Federal State Autonomous
Educational Institution for the Higher Education
National Research University
"Higher School of Economics"
Faculty of Mathematics

Bobrin Maxim Sergeevich

Stability of Gradient flows in Wasserstein spaces

Term paper for the 3rd year of study

Degree programme: bachelor's educational programme "Mathematics"

Scientific supervisor:
Doctor of Sciences, professor
Mauro Mariani

Moscow 2021

1. Introduction

A usual algorithm for sampling from a distribution $e^{-V}dx$ on $\mathbb{R}^d$ is to run a MonteCarlo Markov Chain (MCMC) which admits $e^{-V}dx$ as invariant measure. The classical MCMC approach is based on a Langevin dynamics. It is known, based on results of [3], that such (reversible) Langevin dynamics induces a gradient flow evolution on the space of probability measures endowed with Wasserstein-2 metric.

In many modern applications, the potential V is characterized by a complex landscape, featuring several local minima, to be interpreted as small traps where the deterministic gradient flow dynamics $\dot{X} = -\nabla V(X)$ gets stuck. This may often be modeled as a small perturbation V^ε of a more regular potential V. A simple but clear example to keep in mind is $V^\varepsilon(x) = x^2 + \varepsilon^\alpha \sin(\varepsilon x)$, where the (Sobolev) nature of the convergence $V^\varepsilon \to V$ depends on α.

The main focus of this work is to study conditions which guarantee the convergence of the Langevin dynamics as $\varepsilon \to 0$ under oscillations of V as in the example above, namely assuming a strong convergence of V but not of its derivative.

So we introduce a sequence V^ε that should be thought as fast oscillating. It is clear that, if the dynamics is deterministic $\dot{X}^\varepsilon = -\nabla V^\varepsilon(X^\varepsilon)$, even if the potential converges uniformly to some V, the solution will not converge to the limiting solution, as it may remain 'trapped' in the local minimizers of V^ε.

Here we want to show that this non-converging phenomenon is mitigated by the presence of noise. Namely, the solution to $\dot{X}^\varepsilon = -\nabla V^\varepsilon(X_\varepsilon) + \sqrt{2c}\dot{W}$ will converge to the limiting stochastic equation if V^ε converges (but not its gradient). The case when $c = c_\varepsilon \to 0$, should let us find quantitative estimates, to guarantee that this well-behaved limits also take place, even in presence of vanishing noise. In other words, we want to estimate the 'minimal size' of noise, that guarantees the process not to be trapped in the minimizers of V^ε. It should be possible to achieve this estimates by the same methods explained in this paper in the non-vanishing noise case, but this will be subject of later investigation.

2. Main result

Let V^ε be a sequence of smooth potentials, and consider the stochastic equation on $\mathbb{R}^d$

$$\dot{X} = -\nabla V^\varepsilon(X) + \sqrt{2c}\dot{W}$$

with random initial condition X_0^ε having distribution μ_0^ε. Let X_t^ε be its solution and $\mu_t^\varepsilon \in \mathcal{P}(\mathbb{R}^d)$ be its distribution. Our main result is the following

Theorem 2.1. *Let $V\colon \mathbb{R}^d \to \mathbb{R}$ be smooth (or more generally semiconvex i.e $x \to f(x) + \frac{1}{2}C|x|^2$ is convex with constant $C > 0$), and such that $\int x^2 e^{-V(x)/c}dx < \infty$. Assume that V^ε converges to V uniformly and that the initial distribution μ_0^ε converges to a limiting distribution μ_0 and that its relative entropy $H^\varepsilon(\mu_0^\varepsilon)$ converges to $H(\mu_0)$ (see Section 3 for the definition of the entropy of the initial data). Then X_t^ε converges in law to X_t solution to the limiting equation*

$$\dot{X} = -\nabla V(X) + \sqrt{2c}\dot{W} \tag{2.1}$$

with initial condition μ_0.

The precise definition of the relative entropy condition on the initial data is explained in detail below. Notice that we do not assume any convergence on the gradients of the potential, or uniform semiconvexity, which would make this results a consequence of more general principles. The parameter c here is purely cosmetic, and while the result can be easily generalized to the case where $c^\varepsilon \to c > 0$, the case $c^\varepsilon \to 0$ is our next target of study.

The idea is to avoid a differential approach to the convergence, and to rather use a variational one to prove it under minimal assumptions.

3. Information-theoretic functions on spaces of probability measures

In order to sample from the target distribution, it is enough to choose and optimize some objective function that will be minimizer at the target distribution.

3.1. **Relative entropy.** If $h\colon [0,\infty) \to [0,\infty)$ is a continuous, convex function such that $h(1) = 0$, and if $m \in \mathcal{P}(E)$ is a reference probability measure, one can define

$$\begin{aligned} &H_h\colon \mathcal{P}(E) \to [0,\infty] \\ &H_h(\mu) := \begin{cases} \int h(\varrho(x))dm(x) & \text{if } \frac{d\mu}{dm} = \varrho \\ +\infty & \text{if } \mu \not\ll m \end{cases} \end{aligned} \tag{3.1}$$

$H_h(\mu)$ called the h-relative entropy of μ w.r.t. m. If $h(u) = u(\log u - 1) + 1$, then we drop the index h and call it just relative entropy. So for $\varrho = d\mu/dm$

$$\begin{aligned} H(\mu) &= \int h(\varrho)\, dm = \int \varrho(x) \log \varrho(x) dm(x) = \int \log \varrho(x)\, d\mu(x) \\ &= \sup_{f \text{ bounded measurable}} \int f\, d\mu - \log \int e^f\, dm \end{aligned}$$

Note that H is lower semicontinuous on $\mathcal{P}(E)$ in the weak topology. Indeed, if $\mu_n \to \mu$ weakly, then $\lim_n H(\mu_n) = \lim_n \sup_{f\text{continuous}} \int f\, d\mu_n - \log \int e^f\, dm \geq \sup_f \lim_n \int f\, d\mu_n - \log \int e^f\, dm = \sup_f \int f\, d\mu - \log \int e^f\, dm = H(\mu)$.

