) + f(x)\mathcal{D}^\star g(x)$$

is satisfied with $\mathcal{D}^\star = \Delta^-$, the backward difference operator and $l = -1$.

Example 3

Let $\mu(x)$ be the $\mathcal{N}(0,1)$ measure on $\mathbb{R}$, with density φ, and take

$$\mathcal{D}_\varphi f(x) = f'(x) - xf(x) = \frac{(f(x)\varphi(x))'}{\varphi(x)},$$

see e.g. *Ledoux, Nourdin, Peccati (2014)*. Then

$$\mathcal{D}_\varphi^{-1} f(x) = \frac{1}{\varphi(x)} \int_\bullet^x f(y)\varphi(y)dy.$$

Also we have the product rule

$$\mathcal{D}_\varphi(gf)(x) = (gf)'(x) - xg(x)f(x) = g(x)\mathcal{D}_\varphi f(x) + f(x)g'(x).$$

Hence our assumption

$$\mathcal{D}(f(x)g(x+l)) = g(x)\mathcal{D}f(x) + f(x)\mathcal{D}^\star g(x)$$

is satisfied with $\mathcal{D}^\star g = g'$ and $l = 0$.

Example 4

Let $\mu(x)$ be the Poisson(λ)measure on $\mathbb{Z}^+$ with pmf γ_λ and

$$\Delta^+_\lambda f(x) = \lambda f(x+1) - xf(x) = \frac{\Delta^+(f(x)x\gamma_\lambda(x))}{\gamma_\lambda(x)}.$$

Then

$$(\Delta^+_\lambda)^{-1} f(x) = \frac{1}{x\gamma_\lambda(x)} \sum_{k=\bullet}^{x-1} f(k)\gamma_\lambda(k)$$

(which is ill-defined at $x = 0$) and

$$\Delta^+_\lambda(g(x-1)f(x)) = g(x)\Delta^+_\lambda f(x) + f(x)x\Delta^- g(x).$$

Hence our assumption

$$\mathcal{D}(f(x)g(x+l)) = g(x)\mathcal{D}f(x) + f(x)\mathcal{D}^\star g(x)$$

is satisfied with $\mathcal{D}^\star g(x) = x\Delta^- g(x)$ and $l = -1$.

Remark

In all examples the choice of $\mathcal{D}$ is, in a sense, arbitrary and other options are available. Less conventional choices of $\mathcal{D}$ can be envisaged (even forward differences in the continuous setting, etc.).

The restriction to dimension 1 is not necessary.

From now for the sake of presentation on we concentrate on the Lebesgue measure and $\mathcal{D}$ the usual derivative.

A canonical Stein operator

Let X be a continuous random variable distribution having pdf p with interval support $\mathcal{I} = [a, b] \subset \mathbb{R}$.

We define the *Stein class* of X as the class $\mathcal{F}(X)$ of functions $f : \mathbb{R} \to \mathbb{R}$ such that
(i) $x \mapsto f(x)p(x)$ is differentiable on $\mathbb{R}$
(ii) $(fp)'$ is integrable and $\int (fp)' = 0$.

To X we associate the *Stein operator* $\mathcal{T}_X$ of X such that

$$\mathcal{T}_X f = \frac{(fp)'}{p}$$

with the convention that $\mathcal{T}_X f = 0$ outside of $\mathcal{I}$.

A useful relationship

We have a distributional characterisation:

$$Y \overset{\mathcal{D}}{=} X \text{ if and only if } (\mathcal{T}_Y, \mathcal{F}(Y)) = (\mathcal{T}_X, \mathcal{F}(X))$$

for all random variables Y which have the same support as X. See *Ley and Swan (2011)* for more details.

By the product rule,

$$\mathbb{E}\left[g'(X)f(X)\right] = -\mathbb{E}\left[g(X)\mathcal{T}_X f(X)\right]$$

for all $f \in \mathcal{F}(X)$ and for all differentiable functions g such that $\int (gfp)' dx = 0$, and $\int |g' fp| dx < \infty$; we say that $g \in dom((\cdot)', X, f)$.

