

# Notes on Stein Method

Carlos C. Rodríguez

March 9, 2020

## 1 Sourav Chatterjee 2004 Trick

Assume  $(X, X')$   $\mu$ -exchangeable variables, i.e.

$$\mu\{X \in A, X' \in B\} = \mu\{X' \in A, X \in B\}$$

every function splits as,

$$h(X) = \frac{1}{2}[h(X) + h(X')] + \frac{1}{2}[h(X) - h(X')]$$

i.e.,  $h = h^+ + h^- = \text{sym.} + \text{antisym.}$  and we have,

$$\mu h = \int h(x) d\mu = \mu h^+ \text{ since } \mu h^- = 0.$$

Given,

$$F(X, X') = -F(X', X) \text{ antisymmetric.}$$

Let

$$f(X) = \mu_X F = E[F(X, X')|X]$$

and thus,

$$\mu h f = \mu h F = \mu h^- F \text{ since } \mu h^+ F = 0.$$

**Theorem 1 (Chatterjee 2004.)** *Just set  $h = f$  above to get,*

1. *The Trick!*

$$\mu f^2 = \mu f^- F.$$

2. *Hoeffding++*

$$\mu_X |f^- F| \leq C \Rightarrow \mu e^{\theta f} \leq e^{C\theta^2/2} \text{ for all } \theta \in \mathbb{R}$$

and

$$\mu\{|f| > t\} \leq 2e^{-t^2/2C} \text{ for all } t \geq 0.$$

*Proof.* It follows the proof of Hoeffding's theorem very closely. Let  $\theta \in R$  and define  $h(X) = e^{\theta f(X)}$ . The m.g.f. of  $f(X)$  is,

$$m(\theta) = \mu e^{\theta f} = \mu h$$

and The Trick gives,

$$|m'(\theta)| = |\mu h f| = |\mu h^- F| \leq \mu |h^- F|.$$

The convexity of  $x \rightarrow e^x$  produces the well known inequality,

$$\left| \frac{e^x - e^y}{x - y} \right| \leq \frac{1}{2}(e^x + e^y)$$

that translates into,

$$|h^-| \leq h^+ |\theta f^-|$$

thus,

$$|m'(\theta)| \leq |\theta| \mu h^+ |f^- F| = |\theta| \mu h |f^- F|$$

since  $h^- |f^- F|$  is antisymmetric. Hence,

$$|m'(\theta)| \leq |\theta| C m(\theta).$$

Just solve for  $m(\theta)$  and optimize (just like in Hoeffding's theorem) to get,

$$m(\theta) \leq e^{C\theta^2/2}$$

and,

$$\mu\{|f| > t\} \leq 2e^{-t^2/2C} \blacksquare$$

Even though the proof is sort of trivial (which is great!), the freedom of choosing the pair  $(X, X')$  of exchangeable variables and the antisymmetric function  $F$ , makes Chatterjee's beautiful trick invaluable in situations involving functions of dependent variables, e.g. Hamiltonians for ferromagnetic models.

## 1.1 Canonical Case

Let  $Y_1, Y_2, \dots, Y_n$  be statistical variables not necessarily independent nor with the same distribution. Suppose  $(Y_i, Y'_i)$  exchangeable for  $i = 1, \dots, n$ . Define the vectors  $Y$  and  $Y'$  by replacing a randomly chosen entry  $Y_I$  from  $Y$  with  $Y'_I$ , i.e.

$$Y = \sum_{i=1}^n Y_i e^i$$

$$Y' = Y + (Y'_I - Y_I) e^I$$

where  $I \sim \text{Unif}\{1, 2, \dots, n\}$  independently of everything and let,

$$\begin{aligned} X &= f(Y) \\ X' &= f(Y') \end{aligned}$$

“any” function  $f$ . Then,  $(X, X')$  are exchangeable and,  $F = u^-$  for “any” function  $u(X)$  will do provided  $\mu_X X' = X + o(1)$  as  $n \rightarrow \infty$ . For example if we demand  $f$  to have  $o(1)$  bounded differences,

$$|f(Y) - f(Y + te^i)| \leq o(1) \text{ for all } i \text{ and all } t \in R$$

the theorem above will produce a generalization of McDiarmid’s without the need for independence.

## 2 Stein Method

Chatterjee’s thesis demystifies Stein method to a high degree by showing that it proves a measure  $\mu$  must be close to a target measure  $\mu_0$  when their *defining* operators  $T$  and  $T_0$  are close to each other. Stein calls  $T$  a defining operator for  $\mu$  by a short exact sequence:

$$0 \longrightarrow L^2(\lambda) \xrightarrow{T} L^2(\mu) \xrightarrow{\mu} 0$$

exact means  $\text{Ker } \mu = \text{Im } T$  so that  $\mu Th = 0$  for  $h \in L^2(\lambda)$ . We assume  $\mu \ll \lambda$ .