3.2. **Fisher Information.** For $E = \mathbb{R}^d$ and $m = e^{-V}/Z$ a Borel measure on $\mathbb{R}^d$ define the Fisher information $I\colon \mathcal{P}(\mathbb{R}^d) \to [0,\infty]$ as

$$I(\mu) = \begin{cases} \int \frac{|\nabla \varrho(x)|^2}{\varrho(x)} dm(x) & \text{if } \frac{d\mu}{dm} = \varrho \\ +\infty & \text{otherwise} \end{cases}$$

Where the first integral means $+\infty$ if ϱ does not have a weak gradient in L_2.

To our aim, it is interesting to notice that the Fisher information can be also obtained as a metric derivative of the relative entropy in the Wasserstein-2 space.

Namely define

$$|DH|(\mu) := \varlimsup_{\nu \to \mu} \frac{\big[H(\mu) - H(\nu)\big]^+}{W_2(\mu,\nu)} \in [0,\infty]$$

where W_2 denotes the Wasserstein distance and a^+ the positive part of the real number a (notice that H is not continuous, so the positive part is needed in place of the absolute value).

Then $I(\mu) = |DH|^2(\mu)$. While we do not need formally this result, it is interesting to give a quick informal derivation of it. Fix μ and take a sequence ν such that $\nu \to \mu$ while minimizing $H(\mu) - H(\nu)$. If we call $\varrho = d\mu/dm$, $\chi = d\nu/dm$

$$\begin{aligned} H(\mu) - H(\nu) &= \int_E \log \varrho(x) d\mu(x) - \int_E \log \chi(y) d\nu(y) \\ &= \int_{E\times E} (\log \chi(x) - \log \chi(y)) d\gamma(x,y) + \int_E \log(\varrho(x)/\chi(x)) d\mu(x) \\ &\leq \int_{E\times E} (\log \chi(x) - \log \chi(y)) d\gamma(x,y) \end{aligned}$$

for any coupling $\gamma \in \Gamma(\mu,\nu)$ between μ and ν. In particular we can take the optimal (for the W_2-distance) coupling γ. By (informally applying) Taylor expansion

$$|\log \chi(y) - \log \chi(x)| \leq \frac{|\nabla \chi(x)|}{\chi(x)} d(x,y) + o(d(x,y))$$

So that, informally assuming that χ converges to ϱ in a strong sense

$$\begin{aligned} &[H(\mu) - H(\nu)]^+ \leq \int \tfrac{1}{\chi(x)} |\nabla \chi(x)| d(x,y) + o(d(x,y)) d\gamma(x,y) \\ &\leq \Big(\int \tfrac{1}{\chi(x)^2} |\nabla \chi(x)|^2) d\gamma(x,y) \Big)^{1/2} \Big(\int d(x,y)^2 d\gamma(x,y) \Big)^{1/2} + o(W_2(\nu,\mu)) \\ &= \big(\int \tfrac{|\nabla \chi(x)|^2 \varrho(x)}{\chi^2(x)} dm(x) \big)^{1/2} W_2(\mu,\nu) + o(W_2(\nu,\mu)) \\ &= \big(\int \tfrac{|\nabla \varrho(x)|^2}{\varrho(x)} dm(x) \big)^{1/2} W_2(\mu,\nu) + o(W_2(\nu,\mu)) \end{aligned}$$

So this suggests the above definition of I.

Remark 3.1. *The Fisher information writes as*

$$I(\mu) = \sup_v 2 \int \operatorname{div}(v) e^V d\mu - \int |v|^2 e^{2V} d\mu \tag{3.2}$$

where the supremum is taken over all smooth, compactly supported $v \colon \mathbb{R}^d \to \mathbb{R}$.

The proof of this useful formula is rather elementary, and just a version of Riesz representation theorem. In particular, if $I(\mu) < \infty$, then $\mu = \varrho e^{-V} dx$, where we drop the normalization constant Z since it is easily seen that I does not depend on it. Then, since ϱ has a weak gradient, we can integrate by parts the $\operatorname{div}(v)$ in the left hand side of (3.2) to obtain $\sup_v \int -2v \cdot \nabla \varrho \, dx - |v|^2 e^V \varrho \, dx$. This is the supremum of a quadratic form immediately yielding (3.2).

Other characterizations of the Fisher information, as for instance the Dirichlet form of the process (2.1) on the square root of ϱ are also interesting but will not play a role here.

4. Langevin dynamics and gradient flows

Notice that any probability measure with strictly positive density on $\mathbb{R}^d$ can be written in the form $m = e^{-V} dx$. We consider the evolution:

$$\dot{X} = -\nabla V(X) + \sqrt{2c} \dot{W} \tag{4.1}$$

and notice that, it has

$$r_c(x) = \frac{1}{Z} e^{-V(x)/c}, \quad Z_c = \int_{\mathbb{R}^d} e^{-V(x)/c} dx$$

as stationary distribution (unique invariant measure) [6].

That is, if we start with initial density r_c, then it will be preserved by the flow of (4.1). The law p_t of the solution to (4.1) solves the linear backward *Kolmogorov-Fokker-Planck* equation on $\mathbb{R}^d$

$$\partial_t p_t = c\Delta p_t + \nabla \cdot (p_t \nabla V)$$

where here and in the following we will always identify an absolutely continuous measure with its density.

Notice that the relative entropy with reference measure r_c decomposes as a sum of two terms:

$$H(p) = \tfrac{1}{c} \int V(x) p(x)\, dx + \int p(x) \log p(x) dx$$

The distribution X_t with density p_t will converge to the target measure $m = e^{-V/c}$ along Langevin dynamics, as we state quickly in the following remark.

