Estimating transport distances via Stein's method

Max Fathi

LJLL & LPSM, Université de Paris

2 décembre 2021

General framework of Stein's method (1975-1982) : bounding a distance between a probability measure μ and a target measure ν in the form

$$d(\mu, \nu) \leq \sup_{f \in \mathcal{F}} \int Lfd\mu$$

with L a linear operator depending only on ν.

A first example

Consider γ the standard Gaussian measure on $\mathbb{R}$, with density $(2\pi)^{-1/2}\exp(-x^2/2)$. It satisfies the integration by parts formula

$$\int f' - xfd\gamma = 0 \quad \forall f \in C_b^1.$$

The L^1 Wasserstein distance satisfies the duality formula

$$W_1(\mu,\gamma) = \sup_{g1-lip} \int fd\mu - \int fd\gamma.$$

If we solve the ODE

$$f_g' - xf_g = g - \int gd\gamma$$

then

$$W_1(\mu,\gamma) \leq \sup_{g1-lip} \int f_g' - xf_g d\mu.$$

Analyzing solutions to the ODE shows there is a solution f_g which is twice-differentiable, with

$$\max(||f_g||_\infty, ||f_g'||_\infty, ||f_g''||_\infty) \leq 2$$

Therefore we get Stein's lemma for the Gaussian distribution

$$W_1(\mu, \gamma) \leq \sup\left\{ \int f' - xfd\mu ; \max(||f||_\infty, ||f'||_\infty, ||f''||_\infty) \leq 2 \right\}.$$

Application : rate of convergence in the CLT

Let (X_i) be a sequence of independent centered reduced r.v. and μ_n the law of $S_n = n^{-1/2}\sum X_i$. Let I be a uniform random variable in $\{1, ..., n\}$ and $S_n' = S_n - (X_I - X_I')/\sqrt{n}$, where the X_i' are independent copies of the X_i.

Let F be such that $F' = f$. Then by a Taylor expansion of the identity $\mathbb{E}[F(S_n) - F(S_n')] = 0$ we have

$$\left|\mathbb{E}\left[\frac{1}{\sqrt{n}}(X_I' - X_I)f(S_n) + \frac{1}{2n}(X_I' - X_I)^2 f'(S_n)\right]\right| \leq \frac{||f''||_\infty}{2n^{3/2}}.$$

Since $\mathbb{E}[X_I] = \frac{1}{\sqrt{n}} S_n$ and X_I' is independent of the X_i, using $\mathbb{E}[|\sum X_i^2 - n|] \leq \sqrt{\sum \mathbb{E}[X_i^4]}$ we get

$$\mathbb{E}[f'(S_n) - S_n f(S_n)] \leq \frac{||f''||_\infty}{2n^{3/2}} \sum \mathbb{E}[|X_i|^3] + \frac{||f'||_\infty}{n} \sqrt{\sum \mathbb{E}[X_i^4]}.$$

Hence if the X_i have a common bound $\beta \geq 1$ on their fourth moment, we get a Berry-Esseen type bound

$$W_1(\mu_n, \gamma) \leq \frac{3\beta^{3/4}}{\sqrt{n}}.$$

Approximate normality for eigenfunctions

Theorem (E. Meckes 2009)

Let (M, g) be a compact Riemannian manifold with normalized volume measure μ, and let f be an eigenfunction of the Laplacian, with eigenvalue $-\lambda$, and $||f||_{L^2(\mu)} = 1$. Then

$$W_1(\mu \circ f^{-1}, \gamma) \leq \frac{2}{\lambda} \mathrm{Var}(|\nabla f|^2)^{1/2}.$$

Proof : if g is a smooth test function

$$\int fg(f) d\mu = -\frac{1}{\lambda} \int (\Delta f) g(f) d\mu = \frac{1}{\lambda} \int g'(f) |\nabla f|^2 d\mu$$

and $\int |\nabla f|^2 d\mu = \lambda ||f||^2_{L^2(\mu)}$.

