Wasserstein Gradient Flows, χ^2 divergence, and Stein Variational Gradient Descent

arXiv:2006.02509v1 and arXiv:2005.09669v2

Andrew Warren

December 1 2020

General themes

We look at two papers, arXiv:2006.02509v1 and arXiv:2005.09669v2, published simultaneously in Neurips this year, by Sinho Chewi et al. What is in the two papers? (non-exhaustive, especially not for arXiv:2005.09669v2)

- measuring the rate of convergence of a KL gradient flow according to χ^2 divergence allows for exponential convergence under weaker assumptions
- dually, a χ^2 gradient flow converges exponentially quickly, according to KL, under rather general assumptions
- Stein Variational Gradient Flow (as seen in Wony's talk) can be viewed as a χ^2 gradient flow in the "kernelized Benamou-Brenier" sense
- this might suggest some interesting numerical procedures, selection of favorable kernels, etc

But there are also some results one might expect to see that are conspicuously absent.

Review of KL and χ^2 divergences

Let $\mu, \pi \in \mathcal{P}(\mathbb{R}^d)$. Recall the *KL* divergence from μ to π, with $\mu \ll \pi$:

$$KL(\mu \mid \pi) := \int_{\mathbb{R}^d} \frac{d\mu}{d\pi} \log \frac{d\mu}{d\pi} d\pi.$$

Whereas, the χ^2 divergence from μ to π, with $\mu \ll \pi$ is defined by

$$\chi^2(\mu \mid \pi) := \mathrm{var}_\pi \frac{d\mu}{d\pi} = \int_{\mathbb{R}^d} \left(\frac{d\mu}{d\pi}\right)^2 d\pi - 1.$$

A standard fact is that $KL(\mu \mid \pi) \leq \chi^2(\mu \mid \pi)$. (See Tsybakov nonparametric statistics book) Both are examples of *Cziszar divergences* aka *internal energies*, namely functionals of the form

$$F_\pi(\mu) = \int f\left(\frac{d\mu}{d\pi}\right) d\pi.$$

Review of some Wasserstein gradient flow facts

Recall also the Benamou-Brenier formulation of the 2-Wasserstein distance:

$$W_2^2(\mu,\nu) = \inf_{(\rho,\nu)} \left\{ \int_{\mathbb{R}^d} \|v_t\|^2_{L^2(\rho_t)} dt : \rho_0 = \mu, \rho_1 = \nu, \partial_t \rho_t + \text{div}(\rho v) = 0 \right\}$$

If $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is some "reasonable" functional, then this presentation of W_2 allows us to write down the gradient of F, namely,

$$\text{grad}_{W_2} F(\rho) = -\text{div}\left(\rho \nabla \frac{\delta F}{\delta \rho}\right)$$

so that the gradient flow of F satisfies $\partial_t \rho_t = \text{div}\left(\rho \nabla \frac{\delta F}{\delta \rho}\right)$. Of classical interest is the case where $F(\rho) = KL(\rho \mid \pi)$ where $\pi = e^{-V} dVol$; in this case the gradient flow equation is the Fokker-Planck equation $\partial_t \rho_t = \text{div}(\rho \nabla V) + \Delta \rho$.

Review of some Wasserstein gradient flow facts

If $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is some "reasonable" functional, and μ_t is the Wasserstein gradient flow of F, then we have the "entropy production" formula

$$\frac{d}{dt} F(\mu_t) = -\int_{\mathbb{R}^d} |\nabla \frac{\delta F}{\delta \mu_t}|^2 d\mu_t.$$

For instance, when F is an internal energy of the form $\int_{\mathbb{R}^d} f\left(\frac{d\mu_t}{d\pi}\right) d\mu_t$, we have that $\frac{\delta F}{\delta \mu_t} = f'(\frac{d\mu_t}{d\pi})$. For $KL(\cdot \mid \pi)$, this means that along a KL gradient flow, we have that

$$\frac{d}{dt} KL(\mu_t \mid \pi) = -\int |\nabla \log \frac{d\mu_t}{d\pi}|^2 d\mu_t = -4 \int |\nabla \sqrt{\frac{d\mu_t}{d\pi}}|^2 d\pi.$$

Similarly for $\chi^2(\cdot \mid \pi)$, we have that

$$\frac{d}{dt} \chi^2(\mu_t \mid \pi) = -\int |\nabla \frac{d\mu_t}{d\pi}|^2 d\mu_t.$$

