MATH 697 report

Ayoub El Hanchi

July 2020

1 Introduction

A fundamental algorithmic task in statistics and machine learning is that of sampling from an absolutely continuous probability distribution $\pi(dx) = e^{-f}dx$ on $\mathbb{R}^d$ given its (unnormalized) density. The Markov chain Monte Carlo (MCMC) solution to this problem is to use an easily simulable Markov Kernel $\kappa(x, B)$ to construct a time-homogeneous Markov chain $(X_n)_{n=1}^{\infty}$ such that the distribution π_n of X_n converges to π in some metric on the space $\mathscr{P}(\mathbb{R}^d)$ of probability measure on $\mathbb{R}^d$. One can easily construct such kernels by, for example, composing a proposal distribution with a Metropolis-Hastings filter. The challenge lies in finding a Markov kernel that generates a rapidly converging chain while still being easy to simulate.

Let us consider the following variational formulation of the problem of sampling from π. Let $\mathscr{F}_\pi : \mathscr{P}(\mathbb{R}^d) \to \mathbb{R}$ be a functional minimized at the target distribution π. Sampling from π through a Markov chain starting at $X_1 \sim \pi_1$ can now be seen to be equivalent to minimizing $\mathscr{F}_\pi$ in $\mathscr{P}(\mathbb{R}^d)$ starting from π_1. Seen through this lens, one might wonder if an equivalent of gradient descent on $\mathscr{F}_\pi$ exists in $\mathscr{P}(\mathbb{R}^d)$. Intuitively, this will give us a rapidly converging sequence, and with a bit of luck, we can hope to find an easily simulable Markov kernel that reproduces it. Going to continuous time, we can pose the same question about the existence of a gradient flow of $\mathscr{F}_\pi$ in $\mathscr{P}(\mathbb{R}^d)$ and whether it can be identified with any of the known continuous-time Markov processes. Of course, the first step to answering these questions is defining what we mean by a gradient flow of a functional in $\mathscr{P}(\mathbb{R}^d)$ given that it has no vector space structure.

It turns out that there is a reasonable way to extend the definition of the gradient flow of a functional defined on the metric space $(\mathscr{P}_2(\mathbb{R}^d), W_2)$, where $\mathscr{P}_2(\mathbb{R}^d)$ is the space of probability measures on $\mathbb{R}^d$ with finite second moment, and W_2 is the 2-Wasserstein distance. Remarkably, the overdamped Langevin diffusion process can be identified with the gradient flow of the relative entropy in this formulation. This is the subject of this report. Our goal in this report will be to give an overview of the main ideas needed to define gradient flows of functionals in $(\mathscr{P}_2(\mathbb{R}^d), W_2)$, paying special attention to the case of relative entropy.

We start our discussion by recalling the notion of a gradient flow of a functional in $\mathbb{R}^d$. Almost the same theory holds for a general Hilbert space. In preparation for subsequent sections, we then review results from analysis on metric spaces, as well as from optimal transport theory. Finally, we discuss how the space $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ can be endowed with a differential structure leading to a definition of a gradient flow of a functional. Finally, we discuss the special case of the gradient flow of relative entropy, and give explicit convergence rates in continuous time.

2 Background on gradient flows in $\mathbb{R}^d$

We work in $\mathbb{R}^d$ equipped with the usual inner product $\langle \cdot, \cdot \rangle$.

2.1 The differentiable case

Definition 2.1 (Euclidean gradient flow)**.** Let $F : \mathbb{R}^d \to \mathbb{R}$ be a differentiable functional with Lipschitz gradient. Then the gradient flow of F starting at $x_0 \in \mathbb{R}^d$ is defined to be the unique differentiable curve $x(t) : [0, \infty) \to \mathbb{R}^d$ satisfying the ordinary differential equation:

$$\begin{aligned} \frac{dx(t)}{dt} &= -\nabla F(x(t)) \\ x(0) &= x_0 \end{aligned} \tag{1}$$

From an optimization point of view, the main interest in gradient flow curves comes from the following result.