The generator approach to Stein's method (Barbour 1991)

We can extend the framework by noting that, up to adding a derivative, we can view the linear operator L as

$$Lf = f'' - xf'$$

which is precisely the generator of the Ornstein-Uhlenbeck process

$$dX_t = -X_t dt + \sqrt{2} dB_t$$

which has invariant reversible measure γ.
If we have a probability measure $\nu_V = e^{-V} dx$, it is the reversible measure for

$$dX_t = -\nabla V(X_t) dt + \sqrt{2} dB_t.$$

We can seek a bound of the form

$$W_1(\mu, \nu_V) \leq \sup_{f \in \mathcal{F}} \int \Delta f - \nabla V \cdot \nabla f d\mu.$$

To implement Stein's method for this problem, we must analyze regularity of solutions to Poisson equations

$$g - \int g d\mu = \Delta f - \nabla V \cdot \nabla f.$$

Many techniques available, either analytic (elliptic regularity, maximum principle...) or probabilistic, via the representation

$$f(x) = -\int_0^\infty \mathbb{E}[g(X_s)|X_0 = x]ds.$$

Typically, regularity properties of f are connected to strong ergodic properties of the flow.

Let's consider the case Hess $V \geq \rho$ with $\rho > 0$. Then if we couple two solutions X_t and Y_t with same Brownian motion and different initial data x and y, we have

$$\frac{d}{dt}|X_t - Y_t|^2 = -2\langle (X_t - Y_y), (\nabla V(X_t) - \nabla V(Y_t))\rangle \leq -2\rho|X_t - Y_t|^2.$$

Hence if g is 1-lipschitz, $x \longrightarrow \mathbb{E}[g(X_t)|X_0 = x]$ is $\exp(-\rho t)$ lipschitz, and the solution f is ρ^{-1}-lipschitz.

Typically, also want second-order regularity, eg. Fang, Shao & Xu 2019.

An example of motivation

Say we are attempting to approximate a probability measure $\mu = \exp(-V)dx$ with a sample $X_1, ..., X_n$. How to decide if the sample is a good approximation ?

If we can successfully implement Stein's method, then

$$W_1(n^{-1} \sum \delta_{X_i}, \mu) \leq \sup_f \frac{1}{n} \sum \Delta f(X_i) - \nabla V(X_i) \cdot \nabla f(X_i).$$

Does not require a priori computations of averages w.r.t. μ. May know V only up to an additive constant.

Of course, cannot use all functions f, so should restrict to a well-chosen finite family.

First implementations by Gorham and Mackey (2015). Survey of recent developments by Anastasiou et. al. (2021).

Many other choices of linear operator via Markov processes. Can also use jump processes (eg. discrete spaces, stable laws...)
Even for a given target distribution different choices of operator may be useful in various situations.
Can also use operators that are not Markov genrators, eg.

$$\int (x^2 - 1)f(x)d\gamma = \int xf'(x)d\gamma.$$

Examples of applications

- Limit laws for statistics of random matrices (spectrum, traces, powers...)
- Statistical physics (spin glasses, Ising model...)
- Poisson approximation (law of small numbers...)
- Concentration inequalities

Surveys by Ross (2011), Chatterjee (2014), Azmoodeh, Peccati & Yang (2021).

What about other transport distances ?

Consider the L^p Wasserstein distance

$$W_p(\mu,\nu)^p = \inf\left\{\mathbb{E}[|X-Y|^p]; X\sim\mu, Y\sim\nu\right\}.$$

In general, it cannot be written as

$$\sup_f \mathbb{E}[f(X)] - \mathbb{E}[f(Y)]$$

but requires optimizing over pairs of functions. So Stein's approach does not apply as is.

The Gaussian setting

Definition

Let $\mu \in \mathcal{P}(\mathbb{R}^d)$. A matrix-valued function τ is a Stein kernel for μ if

$$\int x \cdot \nabla f d\mu = \int \langle \tau, \nabla^2 f \rangle_{HS} d\mu \quad \forall f \text{ smooth.}$$

The Stein discrepancy of ν is $S(\mu)^2 := \inf_\tau \int ||\tau - \mathrm{Id}\,||_{HS}^2 d\mu$.