Review of some functional inequalities

Let $\pi \in \mathcal{P}_2(\mathbb{R}^d)$. We say π satisfies a:

- Poincaré inequality provides that for all test functions f,

$$\mathrm{var}_\pi[f] := \|f - \mathbb{E}_\pi[f]\|^2_{L^2(\pi)} \leq C_P \|\nabla f\|^2_{L^2(\pi)}$$

- log-Sobolev inequality provided that for all test functions f,

$$\int_{\mathbb{R}^d} f^2 \ln(f^2) d\pi - \left(\int_{\mathbb{R}^d} f^2 d\pi\right) \ln\left(\int_{\mathbb{R}^d} f^2 d\pi\right) \leq 2C_{LS} \int_{\mathbb{R}^d} |\nabla f|^2 d\pi.$$

If we put $f^2 = \frac{d\mu}{d\pi}$, the latter reduces to

$$KL(\mu \mid \pi) \leq 2C_{LS} \int_{\mathbb{R}^d} |\nabla \sqrt{\frac{d\mu}{d\pi}}|^2 d\pi.$$

Since $\int_{\mathbb{R}^d} |\nabla \sqrt{\frac{d\mu}{d\pi}}|^2 d\pi = \frac{1}{4} \int |\nabla \log \frac{d\mu}{d\pi}|^2 d\mu$, this actually means that

$$\partial_t KL(\mu_t \mid \pi) \leq -(C_{LS}/2) KL(\mu_t \mid \pi)$$

Which, by Gronwall's inequality, implies

$$KL(\mu_t \mid \pi) \leq e^{-2t/C_{LS}} KL(\mu_0 \mid \pi).$$

KL gradient flow according to χ^2 divergence

Now, we compute formally as follows. (Throughout, μ and π are a.c. with respect to the Lebesgue measure on $\mathbb{R}^d$. Let $\mu(x)$ denote the density of μ wrt Lebesgue and similarly for $\pi(x)$.)
Given that $(\mu_t)_{t\geq 0}$ is the W_2 gradient flow of $KL(\cdot \mid \pi)$, we have that $\partial \mu_t(x) = \mathrm{div}(\mu_t(x)\nabla \ln \frac{d\mu_t}{d\pi})$. So,

$$\frac{1}{2}\partial_t \chi^2(\mu_t \mid \pi) = \frac{1}{2}\int_{\mathbb{R}^d} \partial_t \left(\frac{\mu_t(x)}{\pi(x)}\right)^2 \pi(x)dx = \int_{\mathbb{R}^d} \frac{\mu_t(x)}{\pi(x)}\partial_t \mu_t dx$$

$$= \int_{\mathbb{R}^d} \frac{\mu_t(x)}{\pi(x)}\mathrm{div}\left(\mu_t \nabla \ln \frac{d\mu_t}{d\pi}\right) dx = -\int_{\mathbb{R}^d} \left\langle \nabla \frac{\mu_t(x)}{\pi(x)}, \mu_t(x)\nabla \ln \frac{\mu_t(x)}{\pi(x)} \right\rangle dx$$

$$= -\int |\nabla \frac{d\mu_t}{d\pi}|^2 d\pi$$

where we have used the fact that $\nabla \ln \frac{\mu_t(x)}{\pi(x)} = \left(\frac{\mu_t(x)}{\pi(x)}\right)^{-1} \nabla \frac{\mu_t(x)}{\pi(x)}$.

KL gradient flow according to χ^2 divergence

Now, if π satisfies a Poincaré inequality, we have that $C_P \int |\nabla \frac{d\mu_t}{d\pi}|^2 d\pi \geq \mathrm{var}_\pi \left(\frac{d\mu_t}{d\pi} \right)$. But $\mathrm{var}_\pi \left(\frac{d\mu_t}{d\pi} \right)$ is none other than $\chi^2(\mu_t \mid \pi)$, so we actually have

$$\frac{1}{2} \partial_t \chi^2(\mu_t \mid \pi) \leq -\frac{1}{C_P} \chi^2(\mu_t \mid \pi).$$

Therefore, Gronwall's inequality implies that

$$\chi^2(\mu_t \mid \pi) \leq e^{-2t/C_P} \chi^2(\mu_0 \mid \pi).$$

In other words, if we measure how close μ_t is to equilibrium using χ^2 rather than KL, we only need π to satisfy a Poincaré inequality rather than a log-Sobolev inequality.

