

AN OPTIMIZATION PERSPECTIVE ON SAMPLING USING OPTIMAL TRANSPORT

PHILIPPE RIGOLLET[†]

ABSTRACT. These are the notes for a series of lectures given by [Philippe Rigollet](#) at Sorbonne Université between May 31 and June 9, 2022. Information on the lectures is available at <https://math.mit.edu/~rigollet/FSMP/>. These notes were taken by [Théo Dumont](#). For any remark/correction/suggestion, please send an email to theo.dumont@univ-eiffel.fr.


CONTENTS

1	Optimization	2
1.1	Strongly convex functions	2
1.2	Polyak-Lojasiewicz	3
1.3	Smoothness	3
2	Sampling	4
2.1	Markov semigroups	4
2.2	Functional inequalities and rates of convergence	6
3	Optimal Transport (OT)	10
3.1	The OT problem	10
3.2	Fundamental theorem of OT	11
3.3	Curves in the Wasserstein space	13
4	Wasserstein gradient flows	18
4.1	Wasserstein gradient	18
4.2	Langevin diffusion as a Wasserstein gradient flow	19
4.3	Rates of convergence	19
5	Applications	22
5.1	Stein Variational gradient descent (SVGD)	22
5.2	Variational Inference (VI)	24
	References	28

INTRODUCTION

💡 Sampling is a fundamental question in statistics and machine learning, most notably in Bayesian methods. Sampling and optimization present many similarities, some obvious, others more mysterious. In particular, the seminal work of Jordan, Kinderlehrer and Otto [JKO98] has unveiled a beautiful connection between the Brownian motion and the heat equation on the one hand, and optimal transport on the other. They showed that certain stochastic processes may be viewed as gradient descent over the Wasserstein space of probability distributions. This connection opens the perspective of a novel approach to sampling that leverages the rich toolbox of optimization to derive and analyze sampling algorithms. The goal of this course is to bring together the many ingredients that make this perspective possible starting from the basics and building to some of the most recent advances in sampling.

[†]Notes taken by [Théo Dumont](#). Revised July 26, 2022.

Motivation.  The course is largely motivated by a common goal: compute summary statistics from posterior distributions. We primarily use the Langevin diffusion as a viable strategy to sample from a posterior, and ultimately compute such summary statistics. Indeed, it does not require knowledge of the normalizing constant of the target distribution.

Bayesian statistics. In the field of Bayesian statistics, we have a posterior

$$p(\theta | x) = \frac{\ell(x, \theta)p(\theta)}{\int \ell(x, \theta)p(\theta) d\theta}$$

from which we want to sample; but very often, the normalizing constant $Z = \int \ell(x, \theta)p(\theta) d\theta$ is unknown. One particular case is when we want to sample from the posterior

$$\pi(x) = \frac{e^{-V(x)}}{\int e^{-V(y)} dy} \propto e^{-V(x)},$$


where V is known. In order to remove the normalizing constant, an idea is to write that $\log \pi(x) = -V(x) - \log Z$, and therefore that $\nabla_x \log \pi(x) = -\nabla_x V(x)$, which we know. Now, we want to sample from π using only its gradient/Hessian.

Langevin diffusion. The Langevin diffusion equation reads

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

and we know that $\text{Law}(X_t) \xrightarrow{t \rightarrow \infty} \pi$. This comes handy in a lot of applications, such as numerical integration, statistical physics – where our posterior is $\pi(x) \propto e^{-V(x)}$ where V is the *free energy* –, or uncertainty quantification.

1. OPTIMIZATION

 We cover a quick introduction to optimization of strongly convex functions using gradient flows. Then we relax strong convexity to a Polyak-Łojasiewicz condition. We also discuss that in continuous time, which is the focus of this course, smoothness does not appear like it does for gradient descent. Optimization focuses on problems on the form

$$\min_x f(x),$$

and while for long these problems were separated between linear and non-linear, it is now commonly acknowledged that the good classification is convex and non-convex – the latter being arbitrarily difficult. We will focus here on convex optimization problems, *i.e.* when the function f is convex. We consider first order optimization methods, *i.e.* that only involve the gradient of the function f , and in particular gradient descent and gradient flow:

$$X_{t+1} = X_t - \eta_t \nabla f(X_t) \quad (\text{gradient descent})$$

$$\partial_t X_t := \dot{X}_t = -\nabla f(X_t). \quad (\text{gradient flow})$$

1.1. Strongly convex functions.

Definition 1.1 (Strong convexity). f is said to be α -strongly convex (SC) if it satisfies the following inequality:

$$\text{for all } x \text{ and } y, \quad f(x) - f(y) \geq \langle \nabla f(y), x - y \rangle + \frac{\alpha}{2} \|y - x\|^2. \quad (\text{SC})$$

Given that f is strongly-convex, one has:

$$\begin{aligned} \frac{d}{dt} \|x_t - y_t\|^2 &= 2 \langle x_t - y_t, \dot{x}_t - \dot{y}_t \rangle \\ &= -2 \langle x_t - y_t, \nabla f(x_t) - \nabla f(y_t) \rangle \\ &\leq -2\alpha \|x_t - y_t\|^2. \end{aligned} \quad \text{by (SC)}$$

Grönwall's lemma then gives that $\|x_t - y_t\|^2 \leq e^{-2\alpha t} \|x_0 - y_0\|^2$. In particular, if $y_0 = x^* := \arg \min f(x)$, then $\dot{y}_t = -\nabla f(x^*) = 0$, hence $\|x_t - x^*\|^2 \leq e^{-2\alpha t} \|x_0 - x^*\|^2$.

Proposition 1.2 (Grönwall's lemma, elementary version). *Let $\varphi : [0, T] \rightarrow \mathbb{R}$ be a nonnegative differentiable function for which there exists a constant C such that*

$$\varphi'(t) \leq C\varphi(t) \quad \text{for all } t \in [0, T].$$

Then

$$\varphi(t) \leq e^{Ct}\varphi(0) \quad \text{for all } t \in [0, T].$$

1.2. Polyak-Łojasiewicz. One could also try to bound the gap between $f(x_t)$ and $f(x^*)$. In order to do that, we will only need f to satisfy the Polyak-Łojasiewicz inequality, and not necessarily (SC):

Definition 1.3 (Polyak-Łojasiewicz inequality). *f is said to satisfy the Polyak-Łojasiewicz (PL) inequality if there exists a constant C_{PL} such that*

$$\text{for all } x, \quad f(x) - f(x^*) \leq C_{PL}\|\nabla f(x)\|^2. \quad (\text{PL})$$

Now, given that f satisfies the (PL) inequality, we can proceed:

$$\begin{aligned} \frac{d}{dt}[f(x_t) - f(x^*)] &= \langle \nabla f(x_t), \dot{x}_t \rangle \\ &= -\|\nabla f(x_t)\|^2 \\ &\leq \frac{-1}{C_{PL}}[f(x_t) - f(x^*)]. \end{aligned} \quad \text{by (PL)}$$

Grönwall's lemma then gives that

$$f(x_t) - f(x^*) \leq e^{-t/C_{PL}}[f(x_0) - f(x^*)].$$

Proposition 1.4. (SC) implies (PL).

Proof. Let f be a α -strongly convex function. Then:

$$\begin{aligned} f(x^*) - f(x) &\geq -\langle \nabla f(x), x - x^* \rangle + \frac{\alpha}{2}\|x - x^*\|^2 \\ &\geq -\frac{1}{2} \left[\delta \|\nabla f(x)\|^2 + \frac{1}{\delta} \|x - x^*\|^2 \right] + \frac{\alpha}{2}\|x - x^*\|^2 \quad \text{for all } \delta \text{ by Young's inequality} \\ &= -\frac{1}{2\delta} \|\nabla f(x)\|^2, \quad \text{by taking } \delta = \frac{1}{\alpha} \end{aligned}$$

which shows that f satisfies (PL) with constant $C_{PL} = \frac{1}{2\alpha}$, giving the same constant in the exponential as before ($-2\alpha t$). \square

Proposition 1.5 (Young's inequality). *For $\delta > 0$, developing $\|a\sqrt{\delta} + b/\sqrt{\delta}\|^2$ gives*

$$2\langle a, b \rangle \leq \delta \|a\|^2 - \frac{1}{\delta} \|b\|^2.$$

For more information on the (PL) inequality as well as its link with strong convexity and other linear convergence rate conditions, see [KNS16].