Remark 4.1. *For H and I the relative entropy and Fisher information with reference measure r_c, p_t satisfies*

$$\frac{d}{dt} H(p_t) = -I(p_t)$$

Proof. We give a quick prove since this is not needed in the formal proof of the main result. Recalling the notation of (3.1), here we have $p_t = r_c \varrho$. So

$$\begin{aligned} \frac{d}{dt} H(p_t) &= \frac{d}{dt} \int p_t \log(\frac{p_t}{r_c}) dx = \int \frac{\partial p_t}{\partial t} \log \frac{p_t}{r_c} dx \\ &= -\int \nabla(p_t \nabla \log \frac{p_t}{r_c}) \log \frac{p_t}{r_c} dx = -\int p_t ||\nabla \log \frac{p_t}{r_c}||^2 dx = -I(p_t) \end{aligned}$$

□

In particular, since $I(\mu) \geq 0$ we have that relative entropy is indeed decreasing.

This, together with the fact that $I = |DH|^2$, suggests that μ_t satisfies a formal gradient flow equation $\partial_t \mu_t = -DH(\mu_t)$ in a space of probability measures. We recall how to rigorously formulate gradient flow evolutions in metric spaces in the appendix, and here we just recall a formal result by [3].

Theorem 4.2 (Jordan)**.** *Let H and I be the relative entropy and Fisher information with reference measure r_c. Assume that $H(\mu_0) < \infty$. Then the Kolmogorov evolution equation (see the appendix) for the law μ_t of the solution X_t to* (2.1) *is equivalent to the inequality*

$$H(\mu_0) \geq H(\mu_t) + \tfrac{1}{2} \int_0^t I(\mu_s)\, ds + \tfrac{1}{2} \int_0^t \|\partial_s \mu_s\|^2\, ds, \qquad \forall t \geq 0 \tag{4.2}$$

where $\|\partial_s \mu_s\| = \overline{\lim}_{t \to s} \frac{W_2(\mu_t, \mu_s)}{|t-s|}$ is the metric derivative of μ.

5. Convergence under constant noise

In this section we prove the main Theorem 2.1.

We need to prove that for each $t > 0$ and each smooth, compactly supported function f, it holds

$$\lim_{\varepsilon \to 0} \mathbb{E}_x[f(X_t^\varepsilon)] = \mathbb{E}_x[f(X_t)]$$

Then, define $u^\varepsilon(t,x) = \mathbb{E}_x[f(X_t^\varepsilon)]$, $u(t,x) = \mathbb{E}_x[f(X_t)]$ and recall Kolmogorov equation $\partial_t u^\varepsilon = A^\varepsilon u^\varepsilon$. We need to check that it converges to the solution to $\partial_t u = Au$.

Lemma 5.1. *Suppose that we have a curve μ_t^ε is the Wasserstein space $\mathcal{P}(\mathbb{R}^d)$. Suppose that we know that, for every $t \geq 0$, μ_t^ε is precompact in $\mathcal{P}(\mathbb{R}^d)$, meaning that it admits limit points μ_t (in the weak topology of probabilities) along subsequences. Suppose that for some V^ε, μ_t^ε satisfies*

$$H^\varepsilon(\mu_0^\varepsilon) \geq H^\varepsilon(\mu_t^\varepsilon) + \tfrac{1}{2}\int_0^t I^\varepsilon(\mu_s^\varepsilon)ds + \tfrac{1}{2}\int_0^t |\dot{\mu}_s^\varepsilon|^2 ds \tag{5.1}$$

with H^ε and I^ε the relative entropy and Fisher information built with V^ε.

Suppose that one can prove that for each ν in $\mathcal{P}(\mathbb{R}^d)$ and for each sequence ν^ε converging to ν it holds

$$\varliminf_\varepsilon H^\varepsilon(\nu^\varepsilon) \geq H(\nu) \tag{5.2}$$

$$\varliminf_\varepsilon I^\varepsilon(\nu^\varepsilon) \geq I(\nu) \tag{5.3}$$

where H, I are the relative entropy and Fisher information built with V. Then every limit point μ_t is in $\mathcal{P}_2(\mathbb{R}^d)$ of μ_t^ε is a solution to (4.2), provided $\lim_\varepsilon H^\varepsilon(\mu_0^\varepsilon) = H(\mu_0)$.

Proof. Let us take the limit $\varepsilon \downarrow 0$ in (5.1) along a converging subsequence. Then by hypotheses

$$\begin{aligned} H(\mu_0) = \varliminf_\varepsilon H^\varepsilon(\mu_0^\varepsilon) &\geq \varliminf_{\varepsilon\to 0} H(\mu_t^\varepsilon) + \tfrac{1}{2}\varliminf_\varepsilon \int_0^t I^\varepsilon(\mu_s^\varepsilon)ds + \varliminf_\varepsilon \tfrac{1}{2}\int_0^t \|\partial_s \mu_s^\varepsilon\|^2 ds \\ &\geq H(\mu_t) + \tfrac{1}{2}\int_0^t I(\mu_s)ds + \tfrac{1}{2}\int_0^t \|\partial_s \mu_s\|^2 ds \end{aligned}$$

where we used (5.2) for the term involving H, we used Fatou's lemma and (5.3) and for the term involving I. For the last term, notice

$$\tfrac{1}{2}\int_0^t \|\partial_s \mu_s^\varepsilon\|^2 ds = \sup_f \mu_t^\varepsilon(f_t) - \mu_0^\varepsilon(f_0) - \int_0^t \mu_s^\varepsilon(\partial_s f_s)\, ds - \tfrac{1}{2}\int_0^t \mu_s^\varepsilon(|\nabla f_s|^2)\, ds$$

where the supremum is taken over smooth functions $f(s,x) \equiv f_s(x)$. So that the inequality is achieved exchanging the limit and the supremum. □

Lemma 5.2. *Under the same hypothesis as in Theorem 2.1, μ_t^ε is tight, and thus precompact, for every $t \geq 0$.*

Proof. By Prohorov theorem, one needs to check that for each $\delta > 0$ there is a compact K_δ in $\mathbb{R}^d$ such that $\mu_t^\varepsilon(K_\delta) \geq 1 - \delta$.