Stein characterisations

Let Y be continuous with density q, and same support as X.

1. $$Y \overset{\mathcal{D}}{=} X \text{ if and only if } \mathbb{E}\left[f(Y)g'(Y)\right] = -\mathbb{E}\left[g(Y)\mathcal{T}_X f(Y)\right]$$

   for all $f \in \mathcal{F}(X)$ and for all $g \in dom((\cdot)', X, f)$.

2. Suppose that $\frac{q}{p}$ is differentiable. Take $g \in \cap_{f \in \mathcal{F}(X)} dom((\cdot)', X, f)$ such that g is X-a.s. never 0 and $g\frac{q}{p}$ is differentiable. Then

   $$Y \overset{\mathcal{D}}{=} X \text{ if and only if } \mathbb{E}\left[f(Y)g'(Y)\right] = -\mathbb{E}\left[g(Y)\mathcal{T}_X f(Y)\right]$$

   for all $f \in \mathcal{F}(X)$.

3. Let $f \in \mathcal{F}(X)$ be X-a.s. never zero and assume that $dom((\cdot)', X, f)$ is dense in $L^1(X)$. Then

   $$Y \overset{\mathcal{D}}{=} X \text{ if and only if } \mathbb{E}\left[f(Y)g'(Y)\right] = -\mathbb{E}\left[g(Y)\mathcal{T}_X f(Y)\right]$$

   for all $g \in dom((\cdot)', X, f)$.

Some special cases

Take $g \equiv 1$ (this is always permitted) to obtain the Stein characterization

$$Y \overset{\mathcal{D}}{=} X \text{ if and only if } \mathbb{E}\left[\mathcal{T}_X f(Y)\right] = 0 \text{ for all } f \in \mathcal{F}(X).$$

If $f \equiv 1$ is in $\mathcal{F}(X)$ then we obtain the Stein characterization

$$Y \overset{\mathcal{D}}{=} X \iff \mathbb{E}[g'(Y)] = -E\left[\frac{p'(Y)}{p(Y)} g(Y)\right] = 0 \text{ for all } g \in dom((\cdot)', X, 1).$$

A connection to couplings: an equation

Let X be a mean zero random variable with finite, nonzero variance σ^2. We say that X^* has the X-zero biased distribution if for all differentiable f for which $\mathbb{E}Xf(X)$ exists,

$$\sigma^2 \mathbb{E} f'(X^*) - \mathbb{E} X f(X) = 0;$$

$\mathcal{N}(0, \sigma^2)$ is the unique fixed point of the zero-bias transformation.

More generally, if X is a random variable with differentiable density $p_X =$ then for all differentiable f,

$$p_X(x) \mathcal{T}_X(f)(x) = (f(x) p_X(x))' = p_X(x) f'(x) + f(x) p_X'(x)$$

and so

$$\mathbb{E}\left\{f'(X)\right\} + \mathbb{E}\left\{f(X) \frac{p_X'(X)}{p_X(X)}\right\} = 0.$$

A connection to couplings: a transformation

The equation

$$\mathbb{E}\left\{f'(X)\right\} + \mathbb{E}\left\{f(X)\frac{p_X'(X)}{p_X(X)}\right\} = 0$$

could lead to a transformation which maps a random variable Y to $Y^{(X)}$ such that for all differentiable $f \in$ for which the expressions exist,

$$\mathbb{E}f'(Y^{(X)}) = -\left\{\mathbb{E}f(Y)\frac{p_X'(Y)}{p_X(Y)}\right\}.$$

A connection to couplings: unique fixed points

Now assume that $f \in \mathcal{F}(X) \cap dom(\mathcal{D})$ is dense in $L^1(X)$. and that $Y^{(X)}$ is well-defined. To see that $Y =_d X$ if and only if $Y^{(X)} =_d Y$:

As for all $f \in \mathcal{F}(X)$,

$$\mathbb{E}\left\{f'(X)\right\} + \mathbb{E}\left\{f(X)\frac{p_X'(X)}{p_X(X)}\right\} = 0$$

and

$$\mathbb{E}f'(Y^{(X)}) = -\left\{\mathbb{E}f(Y)\frac{p_X'(Y)}{p_X(Y)}\right\},$$

if $Y =_d X$ then $Y^{(X)} =_d Y$.

