Stein's method for stability of variational problems over spaces of probability measures.

Max Fathi

LJLL & LPSM, Université Paris Cité

12 avril 2022

Goal : use Stein's method to study stability of optimizers in variational problems over spaces of probability measures.
Consider

$$F^* := \inf F(\mu), \quad \mu \in A \subset \mathcal{P}(E).$$

Optimizer : measure μ such that $F(\mu) = F^*$. $\mathcal{F}^*$: set of minimizers
Near-optimizer : $F(\mu) \leq F_* + \epsilon$.
Question : when are near-optimizers close to actual minimizers ?

Goal : estimates of the form

$$d(\mu, \mathcal{F}^*) \leq C(F(\mu) - F^*)^\alpha.$$

d is a distance on the space of probaility measures. It will typically be here the L^1 Wasserstein distance

$$W_1(\mu, \nu) = \inf_{f 1-lip} \int f d\mu - \int f d\nu.$$

C and α are constants, which hopefully behave nicely with respect to the parameters of the problem. Here, we will often care about dimension-free estimates.

The general philosophy in what follows is that for many variational problems over spaces of probability measures, the Euler-Lagrange equation takes the form of an integration by parts formula.
We can then expect near-minimizers to *almost* satisfy the same formula.
If yes, can try to use Stein's method to compare near-minimizers to minimizers. Idea appears in works of Utev (1989).

If we look at a function of the form

$$F(\mu) = \sup_{f \in \mathcal{H}} \int H(f, \nabla f) d\mu$$

the Euler Lagrange equation for an optimal function f_0 is

$$\int h \partial_1 H(f_0, \nabla f_0) + \nabla h \cdot \nabla_2 H(f_0, \nabla f_0) d\mu = 0$$

for all $h \in \mathcal{H}$.
Can derive for optimal measures an integration by parts formula, that depends on the optimal function f_0. Need information on f_0 to characterize the measure, will be possibel for the results presented today.

An artificial example

Consider the SDE

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

Markov process, with generator

$$\mathcal{L}f = \Delta f - \nabla V \cdot \nabla f$$

and invariant probability measure $\mu = e^{-V}dx$.
Expect ν to be close to μ if $\int \mathcal{L}f d\nu \approx 0$ for a large enough class of test functions.

The relative Fisher information of a probability measure $\nu = \rho\mu$ is $I(\nu) = \int |\nabla \log \rho|^2 d\nu$.

μ is trivially the unique global mnimizer of the Fisher information. What about near-minimizers? Can we control $W_1(\mu, \nu)$ by the Fisher information?

Variational viewpoint :

$$I(\nu) = \left(\sup\left\{\int \nabla \log \rho \cdot \nabla g d\nu; \int |\nabla g|^2 d\nu \leq 1\right\}\right)^2$$

$$= \left(\sup\left\{\int \mathcal{L} g d\nu; \int |\nabla g|^2 d\nu \leq 1\right\}\right)^2$$

$$\geq \left(\sup\left\{\int \mathcal{L} g d\nu; ||\nabla g||_\infty \leq 1\right\}\right)^2.$$

This is the kind of quantities that we use to control distances when applying Stein's method.

Rigorous result :

Theorem (Guillin, Leonard, Wu and Yao 2009, Mijoule, Reinert and Swan 2019)

Assume that solutions to the Poisson equation $\mathcal{L} f = g$ with g a 1-Lipschitz function are α-Lipschitz. Then for any ν we have

$$W_1(\nu, \mu)^2 \leq \alpha^2 I(\nu).$$

Can be adapted to other types of Markov generators (discrete spaces, non-constant diffusion matrices, Riemannian manifolds...).
This inequality implies Gaussian concentration for μ.

An aside : why Fisher information ?

The evolution of the law of X_t can be viewed as the gradient descent of the relative entropy $\text{Ent}_\mu(\nu) = \int \rho \log \rho d\mu$ with respect to the Wasserstein distance W_2.
The Fisher information, as the derivative of the entropy along the flow, can be re-interpreted as the squared norm of the gradient.
Transport-information inequality interpreted as

$$d(\nu, \mu)^2 \leq C|\nabla F(\nu)|^2.$$

Classical tool for gradient descent (Lojasiewicz-type inequality).

Outcome 1

Consider a uniformly log-concave measure on $\mathbb{R}^d$, that is $\mu = \exp(-V)dx$ with Hess $V \geq I_d$.
Gaussian concentration : for any 1-lipchitz funciton f,

$$\int e^{\lambda f} d\mu \leq \exp\left(\lambda^2/2 + \lambda \int f d\mu\right).$$

Implies deviation estimates via Chernoff's inequality.
Applications in statistics, geometry, information theory...

Theorem (Courtade & F., 2020)

Assume that the convexity condition holds, and that there exists f 1-lipschitz and $\lambda > 0$ such that

$$\int e^{\lambda f} d\mu \geq \exp\left((1-\epsilon)\lambda^2/2 + \lambda \int f d\mu\right).$$

Then, up to a translation and rotation,

$$W_1(\mu, \gamma_1 \otimes \mu') \leq C(\lambda)\sqrt{\epsilon}$$

where γ_1 is a one-dimensional standard gaussian measure.

Outcome 2

Classical topic in geometry : optimizing a geometric quantity subject to a constraint.

Eg. : Isoperimetric problem. Among all shapes with fixed volume the sphere minimizes the perimeter.
We consider a smooth N-dimensional Riemannian manifold (M, g) whose Ricci curvature tensor satisfies

$$\mathrm{Ric} \geq (N-1)g.$$

The constant is chosen so that the sphere with unit radius satisfies this bound.