Fix a measurable set A in $\mathbb{R}^d$. Since for every f bounded measurable we have $\mu(f) \le H(\mu) + \log(m(e^f))$, taking $f(x) = (\log(1 + 1/m(A)))\mathbf{1}_A(x)$ we get

$$\mu_t^\varepsilon(A) \le \frac{H(\mu_t^\varepsilon) + \log 2}{\log(1 + 1/m(A))} \le \frac{H(\mu_0^\varepsilon) + \log 2}{\log(1 + 1/m(A))}$$

Since $H(\mu_0^\varepsilon)$ is bounded uniformly in ε. Then take $A = ([-L, L]^d)^c$, and notice that $\lim_L m(A) = 0$, so that we can take L large enough to have $\mu_t^\varepsilon([-L, L]^d) \ge 1-\delta$. □

Lemma 5.3. *Suppose that $\mu_t \in \mathcal{P}(\mathbb{R}^d)$ satisfies (4.2) and that $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Then $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ for every $t \ge 0$.*

Proof. We need to prove $\int |x|^2 d\mu_t(x) < \infty$ for every $t \ge 0$.

We have for every $t \ge 0$ and smooth $f(x)$ with bounded derivative

$$H(\mu_0) \ge \tfrac{1}{2} \int_0^t \|\partial_s \mu_s\|^2 ds \ge \ \mu_t(f) - \mu_0(f) - \tfrac{1}{2} \int_0^t \mu_s(|\nabla f|^2)\, ds$$

For $n \ge 2$, take $f = f_n$ defined as

$$f_n(x) = \begin{cases} |x|^2/2 & \text{if } |x| \le n \\ \chi_n(|x|) & \text{if } n \le |x| \le n+1 \\ (n+1)^2/2 & \text{if } |x| \ge n+1 \end{cases}$$

with $\chi_n(r)$ and increasing smooth function in $r \in [n, n+1]$ such that f_n so defined is smooth, $\chi_n(x) \le x^2/2 + 1$ and $|\chi_n'(r)|^2 \le 6n^2$ (where the constant 6 is choses as to have such a χ to exist). Then, since $\chi_n(|x|) \ge n^2$, we have that $|\nabla f_n|^2 \le 6 f_n$. Taking also $f(x) = x^2/2$, so that $f_n \le f + 1$

$$\mu_t(f_n) \le H(\mu_0) + \mu_0(f_n) + \tfrac{1}{2} \int_0^t \mu_s(|\nabla f_n|^2)\, ds \le H(\mu_0) + \mu_0(f+1) + 3 \int_0^t \mu_s(f_n)\, ds$$
$$= C + 3 \int_0^t \mu_s(f_n)\, ds$$

So that by Gronwall lemma:

$$\mu_t(f_n) \le e^{3t} C$$

uniformly in n. In particular, by Fatou's lemma

$$\mu_t(f) \le \lim_n \mu_t(f_n) \le e^{3t} C < \infty$$

□

Lemma 5.4. *Suppose that V^ε and V are as in Theorem 2.1, then (5.2) holds.*

Proof. Notice that $r_c^\varepsilon := e^{-V^\varepsilon/c} dx / Z^\varepsilon$ converges weakly to $r_c := e^{-V/c} dx / Z$. Then if $\nu^\varepsilon \to \nu$ and f is continuous and bounded on $\mathbb{R}^d$

$$\varliminf_\varepsilon H^\varepsilon(\nu^\varepsilon) \ge \varliminf_\varepsilon \int f d\nu_\varepsilon - \log \int e^f r_c^\varepsilon\, dx = \int f d\nu - \log \int e^f\, r_c\, dx$$

Taking the supremum over f we get (5.2). □

Lemma 5.5. *Suppose that V^ε and V are as in Theorem 2.1, then (5.3) holds.*

Proof. Let $\nu^\varepsilon \to \nu$ and let us consider the subsequence along which the $\underline{\lim}$ is actually a limit. If $I(\nu^\varepsilon) = +\infty$, there is nothing to prove. So we can assume $d\nu^\varepsilon = \varrho^\varepsilon dr_c^\varepsilon$. Then using Remark 3.1 with V replaced by V^ε

$$\begin{aligned} I^\varepsilon(\nu^\varepsilon) &= \varliminf_\varepsilon \sup_v 2\int \operatorname{div}(v)e^{V^\varepsilon}d\nu^\varepsilon - \int |v|^2 e^{2V^\varepsilon}d\nu^\varepsilon \\ &\geq \sup_v \varliminf_\varepsilon 2\int \operatorname{div}(v)e^{V^\varepsilon}d\nu^\varepsilon - \int |v|^2 e^{2V^\varepsilon}d\nu^\varepsilon \end{aligned}$$

Now, since v is smooth and compactly supported. and V^ε converges uniformly, the integrals in the last line converge to the limiting ones, by a strong-weak convergence. So we get $\varliminf_\varepsilon I^\varepsilon(\nu^\varepsilon) \geq \sup_v 2\int \operatorname{div}(v)e^V d\nu - \int |v|^2 e^{2V} d\nu$. □

Proof of Theorem 2.1. Let μ_t^ε be the law of X_t^ε and let μ_t be a limit point, which exists by the previous lemma. We know from the lemmas above that μ_t satisfies (5.1), namely μ_t is the law of X_t. Since it satisfies (4.2), $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ by Lemma 5.3. □

Appendix

Gradient flows and corresponding energy functionals

Gradient flows. We will study gradient flows on the space of measures endowed with Wasserstein metric. Let us first review the classical case of gradient flows on $\mathbb{R}^d$ to get a hint about their generalizations on metric spaces.