If $Y^{(X)} =_d Y$, then $\mathbb{E}\mathcal{T}_X(f)(Y) = 0$ for all differentiable $f \in \mathcal{F}(X)$, and the assertion follows from the density assumption and using $g = 1$ in

$$Y \stackrel{\mathcal{D}}{=} X \text{ if and only if } \mathbb{E}\left[f(Y)g'(Y)\right] = -\mathbb{E}\left[g(Y)\mathcal{T}_X f(Y)\right]$$

for all $f \in \mathcal{F}(X)$.

The inverse Stein operator

With X as above we define the class

$$\mathcal{F}^{(0)}(X) = \left\{h : \mathbb{R} \to \mathbb{R} \text{ such that } E\left[h(X)\right] = 0\right\};$$

and the *inverse Stein operator* $\mathcal{T}_X^{-1} : \mathcal{F}^{(0)}(X) \to \mathcal{F}(X)$ as

$$\mathcal{T}_X^{-1} h(x) = -\frac{1}{p(x)} \int_a^x p(y)h(y)dy = \frac{1}{p(x)} \int_x^b p(y)h(y)dy$$

for all $h \in \mathcal{F}^{(0)}(X)$.

Stein equations

Let $h \in L^1(X)$. The equation

$$h(x) - \mathbb{E}h(X) = f(x)g'(x) + g(x)\mathcal{T}_X f(x), \quad x \in \mathcal{I},$$

is a *Stein equation for the target* X.

Solutions of this equation are *pairs* of functions (f, g) such that

$$fg = \mathcal{T}_X^{-1}(h - \mathbb{E}_p h).$$

Although fg is unique, the individual f and g are not (just consider multiplication by constants).

Stein equations

Let $h \in L^1(X)$. The equation

$$h(x) - \mathbb{E}h(X) = \mathcal{T}_X\left(fg\right)(x) = f(x)g'(x) + g(x)\mathcal{T}_X f(x), \quad x \in \mathcal{I},$$

is a *Stein equation for the target* X.

Solutions of this equation are *pairs* of functions (f, g) such that

$$fg = \mathcal{T}_X^{-1}(h - \mathbb{E}_p h).$$

Although fg is unique, the individual f and g are not (just consider multiplication by constants).

Special Stein operators

Our general Stein operator is an operator on pairs of functions (f,g);

$$\mathcal{A}(f,g)(x) = \mathcal{T}_X(fg)(x) = f(x)g'(x) + g(x)\mathcal{T}_X f(x).$$

A second particular Stein operator fixes a differentiable g and uses

$$\mathcal{A}_X f = \mathcal{T}_X\left(fg\right) = fg' + g\mathcal{T}_X f$$

and $f \in \mathcal{F}_{\mathcal{A}}(X) \subset \mathcal{F}(X)$.

A particular Stein operator is given by fixing $f = c \in \mathcal{F}(X)$ and using

$$\mathcal{A}_X g(x) = c(x)g'(x) + g(x)\mathcal{T}_X c(x).$$

Sometimes we call this the c-operator (see *Goldstein and R. (2013)*).

The score function

Suppose that X is such that the constant function $1 \in \mathcal{F}(X)$ (this is no small assumption). Then taking $c = 1$ in

$$\mathcal{A}_X g(x) = c(x) g'(x) + g(x) \mathcal{T}_X c(x).$$

we get

$$\mathcal{A}_X g(x) = g'(x) + g(x) \rho(x)$$

with

$$\rho(x) = \mathcal{T}_X 1(x) = \frac{p'(x)}{p(x)}$$

the so-called "score function" of X; see for example *Stein (2004)*.