Translation of Stein's lemma : we ca take $\tau = \mathrm{Id}$ iff $\nu = \gamma$.

Theorem (Ledoux, Nourdin & Peccati 2015)

$$W_2(\mu, \gamma)^2 \leq S(\mu)^2.$$

Proof : consider the Ornstein-Uhlenbeck dynamic

$$dX_t = -X_t dt + \sqrt{2} dB_t$$

It is the gradient flow of the entropy $\int \rho \log(\rho/\gamma)$ with respect to the W_2 distance (Jordan, Kinderlehrer & Otto 1998, Otto 2001). The entropy dissipation along the flow is the Fisher information $I(\rho) = \int |\nabla \log(\rho/\gamma)|^2 d\rho$.

Hence $W_2(\mu, \gamma) \leq \int_0^\infty \sqrt{I(\mu_t)} dt$ if μ_t is the law at time t along the OU flow.

Can relate Fisher information to Stein kernel by integration by parts, and get

$$I(\mu_t) = \int \langle \tau - \mathrm{Id}, \nabla^2 P_t \log(\rho_t/\gamma) \rangle d\mu$$

and then use decay estimates for $\nabla^2 P_t \log \rho_t$.

Decay estimates for $\nabla^2 P_t f$ use third-order Bakry-Emery Gamma calculus.

Approach can be extended to other target measures (eg. Gamma distributions), but strong algebraic requirements.
Variant for L^p Wasserstein distance

$$W_p(\mu,\gamma) \leq C_p \left(\int ||\tau - \mathrm{Id}\,||_{HS}^p d\mu \right)^{1/p}.$$

Several possible constructions of Stein kernels, including two via optimal transport maps.
Chatterjee (2008), Nourdin, Peccati & Reveillac (2009) : if $X \sim \gamma$ and $\mu \sim F(X)$ with $\mathbb{E}[F(X)] = 0$ then

$$\tau(x) = \int_0^\infty e^{-t}\,\mathbb{E}[P_t \nabla F(X)(\nabla F(X))^t | F(X) = x]dt$$

is a Stein kernel for μ.
If ∇F satisfies uniform bounds, we get good bounds on τ. Can use the Brenier map. Caffarelli (2001) : if μ is uniformly log-concave, the Brenier map is globally lipschitz.
Mikulincer (2020) : CLT for fluctuations of sample moment tensors with iid uniformly log-concave data.

Second construction : consider the Monge-Ampère PDE

$$e^{-\varphi} = \rho(\nabla\varphi)\,\mathrm{det}(\nabla^2\varphi).$$

$\nabla\varphi$ is the Brenier map from $e^{-\varphi}$ onto ρ.
Wang & Zhu (2004), Berman & Berndtson (2013), Cordero-Erausquin & Klartag (2015) : weak solutions exist if ρ is centered and has finite first moment.
Santambrogio (2015) : arises as Euler-Lagrange equation for

$$\mu \longrightarrow \mathrm{Ent}_\gamma(\mu) - \frac{1}{2}W_2(\mu,\rho)^2.$$

Theorem

If φ is smooth enough, then $(\nabla^2\varphi) \circ \nabla\varphi^*$ is a Stein kernel for ρ, where φ^* is the Legendre transform of φ.

$$\int x \cdot \nabla f(x) d\rho = \int \nabla\varphi \cdot \nabla f(\nabla\varphi) e^{-\varphi} dx$$

$$= \int \mathrm{Tr}(\nabla^2 f(\nabla\varphi)\nabla^2\varphi) e^{-\varphi} dx$$

and apply the inverse transport map $\nabla\varphi^*$.
Good a priori bounds on $\nabla^2\varphi$ when ρ is log-concave (Klartag 2014, Klartag & Kolesnikov 2015). Application to CLT.

Thanks for your attention !