Theorem 2.2. *Suppose that F has a global minimum x^* and satisfies the gradient domination condition (which is weaker than, and implied by strong convexity):*

$$F(x) - F(x^*) \leq \frac{1}{2\alpha} \left\| \nabla F(x) \right\|_2^2 \tag{2}$$

Then we have:

$$F(x(t)) - F(x^*) \leq \exp\left(-2\alpha t\right)\left[F(x_0) - F(x^*)\right]$$

Proof.

$$\begin{aligned}\frac{d}{dt}\left[F(x(t)) - F(x^*)\right] &= \langle \nabla F(x(t)), \frac{d}{dt}x(t)\rangle \\ &= \|\nabla F(x(t))\|_2^2 \\ &\leq 2\alpha\left[F(x(t)) - F(x^*)\right]\end{aligned}$$

The result then follows from Gronwall's inequality. □

2.2 The λ-convex case

One can similarly define the gradient flow of (possibily non-differentiable) λ-convex functions for $\lambda \in \mathbb{R}$. Recall that a functional F is λ-convex if:

$$F(tx + (1-t)y) \leq tF(x) + (1-t)F(y) - \frac{\lambda}{2}t(1-t)\left\|x-y\right\|_2^2 \tag{3}$$

In this case, one defines the subgradient of F at x as the set:

$$\partial F(x) := \left\{ g \in \mathbb{R}^d \mid F(y) \geq F(x) + \langle g, y-x\rangle \ + \frac{\lambda}{2}\left\|y-x\right\|_2^2 \quad \forall y \in \mathbb{R}^d \right\}$$

Assuming that F has a non-empty subgradient for all $x \in \mathbb{R}^d$, one can then consider the following curve.

Definition 2.3 (Euclidean subgradient flow)**.** Let $F : \mathbb{R}^d \to \mathbb{R}$ be a λ-convex functional with non-empty subgradients $\partial F(x)$ for all $x \in \mathbb{R}^d$. Then the subgradient flow of F starting at $x_0 \in \mathbb{R}^d$ is defined to be the unique absolutely continuous curve $x(t) : [0,\infty) \to \mathbb{R}^d$ satisfying the ordinary differential inclusion:

$$\begin{aligned}&\frac{dx(t)}{dt} \in -\partial F(x(t)) \quad \text{for a.e.} \quad t \in [0,\infty) \\ &x(0) = x_0\end{aligned} \tag{4}$$

The main advantage of this definition is that many of the concepts it uses can be generalized to the metric space case. Let us compare definition 2.1 with definition 2.3. For the first, one needs to have a concept of differentiability of the curve, which requires a normed vector space structure on the space. On the other hand, the second definition only requires absolute continuity, which, as we will see, can easily be generalized to metric spaces. Similarly, in the first definition, one requires differentiability of the functional, which, again, requires the space to be a normed vector space. The second definition only requires the definition of subgradients which, at least intuitively, seem more easily generalizable. Note that one can expand the definition of subgradients and of subgradient flows to more general functional, but we will have no use for such definitions.

3 Analysis on metric spaces

Here we give a few basic definitions that will allow us to generalize some aspects of differentiability to curves on metric spaces. All along, we assume that (X, d) is a complete metric space.

Definition 3.1 (Curve)**.** Let $I \subseteq \mathbb{R}$ be an interval. A continuous function $\gamma : I \to X$ is called a curve.

Definition 3.2 (Metric derivative)**.** Let $\gamma : I \to X$ be a curve. When it exists, we define its metric derivative at $t \in I$ to be:

$$|\gamma'(t)| := \lim_{h \to 0} \frac{d(\gamma(t+h), \gamma(t))}{|h|} \tag{5}$$

Inspired by the characterization of absolutely continuous functions on $\mathbb{R}$ given by the fundamental theorem of calculus, one can define absolutely continuous curves as follows:

Definition 3.3. A curve $\gamma : I \to X$ is said to be absolutely continuous if there exists a $\beta \in L^1(I)$ such that:

$$d(\gamma(s), \gamma(t)) \leq \int_s^t \beta(r) dr \quad \forall s < t \in I \tag{6}$$

The following theorem gives some justification for this definition.

Theorem 3.4. *Let $\gamma : I \to X$ be an absolutely continuous curve. Then for a.e. $t \in I$,γ has a metric derivative, $|\gamma'| \in L^1(I)$, and:*

$$d(\gamma(s), \gamma(t)) \leq \int_s^t |\gamma'|(r)dr \quad \forall s < t \in I$$

Furthermore, for any $\beta \in L^1(I)$ satisfying (6):

$$|\gamma'|(t) \leq \beta(t) \quad \textit{for a.e.} \quad t \in I$$

To see why this justifies the above definition of absolute continuity, consider the following: if X was $\mathbb{R}$ and γ was absolutely continuous (in the usual sense), then it would be a.e. differentiable and we would have equality in the integral (6). The above theorem says that a version of this holds on general metric spaces with the above definition of absolute continuity.