The Stein kernel

If X has finite mean ν we can take $c = \mathcal{T}_X^{-1}(\nu - Id)$ with Id the identity function (this is always allowed) in

$$\mathcal{A}_X g(x) = c(x)g'(x) + g(x)\mathcal{T}_X c(x).$$

This yields

$$\mathcal{A}_X g(x) = \tau(x)g'(x) + (\nu - x)g(x)$$

with

$$\tau = \mathcal{T}_X^{-1}(\nu - Id)$$

what we call the “Stein kernel of X”. This approach was used for example in *Stein (1986, Lesson 6)* and *Cacoullos et al. (1992)*.

Outline

1 Stein's method

2 A canonical Stein operator

3 Examples

4 Distances between expectations

5 Distance between posteriors

6 Last words

Example: Normal

In the example of a $\mathcal{N}(0, \sigma^2)$ random variable, our operator translates to

$$\mathcal{T}_N f(x) = -f'(x) + \frac{1}{\sigma^2} x f(x)$$

which contrasts with

$$\sigma^2 f'(x) - x f(x),$$

the standard Stein operator for this case. The score function is $-\frac{x}{\sigma^2}$. We compute the Stein kernel

$$\tau(x) = \sigma^2.$$

The c-Stein operator is the standard operator.

Example: Beta

Consider beta distributions with density

$$p(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}\mathbf{1}_{\{x\in[0,1]\}}.$$

Here

$$\mathcal{T}_B f(x) = -f'(x) - \frac{1}{x(1-x)}f(x)((\alpha-1)x-(\beta-1)(1-x)).$$

The standard Stein operator for this case is

$$\mathcal{A}f(x) = x(1-x)f'(x) + (\alpha(1-x)-\beta x)f(x),$$

see *Doebler (2012)*. The score function, defined when $\alpha > 1$ and $\beta > 1$, is

$$\rho(x) = \frac{\alpha}{x} - \frac{\beta-1}{x-1}.$$

The beta Stein kernel is

$$\tau(x) = \frac{x(1-x)}{\alpha+\beta}.$$

Outline

1. Stein's method
2. A canonical Stein operator
3. Examples
4. Distances between expectations
5. Distance between posteriors
6. Last words

Comparison of expectations

Let X_1 and X_2 be such that their densities p_1 and p_2 have interval support.

Denote by $\mathcal{T}_1$ and $\mathcal{T}_2$ the Stein operators associated with X_1 and X_2, acting on Stein classes $\mathcal{F}_1 = \mathcal{F}(X_1)$ and $\mathcal{F}_2 = \mathcal{F}(X_2)$. Let $\mathbb{E}_i h = \mathbb{E}h(X_i), i = 1, 2$.

Let h be such that $\mathbb{E}_i|h| < \infty$ for $i = 1, 2$. Fix $f_1 \in \mathcal{F}_1$ and define

$$g_h := \frac{1}{f_1}\mathcal{T}_1^{-1}(h - \mathbb{E}_1 h).$$

Then, for all $f_2 \in \mathcal{F}_2$ such that $|\mathbb{E}_2 f_2 g_h|$ exists,

$$\mathbb{E}_2 h - \mathbb{E}_1 h = \mathbb{E}_2\left[f_1 g_h' + g_h \mathcal{T}_1 f_1\right] = \mathbb{E}_2\left[(f_1 - f_2)g_h' + g_h\left\{\mathcal{T}_1 f_1 - \mathcal{T}_2 f_2\right\}\right].$$

An obvious choice is $f_1 = f_2$ if permitted, leading to ...

A Corollary

Assume that $\mathcal{X}_1 = \mathcal{X}_2$. Let $\mathcal{H} \subset L^1(X_1) \cap L^1(X_2)$. Take $f \in \mathcal{F}_1 \cap \mathcal{F}_2$ and suppose that for all $h \in \mathcal{H}$ we have that $g_h = (1/f)\mathcal{T}_1^{-1}(h - \mathbb{E}_1 h)$ is such that $|\mathbb{E}_i f g_h|$ exists, $i = 1, 2$. Then

$$\sup_{h \in \mathcal{H}} |\mathbb{E}_1 h - \mathbb{E}_2 h| \leq \kappa_{\mathcal{H},1}(f) \mathbb{E}_2 |\mathcal{T}_1 f - \mathcal{T}_2 f|$$

with $\kappa_{\mathcal{H},1}(f) = \sup_{h \in \mathcal{H}} \|(1/f)\, \mathcal{T}_1^{-1}(h - \mathbb{E}_1 h)\|_\infty$.