1.3. Smoothness. The notion of smoothness is linked to inequalities of the form

$$f(x) - f(y) \leq \langle \nabla f(y), y - x \rangle + \frac{\beta}{2\|x - y\|^2}.$$

Given that f is smooth, let us now try to bound the gap between $f(x_{t+1})$ and $f(x^*)$ for the gradient descent $x_{t+1} = x_t - \eta \nabla f(x_t)$:

$$\begin{aligned} f(x_{t+1}) - f(x^*) &= f(x_t - \eta \nabla f(x_t)) - f(x^*) \\ &\leq f(x_t) - \langle \nabla f(x_t), \eta \nabla f(x_t) \rangle + \frac{\beta}{2} \|\eta \nabla f(x_t)\|^2 - f(x^*) \quad (\text{smoothness}) \\ &= f(x_t) - f(x^*) + \underbrace{\left(\frac{\beta\eta^2}{2} - \eta \right)}_{\text{if } \leq 0 \text{ then (PL)}} \|\nabla f(x_t)\|^2 \end{aligned}$$

$$\leq \left(1 - \frac{1}{2C_{\text{PL}}\beta}\right) [f(x_t) - f(x^*)] \quad \text{with the optimal } \eta = 1/\beta.$$

Hence

$$\begin{aligned} f(x_t) - f(x^*) &\leq \left(1 - \frac{1}{2C_{\text{PL}}\beta}\right)^t [f(x_0) - f(x^*)] \\ &\leq e^{-\frac{\alpha}{\beta}t} [f(x_0) - f(x^*)] \quad \text{as } \left(1 - \frac{\alpha}{\beta}\right)^t \leq e^{-\frac{\alpha}{\beta}t}. \end{aligned}$$


2. SAMPLING

We recall the Langevin diffusion equation:

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

The questions we are interested in:

- do we have $\text{Law}(X_t) \xrightarrow[t \rightarrow \infty]{} \pi \propto e^{-V}$?
- what is the rate of convergence?

2.1. Markov semigroups.  Markov semigroups are a canonical tool to study the convergence of Markov processes. We introduce the main notions: semigroup, infinitesimal generator, Kolmogorov's backward equation and Kolmogorov's forward equation (called Fokker–Planck in this course), and studied the stationary distribution from Fokker–Planck. All notions are computed for Langevin as a concrete example.

The idea: to a Markov process $(X_t)_{t \geq 0}$, associate a family $(P_t)_{t \geq 0}$ of operators acting on functions. We refer to [BGL+14; Van16] for details.

2.1.1. Markov semigroups and generators.

Definition 2.1 (Markov semigroup). *The Markov semigroup $(P_t)_{t \geq 0}$ associated to a Markov process $(X_t)_{t \geq 0}$ is defined by*

$$P_t f(x) = \mathbb{E}[f(X_t) \mid X_0 = x].$$

Lemma 2.2. *We have the following properties:*

- $P_0 = \text{id}$
- $P_{s+t} = P_s P_t = P_t P_s$

Proof.

$$\begin{aligned} P_{s+t} f(x) &= \mathbb{E}[f(X_{t+s}) \mid X_0 = x] \\ &= \mathbb{E}\left[\underbrace{\mathbb{E}[f(X_{s+t}) \mid \{X_r\}_{r \leq t}]}_{P_s f(X_t)} \mid X_0 = x\right] \\ &= P_t P_s f(x). \quad \square \end{aligned}$$

Definition 2.3 (Infinitesimal generator). *The infinitesimal generator \mathcal{L} associated to a Markov semigroup $(P_t)_{t \geq 0}$ is the operator defined by*

$$\mathcal{L}f(x) = \lim_{t \searrow 0} \frac{P_t f - f}{t}.$$

Example 2.4 (Langevin). $B_t \sim \mathcal{N}(0_d, tI_d)$

$$\begin{aligned} X_t &= X_0 + \int_0^t -\nabla V(x_s) ds + \sqrt{2} B_t \\ &= X_0 - t \nabla V(X_0) + \sqrt{2} \underbrace{B_t}_{\text{order } \sqrt{t}} + o(t). \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{E}[f(X_t) \mid X_0 = x] &= \mathbb{E}\left[f(X_0) - \langle \nabla V(X_0), t\nabla f(X_0) + \sqrt{2}B_t \rangle \right. \\ &\quad \left. + \frac{1}{2} \langle -t\nabla V(X_0) + \sqrt{2}B_t, \nabla^2 f(X_0)(-t\nabla V(X_0) + \sqrt{2}B_t) \rangle + o(1) \mid X_0 = x\right] \\ &= f(x) - t\langle \nabla V(x), \nabla f(x) \rangle + \mathbb{E}[\langle B_t, \nabla^2 f(x)B_t \rangle] + o(t), \end{aligned}$$

and as $\langle B_t, \nabla^2 f(x)B_t \rangle = \text{Tr}(\nabla^2 f(x)B_t B_t^\top) = t \text{Tr}(\nabla^2 f(x)) = t\Delta f(x)$, we get that

$$P_t f(x) = f(x) - t\langle \nabla V(x), \nabla f(x) \rangle + t\Delta f(x) + o(t),$$

and finally

$$\mathcal{L}f(x) = -\langle \nabla V(x), \nabla f(x) \rangle + \Delta f(x).$$

2.1.2. *Dynamics.*  What about $\partial_t P_t f$?

Proposition 2.5 (Kolmogorov Backward Equation). *The Kolmogorov Backward Equation (KBE) is $\partial_t P_t f = \mathcal{L}P_t f = P_t \mathcal{L}f$.*

Proof.

$$\frac{P_{t+h}f - P_t f}{h} = \underbrace{\frac{P_h - \text{id}}{h}}_{\xrightarrow{t \rightarrow \infty} \mathcal{L}} P_t f = P_t \underbrace{\frac{P_h - \text{id}}{h}}_{\xrightarrow{t \rightarrow \infty} \mathcal{L}} f. \quad \square$$

And its dual version:

Proposition 2.6 (Fokker–Planck). *The Kolmogorov Forward Equation, or Fokker–Planck equation, is for π_0 the density of X_0 , $\partial_t P_t^* \pi_0 = \mathcal{L}^* P_t^* \pi_0 = P_t^* \mathcal{L}^* \pi_0$.*

Proof. One has

$$\begin{aligned} \mathbb{E}f(X_t) &= \mathbb{E}\left[\underbrace{\mathbb{E}[f(X_t) \mid X_0]}_{P_t f(X_0)}\right] \\ &= \int P_t f(x) \pi_0(x) \, dx \\ &= \langle P_t f, \pi_0 \rangle_{L^2(dx)} \\ &= \langle f, P_t^* \pi_0 \rangle \\ &= \int f(x) P_t^* \pi_0(x) \, dx. \end{aligned}$$

Hence the density of X_t is $P_t^* \pi_0$. Then, one has

$$\begin{aligned} \partial_t \int f P_t^* \pi_0 &= \partial_t \int P_t f \pi_0 && \text{(duality)} \\ &= \int P_t \mathcal{L}f \pi_0 && \text{(KBE)} \\ &= \int \mathcal{L}f P_t^* \pi_0 && \text{(duality)} \\ &= \int f \mathcal{L}^* P_t^* \pi_0. && \text{(duality)} \quad \square \end{aligned}$$

To sum up:

$$\begin{cases} u_t = P_t f & ; & \partial u_t = \mathcal{L}u_t & \text{(backward)} \\ \pi_t = P_t^* \pi_0 & ; & \partial \pi_t = \mathcal{L}^* \pi_t & \text{(Fokker–Planck)} \end{cases}$$

Proposition 2.7 (Stationarity). *The following statements are equivalent:*

(i) π is a stationary distribution for $(X_t)_{t \geq 0}$;

- (ii) $\mathcal{L}^*\pi = 0$;
 (iii) $\mathbb{E}_\pi[\mathcal{L}f] = 0$ for all f .

Example 2.8 (Langevin).

$$\begin{aligned}
 \langle f, \mathcal{L}g \rangle &= \int f [-\langle \nabla V, \nabla g \rangle + \Delta g] && \text{(infinitesimal generator for Langevin)} \\
 &= - \int \langle f \nabla V, \nabla g \rangle + \int f \operatorname{div}(\nabla g) && \text{since } \Delta g = \operatorname{div}(\nabla g) \\
 &= \int \operatorname{div}(f \nabla V) g - \int \langle \nabla f, \nabla g \rangle && \text{(integration by parts)} \\
 &= \int \operatorname{div}(f \nabla V) g + \int (\Delta f) g && \text{(integration by parts)} \\
 &:= \langle \mathcal{L}^* f, g \rangle.
 \end{aligned}$$

Hence $\mathcal{L}^* f = \operatorname{div}(f \nabla V) + \Delta f$ for Langevin, and the Fokker–Planck equation for Langevin is

$$\partial_t \pi_t = \operatorname{div}(\pi_t \nabla V) + \Delta \pi_t. \quad (2.1)$$


(Remark that when $\nabla V = 0$, we recover the heat equation / a Brownian motion.) With this, we can determine the stationary distributions for $(X_t)_{t \geq 0}$ in the Langevin case:

$$\begin{aligned}
 0 &= \mathcal{L}^* \pi \\
 &= \operatorname{div}(\pi \nabla V) + \Delta \pi \\
 &= \operatorname{div}(\pi \nabla V + \nabla \pi) \\
 &= \operatorname{div}(\underbrace{\pi(\nabla V + \nabla \log \pi)}_{\text{hence } =0}),
 \end{aligned}$$

which means that

$$\begin{aligned}
 -\nabla V &= \nabla \log \pi \\
 \log \pi &= -V + \text{cst} \\
 \pi &\propto e^{-V}.
 \end{aligned}$$

2.2. Functional inequalities and rates of convergence.

2.2.1. Poincaré inequalities.  Focusing on reversible Markov processes, we define the Dirichlet form and the Dirichlet energy and show that reversible Markov processes can be seen as a gradient flows of the Dirichlet energy. Then we discuss Poincaré and log-Sobolev inequalities and how they lead to exponential convergence to stationarity in chi-squared and KL respectively. We show that log-Sobolev implies Poincaré and briefly discussed connections to concentration and pointing to van Handel’s notes for more details.