These definitions give us a way to extend the concept of differentiabiliy to general metric spaces in two ways: first through the metric derivative which gives a definition of "speed" of a curve (but no direction), and through the definition of absolutely continuous curves, which we can view as the generalization of differentiable curves in Euclidean space. Surprisingly, as we will see, on the space $(\mathscr{P}_2(\mathbb{R}^d), W_2)$, these are enough to fully specify a notion of derivative for an absolutely continuous curve.

4 Optimal transport theory

In this section, we review relevant results from optimal transport theory. We will restrict our attention to the case of Euclidean space $\mathbb{R}^d$. Given a cost function $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, and two probability measures μ, $\nu \in \mathscr{P}(\mathbb{R}^d)$, the optimal transport problem is given by:

$$\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x, y) d\gamma(x, y) \tag{7}$$

where $\Gamma(\mu, \nu)$ is the set of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν respectively. One can show existence of solutions under mild conditions on $c(x, y)$. Here we will restrict ourselves to the quadratic case $c(x, y) = \|x - y\|_2^2$ and where $\mu, \nu \in \mathscr{P}_2^{ab}(\mathbb{R}^d)$, the space of absolutely continuous probability measures (with respect to Lebesgue measure) on $\mathbb{R}^d$ with finite second moment. For a measurable function T, we define the pushforward measure $T_\#\mu$ to be:

$$T_\#\mu(B) := \mu(T^{-1}(B))$$

With this notation in place, we state the main existence and uniqueness result in our case of interest:

Theorem 4.1 (Existence and uniqueness of optimal transport map)**.** *Let $\mu, \nu \in \mathscr{P}_2^{ab}(\mathbb{R}^d)$ and let $c(x, y) = \|x - y\|_2^2$. Then there is a unique optimal transport plan $\gamma^* \in \Gamma(\mu, \nu)$ minimizing (7). Furthermore, there is a unique optimal transport map t_μ^ν such that $t_{\mu\#}^\nu \mu = \nu$, and $\gamma^* = (Id, t_\mu^\nu)_\# \mu$ where Id is the identity map.*

Using this existence result, it is not hard to show that the following function:

$$W_2(\mu, \nu) := \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 \, d\gamma(x, y) \right)^{1/2} \qquad \forall \mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$$

defines a metric on $\mathscr{P}_2(\mathbb{R}^d)$. In fact $(\mathscr{P}_2(\mathbb{R}^d),\, W_2)$ is a complete separable metric space, so that the results and definitions from the previous section apply. In particular, we can ask if there is a way to characterize absolutely continuous curves in this metric space. Surprisingly, there is, and this will turn out to be crucial in defining a notion of derivative for such curves. The result is the following:

Theorem 4.2. *Let I be an open interval in $\mathbb{R}$. Let $(\mu_t)_{t \in I}$ be a curve in $\mathscr{P}_2(\mathbb{R}^d)$. Then $(\mu_t)_{t \in I}$ is absolutely continuous if and only if there exists for each $t \in I$ a measurable vector field $v_t : \mathbb{R}^d \to \mathbb{R}^d$ such that:*

- *$v_t \in L^2(\mu_t, \mathbb{R}^d)$ for a.e. $t \in I$ $\left(\int_{\mathbb{R}^d} \|v_t\|_2^2 \, d\mu_t < \infty\right)$.*
- *$\|v_t\|_{L^2(\mu_t, \mathbb{R}^d)} = |\mu'|(t)$ for a.e. $t \in I$.*
- *the continuity equation:*

$$\partial \mu_t + \nabla \cdot (v_t \mu_t) = 0 \tag{8}$$

holds in the sense of distributions with compactly supported smooth test functions.

5 Subgradients in $\mathscr{P}_2(\mathbb{R}^d)$ and gradient flows

Let us know view the results of the previous section from the perspective of trying to build gradient flows in $\mathscr{P}_2(\mathbb{R}^d)$. In light of theorem 4.2, it seems natural to identify the "derivative" at time t of an absolutely continuous curve at in $\mathscr{P}_2(\mathbb{R}^d)$ with the vector field v_t. This is because, on the one hand, the L^2 norm of v_t matches the metric derivative, and on the other, $(v_t)_{t \in I}$ completely characterizes the curve through the continuity equation, just like the derivative of a real-valued absolutely continuous function characterizes it. We make therefore the heuristic association "$\frac{d}{dt}\mu_t$" $\equiv v_t$.