χ^2 gradient flow according to KL divergence

What happens if we switch things around? Well, the preceding calculation looks very similar, but gradient vectors of KL and χ^2 trade places. Indeed,

$$\frac{1}{2}\partial_t KL(\mu_t \mid \pi) = \frac{1}{2}\int_{\mathbb{R}^d} \partial_t \left(\frac{\mu_t(x)}{\pi(x)} \ln \frac{\mu_t(x)}{\pi(x)}\right)\pi(x)dx = \int_{\mathbb{R}^d} \nabla \ln \frac{\mu_t(x)}{\pi(x)} \partial_t \mu_t dx$$

$$= \int_{\mathbb{R}^d} \frac{\mu_t(x)}{\pi(x)} \mathrm{div}\left(\mu_t \nabla \frac{d\mu_t}{d\pi}\right) dx = -\int_{\mathbb{R}^d} \left\langle \nabla \ln \frac{\mu_t(x)}{\pi(x)}, \mu_t(x) \nabla \frac{\mu_t(x)}{\pi(x)} \right\rangle dx$$

$$= -\int |\nabla \frac{d\mu_t}{d\pi}|^2 d\pi.$$

Consequently,

$$\frac{1}{2}\partial_t KL(\mu_t \mid \pi) = -\int |\nabla \frac{d\mu_t}{d\pi}|^2 d\pi. \leq -\frac{1}{C_P}\chi^2(\mu_t \mid \pi) \leq -\frac{1}{C_P} KL(\mu_t \mid \pi)$$

So here also, Gronwall implies that $KL(\mu_t \mid \pi) \leq e^{-2/C_P} KL(\mu_0, \pi)$. (This type of dualization works in great generality, as observed by Matthes-McCann-Savaré.)

A χ^2 Lojasiewicz inequality

Suppose that π satisfies a Poincaré inequality with constant $C_P > 0$. Then $\forall \mu \ll \pi$,

$$\chi^2(\mu \mid \pi)^{3/2} \leq \frac{9C_P}{4} \int |\nabla \frac{d\mu}{d\pi}|^2 d\mu.$$

Proof.

First note that

$$\int |\nabla \frac{d\mu}{d\pi}|^2 d\mu = \int |\nabla \frac{d\mu}{d\pi}|^2 \frac{d\mu}{d\pi} d\pi = \frac{4}{9} \int |\nabla \left(\frac{d\mu}{d\pi}\right)^{3/2}|^2 d\pi$$

$$\text{(by Poincaré)} \geq \frac{4}{9C_P} \text{var}_\pi \left(\left(\frac{d\mu}{d\pi}\right)^{3/2} \right).$$

□

A χ^2 Lojasiewicz inequality

(cont'd)

Proof.

It then suffices to argue that $\chi^2(\mu \mid \pi)^{3/2} \leq \mathrm{var}_\pi\left(\left(\frac{d\mu}{d\pi}\right)^{3/2}\right)$. This is essentially just Jensen's inequality. Explicitly:

$$\chi^2(\mu \mid \pi) = \mathrm{var}_\pi\left(\frac{d\mu}{d\pi}\right) \leq \mathbb{E}_\pi\left[\left(\frac{d\mu}{d\pi} - \mathbb{E}_\pi\left[\left(\frac{d\mu}{d\pi}\right)^{3/2}\right]^{2/3}\right)^2\right]$$

$$(x^{2/3} \text{ is Hölder cts}) \leq \mathbb{E}_\pi\left[\left|\left(\frac{d\mu}{d\pi}\right)^{3/2} - \mathbb{E}_\pi\left[\left(\frac{d\mu}{d\pi}\right)^{3/2}\right]\right|^{4/3}\right]$$

$$(\text{Jensen}) \leq \mathbb{E}_\pi\left[\left|\left(\frac{d\mu}{d\pi}\right)^{3/2} - \mathbb{E}_\pi\left[\left(\frac{d\mu}{d\pi}\right)^{3/2}\right]\right|^2\right]^{2/3}$$

$$= \left(\mathrm{var}_\pi\left(\left(\frac{d\mu}{d\pi}\right)^{3/2}\right)\right)^{2/3}.$$

Convergence of χ^2 gradient flow

Suppose that π satisfies a Poincaré inequality with constant C_P. Let $\chi^2(\mu_0 \mid \pi) < \infty$ and let $(\mu_t)_{t \geq 0}$ be the χ^2 Wasserstein gradient flow starting from μ_0. Then,

$$\chi^2(\mu_t \mid \pi) \leq \chi^2(\mu_0 \mid \pi) \wedge \left(\frac{9C_P}{8t}\right)^2.$$

Proof.