Definition 2.9 (Reversible Markov process). *The Markov semigroup $(P_t)_{t \geq 0}$ is reversible w.r.t. π (a stationary distribution) if*

$$\forall f, g \in L^2(\pi), \quad \int P_t f g \, d\pi = \int f P_t g \, d\pi,$$

or equivalently

$$\forall f, g \in L^2(\pi), \quad \int \mathcal{L} f g \, d\pi = \int f \mathcal{L} g \, d\pi,$$

which means that \mathcal{L} is self-adjoint in the sense of $L^2(\pi)$.

Example 2.10. $X_0 \sim \pi$, $f = \mathbb{1}_A$, $g = \mathbb{1}_B$. Then $\mathbb{P}(X_0 \in A, X_t \in B) = \mathbb{P}(X_0 \in B, X_t \in A)$.

Remark 2.11. P_t and \mathcal{L} are symmetric operators and they therefore have a real spectrum. Moreover, P_t is a positive operator as

$$\begin{aligned} \int f P_t f \, d\pi &= \int f P_{t/2} P_{t/2} f \, d\pi \\ &= \int (f P_{t/2})^2 \, d\pi \geq 0. \end{aligned}$$

Since

$$\begin{aligned} (P_t f(x))^2 &= \mathbb{E}[f(X_t) \mid X_0 = x]^2 \\ &\leq \mathbb{E}[f^2(X_t) \mid X_0 = x] \quad \text{by Jensen's inequality,} \end{aligned}$$

one has that

$$\int P_t f(x)^2 \, d\pi \leq \int P_t f^2(x) \, d\pi = \int f^2(x) \, d\pi.$$

As $P_t = e^{t\mathcal{L}}$ and P_t contracts, one has that $\mathcal{L} \leq 0$.

Definition 2.12 (Dirichlet form). *Given a reversible Markov process / Markov semigroup with generator \mathcal{L} and stationary measure π , the corresponding Dirichlet form is defined as*

$$\mathcal{E}(f, g) = -\langle f, \mathcal{L}g \rangle_{L^2(\pi)} = -\langle \mathcal{L}f, g \rangle_{L^2(\pi)}.$$

Example 2.13 (Langevin). Since $\mathcal{L}f = -\langle \nabla V, \nabla f \rangle + \Delta f$, we have that

$$\begin{aligned} \mathcal{E}(f, g) &= - \int f \mathcal{L}g \, d\pi \\ &= \int f \langle \nabla V, \nabla g \rangle \, d\pi - \int f \Delta g \, d\pi \quad (\text{not allowed to do an IBP in } L^2(\pi)) \\ &= \int \langle f \nabla V \cdot \pi, \nabla g \rangle - \int (f \pi) \Delta g \\ &= \int \langle f \nabla V \cdot \pi, \nabla g \rangle + \int \langle \nabla(f \pi), \nabla g \rangle \quad (\text{integration by parts}) \\ &= \int \langle f \pi \nabla V + f \underbrace{\nabla \pi}_{=-\pi \nabla V} + \nabla f \pi, \nabla g \rangle \\ &= \int \langle \nabla f, \nabla g \rangle \, d\pi. \end{aligned}$$

Markov process as a gradient flow of the Dirichlet energy. $\nabla_{L^2(\pi)} \mathcal{E}(f)$ is the element of $L^2(\pi)$ such that for all curve $t \mapsto u_t \in L^2(\pi)$ starting at $u_0 = f$,

$$\partial_t \Big|_{t=0} \mathcal{E}(u_t) = \int u_0 \nabla \mathcal{E}(f) \, d\pi.$$

We then have that

$$\partial_t \Big|_{t=0} \mathcal{E}(u_t) = \partial_t \Big|_{t=0} - \int u_t \mathcal{L}u_t \, d\pi = -2 \int u_0 \mathcal{L}u_0 \, d\pi,$$

hence $\nabla \mathcal{E}(f) = -2\mathcal{L}f$, and the gradient flow of the Dirichlet energy becomes $\dot{u}_t = -\nabla \mathcal{E}(u_t) = 2\mathcal{L}u_t$. We recover the KBE (with a factor 2, which still draws the same curve but with twice the speed).

It appears that $\mathcal{E}(u_t) - \mathcal{E}(u^*)$ is not a relevant quantity to examine. We are looking for a d such that it is relevant to study $d(\pi_t, \pi) \leq e^{-\alpha t}$. We therefore define the χ^2 -divergence:

Definition 2.14 (f -divergence). *f -divergences are functions of the form*

$$(\mu, \nu) \mapsto \int \varphi \left(\frac{d\mu}{d\nu} \right) \, d\nu,$$

where φ is convex such that $\varphi(1) = 0$.

Definition 2.15 (χ^2 -divergence). *The χ^2 -divergence between μ and ν is defined as*

$$\chi^2(\mu \parallel \nu) = \int \left(\frac{d\mu}{d\nu} \right)^2 d\nu - 1.$$

It is a f -divergence, obtained for $\varphi(x) = x^2 - 1$.

Remark 2.16 (Link between χ^2 and KL). The KL divergence is also a f -divergence, obtained for $\varphi(x) = x \log x$. One has that

$$\text{KL}(\mu \parallel \nu) = \int \log \left(\frac{d\mu}{d\nu} \right) d\mu \stackrel{(\text{Jensen})}{\leq} \log \int \frac{\mu^2}{\nu^2} d\nu = \log(\chi^2(\mu \parallel \nu) + 1) \leq \chi^2(\mu \parallel \nu). \quad (2.2)$$

Actually, one has that

$$\begin{cases} \text{KL} \approx \log \chi^2 & \text{when the distributions are far;} \\ \text{KL} \approx \chi^2 & \text{when the distributions are close.} \end{cases}$$

Now, let us upper bound $\partial_t \chi^2(\pi_t \parallel \pi)$. In order to do so, we introduce $\rho_t := \pi_t / \pi$, for which we derive the new Fokker–Planck equation:

$$\begin{aligned} \partial_t \pi_t &= \mathcal{L}^* \pi_t \\ \int f \partial_t \pi_t &= \int f \mathcal{L}^* \pi_t = \int \mathcal{L} f \pi_t. \end{aligned}$$

Dividing and multiplying by π to make ρ_t appear gives

$$\int f \dot{\rho}_t \pi = \int \mathcal{L} f \rho_t \pi \stackrel{(*)}{=} \int f \mathcal{L} \rho_t \pi,$$

where $(*)$ comes from the fact that \mathcal{L} is self-adjoint w.r.t π . The Fokker–Planck equation for ρ_t is therefore $\dot{\rho}_t = \mathcal{L} \rho_t$, and we have:

$$\partial_t \chi^2(\pi_t \parallel \pi) = \partial_t \int (\rho_t)^2 d\pi = 2 \int \rho_t \dot{\rho}_t d\pi = 2 \int \rho_t \mathcal{L} \rho_t d\pi = -2\mathcal{E}(\rho_t).$$

💡 If we could get something like $\mathcal{E}(\rho_t) \geq c\chi^2(\pi_t \parallel \pi)$, then we would be able to apply Grönwall’s lemma and we would be good.

Definition 2.17 (Poincaré inequality). *A Markov process $(X_t)_{t \geq 0}$ is said to satisfy the Poincaré inequality if*

$$\text{for all } f, \quad \text{Var}_\pi(f) \leq C_P \mathcal{E}(f). \quad (\text{Poincaré})$$

Example 2.18. For ρ_t ,

$$\text{Var}_\pi(\rho_t) = \int \rho_t^2 d\pi - \left(\int \rho_t d\pi \right)^2 = \int \left(\frac{d\pi_t}{d\pi} \right)^2 d\pi - 1 = \chi^2(\pi_t \parallel \pi).$$

Therefore,

$$\partial_t \chi^2(\pi_t \parallel \pi) \leq -\frac{2}{C_P} \chi^2(\pi_t \parallel \pi),$$

and then

$$\chi^2(\pi_t \parallel \pi) \leq \chi^2(\pi_0 \parallel \pi) e^{-\frac{2}{C_P} t}.$$

Remark 2.19. For Langevin, the Poincaré inequality is

$$\text{for all } f, \quad \text{Var}_\pi(f) \leq C_P \int \|\nabla f\|^2 d\pi,$$

and this property will often be simply denoted as “ π satisfies the Poincaré inequality”, omitting that it is the Poincaré inequality *for the Langevin process*. If $\pi \propto e^{-V}$ and is strongly log-concave (V is α -strongly convex), then π satisfies the Poincaré inequality with parameter $C_P = \frac{1}{\alpha}$.