Definition (Euclidean case)**.** *A dynamics on $E = \mathbb{R}^d$ is specified by the following differential equation:*

$$\dot{X}_t = F(X_t)$$

where $F\colon E \to \mathbb{R}^d$ is a Lipschitz vector field, that assigns the velocity vector to every point in space. Then, gradient flow dynamics of a smooth potential V in continuous time is the curve evolving according to the differential equation:

$$\dot{X}_t = -\nabla V(X_t)$$

Remark 5.6. *If $F(x) = -\nabla V(x)$, then we can write it as a solution to the following optimization problem:*

$$-\nabla V(x) = \arg\min_{v\in E}\left\{\langle \nabla V(x), v\rangle + \frac{1}{2}||v||^2\right\}$$

To generalize this definitions to metric spaces, we miss the notion of gradient of a function in a metric space. However, the modulus of the gradient of a lower semicontinuous function V is well defined on a metric space as well, as

$$|DV|(x) := \varlimsup_{y\to x} \frac{\big[V(x) - V(y)\big]^+}{d(x,y)} \in [0,\infty]$$

Hereafter E is a complete metric space and $V\colon E \to (-\infty,\infty]$ lower semicontinuous. We say that a curve $X \in C([0,\infty); E)$ is a gradient flow for V starting at X_0 if

$$V(X_0) \geq V(X_t) + \tfrac{1}{2}\int_0^t |DV|^2(X_s)ds + \tfrac{1}{2}\int_0^t |\dot{X}_s|^2 ds$$

As an example, let us consider the following statement:

Example 5.7. *The heat equation can be viewed as the gradient flow for the Dirichlet form on* $L_2(S^1)$

$$\mathcal{E}(f) = \frac{1}{2}\int_{S^1} |\nabla f|^2 dx$$

That is Δf *is the negative gradient of Dirichlet energy* $D(f,f)$*, which is understood to be* $+\infty$ *if* f *does not have a distributional derivative in* L_2 *(i.e there is no function* $g \in L_2$ *such that* $\int \nabla f g = -\int \nabla g f$*).*

Note. Heat equation can be understood as gradient flow in two senses: as gradient flow of the Dirichlet enertgy in L_2 space and also of the relative entropy in the Wasserstein space.

Proof. Let us consider $\mathcal{F} = L^2(\mathcal{X}, dx)$, i.e the vector space if square-integrable functions $f : \mathcal{X} \to \mathbb{R}$. with $\int_{\mathcal{X}} |f(x)|^2 < \infty$. Then, given any functional $\varepsilon : \mathcal{F} \to \mathbb{R}$ we define its gradient as a function $\partial\varepsilon(f) : \mathcal{X} \to \mathbb{R}$ with the property that it gives best linear approximation, that is:

$$\varepsilon(f + hg) = \varepsilon(f) + \langle \partial\varepsilon(f), g\rangle h + O(h^2)$$

Where g is an element from tangent vector space with base point f. Now, let us take ε as Dirichlet energy:

$$\varepsilon(f) = D(f,f) = \frac{\eta}{2}\int_{\mathcal{X}} ||\nabla f(x)||^2 dx,\ \eta > 0$$

Thus

$$\begin{aligned}\mathcal{E}(f+u) &= \frac{1}{2}\int_{S^1} \|\nabla f + \nabla u\|^2 dx \\ &= \frac{1}{2}\int_{\mathcal{X}} \|\nabla f\|^2 dx + \int_{\mathcal{X}} \langle \nabla f, \nabla u\rangle dx + \frac{1}{2}\int_{\mathcal{X}} \|\nabla u\|^2 dx \\ &= \mathcal{E}(f) - \langle \Delta f, u\rangle + \mathcal{E}(u)\end{aligned}$$

□

Geodesics

We do not use geodesics in this paper, but if instead of $\mathbb{R}^d$ we chose Riemannian manifold, then geodesic curves would be crucial.
Relationship between two tangent spaces of different points on a smooth manifold is determined by *affine connection*. Thus, for a vector field $V(t) \in T_{\gamma(t)}$, corresponding to the path γ between those two points, we can measure the covariant derivative. When $\dot\gamma$ follows affine connection, then covariant derivative is 0, and is called *geodesic*. This expressed via geodesic equation:

$$\ddot\gamma_i(t) + \sum_{j,k} \Gamma^i_{jk}(\gamma(t))\dot\gamma_j(t)\dot\gamma_k(t) = 0$$

There are a lot of different geodesics, but for Riemannian manifold with riemannian metric there is a unique one, known as Levi-Civita connection.

Definition (Geodesic)**.** *A curve* $\gamma(t) : I \to M$ *called a geodesic if it satisfies following conditions:*

(1) It is the shortest path between $\gamma(0)$ *and* $\gamma(1)$*, i.e we have* $\int_0^1 ||\dot\gamma||dt = 0$
(2) $||\frac{d}{dt}\gamma(t)||_\gamma(t)$ *is constant for* $\forall t \in [0,1]$

A curve $t \in [0,1] \to \nu_t \in \mathcal{P}_2(\mathbb{R}^d)$ is called geodesic of constant speed if $W_2(\mu_s, \mu_t) = (t-s)W_2(\mu_0, \mu_1)$ for $1 \geq t \geq s \geq 0$.

6. Markov processes and generators

Distributions of Markov processes, and in particular diffusion processes, will be the main example of curves over spaces of probability measures. We will we interested in covergence of these distributions when V features some kind of 'traps' and the noise is small.
First of all, we will give definition of Markov semigroup, because there is close relationship with Markov processes, that is many aspects of the behavior of Markov processes follow from analysis of corresponding generators and semigroups. For example, basic example of Markov process is of Brownian motion with heat semigroup and generator of Laplacian Δ, which we derived in next section.

6.1. **Markov semigroups.** Consider a family $\mathbf{P} = (P_t)_{t\geq 0}$ of operators which is defined on some measurable set of functions $(E, \mathcal{F})$, then properties of semigroup are:
(i) $P_0 = Id$
(ii) P_t sends bounded measurable functions to bounded measurable functions
(iii) If $f \geq 0$ then $P_t(f) \geq 0$
(iv) (semigroup property) $P_{t+s} = P_t P_s$
Also there is a continuity property at t=0, but it requires so called invariant measure, that is
σ-finite measure μ is said to be invariant for family of operators $(P_t)_{t\geq}$ if for every bounded positive measurable function $f : E \to R$

$$\int_E P_t f d\mu = \int_E f d\mu$$

With this in mind we have last continuity property
(v) For every $f \in L^2(\mu)$ $P_t(f)$ converges to f in $L^2(\mu)$as $t \to 0$

6.2. **Markov evolutions of probability measures on $\mathbb{R}^d$.** Let m be a probability measure on $\mathbb{R}^d$ and A a (possibly unbounded) operator on $L_2(m)$ such that A generates a Markov process with invariant m for the process. In particular $A\mathbf{1} = 0$.