"Entropy production" implies that $\partial_t \chi^2(\mu_t \mid \pi) = -4 \int |\nabla \frac{d\mu_t}{d\pi}|^2 d\mu_t$. With the Lojasiewicz inequality, this implies

$$\partial_t \chi^2(\mu_t \mid \pi) \leq -\frac{16}{9C_P} \chi^2(\mu_t \mid \pi)^{3/2}.$$

This implies that

$$\chi^2(\mu_t \mid \pi) \leq \frac{\chi^2(\mu_0 \mid \pi)}{\left[1 + 8t\sqrt{\chi^2(\mu_0 \mid \pi)}/(9C_P)\right]^2}.$$

The claim follows.

Convergence of χ^2 gradient flow (cont'd)

Much like in the case of the KL divergence, if we further assume that π satisfies a log-Sobolev inequality (or furthermore is log-concave) then in fact we have exponential rather than $1/t^2$ rate of convergence. More precisely,

$$\chi^2(\mu_t \mid \pi) \leq \left(\chi^2(\mu_0 \mid \pi) \wedge 2\right) e^{-t/9C_{LS}} \quad t \geq 7C_{LS}.$$

(This is Theorem 3 in "SVGD as a kernelized Wasserstein..."). Proof is just slightly too involved for this talk.

What about SVGD???

Let's actually get to the connection with Stein Variational Gradient Descent.
Fix a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Define the induced operator $\mathcal{K}$ by

$$\mathcal{K}_\mu : L^2(\mu) \to L^2(\mu)$$

$$\mathcal{K}_\mu f(x) := \int_{\mathbb{R}^d} K(x, y) f(y) d\mu(y).$$

A "cheap" gloss on the continuum formulation of SVGD is that we simply "kernelize the continuity equation" for a gradient flow by inserting $\mathcal{K}_\mu$ like so:

$$\partial_t \mu_t = \operatorname{div}\left(\mu_t \underbrace{\mathcal{K}_\mu}_{\uparrow} \nabla \ln \frac{d\mu_t}{d\pi} \right).$$

(See Duncan et al. Lemma 9)

SVGD as χ^2 gradient flow

Compute as follows:

$$\mathcal{K}_{\mu_t} \nabla \ln \left(\frac{d\mu_t}{d\pi} \right)(x) = \int K(x,y) \nabla \ln \left(\frac{d\mu_t}{d\pi} \right) d\mu_t(y) = \int K(x,y) \nabla \frac{d\mu_t}{d\pi} d\pi$$
$$= \mathcal{K}_\pi \nabla \frac{d\mu_t}{d\pi}(x).$$

Consequently

$$\partial_t \mu_t = \operatorname{div} \left(\mu_t \mathcal{K}_{\mu_t} \nabla \ln \frac{d\mu_t}{d\pi} \right) = \operatorname{div} \left(\mu_t \mathcal{K}_\pi \nabla \frac{d\mu_t}{d\pi} \right).$$

Since $2\nabla \frac{d\mu_t}{d\pi}$ corresponds to the Wasserstein gradient vector for the χ^2 divergence, we conclude that SVGD is recast (up to a constant factor) as a "kernelized Wasserstein gradient flow" of the χ^2 divergence, but with *constant* kernel operator $\mathcal{K}_\pi$.

SVGD as χ^2 gradient flow cont'd

Cool! Where next?
Chewi et al. then point out that if we (daftly) pick $\mathcal{K}_\pi = id$, the preceding analysis of (vanilla) Wasserstein GFs with $\chi^2(\cdot \mid \pi)$ applies. Either this amounts to saying that χ^2 gradient flow is an underexploited way to sample from π, or this is an instance of mathiness to impress the Neurips reviewers (plausibly both).
However, they then sketch the following approach. In general, we have the "entropy production" formula for the KL divergence along the χ^2 GF:

$$\partial_t KL(\mu_t \mid \pi) = -\mathbb{E}_\pi \langle \nabla \frac{d\mu_t}{d\pi}, \mathcal{K}_\pi \nabla \frac{d\mu_t}{d\pi} \rangle.$$

If it were the case that $\mathcal{K}_\pi$ had a spectral gap, this would immediately imply an exponential rate of convergence in KL. However, they note that $\mathcal{K}_\pi$ does not have a spectral gap when e.g. $K \in L^2(\pi \otimes \pi)$ with $\pi \propto e^{-V}$, and so reject this strategy. (But is this another opening for singular kernels SVGD??)