It remains to specify the "subgradient" of a given functional $\mathscr{F} : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R}$. First let us define λ-convex functionals on $\mathscr{P}_2(\mathbb{R}^d)$. Perhaps the obvious way to do so is to notice that $\mathscr{P}_2(\mathbb{R}^d)$ is a convex subset of the vector space $\mathcal{M}(\mathbb{R}^d)$ of finite signed measures, and therefore convex combinations of probability measures are well defined. Looking at the definition of λ-convexity for functional on $\mathbb{R}^d$, it then seems natural to simply replace the term $\|x-y\|_2^2$ by $W_2^2(x,y)$. While this definition seems reasonable, it turns out that it is better to consider convex combination of probability measures along geodesics of the space $(\mathscr{P}_2(\mathbb{R}^d), W_2^2)$. We have the following definition:

Definition 5.1. $\mathscr{F} : \mathscr{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is a λ-convex (along geodesics) functional if for any $\mu_0, \mu_1 \in \mathscr{P}_2(\mathbb{R}^d)$ and any optimal transport plan γ between them, we have:

$$\mathscr{F}(\mu_t) \leq (1-t)\mathscr{F}(\mu_0) + t\mathscr{F}(\mu_1) - \frac{\lambda}{2}t(1-t)W_2^2(\mu_0,\mu_1)$$

where $\mu_t = ((1-t)\pi_1 + t\pi_2)_{\#}\gamma$ and π_1, π_2 are the projections onto the first and second coordinate.

Now, looking at the definition of the subgradient for λ-convex functionals in the Euclidean case, we see that we need a replacement for the last two terms. The term $\|y-x\|_2^2$ can readily be replaced by $W_2^2(\mu,\nu)$. The problematic term is the inner product. In particular, we need a way to replace the term $y-x$. Let $\mu \in \mathscr{P}_2^{ab}(\mathbb{R}^d)$ be the probability measure at which we would like to define a subgradient. By theorem 4.1, we can associate a transport map t_μ^ν to every probability measure $\nu \in \mathscr{P}_2^{ab}$ in a unique way. Therefore, for a reference measure μ, we can make the association "$\nu - \mu$" $\equiv t_\mu^\nu - Id$, and take advantage of the inner product of the Hilbert space $L^2(\mu, \mathbb{R}^d)$ to define the inner product term, and hence the subgradient:

Definition 5.2. Let $\mathscr{F} : \mathscr{P}_2^{ab}(\mathbb{R}^d) \to \mathbb{R}$ be λ-convex. We define the subgradient of $\mathscr{F}$ at $\mu \in \mathscr{P}_2^{ab}(\mathbb{R}^d)$ to be:

$$\partial\mathscr{F}(\mu) := \left\{ \xi \in L^2(\mu,\mathbb{R}^d) \mid \mathscr{F}(\nu) \geq \mathscr{F}(\mu) + \langle \xi, t_\mu^\nu - Id \rangle + \frac{\lambda}{2}W_2^2(\mu,\nu) \ \forall \nu \in \mathscr{P}_2^{ab}(\mathbb{R}^d) \right\}$$

where the inner product is taken in $L^2(\mu, \mathbb{R}^d)$.

Equipped with these definitions, we are now finally ready to formulate the notion of a gradient flow of a functional in $\mathscr{P}_2(R^d)$.

Definition 5.3. The gradient flow of a λ-convex functional $\mathscr{F} : \mathscr{P}_2^{ab}(\mathbb{R}^d) \to \mathbb{R}$ starting at $\mu_0 \in \mathscr{P}_2^{ab}(\mathbb{R}^d)$ is the unique absolutely continuous curve $\gamma : [0, \infty) \to \mathscr{P}_2^{ab}(\mathbb{R}^d)$ such that:

$$\begin{aligned} v_t &\in \partial \mathscr{F}(\gamma(t)) \quad \text{for a.e.} \quad t \in [0, \infty) \\ \gamma(0) &= \mu_0 \end{aligned}$$

In words, what we have been able to achieve is the following. We have been able to define the equivalent of the derivative of an absolutely continuous curve in $(\mathscr{P}_2(\mathbb{R}^d), W_2)$ as a vector field in the Hilbert space $L^2(\mu, \mathbb{R}^d)$. This makes sense of the left-hand side of the gradient flow equation. Then, for a reference probability measure μ, we were able to map all probability measures to the space $L^2(\mu, \mathbb{R}^d)$ through the optimal transport map associated with them. This allows us to leverage the vector-space structure of $L^2(\mu, R^d)$ to define subgradients of functionals, thereby making sense of the right-hand side of the gradient flow equation.

