Notes on Stein Method

Carlos Rodriguez

March 9, 2020

1 Sourav Chatterjee 2004 Trick

Assume (X, X') μ-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., $h = h^+ + h^-$ = sym. + 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,

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

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

1. *The Trick!*
$$\mu f^2 = \mu f^- F$$
2. *Hoeffding++*
$$\mu_X |f^- F| \leq C \Rightarrow \mu e^{\theta f} \leq e^{C\theta^2/2} \text{ for all } \theta \in R$$
and
$$\mu\{|f| > t\} \leq 2e^{-t^2/2C} \text{ for all } t \geq 0$$

Proof. If 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 $x \to e^x$ 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,

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

□

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 Cannonical Case

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

$$Y = \sum_i^n Y_i e^i$$
$$Y' = Y + (Y_I' - Y_I) e^I$$

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

$$X = f(Y)$$
$$X' = f(Y')$$

"any" function f. Then (X, X') exchangeable and, $F = u^-$ for "any" function $u(X)$ will do provided $\mu_X X' = X + o(1)$ as $n \to \infty$. 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 above theorem 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 μ must be close to a target measure μ_0 when their *defining* operators T and T_0 are close to each other.

Stein calls T a defining operator for μ 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 $\mu T h = 0$ for $h \in L^2(\lambda)$. We assume $\mu << \lambda$.

If T fixes μ and T_0 fixes μ_0, T is close to T_0 when for all x

$$|T\,h(x) - T_0\,h(x)| \leq \epsilon_h$$

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

$$T_0\,h = (1 - \mu_0)\,g$$

we have,

$$\begin{aligned} |\mu\,g - \mu_0\,g| &= |\mu\,T_0\,h - \mu\,T\,h| \\ &= |\mu(T_0\,h - T\,h)| \\ &\leq \int |T_0\,h(x) - T\,h(x)|\,d\mu \\ &\leq c\epsilon_n \end{aligned}$$

yes, c since μ is not normalized.

Thus, $T\,h(X) = \mu_X\,h^-$ has what we need: $\mu\,T\,h = 0$. 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 $T_0^\dagger g_0(x) = g_0'(x) + x\, g_0(x) = 0$. Thus, the *adjoint* operator

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

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

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

then integration by parts conjugates the operator,

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

so by letting $h(x) = x\, g(x) + g'(x)$, 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 above T_0 can be thought of a square root of what's needed. Notice that

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

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 method to over 200 years of Fourier, Sturn-Louville, Quantum Mechanics, Banach Algebras etc.

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,

$$\mathrm{dvol} = \sqrt{\det g}\, dx^1 \wedge \ldots \wedge dx^n$$

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

$$\mathrm{vol}(M) = \int_M \mathrm{dvol}(x)$$

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

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

with inner product,

$$< f, g >= \int_M \bar{f(x)} g(x) \mathrm{dvol}(x)$$

and for vector fields X, Y,

$$< X, Y >= \int_M g(X, Y) \mathrm{dvol}$$

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 $-$div of a vector field X,

$$\mathrm{div}\, X = \partial_i X^i$$

since,

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

Finally the Laplacian is defined by,

$$\Delta = -\nabla^2 = -\mathrm{div} \circ \nabla =< \nabla^\dagger, \nabla >= ||\nabla||^2_{L^2(M)}$$

In local coordinates,

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

Priors and Posteriors:

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