"Laplacian adjusted Wasserstein gradient descent"

Instead the authors go by another route. If we can select K carefully so that $\mathbb{E}_\pi \langle \nabla \frac{d\mu_t}{d\pi}, \mathcal{K}_\pi \nabla \frac{d\mu_t}{d\pi} \rangle \geq C \cdot KL(\mu_t \mid \pi)$, then we still have that

$$\partial_t KL(\mu_t \mid \pi) \leq -C \cdot KL(\mu_t \mid \pi)$$

and so Gronwall's inequality still allows for exponential rate of convergence.
This turns out to be achievable with the clever choice of $\mathcal{K}_\pi = \mathcal{L}^{-1}$ (where $\mathcal{L} = -\Delta + \langle \nabla V, \nabla \cdot \rangle$). Indeed, since (this is the integration by parts formula from Markov semigroup theory)

$$\mathbb{E}_\pi \langle \nabla f, \nabla g \rangle = \mathbb{E}[f \mathcal{L} g]$$

we see that

$$\mathbb{E}_\pi \langle \nabla \frac{d\mu_t}{d\pi}, \nabla \mathcal{L}^{-1} \frac{d\mu_t}{d\pi} \rangle = \mathbb{E}_\pi \left[\frac{d\mu_t}{d\pi} \mathcal{L} \mathcal{L}^{-1} \frac{d\mu_t}{d\pi} \right] = \mathbb{E}_\pi \left[\left(\frac{d\mu_t}{d\pi} \right)^2 \right].$$

"Laplacian adjusted Wasserstein gradient descent"

Thus, by choosing $\mathcal{K}_\pi = \mathcal{L}^{-1}$, we actually have that $\mathbb{E}_\pi \langle \nabla \frac{d\mu_t}{d\pi}, \nabla \mathcal{K}_\pi \frac{d\mu_t}{d\pi} \rangle = \chi^2(\mu_t \mid \pi)$. Thus,

$$\partial_t KL(\mu_t \mid \pi) = -\chi^2(\mu_t \mid \pi) \leq -KL(\mu_t \mid \pi)$$

as desired.

Can we implement this? Chewi et al. propose the following. Compute that

$$\begin{aligned} \mathcal{K}_\pi \nabla \frac{d\mu_t}{d\pi}(x) &= \int K(x,y) \nabla \frac{d\mu_t}{d\pi}(y) d\pi(y) \\ &= \int \nabla_1 K(x,y) \frac{d\mu_t}{d\pi}(y) d\pi(y) = \int \nabla_1 K(x,y) d\mu_t(y) \end{aligned}$$

Plug this into the GF equation $\partial \mu_t = \operatorname{div}\left(\mu_t \mathcal{K}_\pi \nabla \frac{d\mu_t}{d\pi}\right)$; then, replace μ_t with an empirical measure and the time derivative with a finite difference, for

$$X_{t+1}^{[i]} \leftarrow X_t^{[i]} - \frac{h}{N} \sum_{j=1}^{N} \nabla_1 K(X_t^{[i]}, X_t^{[j]}).$$

"Laplacian adjusted Wasserstein gradient descent"

It then remains only to extract a kernel function $K(x, y)$ from the desired kernel operator $\mathcal{K}_\pi = \mathcal{L}^{-1}$. For "nice" potential V, one can perform spectral decomposition and "just" write down

$$K(x, y) = \sum_{i=1}^{\infty} \frac{\phi_i(x)\phi_i(y)}{\lambda_i}$$

where (λ_i, ϕ_i) is the ith eigenvalue-eigenfunction pair for $\mathcal{L}$. Therefore, if we get the spectral decomposition of $\mathcal{L}$ from an oracle (!) and handwave on the discretization error for our numerical scheme, we have a gradient descent procedure that converges exponentially in KL to π. Moreover, the rate is independent of the spectral gap of V (!!) – this comes directly from Gronwall and the fact that $\partial_t KL(\mu_t \mid \pi) \leq -KL(\mu_t \mid \pi)$.
Obviously a number of implementation details are missing but there are some compelling ideas here.