Example: Distance between Gaussians via Stein factor

For $X_i \sim \mathcal{N}(0, \sigma_i^2)$, $i = 1, 2$, with $\sigma_1^2 \leq \sigma_2^2$ the usual Stein operator gives

$$\mathbb{E}h(X_1) - \mathbb{E}h(X_2) = (\sigma_1^2 - \sigma_2^2)\mathbb{E}f'_{h,\sigma_2}(X_1)$$

with f_{h,σ_2} the solution of the Gaussian Stein equation, yielding

$$d_{\mathrm{TV}}(X_1, X_2) \leq 2\frac{|\sigma_1^2 - \sigma_2^2|}{\sigma_2^2},$$

see for example *Nourdin and Peccati (2011)*. This is also what we get when we use the Stein kernel approach, with $\tau_i(x) = \sigma_i^2$.
Using the score functions $\rho_i(x) = -x/\sigma_i^2$, $i = 1, 2$ we get

$$d_{\mathrm{TV}}(X_1, X_2) \leq \sigma_1 \sqrt{\frac{\pi}{2}}\mathbb{E}|X_2| \left| \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right| = \frac{|\sigma_1^2 - \sigma_2^2|}{\sigma_1 \sigma_2};$$

if $\sigma_1 < \frac{\sigma_2}{2}$ then this bound beats the first bound.

From Student to Gauss

Set $X_1 = Z$ standard Gaussian and $X_2 = W_\nu$ a Student t random variable with $\nu > 2$ degrees of freedom. The Stein kernels for both distributions are $\tau_1 = 1$ and $\tau_2(x) = \frac{x^2+\nu}{\nu-1}$; we obtain using the Stein kernel approach that

$$d_{\mathrm{TV}}(Z, W_\nu) \leq 2\mathbb{E}\left|\frac{W_\nu^2+\nu}{\nu-1} - 1\right| \leq \frac{4}{\nu-2}.$$

Using the score function approach we obtain

$$d_{\mathrm{TV}}(Z, W_\nu) \leq \sqrt{\frac{\pi}{2}}\frac{-2+8\left(\frac{\nu}{1+\nu}\right)^{(1+\nu)/2}}{(\nu-1)\sqrt{\nu}B(\nu/2, 1/2)},$$

which is of the same order, with a better constant.

Outline

1 Stein's method

2 A canonical Stein operator

3 Examples

4 Distances between expectations

5 Distance between posteriors

6 Last words

Distances between likelihoods

We can apply the inequality

$$\sup_{h \in \mathcal{H}} |\mathbb{E}_1 h - \mathbb{E}_2 h| \leq \kappa_{\mathcal{H},1}(f) \mathbb{E}_2 |\mathcal{T}_1 f - \mathcal{T}_2 f|$$

to bound distances between likelihoods:
Let $\pi_i(\theta)$, $i = 0, 1$, be two absolutely continuous positive functions with common support $\mathcal{I}$ with closure $\bar{\mathcal{I}} = [a, b]$. Assume that π_1 and $\pi_0 \pi_1$ are integrable. Define

$$p_1(\theta) = \kappa_1 \pi_1(\theta) \text{ and } p_2(\theta) = \kappa_2 \pi_0(\theta) \pi_1(\theta)$$

where κ_i, $i = 1, 2$, are normalising constants. For $i = 1, 2$ let $\Theta_i \sim p_i$. Assume that $\mu_1 = \mathbb{E}[\Theta_1] < \infty$.