If  $T$  fixes  $\mu$  and  $T_0$  fixes  $\mu_0$ ,  $T$  is close to  $T_0$  when for all  $x$ ,

$$|Th(x) - T_0h(x)| \leq \epsilon_h$$

with  $\epsilon_h \approx 0$  independent of  $x$ . Hence, if there are  $h$  and  $g$  such that,

$$T_0h = (1 - \mu_0)g$$

we have,

$$\begin{aligned} |\mu g - \mu_0 g| &= |\mu T_0h - \mu Th| \\ &= |\mu(T_0h - Th)| \\ &\leq \int |T_0h(x) - Th(x)| d\mu \\ &\leq c \cdot \epsilon_h \end{aligned}$$

yes,  $c$  since  $\mu$  is not normalized.

Thus,  $Th(X) = \mu_X h^-$  has what we need:  $\mu Th = 0$ . Again, provided  $X'$  is a small perturbation of  $X$ , i.e.  $\mu_X X' = X + o(1)$ . Stein method produces Berry-Esseen type of inequalities for a variety of target distributions including: gaussian, beta, gamma and poisson. Almost all of these results are for univariate (i.e.  $x \in R$ ) distributions.

## 2.1 The Choice of $T_0$

The starting point for Stein was the gaussian:  $g_0(x) = e^{-x^2/2}$  that satisfies  $T_0^\dagger g_0(x) = g_0'(x) + xg_0(x) = 0$ . Thus, the adjoint operator

$$T_0 = x - \frac{d}{dx}$$

works since  $x \in R$ . Notice that  $T_0^\dagger$  fixes  $1/g_0 = \exp(x^2/2)$  but not  $g_0$ . Let  $\mu \ll \lambda$  with density  $g(x)$ . If for any  $h$

$$0 = \mu T_0 h = \int_R (xh(x) - h'(x))g(x) d\lambda$$

then integration by parts conjugates the operator,

$$0 = \int_R (xg(x) + g'(x))h(x) d\lambda$$

so by letting  $h(x) = xg(x) + g'(x)$  in the last equation above, we must have  $T_0^\dagger g = 0$  and thus,  $g = g_0$  which shows that  $T_0$  indeed fixes  $g_0$ . However, this first order differential operator does not generalize to  $x \in R^n$ .

The only meaningful scalar operator on Riemannian manifolds (e.g.  $R^n$  and  $R$  also) is the Laplacian (check out the *little giant* blue book by S. Rosenberg, *The Laplacian on a Riemannian Manifold*).

The  $T_0$  above can be thought of a square root of what's needed. Notice that

$$T_0^\dagger T_0 g(x) = \left( -\frac{d^2}{dx^2} + x^2 - 1 \right) g(x) \neq \left( -\frac{d^2}{dx^2} + x^2 \right) g(x)$$

and now  $T_0^\dagger T_0 g = (H - 1)g = 0$  does fix  $g = g_0$  by finding the eigen function of the ground state of the self-adjoint Hamiltonian  $H = H^\dagger$  of the quantum harmonic oscillator. In this way we hook up Stein to over 200 years of Fourier, Sturm-Louville, Quantum Mechanics and Banach Algebras.

### The Laplacian

Given  $(M, g)$  Riemannian manifold with metric  $g$ , the element of volume is given by the highest dimensional differential form on  $M$ , which in local coordinates equals,

$$d\text{vol} = \sqrt{\det g} dx^1 \wedge \dots \wedge dx^n$$

provided  $(\partial_{x^1}, \dots, \partial_{x^n})$  form a positively oriented basis of vector fields on the tangent bundle. We have,

$$\text{vol}(M) = \int_M d\text{vol}(x).$$

We can then define  $L^2(M)$  the space of scalar functions  $f$  on  $M$  such that,

$$\int_M |f(x)|^2 \operatorname{dvol}(x) < \infty$$

which inner product,

$$\langle f, g \rangle = \int_M f(x)g(x) \operatorname{dvol}(x)$$

and for vector fields  $X, Y$ ,

$$\langle X, Y \rangle = \int_M g(X, Y) \operatorname{dvol}.$$

Integration by parts shows that the gradient of a scalar function  $f$ ,

$$\nabla f = g^{ij} \partial_i f \partial_j$$

is the (formal) adjoint to  $-\operatorname{div}$  of a vector field  $X$ ,

$$\operatorname{div} X = \partial_i X^i$$

since,

$$\langle -\operatorname{div} X, f \rangle = \langle X, \nabla f \rangle.$$

Finally the Laplacian is defined by,

$$\Delta = -\nabla^2 = -\operatorname{div} \circ \nabla = \langle \nabla^\dagger, \nabla \rangle = \|\nabla\|_{L^2(M)}^2.$$

In local coordinates,

$$\Delta = -\frac{1}{\sqrt{\det g}} \partial_j \left( g^{ij} \sqrt{\det g} \partial_i f \right)$$

### Priors and Posteriors

Berry-Esseen and Concentration Inequalities for prior and posterior distributions will require the full Laplacian on the Statistical Manifold with Fisher Information as the metric.