Definition 2.20 (Brascamp-Lieb inequality). For $\pi \propto e^{-V}$ where V strictly convex, i.e. $(\nabla^2 V)^{-1}$ exists, we say that f satisfies a Brascamp-Lieb inequality if

$$\text{Var}_\pi(f) \leq C \int \nabla f^\top (\nabla^2 V)^{-1} \nabla f \, d\pi.$$

More generally, the Mirror Poincaré inequality is

$$\text{Var}_\pi(f) \leq C \int \nabla f^\top (\nabla^2 \varphi)^{-1} \nabla f \, d\pi.$$

A natural question that arises is to ask for which Markov process do we have $\mathcal{E}(f) = \int \nabla f^\top (\nabla^2 \varphi)^{-1} \nabla f \, d\pi$, with φ arbitrary?

Definition 2.21 (Mirror Langevin). The Mirror Langevin process is defined by

$$\begin{cases} X_t &= \nabla \varphi^*(Y_t) \\ dY_t &= -\nabla \varphi(X_t) \, dt + \sqrt{2}[\nabla^2 \varphi(X_t)]^{1/2} \, dB_t. \end{cases}$$

When $\varphi = V$, we get the Newton-Langevin process [Che+20].

2.2.2. Log-Sobolev inequalities.  The idea: getting the same convergence properties but with KL instead of χ^2 .

$$\begin{aligned} \partial_t \text{KL}(\pi_t \parallel \pi) &= \partial_t \int \log(\rho_t) \rho_t \pi \\ &= \int \frac{\dot{\rho}_t}{\rho_t} \rho_t \pi + \int \log(\rho_t) \dot{\rho}_t \pi \\ &= \underbrace{\int \mathcal{L} \rho_t \pi}_{=0} + \int \log(\rho_t) \mathcal{L} \rho_t \pi \\ &= -\mathcal{E}(\rho_t, \log \rho_t), \end{aligned}$$

so we are looking for an inequality that looks like $\text{KL}(\pi_t \parallel \pi) \leq C \mathcal{E}(\rho_t, \log \rho_t)$.

Definition 2.22 (Log-Sobolev inequality). A Markov process is said to satisfy a logarithmic Sobolev inequality (LSI) if for all ρ densities w.r.t. π (and not only ρ_t),

$$\text{KL}(\rho \pi \parallel \pi) \leq \frac{C_{\text{LS}}}{2} \mathcal{E}(\rho, \log \rho). \quad (\text{LSI})$$

With this, using Grönwall's lemma we obtain $\text{KL}(\pi_t \parallel \pi) \leq \frac{C_{\text{LS}}}{2} \mathcal{E}(\rho, \log \rho)$, which is better than the same inequality with χ^2 because of the cold start (π_0 far from π).

Let us recap. We have shown so far:

$$\begin{cases} (\text{Poincaré}) & \iff \chi^2(\pi_t \parallel \pi) \leq \chi^2(\pi_0 \parallel \pi) e^{-\frac{2t}{C_{\text{P}}}} \\ (\text{LSI}) & \iff \text{KL}(\pi_t \parallel \pi) \leq \text{KL}(\pi_0 \parallel \pi) e^{-\frac{2t}{C_{\text{LS}}}}, \end{cases}$$

and the \iff is the important thing [Van16].

Example 2.23 (Langevin).

$$\begin{aligned} \mathcal{E}(\rho, \log \rho) &= \int \langle \nabla \rho, \nabla \log \rho \rangle \, d\pi \\ &= \int \langle \nabla \log \rho, \nabla \log \rho \rangle \underbrace{\rho \pi}_{:=\mu} \\ &= \int \left\| \nabla \log \frac{\mu}{\pi} \right\|^2 \, d\mu, \end{aligned} \quad (\text{FI})$$

which is the Fisher information of μ w.r.t. π (but not the same Fisher information that the one in statistics!).

The logarithmic-Sobolev inequality can actually be re-written as

$$\text{KL}(\mu \parallel \nu) \leq \frac{1}{2\rho} \text{FI}(\mu \mid \nu).$$

Other relevant inequalities are *Talagrand's inequality*

$$W_2(\mu, \nu) \leq \sqrt{\frac{2\text{KL}(\mu \parallel \nu)}{\rho}}$$

and the HWI inequality

$$\text{KL}(\mu \parallel \nu) \leq W_2(\mu, \nu) \sqrt{\text{FI}(\mu \mid \nu)} - \frac{K}{2} W_2(\mu, \nu)^2.$$

See [OV00] for more information.

Proposition 2.24. (LSI) \implies (Poincaré).

Proof. $\rho = 1 + \varepsilon f$ with $\int f \, d\pi = 0$.

$$\begin{aligned} \text{KL}(\rho\pi \parallel \pi) &= \int \log(1 + \varepsilon\rho) \rho\pi = \frac{\varepsilon^2}{2} \int f^2 \, d\pi + o(\varepsilon^2) \\ \mathcal{E}(\rho, \log \rho) &= - \int (\mathcal{I} + \varepsilon f) \mathcal{L} \log(1 + \varepsilon f) \, d\pi \\ &= \varepsilon^2 \underbrace{\int f \mathcal{L} f \, d\pi}_{=\mathcal{E}(f)} + o(\varepsilon^2). \end{aligned}$$

Then, LSI implies that $\frac{1}{2} \int f^2 \, d\pi \leq \frac{C_{\text{LS}}}{2} \mathcal{E}(f) + o(1)$, i.e. $\int f^2 \, d\pi \leq C_{\text{LS}} \mathcal{E}(f)$, which is the Poincaré inequality with constant $C_{\text{P}} = C_{\text{LS}}$. \square

This means we can have concentration inequalities for the stationary process:

Theorem 2.25 (Concentration inequalities, [BGL+14]). *Let π be the stationary distribution of the Langevin distribution and let $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be 1-Lipschitz. Then*

(i) *if π satisfies a (Poincaré) inequality with constant C_{P} , then*

$$\pi(f - \mathbb{E}_{\pi} f > t) \leq 3e^{-t/\sqrt{C_{\text{P}}}}.$$

(ii) *if π satisfies a (LSI) with constant C_{LS} , then*

$$\pi(f - \mathbb{E}_{\pi} f > t) \leq 3e^{-t^2/2C_{\text{LS}}},$$

which is great for tensorization.

3. OPTIMAL TRANSPORT (OT)

For references, see [Vil03; Vil09; San15; AG13].

3.1. The OT problem. ¹ The Monge problem is

$$\min_{T_{\#}\mu=\nu} \int c(x, T(x)) \, d\mu(x), \tag{Monge}$$

and often we will be using $c(x, T(x)) := d^2(x, T(x)) = \|x - T(x)\|^2$. But in the general case, the optimal T will not be a map, but more generally a coupling. For instance, in the case where μ is one dirac and ν two diracs (Figure 1), we would like the optimal T to be

$$T(x) = \begin{cases} y_1 & \text{with probability } 1/2, \\ y_2 & \text{with probability } 1/2, \end{cases}$$

¹  I did not take notes for this part of the course. If you did, it would be great if you could send them to me by email at theo.dumont@univ-eiffel.fr!

which is not a function but rather a Markov kernel.

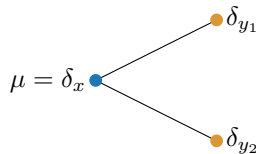


FIGURE 1. Optimal transport between one dirac and two diracs.

Example 3.1 (Couplings).

- $\gamma(A \times B) = \gamma(A)\gamma(B)$, the independent coupling
- for $X \sim \mathcal{N}(0, 1)$ and $Y \sim \mathcal{N}(0, 1)$, define $X, Y \sim \gamma \iff (X, Y) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$, the Gaussian coupling
- for $X \sim \mathcal{N}(0, 1)$ and $Y \sim \text{Law}(X^2)$, define $X, Y \sim \gamma \iff Y = X^2$ a.s.. This can be written $\gamma(dx, dy) = \mu(dx)\delta_{y=x^2}$, or $(X, Y) \sim (\text{id}, x \mapsto x^2)_{\#}\mu$.

Theorem 3.2 (Existence of a minimizer). *The minimum is achieved (by a $\bar{\gamma}$).*

Proof. Use the l.s.c. and the fact that $\Gamma_{\mu, \nu}$ is compact (Prokhorov's theorem). \square

Distances between probability measures. Quantifying the distance between two probability measures can be done using e.g. the Total Variation (TV) distance $\text{TV} = \frac{1}{2} \int |p - q|$, the L^p distance $L^p = \|p - q\|_{L^p}$, the χ^2 , KL, the Hellinger distance... But all do not behave the same way: for instance, $\text{TV}(\delta_x, \delta_y) = \mathbb{1}_{x \neq y}$, which is not nice in our context. We would rather prefer something like the Wasserstein distance, namely $W_2(\delta_x, \delta_y) = \|x - y\|$.