Further we assume that

$$\lim_{\theta \to b} \pi_0(\theta) \int_\theta^b (\mu_1 - u) \pi_1(u) du = \lim_{\theta \to a} \pi_0(\theta) \int_a^\theta (\mu_1 - u) \pi_1(u) du = 0.$$

A Wasserstein bound

Using the Stein kernel $\tau_1 = \mathcal{T}_X^{-1}(\mu_1 - Id)$ we get

$$\mathcal{T}_2\tau_1(\theta) = \frac{\pi_0'(\theta)}{\pi_0(\theta)}\tau_1(\theta) + \mathcal{T}_1\tau_1(\theta)$$

and so

$$\mathcal{T}_2\tau_1(\theta) - \mathcal{T}_1\tau_1(\theta) = \frac{\pi_0'(\theta)}{\pi_0(\theta)}\tau_1(\theta).$$

We obtain

$$d_{\mathcal{W}}(\Theta_1, \Theta_2) \leq \frac{\kappa_2}{\kappa_1}\mathbb{E}\left|\pi_0'(\Theta_1)\tau_1(\Theta_1)\right|.$$

The Bayesian approach in a nutshell

Given observations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ which are seen as realisations of random variables $X_1, \ldots, X_n$ with joint distribution (density)

$$\pi_1(x_1, x_2, \ldots, x_n | \theta),$$

which depends on an unknown parameter θ, we would like to draw inference on θ.

The parameter Θ is viewed as a random element. Before any observation has been made (a priori) we think that Θ has the (prior) distribution p_0. We update our belief on Θ in light of the observations by applying Bayes' formula, so that the posterior density of Θ, given the observations $\mathbf{y}$, is

$$p_2(\theta|\mathbf{x}) = \pi_1(\mathbf{x}|\theta) p_0(\theta) = \kappa_1(\mathbf{x}) p_1(\theta, \mathbf{x}) p_0(\theta).$$

Here $p_1(\theta, \mathbf{x})$ is a probability density for θ.

The choice of prior

Some typical choices of prior are

- Priors which are elicited from experts or from previous experiments;
- Conjugate priors, where the posterior belongs to the same distributional family as the prior, making updating easy;
- Uniform distribution (when it exist) to reflect no information;
- Jeffreys' prior $p_0(\theta) \propto |I(\theta)|^{\frac{1}{2}}$, the determinant of the Fisher information matrix;
- priors which are adapted to particular problems.

The choice of prior affects the inference, but the hope is that the effect of the prior wanes with increasing number of observations. We can quantify this effect, using Stein's method.

Bayesian interpretation

We observe data points $x := (x_1, x_2, \ldots, x_n)$ with sampling distribution $\pi_1(x \,|\, \theta)$. We take θ, the one dimensional parameter, to be distributed according to some (possibly improper) prior $p_0(\theta)$, and let the posterior be given by $p_2(\theta; x) \propto p_0(\theta) p_1(\theta; x)$. Set

$$p_1(\theta; x) = \kappa_1(x) \pi_1(x; \theta)$$

and

$$p_2(\theta; x) = \kappa_2 \pi_0(\theta) \pi_1(x, \theta).$$

Then our theorem applies and we can assess the influence of the prior on the posterior.

Example: Normal model, normal prior

Assume that $X_1, \ldots, X_n \sim \mathcal{N}(\theta, \sigma^2)$, conditionally independent given θ, where σ^2 is known, and assume that the prior is normal, $\pi(\theta) \sim \mathcal{N}(\mu, \delta^2)$, where μ and δ^2 are known. Then

$$\begin{aligned}\pi_1(x_1, \ldots, x_n, \theta) &= (2\pi\sigma^2)^{-\frac{n}{2}} exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2}\right\};\\ p_1(\theta) &\sim \mathcal{N}\left(\frac{1}{n}\sum_{i=1}^{n}x_i, \frac{\sigma^2}{n}\right).\end{aligned}$$

It is a standard calculation that the posterior is normal,

$$p_2(\theta, x) \sim \mathcal{N}\left(\frac{b(x)}{a}, \frac{1}{a}\right).$$

with

$$a = \frac{n}{\sigma^2} + \frac{1}{\delta^2}, \text{ and } b(x) = \frac{1}{\sigma^2}\sum x_i + \frac{\mu}{\delta^2}.$$