We consider the evolution in $L_2(m)$

$$\partial_t \varrho = A^\dagger \varrho$$

If $X \equiv (X_t)_{t\geq 0}$ is a stochastic process, and f is a function, we write

$$(S_t f)(x) = \mathbb{E}[f(X_t)|X_0 = x], \qquad \text{for any } t \geq 0$$

For instance $S_0 f = f$. The Markov property means $S_{t+s} = S_s S_t$. This means

$$S_{t+s} f(x) = \mathbb{E}[f(X_{t+s})|X_0 = x]$$

while $S_s S_t f$

$$g(y) = \mathbb{E}[f(X_t)|X_0 = y] = (S_t f)(y)$$
$$(S_s S_t f)(x) = \mathbb{E}[g(X_s)|X_0 = x]$$

Then the generator of the process is defined as

$$Af = \lim_{t \downarrow 0} \frac{S_t f - S_0 f}{t}|_{t=0}; \qquad \text{for in } D(A)$$

Example: A Markov which is not random is the following

$$\dot{X} = b(X)$$
$$X_0 = x$$

$S_t f(x) = \mathbb{E}[f(X_t)|X_0 = x] = f(X_t^x)$. Then $S_{t+s}f(x) = f(X_{t+s}^x) = f(X_s^{X_t^x}) = S_s(S_t f)(x)$. Its generator is

$$\begin{aligned} Af(x) :=& \lim_{t\downarrow 0} \frac{f(X_t^x) - f(x)}{t} \\ =& (\dot{X}_t^x \cdot df(X_t^x))_{|t=0} = (b(X_t^x) \cdot df(X_t^x))_{t=0} = b(x) \cdot df(x) \end{aligned}$$

6.3. **Examples of generators.**

6.3.1. *Poisson processes.* Let us consider a Poisson process. We start in $x \in \mathbb{N}$, and wait for an exponential time τ_1 of parameter λ. At this time we add 1 and start again (wait τ_2, add 1 etc).

It means that we are given a sequence (τ_i) of i.i.d. expoential random variables $\tau_i \sim \exp(\lambda)$, $\mathbb{P}(\tau_i > t) = e^{-\lambda t}$. Then $N_t := \inf\{n \geq 0 : \sum_{i=1}^{n+1} \tau_i \geq t\}$. This means

$$N_t = \begin{cases} x & \text{if } t \in [0, \tau_1) \\ x+1 & \text{if } t \in [\tau_1, \tau_1 + \tau_2) \\ x+2 & \text{etc} \end{cases}$$

Then writing $\mathbb{E}_x = \mathbb{E}[\cdot|N_0 = x]$

$$\begin{aligned} \mathbb{E}_x[f(N_t)] =& \mathbb{E}_x[f(N_t)|\tau_1 > t]\mathbb{P}_x(\tau_1 > t) \\ &+ \mathbb{E}_x[f(N_t)|\tau_1 \leq t, \tau_1 + \tau_2 > t]\mathbb{P}_x(\tau_1 \leq t, \tau_1 + \tau_2 > t) \\ &+ \mathbb{E}_x[f(N_t)|\tau_1 + \tau_2 \leq t]\mathbb{P}_x(\tau_1 + \tau_2 \leq t) \\ &= f(x)e^{-\lambda t} + f(x+1)(1 - e^{-\lambda t} - o(t)) + o(t) \end{aligned}$$

To compute the generator

$$\begin{aligned} (Af)(x) &= \lim_{t\to 0} \frac{\mathbb{E}_x[f(N_t) - f(x)]}{t} = \lim_{t\to 0} \frac{f(x)e^{-\lambda t} + f(x+1)(1 - e^{-\lambda t} - o(t)) + o(t) - f(x)]}{t} \\ &= \lim_{t\to 0} \frac{(f(x+1) - f(x))(1 - e^{-\lambda t})}{t} = \lambda(f(x+1) - f(x)) \end{aligned}$$

6.3.2. *Pure jump processes.* We are on a measurable space E, and we build a process as follows. We start in a point $X_0 = x_0$. Then the process waits an exponential time τ_1 with rate $\lambda(x_0)$. Then the process jumps to a new point, sampling it with probability $p(x_0, dy)$. So we go to point x_1 and so on. So X_t will be

$$X_t = \begin{cases} x_0 & \text{if } t \in [0, \tau_1) \\ x_1 & \text{if } t \in [\tau_1, \tau_1 + \tau_2) \\ x_2 & \text{etc} \end{cases}$$

With the same reasoning, $(Af)(x) = \int \lambda(x)p(x, dy)(f(y) - f(x))$. The Poisson case corresponds to $E = \mathbb{N}$, $\lambda(x) = 1$ and $p(x, dy) = \delta_{x+1}(dy)$.

If $X_t = \sum_{i=0}^{N_t} Z_i$ with Z_i i.i.d. and N_t the simple Poisson process, then $\lambda(x) = 1$, $p(x, dy) = \mu(x + dy)$ where μ is the law of Z_i so

$$Af(x) = \lambda \int (f(x + y) - f(x))d\mu(y)$$

6.3.3. *Brownian motion on* $\mathbb{R}^d$. Not all Markov processes are pure-jump. As an example, we can take Brownian motion. This has $Af = \frac{1}{2}\Delta$.

One can make a more explicit construction. Up to a.e. equivalence, there is a unique probability measure $\mathbb{P}_x$ on $C([0, T]; \mathbb{R})$ such that

- $\mathbb{P}_x(B_0 = x) = 1$.
- For $t \geq s$, $B_t - B_s \sim \mathcal{N}(0, t - s)$ is independent of $(B_r)_{r \leq s}$.

In particular, as a consequence

- For each t, $B_t \sim \mathcal{N}(x, t)$.
- B_t is a martingale and $(dB(t))^2 = dt$.