3.2. Fundamental theorem of OT. We focus in the rest of the lecture on the case where $c(x, y) = \frac{1}{2}\|x - y\|^2$. *Duality.*

$$\begin{aligned} \frac{1}{2} W_2^2(\mu, \nu) &= \inf_{\gamma \in \mathcal{M}_+} \sup_{f, g \in L^1} \left[\int \frac{\|x - y\|^2}{2} d\gamma(x, y) + \int f d\mu - \int f(x) d\gamma(x, y) + \int g d\nu - \int g(y) d\gamma(x, y) \right] \\ &\geq \sup_{f, g \in L^1} \inf_{\gamma \in \mathcal{M}_+} \left[\int f d\mu + \int g d\nu + \underbrace{\int \left(\frac{\|x - y\|^2}{2} - f(x) - g(y) \right) d\gamma(x, y)}_{\text{have to be } \geq 0} \right] \\ &= \sup_{f, g \in D(\mu, \nu)} \int f d\mu + \int g d\nu, \end{aligned} \tag{DP}$$

where we denoted

$$D(\mu, \nu) := \left\{ (f, g) \in L^1(\mu) \times L^1(\nu) : f(x) + g(y) \leq \frac{\|x - y\|^2}{2} \text{ for all } (x, y) \right\}.$$

This inequality is already quite nice, since it allows one to give a bound on $\int f + \int g$ by upper bounding W_2^2 , which is easy to do (one just has to exhibit a non-optimal γ). But one can actually do much better:

Theorem 3.3 (Fundamental theorem of OT). *For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\int \|x\|^2 d\mu(x) < +\infty$ (same for ν), one has*

- (strong duality) $\frac{1}{2} W_2^2(\mu, \nu) = \sup_{f, g} \int f d\mu + \int g d\nu$.
- (existence of dual potentials) There exists \bar{f}, \bar{g} maximizing (DP). Moreover, we can write $\bar{f} = \frac{\|\cdot\|^2}{2} - \varphi$ and $\bar{g} = \frac{\|\cdot\|^2}{2} - \varphi^*$ with φ proper l.s.c. and $\varphi(x) + \varphi^*(y) = \langle x, y \rangle$ $\bar{\gamma}$ -a.s.

(iii) (Brenier's theorem) If, in addition, μ has a density w.r.t. the Lebesgue measure, then $\bar{\gamma}$ is unique and induced by a deterministic transport map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$, characterized by the (μ -a.s.) unique gradient of a proper l.s.c. convex function: $(\nabla\varphi)_\# \mu = \nu$.

Remark 3.4. $\varphi^*(y) = \sup_x \{\langle x, y \rangle - \varphi(x)\}$, therefore $\varphi(x) + \varphi^*(y) \geq \langle x, y \rangle$, and the previous theorem is actually the equality case, given by the first order condition $y = \nabla\varphi(x)$.

Proof of Theorem 3.3. The proof will require the introduction of the notion of cyclical monotonicity.

(1) **Cyclical monotonicity.** For $g : \mathbb{R} \rightarrow \mathbb{R}$ convex, one has

$$\begin{aligned} g(x) - g(y) &\leq (x - y)g'(x) \\ + g(y) - g(x) &\leq (y - x)g'(y) \\ \hline (g'(x) - g'(y))(x - y) &\geq 0, \end{aligned}$$

which is just a writing of the fact that g' is increasing. Similarly, in \mathbb{R}^d , a monotone g will satisfy $\langle \nabla g(x) - \nabla g(y), x - y \rangle \geq 0$, but this will not be enough in our case. For a family of points $(x_i)_{1 \leq i \leq n}$, summing the inequalities

$$\frac{+i \quad g(x_i) - g(x_{i+1}) \leq \langle x_i - x_{i+1}, \nabla g(x_i) \rangle}{\sum_{i=1}^n \langle \nabla g(x_i), x_i - x_{i+1} \rangle \geq 0 \text{ with } x_{n+1} := x_1.}$$

Definition 3.5 (Cyclical monotonicity). A set $A \in \mathbb{R}^d \times \mathbb{R}^d$ is cyclically monotone (CM) if $\forall k \geq 2$ and any $(x_1, y_1), \dots, (x_k, y_k)$,

$$\sum_{i=1}^k \langle x_i - x_{i+1}, y_i \rangle \geq 0, \quad \text{with } x_{k+1} := x_1.$$

We have shown that the set $\{(x, \nabla g(x))\}$ is CM for g convex. The reverse implication will be of particular use for us:

Theorem 3.6 ([Roc70]). *A is cyclically monotone if and only if there exists a closed (proper l.s.c.) convex φ such that $A \subset \{(x, \partial\varphi(x))\}$.*

(2) **OT plans have CM support.** We have the following equivalences:

$$\begin{aligned} (x_1, y_1), \dots, (x_n, y_n) \in \text{supp } \bar{\gamma} &\implies \sum_{i=1}^n \|x_i - y_i\|^2 \leq \sum_{i=1}^n \|x_{\sigma(i)} - y_i\|^2 && \text{for all } \sigma \\ &\iff \sum_{i=1}^n \langle x_{\sigma(i)}, y_i \rangle \leq \sum_{i=1}^n \langle x_i, y_i \rangle && \text{for all } \sigma \\ &\iff \sum_{i=1}^n \langle x_i - x_{\sigma(i)}, y_i \rangle \geq 0 && \text{for all } \sigma \\ &\implies \text{cyclical monotonicity.} \end{aligned}$$

Hence $\text{supp } \bar{\gamma} \subset \partial\varphi = \{\nabla\varphi\}$ by Theorem 3.6, this last equality being Lebesgue a.e., since μ has a density.

(3) **Dual optimality.** We say that $(f, g) \in D(\mu, \nu)$ are dual feasible. Let us now fix f . For all g s.t. (f, g) dual feasible,

$$\begin{aligned} f(x) + g(y) &\leq \frac{\|x - y\|^2}{2} && \text{for all } (x, y) \\ g(y) &\leq \frac{\|x - y\|^2}{2} - f(x) && \text{for all } (x, y) \end{aligned}$$

Let us then take

$$\begin{aligned} g(y) &:= \inf_x \frac{\|x - y\|^2}{2} - f(x) \\ &= \frac{\|y\|^2}{2} - \sup_x \left(\langle x, y \rangle - \left(\frac{\|x\|^2}{2} - f(x) \right) \right) \end{aligned}$$

$$= \frac{\|y\|^2}{2} - \sup_x (\langle x, y \rangle - \varphi(x))$$

we then have $\varphi^*(y) = \frac{\|y\|^2}{2} - g(y)$. So, for fixed $f = \frac{\|\cdot\|^2}{2} - \varphi$, $g = \frac{\|\cdot\|^2}{2} - \varphi^*$. Similarly, for fixed $g = \frac{\|\cdot\|^2}{2} - \varphi^*$, $f = \frac{\|\cdot\|^2}{2} - \varphi^{**}$. If (\bar{f}, \bar{g}) do exist, then they must satisfy

$$\begin{cases} \bar{f} = \frac{\|\cdot\|^2}{2} - \varphi \\ \bar{g} = \frac{\|\cdot\|^2}{2} - \varphi^*, \end{cases}$$

since $\varphi = \varphi^{**}$.

- (4) **Poof of strong duality.** Let $\bar{\gamma}$ be an optimal transport plan. We have $\text{supp } \bar{\gamma} \subset \{(x, \partial\varphi(x))\}$. Let's define $\bar{f} = \|\cdot\|^2/2 - \varphi$ and $\bar{g} = \|\cdot\|^2/2 - \varphi^*$. Are they dual feasible?

$$\begin{aligned} \bar{f}(x) + \bar{g}(y) &= \frac{\|x\|^2}{2} + \frac{\|y\|^2}{2} - (\varphi(x) + \varphi^*(y)) \\ &= \frac{\|x\|^2}{2} + \frac{\|y\|^2}{2} - \langle x, y \rangle \quad \text{on } \text{supp } \bar{\gamma} \quad \text{as } y \in \partial\varphi(x) \\ &= \frac{\|x - y\|^2}{2} \quad \text{on } \text{supp } \bar{\gamma}. \end{aligned}$$

Then $\frac{1}{2} \int \|x - y\|^2 d\bar{\gamma}(x, y) = \int \bar{f} d\mu + \int \bar{g} d\nu$, hence the strong duality.