The resulting bound

Wit h $\tau_1 = \frac{\sigma^2}{n}$ we find

$$\begin{aligned} d_{\mathcal{W}}(\Theta_1, \Theta_2) &\leq \mathbb{E}_2 \left| \frac{\pi_0'(\theta)}{\pi_0(\theta)} \tau_1(\theta) \right| \\ &= \frac{\sigma^2}{n\delta^2} \mathbb{E} \left| \Theta_2 - \mu \right| \\ &\leq \frac{\sigma^2}{n\delta^2} \left(\mathbb{E} \left| \Theta_2 - \frac{b(x)}{a} \right| + \left| \frac{b(x)}{a} - \mu \right| \right) \\ &= \frac{\sqrt{2}}{\sqrt{\pi}} \frac{\sigma^3}{n\delta\sqrt{\delta^2 n + \sigma^2}} + \frac{\sigma^2}{n\delta^2 + \sigma^2} \left| \bar{x} - \mu \right| . \end{aligned}$$

The first term is order $O(n^{-1})$ whereas the second term reflects the influence of the data. The bound decreases when δ increases. The better the guess of μ, the smaller the bound.

Example: Binomial model, Beta prior

Here we have one observation $x \sim \mathit{Binomial}(n, \theta)$, with known n. The prior is

$$\pi_0 = \kappa_0 \theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \theta \in [0,1],$$

with $\alpha > 0$ and $\beta > 0$. Then $\tau_1(\theta) = \frac{\theta(1-\theta)}{n+2}$. A direct computation gives

$$d_{\mathcal{W}}(\Theta_1, \Theta_2) \leq \frac{1}{n+2}\left(|2-\beta-\alpha|\frac{\alpha+x}{\alpha+\beta+n} + |\alpha-1|\right).$$

- Unless $\alpha = 1$ the bound will be of order $1/n$ no matter how favourable x is.
- If $\alpha = 1$ but $\beta \neq 1$ then the bound is smallest when $x = 0$, and is then of order $1/n^2$.
- If $\alpha = 1 = \beta$ then the bound is zero, as it should be as then $p_1 = p_2$, the prior is uniform.

Example: Binomial model, non-informative prior

Using the Haldane prior $p_0(\theta) = \kappa_0(\theta(1-\theta))^{-1}$, direct computation gives

$$d_{\mathcal{W}}(\Theta_1, \Theta_2) \leq \frac{2}{n+2}\left(\left|\frac{x}{n} - \frac{1}{2}\right| + \sqrt{\frac{x(n-x)}{n^2(n+1)}}\right).$$

If $x = \frac{n}{2}$ then the bound is of order $n^{-\frac{3}{2}}$.

Using Jeffreys' prior $p_0(\theta) = \kappa_0(\theta(1-\theta))^{-\frac{1}{2}}$, direct computation gives

$$d_{\mathcal{W}}(\Theta_1, \Theta_2) \leq \frac{1}{n+2}\left(\left|\frac{x+\frac{1}{2}}{n+1} - \frac{1}{2}\right| + \sqrt{\frac{(x+\frac{1}{2})(n-x+\frac{1}{2})}{(n+1)^2(n+2)}}\right).$$

Again if $x = \frac{n}{2}$ then the bound is of order $n^{-\frac{3}{2}}$.

Outline

1. Stein's method
2. A canonical Stein operator
3. Examples
4. Distances between expectations
5. Distance between posteriors
6. Last words

Last remarks

Stein (1964) gives bounds between posteriors in the Kakutani distance, using an algebraic approach.

When $\mathcal{D} \neq \mathcal{D}^*$ the Stein operator becomes

$$\mathcal{A}(f, g)(x) = f(x)\mathcal{D}^* g(x) + g(x)\mathcal{T}_X f(x).$$

The flexibilty in having a pair of functions rather than just one function can be useful for getting bounds; for example we could choose $g(x) = x^\alpha$ and then mimimise our bounds with respect to α.

We are thinking about the multivariate case, too.