Bonnet-Myers Theorem : the diameter is maximized by the sphere.

Obata ('62) : this bound is *rigid* : among all smooth N-manifolds with $\mathrm{Ric} \geq N-1$, the sphere is the only equality case.
Anderson ('90) : this characterization is unstable.
Many works in geometry (Cheeger & Colding, Cheng, Croke, Ketterer, Petersen, Aubry, Cavaletti, Mondino & Semola...)

Theorem (F., Gentil & Serres, 2021)

Assume the curvature condition holds, and that the diameter is greater than $\pi - \epsilon$ for some ϵ small enough. Then there is an eigenfunction of the Laplacian f such that $W_1(f\#Vol, Z_N^{-1}(1-x^2)^{N/2-1}) \leq C(N)\epsilon^{1/N}$.

The symmetrized beta distributions is the distribution of a coordinate on a sphere, which is an eigenfunction.
Fully quantitative statement for almost minimal spectral gap, with sharp dimension-free exponent. Cheng, Croke : spectral gap almost minimal iff diameter almost maximal.

Stability of Poincaré inequalities

Consider an isotropic centered probability measure μ on $\mathbb{R}^d$. Its Poincaré constant is the smallest constant C_P such that

$$\forall f,\ \ \mathrm{Var}_\mu(f) \leq C_P \int |\nabla f|^2 d\mu.$$

Testing a linear function, we see that $C_P \geq 1$.
For the standard Gaussian measure, $C_P = 1$. Simplest proof : L^2 decomposition along Hermitte polynomials.
Chen & Lou (1987) : $C_P = 1$ iff μ is a standard Gaussian measure.

Theorem (Utev 1989, Courtade, F. & Pananjady 2019)

For an isotropic centered probability measure μ*, we have*

$$C_P \geq 1 + \frac{W_2(\mu,\gamma)^2}{d}.$$

Scheme of proof in dimension one : expanding

$$C_P \int (f')^2 d\mu - \mathrm{Var}_\mu(f)$$

for $f = x + \epsilon h$ and h centered, get

$$2\epsilon \int (C_P h' - xhd\mu + \epsilon^2 \int C_P(h')^2 - h^2 d\mu \geq 0.$$

Considering bounded lipschitz test functions and optimizing in ϵ gives

$$\sup_{||h||_\infty, ||h'||_\infty \leq 1} \int h' - xhd\mu \leq \sqrt{C_P - 1}.$$

Applying Stein's lemma concludes the proofs.

Other results

- Stability of Poincaré constants : Poisson distributions (Utev), stable laws (Arras-Houdré), general targets in dimension one (Serres), free probability (Cébron, F. & Mai).
- Higher eigenvalues (Serres)
- Log-Sobolev constants (Courtade & F.)
- Generalized Cauchy distributions for geometric problems (F., Gentil & Serres)

Question 1 : stability of the optimizer for infinite-width two-layer neural networks

Neural network with two layers : given a target function g, find parameters $(w, A, b) \in (\mathbb{R}^{d+2})^N$ such that

$$f_{w,A,b}(x) = \frac{1}{N}\sum_{i=1}^{N} w_i \rho(A_i x + b_i) \approx g.$$

Loss function $R(w, A, b) = \mathbb{E}[(f_{w,A,b}(X) - g(X))^2]$.
Can run gradient descent to approximate optimal parameters. Problem : many local minimizers.
Chizat & Bach 2018 (and many others) : Embed $(w, A, b) \longrightarrow \frac{1}{N}\sum \delta_{w_i,A_i,b_i} \in \mathcal{P}(\mathbb{R}^{d+2})$. $f_{w,A,b}$ can be written as an integral w.r.t. this measure, so R extends to a function over $\mathcal{P}(\mathbb{R}^{d+2})$.

Lift the gradient descent to the gradient descent of R in $\mathcal{P}(\mathbb{R}^{d+2})$ with respect to W_2. At most one local minimizer.
Minimizer might be at infinity, so add a penalization $R(mu) + \epsilon\,\mathrm{Ent}_{dx}(\mu)$.
Nice effect on the gradient descent. Other types of penalizations, such as Renyi entropy or Dirichlet forms.
Is the minimizer stable ? Or when viewed as a minimizer of the energy dissipation along the gradient descent ?

Question 2 : Stein's method as a tool for Lojasiewicz inequalities ?

Gradient Lojasiewicz inequality : if f is an analytic function on a compact set, for any critical point x_0 there are constants C, θ such that

$$(f(x) - f(x_0))^\theta \leq C|\nabla f(x)|.$$

Equivalent to $\mathrm{dist}(x, Z_f)^\alpha \leq C|f(x)|$. Applications to convergence to equilibrium of gradient descent.

Can we use Stein's method as a tool to prove such inequalities over spaces of probability measures ?

Question 3 : Stein's method for shapes ?

Many geometric problems take the form of optimizing a geometric quantity over sets of fixed volume (isoperimetry,...)
Stability : if A is a shape that minimizes some functional F, do we have

$$|A \Delta B|^{\alpha} \leq C(F(B) - F(A))?$$

If A an B have same volume,

$$|A \Delta B| = d_{TV}(\mathbb{1}_A, \mathbb{1}_B).$$

Examples : stability for isoperimetric inequalities, Faber-Krahn inequality, etc...
Can Stein's method find a use here ? Problem : natural integration by parts formulas have boundary terms.

Thanks !