- (5) **Uniqueness for Brenier.** Let us assume another OT plan $\bar{\pi} = (\text{id}, \nabla\varphi_\pi)_\# \mu$, associated with the dual potentials $\bar{f}_\pi = \|\cdot\|^2/2 - \varphi_\pi$ and $\bar{g}_\pi = \|\cdot\|^2/2 - \varphi_\pi^*$. Then:

$$\begin{aligned} \int (\varphi_\pi(x) + \varphi_\pi^*(y)) d\bar{\gamma}(x, y) &= \int \varphi_\pi d\mu + \int \varphi_\pi^* d\nu \\ &= \int \varphi d\mu + \int \varphi^* d\nu \quad \text{as both couples are dual optimal} \\ &= \int \langle x, y \rangle d\bar{\gamma}(x, y), \end{aligned}$$

hence $\int (\varphi_\pi(x) + \varphi_\pi^*(y) - \langle x, y \rangle) d\bar{\gamma}(x, y) = 0$, then $\varphi_\pi(x) + \varphi_\pi^*(y) = \langle x, y \rangle$ $\bar{\gamma}$ -a.s., which means that $y = \nabla\varphi_\pi(x)$ $\bar{\gamma}$ -a.s., and finally $\nabla\varphi(x) = \nabla\varphi_\pi(x)$ μ -a.s. \square

Remark 3.7. φ and φ^* are called Kantorovitch potentials, and are different from the dual potentials.

Remark 3.8. $\nabla\varphi^* = (\nabla\varphi)^{-1}$; if $\nabla\varphi = T_{\mu \rightarrow \nu}$, then $\nabla\varphi^* = T_{\nu \rightarrow \mu}$.

From now on, all measures have a density, *i.e.* are in $\mathcal{P}_{2,ac}(\mathbb{R}^d)$.

3.3. Curves in the Wasserstein space.

3.3.1. The Wasserstein space as a metric space.

Proposition 3.9. $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a complete separable metric space.

Proof. We check the axioms:

- $W_2(\mu, \nu) \geq 0$
- $W_2(\mu, \mu) = \text{cost}(\text{id}) = 0$
- $W_2(\mu, \nu) = W_2(\nu, \mu)$
- triangular inequality: we will need the gluing lemma:

Lemma 3.10 (Gluing lemma). *Let X, Y, Z be three Polish spaces and let $\gamma_{XZ} \in \mathcal{P}(X \times Z), \gamma_{YZ} \in \mathcal{P}(Y \times Z)$ be such that $\pi_{\#}^Z \gamma_{XZ} = \pi_{\#}^Z \gamma_{YZ}$. Then there exists a measure $\gamma \in \mathcal{P}(X \times Y \times Z)$ such that*

$$\begin{aligned} \pi_{\#}^{X,Z} \gamma_{XYZ} &= \gamma_{XZ}, \\ \pi_{\#}^{Y,Z} \gamma_{XYZ} &= \gamma_{YZ}. \end{aligned}$$

Back to the triangle inequality:

$$\begin{aligned}
W_2(\mu, \nu) &\leq \left(\int \|x - y\|^2 d\gamma_{XYZ}(x, y, z) \right)^{1/2} && \text{by sub-optimality of } \gamma_{XYZ} \\
&\leq \left(\int \|x - z\|^2 d\gamma_{XZ}(x, z) \right)^{1/2} + \left(\int \|z - y\|^2 d\gamma_{YZ}(y, z) \right)^{1/2} && \text{by triangular inequality in } L^2 \\
&= W_2(\mu, \rho) + W_2(\rho, \nu). && \square
\end{aligned}$$

In fact, $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a length space! We can use Alexandrov geometry to measure some bounds on the curvature of the space (it is non-negatively curved, and flat in the subspace of the Gaussians as well of the diracs).

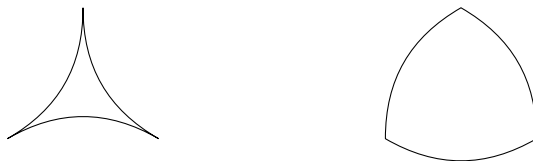


FIGURE 2. Alexandrov geometry.

3.3.2. *The continuity equation.* Let's make a fluid dynamics analogy, where μ_t is the density of the fluid at time t . There are two perspectives:

- the Lagrangian perspective: modelling the movement of each particle $\dot{X}_t = v_t(X_t)$
- the Eulerian perspective: modelling the behaviour of the whole density, $\dot{\mu}_t = ?$

The differential equation satisfied by μ_t will be called the continuity equation.

Theorem 3.11 (Continuity Equation). *Let $(v_t)_{t \geq 0} : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be a time dependent vector field and suppose that particles evolve according to the ODE $\dot{X}_t = v_t(X_t), t \geq 0$. Then $X_t \sim \mu_t$ where μ_t evolves according to the continuity equation (CE)*

$$\partial_t \mu_t + \operatorname{div}(\mu_t v_t) = 0. \quad (\text{CE})$$

Proof. Let f be a test function.

$$\begin{aligned}
\int f \partial_t \mu_t &= \partial_t \mathbb{E}[f(X_t)] \\
&= \mathbb{E}[\langle \nabla f(X_t), \dot{X}_t \rangle] \\
&= \mathbb{E}[\langle \nabla f(X_t), v_t(X_t) \rangle] \\
&= \int \langle \nabla f(x), \mu_t(x) v_t(x) \rangle dx \\
&= - \int f(x) \operatorname{div}(\mu_t(x) v_t(x)) dx. && \text{(integration by parts)} \quad \square
\end{aligned}$$

In fact, every “nice” curve of probability measures can be interpreted as a fluid flow along a time-varying vector field. And by nice, we mean:

Definition 3.12 (a.c. curve). *A curve $t \mapsto \mu_t$ in $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ is said to be absolutely continuous (a.c.) if at every time the metric derivative is finite, i.e. if*

$$\text{for all } t, \quad |\dot{\mu}|(t) := \lim_{s \rightarrow t} \frac{W_2(\mu_s, \mu_t)}{|s - t|} < \infty.$$

This vector field will not be unique: if we take w_t such that $\operatorname{div}(\mu_t w_t) = 0$, then we still have $\partial_t \mu_t = -\operatorname{div}(\mu_t(v_t)) = -\operatorname{div}(\mu_t(v_t + w_t))$. For instance, on the open ball $B_r(\mathbb{R}^2)$ with $r > 0$, equipped with the uniform density, any vector field v_t that consists in a rotation around $(0,0)$ satisfies the continuity equation, but one would prefer to choose the null vector field (Figure 3). Then, how to select v_t in a canonical way? We will either try to minimize the energy $\int \|v_t\|_{L^2(\mu_t)}^2 dt$ or force v_t to be the gradient of a function ψ_t – these two conditions being actually equivalent.

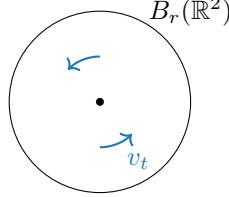


FIGURE 3. Non-uniqueness of the representing vector field: any rotating vector field satisfies (CE).

Theorem 3.13 (Curves of measures as fluid flows). *Let $t \mapsto \mu_t$ be an a.c. curve of measures. Then*

(i) *For any vector field $(\tilde{v}_t)_{t \geq 0}$ s.t. (CE) holds, we have*

$$|\dot{\mu}|(t) \leq \|\tilde{v}_t\|_{L^2(\mu_t)} \quad \text{“for all } t\text{”}. \quad (3.1)$$

(ii) *Conversely, there exists a unique choice of vector field $(v_t)_{t \geq 0}$ that satisfies (CE) s.t.*

$$\|v_t\|_{L^2(\mu_t)} \leq |\dot{\mu}(t)| \quad \text{“for all } t\text{”}.$$

Moreover, $v_t = \nabla \psi_t$ for $\psi_t : \mathbb{R}^d \rightarrow \mathbb{R}$ and $v_t = \lim_{\delta \rightarrow 0} \frac{T_{\mu_t \rightarrow \mu_{t+\delta}} - \text{id}}{\delta}$ ($T_{\mu_t \rightarrow \mu_{t+\delta}} - \text{id}$ is also called the displacement map).

Proof. (i) $\dot{X}_t = \tilde{v}_t(X_t)$. We define a flow map $\tilde{F}_{s,t}$ that maps X_s to X_t by any $(F_{t,t+\delta})_{\#} \mu_t = \mu_{t+\delta}$, or equivalently by $X_t \sim \mu_t \implies \tilde{F}_{t,t+\delta}(X_t) \sim \mu_{t+\delta}$. By sub-optimality, we have

$$\begin{aligned} \frac{W_2^2(\mu_t, \mu_{t+\delta})}{\delta^2} &\leq \int \frac{\|\tilde{F}_{t,t+\delta}(x) - x\|^2}{\delta^2} d\mu_t(x) \\ &= \int \|\tilde{v}_t\|^2 d\mu_t + o(1) && \text{as } \tilde{F}_{t,t+\delta}(x) - x = \delta \tilde{v}_t(x) + o(\delta) \end{aligned}$$

and since $W_2^2(\mu_t, \mu_{t+\delta})/\delta^2 \xrightarrow{\delta \rightarrow 0} |\dot{\mu}|(t)^2$, we obtain the upper bound.