This is equivalent to the characterization of Levy or, as we see in a moment, the characterization as a Markov process with generator $\frac{1}{2}\Delta$. It should be noted that Brownian motion has infinite variation, that is

$$\sup_P \{\sum_{k=0}^{n-1} |B_{t_{k+1}} - B_{t_k}|\} = \infty$$

However, it has finite quadratic variation:

$$\lim_{n\to\infty} \sum_{k=0}^{n-1} \left(W_{t_{k+1}} - W_{t_k}\right)^2 = T$$

Convergence should be understood in L^2 sense, i.e $\mathbb{E}[(Y_n - T)^2] \to 0$

6.4. **Geometric Brownian motion.**

$$dX_t = \theta X_t dt + aX_t dB_t,\ X_0 = x > 0$$

Solution to this SDE is $X_t = X_0 \exp[(\theta - \frac{a^2}{2})t + aB_t]$, so that

$$\mathbb{E}_x(X_t - X_0) = x\mathbb{E}[\exp(aB_t)] \exp(\theta - \frac{a^2}{2})t - x = x \exp(\theta t) - x = x(\theta t + o(t^2))$$

Now take $r = \theta - \frac{a^2}{2}$, then:

$\mathbb{E}_x(X_t - X_0)^2 = \mathbb{E}[xe^{rt+aB_t} - x]^2 = E[x^2e^{2rt}e^{2aB_t} - 2x^2e^{rt}e^{aB_t}] + x^2 = x^2[a^2t + o(t^2)]$

Thus, dividing by t and taking limit t $\to$0 we obtain $Af(x) = x\theta f'(x) + \frac{f''(x)a^2x^2}{2}$

Remark 6.1. *The Brownian motion as defined above admits the generator* $\frac{1}{2}\Delta$.

Let us prove it in $d = 1$. Then $B_t \sim \mathcal{N}(x, t)$, so it has density $\varrho_t(y) = \exp(-\frac{(y-x)^2}{2t})(2\pi t)^{-1/2}$, therefore

$$\begin{aligned}\frac{\mathbb{E}_x[f(B_t)] - f(x)}{t} &= \int (f(y) - f(x))\varrho_t(y)\, dy \\ &= \int (f(y) - f(x)) \exp(-\frac{(y - x)^2}{2t})(2\pi t)^{-1/2}/t\, dy\end{aligned}$$

Since we want to take the limit $t \to 0$, this suggests the substitution $y = x + \sqrt{t}z$

$$\frac{\mathbb{E}_x[f(B_t)] - f(x)}{t} = \int \frac{f(x+\sqrt{t}z) - f(x)}{t} e^{-\frac{z^2}{2}} (2\pi)^{-1/2}\, dz$$
$$= \int \Big(\frac{f(x+\sqrt{t}z) - f(x) - f'(x)\sqrt{t}z}{t}\Big) \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}\, dz$$

where in the last line we used that $z\exp(-z^2)$ integrates to 0. Now, let assuming for a moment that f has many bounded derivatives, $f(x+t\sqrt{z}) = f(x)+f'(x)\sqrt{t}z+\frac{1}{2}f''(x)tz^2 + O(t^{3/2})$. Using Lebesgue uniform integrability

$$\lim_{t\to 0} \frac{\mathbb{E}_x[f(B_t)] - f(x)}{t} = \int \lim_{t\to 0} \Big(\frac{f(x+\sqrt{t}z) - f(x) - f'(x)\sqrt{t}z}{t}\Big) \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}}\, dz$$
$$= \int f''(x) z^2 \frac{e^{-\frac{z^2}{2}}}{2\sqrt{2\pi}}\, dz = \tfrac{1}{2} f''(x)$$

6.4.1. *Diffusions on* $\mathbb{R}^d$. The following idea corresponds to put a Brownian motion in a potential V. This means that instead of considering $X_t = B_t$ we consider a sort of differential equations written informally as, for some $c \in \mathbb{R}$

$$\dot{X} = -\nabla V(X) + c\dot{B} \tag{6.1}$$

This can be made formal if we read it in integral form. Namely we say that X solves (6.1) if it satisfies

$$X_t - X_0 = c\, B_t - c\, B_0 - \int_0^t (\nabla V)(X_s) ds \tag{6.2}$$

If V is smooth, this problem admits a solution for each given initial data $x \in \mathbb{R}^d$. Also, with no loss of generality we can take $B_0 = 0$.

Remark 6.2. *The process satisfying* (6.2) *is Markov with generator* $Af = \frac{c^2}{2}\Delta f - \nabla V \cdot df$.

Let us compute infinitesimal generator of SDE defined above:

$$\lim_{t\to 0} \tfrac{1}{t} \left(\mathbb{E}_x[f\Big(x + c\, B_t - \int_0^t (\nabla V)(X_s) ds\Big)] - f(x) \right)$$

Take a smooth, bounded $f\colon \mathbb{R}^d \to \mathbb{R}$,

$$\frac{d}{dt}\mathbb{E}_x[f(X_t)]\Big|_{t=0} = \lim_{t\downarrow 0} \tfrac{1}{t}\mathbb{E}_x[f(X_0 + cB_t - cB_0 - \int_0^t \nabla V(X_s)\,) - f(x)] =$$
$$= \lim_{t\downarrow 0} \tfrac{1}{t}\mathbb{E}_x[f(X_0 + cB_t - cB_0) - f(x) - df(X_0 + cB_t - cB_0) \cdot \int_0^t \nabla V(X_s)\, ds + R_t(X)]$$
$$= \lim_{t\downarrow 0} \tfrac{1}{t}\mathbb{E}_x[f(x + cB_t - cB_0) - f(x)] - \lim_{t\downarrow 0} \tfrac{1}{t}\mathbb{E}_x[df(X_0 + cB_t - cB_0) \int_0^t \nabla V(X_s)\, ds]$$

because $|R_t(X)| \le |\sup_x |\nabla V|(x)\, \sup_x |f''(x)| t^2/2$ by Taylor remainder formula for the points $a = X_0 + cB_t - cB_0$ and $b = X_t$. Therefore from the previous computation for the Brownian motion

$$\frac{d}{dt}\mathbb{E}_x[f(X_t)]\Big|_{t=0} = \tfrac{c^2}{2}(\Delta f)(x) - \lim_{t\downarrow 0} \tfrac{1}{t}\mathbb{E}_x[df(X_0 + cB_t - cB_0) \cdot \int_0^t \nabla V(X_s)\, ds]$$