6 Gradient flow of relative entropy

Given a probability measure with density π that we would like to sample from, we can attempt to minimize the relative entropy functional $\mathcal{H}_\pi : \mathscr{P}_2^{ab}(\mathbb{R}^d) \to \mathbb{R}$ defined by:

$$\mathcal{H}_\pi(\rho) := \int_{\mathbb{R}^d} \log\left(\frac{\pi(x)}{\rho(x)}\right) \pi(x) dx \tag{9}$$

For this, we construct the gradient flow associated with this functional. We will be able to show that the resulting curve will be described by a Fokker-Planck equation, therefore making the connection with overdamped Langevin diffusion. We will also show that under the same gradient domination as in theorem 2.2, we obtain linear convergence to the target measure π.

To proceed, we first need to find the subgradient of the relative entropy. It is a theorem that for integral functionals defined on absolutely continuous probability measures, the subgradient is unique and is given by:

$$\partial \mathcal{H}_\pi(\mu) = \nabla \left(\frac{\delta \mathcal{H}}{\delta \rho}\right)$$

In the case of relative entropy, the first variation is given by:

$$\frac{\delta \mathcal{H}_\pi}{\delta \rho} = -\log \frac{\rho}{\pi}$$

Therefore the vector field v_t defining the gradient flow of $\mathcal{H}_\pi$ is given by:

$$v_t(x) = \partial \mathcal{H}_\pi(\mu) = -\nabla \log \frac{\rho(x)}{\pi(x)}$$

and the resulting continuity equation is:

$$\begin{aligned}
\frac{\partial}{\partial t}\rho &= \nabla \cdot (\rho \nabla \log \rho + \rho \nabla \log \pi) \\
&= \nabla \cdot \left(\rho \frac{\nabla \rho}{\rho}\right) + \nabla \cdot (\rho \nabla \log \pi) \\
&= \nabla \cdot (\rho \nabla \log \pi) + \Delta \rho
\end{aligned}$$

which is precisely the Fokker-Planck equation for the overdamped Langevin diffusion process given by the SDE:

$$dX_t = -\nabla \log \pi dt + \sqrt{2} dW_t$$

Finally, we have the following convergence rate, which mirrors exactly that of gradient flow on Euclidean space:

Theorem 6.1. *Suppose that the following gradient domination condition (know as the log-sobolev inequality) holds:*

$$\mathcal{H}_\pi(\rho) \leq \frac{1}{2\alpha} \left\| \partial \mathcal{H}_\pi(\rho) \right\|_{L^2(\rho, \mathbb{R}^d)} = \frac{1}{2\alpha} \int_{\mathbb{R}^d} \rho \left\| \nabla \log \frac{\rho}{\pi} \right\|_2^2 dx$$

then we have:

$$\mathcal{H}_\pi(\rho(t))) \leq \exp(-2\alpha t) \mathcal{H}_\pi(\rho(0))$$

Proof.

$$\begin{aligned}
\frac{d}{dt}\mathcal{H}_\pi(\rho(t)) &= \int_{\mathbb{R}^d} \frac{\partial \rho}{\partial t} \log \frac{\rho}{\pi} dx + \int_{\mathbb{R}^d} \rho \frac{\partial}{\partial t} \log \frac{\rho}{\pi} dx \\
&= \int_{\mathbb{R}^d} \nabla \cdot (\rho \nabla \log \frac{\rho}{\pi}) \log \frac{\rho}{\pi} + \int_{\mathbb{R}^d} \frac{\partial \rho}{\partial t} dx \\
&= \int_{\mathbb{R}^d} \nabla \cdot \left(\rho \log \frac{\rho}{\pi} \nabla \log \frac{\rho}{\pi}\right) dx - \int_{\mathbb{R}^d} \rho \left\| \nabla \log \frac{\rho}{\pi} \right\|_2^2 dx + \frac{\partial}{\partial t} \int_{\mathbb{R}^d} \rho dx \\
&= -\int_{\mathbb{R}^d} \rho \left\| \nabla \log \frac{\rho}{\pi} \right\|_2^2 dx \\
&\leq -2\alpha \mathcal{H}_\pi
\end{aligned}$$

and the result follows from Gronwall's inequality. □

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures.* 2008.

[2] Grigorios A. Pavliotis. *Stochastic Processes and Applications.* 2014.

[3] Filippo Santambrogio. Optimal transport for applied mathematicians : calculus of variations, PDEs, and modeling. *Book*, 2015.

[4] Filippo Santambrogio. Euclidean, metric, and Wasserstein gradient flows: an overview. *Bulletin of Mathematical Sciences*, 2017.

[5] Cédric Villani. Optimal Transport Old and New. *Media*, 2007.

[6] Cedric Villani. *Topics in optimal transportation.* 2003.