(ii) Let (v_t) satisfying (CE) and $\|v_t\|_{L^2(\mu_t)} \leq |\dot{\mu}|(t)$ for all t ((v_t) is a minimizer. Since $g : w_t \mapsto \|v_t + w_t\|_{L^2(\mu_t)}^2$ is minimized at 0 over the convex set $\{w_t \mid \operatorname{div}(\mu_t w_t) = 0\}$ and g is strictly convex, we get the uniqueness. We now show that gradient fields, *i.e.* vector fields of the form $v_t = \nabla \psi_t$, satisfying (CE) are optimal.

• **intuitions:**

(a) We have that

$$\partial_t \int f \mu_t = \partial_t \mathbb{E}[f(X_t)] = \mathbb{E}[\langle \nabla f(X_t), v_t(X_t) \rangle] = \int \langle \nabla f, v_t \rangle d\mu_t,$$

hence if $v_t = v_t^\nabla + v_t^{\nabla^\perp}$, then it will be better to only keep the component v_t^∇ , the projection of v_t on $\{\nabla f, f \in \dots\}$. But this set is not convex and all so this is not rigorous at all.

(b) In the previous proof, we lost information by taking a random flow map. If we take the optimal $T_{\mu_t \rightarrow \mu_{t+\delta}}$ we get $\tilde{v}_t = \lim_{\delta \rightarrow 0} \frac{T_{\mu_t \rightarrow \mu_{t+\delta}} - \text{id}}{\delta}$, and $T_{\mu_t \rightarrow \mu_{t+\delta}}$ is the gradient of a convex function by Brenier’s theorem.

- **actual proof:** Let's take $v_t = \nabla\psi_t$. On one hand,

$$\begin{aligned} A &:= \frac{\int \psi_t d\mu_{t+\delta} - \int \psi_t d\mu_t}{\delta} = \int \psi_t \partial_t \mu_t + o(1) \\ &= - \int \psi_t \operatorname{div}(\mu_t \nabla \psi_t) + o(1) \\ &= \int \|\nabla \psi_t\|^2 d\mu_t + o(1) \\ &= \|\nabla \psi_t\|_{L^2(\mu_t)}^2 + o(1) \end{aligned}$$

and on the other hand

$$\begin{aligned} A &= \frac{\int \psi_t \circ T_{\mu_t \rightarrow \mu_{t+\delta}} d\mu_t - \int \psi_t d\mu_t}{\delta} \\ &= \int \langle \nabla \psi_t, \frac{T_{\mu_t \rightarrow \mu_{t+\delta}} - \operatorname{id}}{\delta} \rangle d\mu_t + o(1) \\ &\leq \|\nabla \psi_t\|_{L^2(\mu_t)} \frac{\|T_{\mu_t \rightarrow \mu_{t+\delta}} - \operatorname{id}\|_{L^2(\mu_t)}}{\delta} + o(1) \\ &\xrightarrow{\delta \rightarrow 0} \|\nabla \psi_t\|_{L^2(\mu_t)} |\dot{\mu}|(t), \end{aligned}$$

which means that $\|\nabla \psi_t\|_{L^2(\mu_t)} \leq |\dot{\mu}|(t)$, hence the optimality. \square

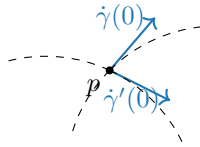
So far, we proved that optimal vector fields are unique and satisfy $\|\nabla \psi_t\|_{L^2(\mu_t)} = |\dot{\mu}|(t)$ (magnitude). Now, we will see the vector fields v_t as a velocity in the tangent space of $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$: this is *Otto calculus*.

3.3.3. The Wasserstein space as a Riemannian manifold: Otto Calculus. Background on Riemannian geometry. For reference on Riemannian geometry, see [Pet06; Do 92; Sch16].

Definition 3.14. A manifold \mathcal{M} of dimension d is a space which is locally homeomorphic to \mathbb{R}^d . At each point $p \in \mathcal{M}$ is a tangent space $\mathcal{T}_p \mathcal{M} = \{\text{velocity of curves through } p\}$. This whole structure (the tangent bundle) has to be smooth. A Riemannian metric is a smoothly varying choice $p \mapsto \langle \cdot, \cdot \rangle_p$, an inner product on $\mathcal{T}_p \mathcal{M}$. Then, for any curve γ , we have $\|\dot{\gamma}(t)\|_{\gamma(t)}^2 = \langle \dot{\gamma}(t), \dot{\gamma}(t) \rangle_{\gamma(t)}$. The distance function is defined as

$$d(p, q) = \inf \left\{ \int_0^1 \|\dot{\gamma}(t)\|^2 dt \mid \gamma(0) = p, \gamma(1) = q \right\}.$$

If the infimum is achieved by a γ^* , then γ^* is called a geodesic between p and q . Geodesics can be reparametrized to have constant speed $\|\dot{\gamma}(t)\| = c$ for all $t \in [0, 1]$. In this case, we get $d(\gamma(s), \gamma(t)) = |s - t|d(\gamma(0), \gamma(1))$ for all $0 \leq s < t \leq 1$. In the following sections, all geodesics considered will be chosen of constant speed.



We denote by $\mathcal{P} := \mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$.

Proposition 3.15. For all $\mu \in \mathcal{P}$, the tangent space of the Wasserstein space at μ is

$$\mathcal{T}_\mu \mathcal{P} = \overline{\{\nabla \psi \mid \psi \in \mathcal{C}_c^\infty(\mathbb{R}^d)\}}^{L^2(\mu)} = \overline{\{\lambda(T - \operatorname{id}) \mid \lambda > 0, T \text{ is an OT map}\}}^{L^2(\mu)},$$

and the inner product on $\mathcal{T}_\mu \mathcal{P}$ is $\langle \nabla \psi, \nabla \psi' \rangle_\mu = \int \langle \nabla \psi, \nabla \psi' \rangle d\mu$.

Recall that every $v_t \in \mathcal{T}_\mu \mathcal{P}$ satisfying (CE) is optimal.

Theorem 3.16. *Let $\mu_0, \mu_1 \in \mathcal{P}$. Then*

$$W_2(\mu_0, \mu_1) = \inf \left\{ \int_0^1 \|v_t\|_{L^2(\mu_t)} dt \mid (\text{CE}) \text{ holds} \right\},$$

the Benamou-Brenier formula. Moreover, the infimum is attained (there exists geodesics) as follows: let $\bar{\gamma}$ be an optimal coupling and $(X_0, X_1) \sim \bar{\gamma}$. Then μ_t is the unique constant speed geodesic between μ_0 and μ_1 , where $X_t = (1-t)X_0 + tX_1 \sim \mu_t$. We can also say that

$$\mu_t = [(1-t)\text{id} + tT_{\mu_0 \rightarrow \mu_1}]_{\#}\mu_0 = [\text{id} + t(T_{\mu_0 \rightarrow \mu_1} - \text{id})]_{\#}\mu_0.$$

Proof. • **equality:** Let's take a partition $0 = t_0 < t_1 < \dots < t_k = 1$.

$$\begin{aligned} W_2(\mu_0, \mu_1) &\leq \sum_{i=1}^k \underbrace{\frac{W_2(\mu_{i-1}, \mu_i)}{t_i - t_{i-1}}}_{\rightarrow |\dot{\mu}|(t_{i-1})} (t_i - t_{i-1}) && \text{(triangle inequality)} \\ &\rightarrow \int_0^1 |\dot{\mu}|(t) dt \\ &\leq \inf \left\{ \int \|v_t\| dt \mid (\text{CE}) \text{ holds} \right\} && \text{by integrating (3.1).} \end{aligned}$$

To show the reverse inequality, we take $X_t = X_0 + t(X_1 - X_0)$, for which $\dot{X}_t = X_1 - X_0$. Since $\dot{X}_t = v_t(X_t)$, $\mathbb{E}\|\dot{X}_t\|^2 = \|v_t\|_{L^2(\mu_t)}^2$. But we also have

$$\mathbb{E}\|\dot{X}_t\|^2 = \mathbb{E}_{\bar{\gamma}}\|X_0 - X_1\|^2 = \int \|x_0 - x_1\|^2 d\bar{\gamma}(x_0, x_1) = W_2^2(\mu_0, \mu_1),$$

which means we have equality for this choice of (v_t) .