On the other the random variable $Y_t = df(X_0 + cB_t - cB_0) \cdot \int_0^t \nabla V(X_s)\, ds/t$ is bounded uniformly in t and converges almost surely to $\nabla V(X_0) \cdot df(X_0)$. So

$$\frac{d}{dt}\mathbb{E}_x[f(X_t)]\Big|_{t=0} = \tfrac{c^2}{2}(\Delta f)(x) - \mathbb{E}_x[\nabla V(X_0) \cdot df(X_0)] = \tfrac{c^2}{2}(\Delta f)(x) - \nabla V(x) \cdot df(x)$$

7. Quick review of Wasserstein space

7.1. **Wasserstein distance.**

Definition (Wasserstein distance)**.** *Let E be a Polish space. Then for any two probability measures ν and μ the Wasserstein distance of order $p = 2$ defined as*

$$W_2(\nu, \mu) = \inf\{\mathbb{E}[d(X, Y)^2] : law(X) = \mu, law(Y) = \nu\} \tag{7.1}$$

Proof, that (7.1) satisfies distance axioms can be found in [Villani].

Definition (Wasserstein space)**.** *The Wasserstein space of order $p = 2$ is defined as*

$$\mathcal{P}_2(E) = \{\mu \in \mathcal{P}(E) | \int_E d(x_0, x)^2 \mu(dx) < \infty\}$$

Then W_2 defines distance on $\mathcal{P}_2(E)$

Next inequalities would be useful:

Definition (Talagrand inequality)**.** *Measure ν with constant $\alpha > 0$ statsfies a Talagrand inequality if for all probability measures ρ such that $\rho << \nu$*

$$W_2(\rho, \nu)^2 \leq \frac{2}{\alpha} H_p \nu$$

Definition (Log-Sobolev inequality)**.** *Probablity measure m on $\mathbb{R}^d$ satisfies logarithmic Sobolev inequality with constant $\alpha > 0$ if*

$$\int_M g^2 \log g^2 dm - (\int_M g^2 dm) \log(\int_M g^2 dm) \leq \frac{2}{\alpha} \int_M ||\nabla g||^2 dm$$

for all smooth functions $g : M \to \mathbb{R}$ with finite second moment. The largest possible constant $\alpha > 0$ is called logarithmic Sobolev constant.

In other words, if we take $g = \sqrt{\varrho}$, where $\varrho = \frac{d\mu}{dm}$, then $H(\mu) \leq \frac{1}{\alpha} I(\mu)$. If V satisfies Log-Sobolev with α we get

$$\tfrac{d}{dt} H(\mu_t) = -I(\mu_t) \leq -\alpha H(\mu_t)$$

so that $H(\mu_t) \leq e^{-\alpha t} H(\mu_0)$.

More details can be see in 7.

7.2. **Geometry on the Wasserstein Space.**

Consider $\mathcal{P}_2(\mathbb{R}^d)$ and two probability measures μ, ν, such that there exists optimal transport plan between them. We would like to construct paths from μ to ν. A naive solution would be to use probability measures defined by $\mu_t = (1 - t)\mu + t\nu$ for $t \in [0, 1]$ but this approach completely ignores Wasserstein distance. Another approach would be to consider path from $[0, 1]$ into $L^2(\Omega, \mathcal{F}, \mathbb{P})$, which goes from random variable X to $\phi(X)$ with distribution ν given that ϕ is a transport map from μ to ν.

$$X_t = (1 - t)X + t\phi(X), \quad t \in [0, 1]$$

Then consider path $\mu = (\mu_t) \in \mathcal{P}_2(\mathbb{R}^d)$ given by distribution of X_t and thus

$$\dot{X} = \phi(X) - X$$

Now, due to Breiner's [1] theorem, if μ is a.c then we know that optimal transport plan satisfies $\phi = \nabla\psi$ for some convex function ψ, take $\psi = \frac{1}{2}|x|^2 + \epsilon\theta(x)$(which is smooth strongly convex function), where θ is smooth real valued function. Then substitute into equation above:

$$X_t = (I + t\epsilon\nabla\theta)(X) \implies \dot{X} = \epsilon\nabla\theta(X)$$

So to go from X to $\phi(X)$ we simply need to follow flow of ODE in linear vector space. In general, we look at probability measures that are transported along the flow of an ODE induced by vector field $v : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$:

$$\dot{\psi_t^x} = v(t, \psi_t^x),\ \psi_0^x = x,\ t \in [0,1],$$

That is in order to transport μ_0 we need to push it forward according to flow solving ODE:

$$\mu_t = \mu_0 \circ (x \to \psi_t^x)^{-1}$$

μ_t is same as probability distribution of the solution X_t with initial condition X_0 with distribution μ_0, so $\dot{X}_t = v(t, X_t)$. Integration gives dynamics of (μ_t) given by continuity equation:

$$\partial_t\mu_t + div(v(t,.)\mu_t) = 0$$

In particular we have that the metric derivative in Wasserstein can be represented as

$$|\partial_t\mu_t|^2 = \int |v(t,x)|^2 d\mu_t(x)$$

References

[1] Villani, Cédric. *Optimal Transport: Old and New*

[2] L.Ambrosio et al. *Gradient Flows in Metric Spaces and in the Space of Probability Measures*

[3] R.Jordan et al. *The variational formulation of the Fokker-Planck equation*

[4] F.Otto and C.Villani *Generalization of an inequality by Talagrand, and links with the Logarithmic Sobolev inequality*

[5] A.Wibisono *Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem*

[6] Roberts, G. O. and Stramer, O. *Langevin Diffusions and Metropolis-Hastings Algorithms. Methodology and Computing in Applied Probability*