• **uniqueness:** Let $\dot{X}_t = \tilde{v}_t(X_t)$ s.t. $\mathbb{E}\|\dot{X}_t\|^2 = \text{cst}$, $X_0 \sim \mu_0$, $X_1 \sim \mu_1$. Then

$$\begin{aligned} W_2(\mu_0, \mu_1) &\leq \mathbb{E}\|X_0 - X_1\|^2 = \mathbb{E}\left\| \int_0^1 \dot{X}_t dt \right\|^2 \\ &\leq \int_0^1 \mathbb{E}\|\dot{X}_t\|^2 dt && \text{by Jensen's inequality} \\ &= \int_0^1 \|\tilde{v}_t\|_{L^2(\mu_t)}^2 dt. \end{aligned}$$

We now have two bound gaps that can be tightened by:

- (1) choosing the optimal coupling;
- (2) applying Jensen over a constant integrand. □

Definition 3.17 (Wasserstein geodesic). *Let $\mu_0, \mu_1 \in \mathcal{P}$, $X_0 \sim \mu_0$, $X_1 \sim \mu_1$ optimally coupled, and $X_t = (1-t)X_0 + tX_1$. Then $\text{Law}(X_t = \mu_t) \iff \mu_t = [\text{id} + t(T_{\mu_0 \rightarrow \mu_1} - \text{id})]_{\#}\mu_0$, and the curve $t \mapsto \mu_t$ is called Wasserstein geodesic between μ_0 and μ_1 , a.k.a. the displacement/McCann interpolation.*

Remark 3.18. This induces a new geometry which is not the one induced by L^2 : if μ_0 and μ_1 have densities p_0 and p_1 , that we define the mixture $p_t = (1-t)p_0 + tp_1$, then $d(\mu_0, \mu_1) = (\int \|p_0 - p_1\|^2)^{1/2} \neq W_2(\mu_0, \mu_1)$.

Remark 3.19. Can we find geometries that share the same geodesics? Yes: recall that

$$W_1(\mu, \nu) = \inf_{\gamma \in \Gamma_{\mu, \nu}} \int \|x - y\| d\gamma(x, y) = \sup_{\|\nabla f\|_{\infty} \leq 1} \int f d(\mu - \nu).$$

W_1 has the same geodesics as L^2 , but it is not a flat space. Indeed,

$$W_1(\mu_s, \mu_t) = \sup_{\|\nabla f\|_{\infty} \leq 1} \int f d(\mu_s - \mu_t) = (t-s) \sup_{\|\nabla f\|_{\infty} \leq 1} \int f d(\mu_0 - \mu_1) = (t-s)W_1(\mu_0, \mu_1).$$

Geometry is carried by the distances, not by the geodesics.

4. WASSERSTEIN GRADIENT FLOWS

💡 We show that the Langevin diffusion can be interpreted as Wasserstein flow of the KL divergence using Otto calculus. We show that Log-Sobolev inequality corresponds to a PL condition in this geometry and we recover exponential convergence of the KL. We also use this new framework to show that such inequalities hold whenever the potential is strongly convex, which, in turn, implies strong convexity of the KL in this geometry.

The goal: Interpret the Langevin diffusion $dX_t = -\nabla V(X_t) dt + \sqrt{2} dB_t$ as a Wasserstein gradient flow of $\text{KL}(\cdot \| \pi)$, with $\pi \propto e^{-V}$, *i.e.* find μ_t such that

$$\partial_t \mu_t + \text{div}(\mu_t \cdot -\nabla_W \text{KL}(\mu_t \| \pi)) = 0.$$

4.1. Wasserstein gradient. We would like to generalize the notion of gradient in a manifold. The form $\nabla f(x) = (\partial_{x_1} f(x), \dots, \partial_{x_n} f(x))$ doesn't seem to be the right thing to generalize.

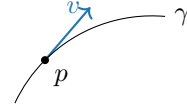
- in the Euclidean space:

$$\frac{f(p+tv) - f(p)}{t} \xrightarrow{t \rightarrow 0} \langle \nabla f(p), v \rangle,$$

but we have to consider $\nabla f(p)$ as an element of the dual of \mathbb{R}^d , and v as an element of the tangent space at $p \in \mathbb{R}^d$.

- in a Riemannian manifold: replacing $p+tv$ by $\gamma(t)$, the Riemannian gradient is defined by:

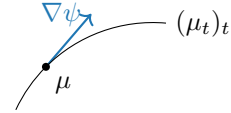
$$\frac{f(\gamma(t)) - f(p)}{t} \xrightarrow{t \rightarrow 0} \langle \nabla_{\mathcal{M}} f(p), v \rangle_p.$$



See [Bou22] for more information.

- in the Wasserstein space: for $F : \mathcal{P} \rightarrow \mathbb{R}$ a functional on the Wasserstein space, the Wasserstein gradient is defined by:

$$\frac{\mathcal{F}(\mu_t) - \mathcal{F}(\mu)}{t} \xrightarrow{t \rightarrow 0} \langle \nabla_W \mathcal{F}(\mu), \nabla \psi \rangle_\mu,$$



and since $\nabla_W \mathcal{F}(\mu)$ belongs to $\mathcal{T}_\mu \mathcal{P}$, it will be the (standard) gradient of some function ψ :

Definition 4.1 (First variation). *Let $\mathcal{F} : \mathcal{P} \rightarrow \mathbb{R}$ be a functional. The first variation of \mathcal{F} , denoted $\delta \mathcal{F}$, is defined by*

$$\lim_{\varepsilon \rightarrow 0} \frac{\mathcal{F}(\mu + \varepsilon \chi) - \mathcal{F}(\mu)}{\varepsilon} = \int \delta \mathcal{F}(\mu) \cdot \chi \quad \text{for all } \chi \text{ s.t. } \int \chi = 0.$$

Now,

$$\begin{aligned} \lim_{t \rightarrow 0} \frac{\mathcal{F}(\mu_t) - \mathcal{F}(\mu)}{t} &= \partial_t \Big|_{t=0} \mathcal{F}(\mu_t) \\ &= \int \delta \mathcal{F}(\mu) \cdot \partial_t \mu_t \, dt \\ &= - \int \delta \mathcal{F}(\mu) \, \text{div}(\mu_t \nabla \psi_t) \\ &= \int \langle \nabla \delta \mathcal{F}(\mu), \nabla \psi_t \rangle \, d\mu_t. \end{aligned}$$

Hence the Wasserstein gradient is $\nabla\delta\mathcal{F}(\mu)$, and it is unique. This also shows that the Wasserstein gradient $\nabla_{\mathbb{W}}\mathcal{F}(\mu)$ is a vector field of \mathbb{R}^d such that for all vector field v ,

$$\mathcal{F}((I + tv)_{\#}\mu) = \mathcal{F}(\mu) + t\langle \nabla_{\mathbb{W}}\mathcal{F}(\mu), v \rangle_{L^2(\mu)} + o(h).$$

Definition 4.2 (Wasserstein gradient). *Let $\mathcal{F} : \mathcal{P} \rightarrow \mathbb{R}$ be a functional. The Wasserstein gradient of \mathcal{F} at μ is $\nabla_{\mathbb{W}}\mathcal{F}(\mu) = \nabla\delta\mathcal{F}(\mu)$.*

Definition 4.3 (Wasserstein gradient flow). *The (negative) Wasserstein gradient flow of \mathcal{F} is a curve of measures $t \mapsto \mu_t$ s.t. its tangent vector at time t is $v_t = -\nabla_{\mathbb{W}}\mathcal{F}(\mu_t)$, or equivalently s.t. $\partial_t\mu_t = \operatorname{div}(-\mu_t\nabla_{\mathbb{W}}\mathcal{F}(\mu_t)) = \operatorname{div}(-\mu_t\nabla\delta\mathcal{F}(\mu_t))$.*

Example 4.4 (First variations and Wasserstein GF of important functionals).

- (1) The *potential energy* $\mathcal{E}(\mu) = \int V d\mu$: since $\frac{\mathcal{E}(\mu+\varepsilon\chi) - \mathcal{E}(\mu)}{\varepsilon} = \int V d\chi$, one has that $\delta\mathcal{E}(\mu) = V$, and therefore that $\nabla_{\mathbb{W}}\mathcal{E}(\mu) = \nabla V(\cdot)$;
- (2) The *entropy* $\mathcal{H}(\mu) = \int \mu \log \mu$: similarly, $\nabla_{\mathbb{W}}\mathcal{H}(\mu) = \nabla \log \mu$;
- (3) The *interaction potential* $\mathcal{W}(\mu) = \frac{1}{2} \int \int W(x-y) d\mu(x) d\mu(y)$: one has that $\nabla_{\mathbb{W}}\mathcal{W}(\mu) = \int \nabla W(\cdot - y) d\mu(y)$.

4.2. Langevin diffusion as a Wasserstein gradient flow. Let's compute the Wasserstein gradient of $\mathcal{F} := \operatorname{KL}(\cdot \| \pi)$, with $\pi \propto e^{-V}$:

$$\begin{aligned} \mathcal{F}(\mu) &= \operatorname{KL}(\mu \| \pi) \\ &= \int \mu \log \frac{\mu}{\pi} \\ &= \int \mu \log \mu - \int \underbrace{\mu \log \pi}_{=-V} \\ &= \mathcal{H}(\mu) + \mathcal{E}(\mu). \end{aligned}$$

Therefore, $\nabla_{\mathbb{W}}\mathcal{F}(\mu) = \nabla \log \mu + \nabla V = \nabla \left(\log \frac{\mu}{\pi} \right)$, and the Wasserstein gradient flow reads

$$\dot{X}_t = -\nabla_{\mathbb{W}}\mathcal{F}(\mu_t)(X_t) = -\nabla \log \frac{\mu_t}{\pi}(X_t) = -\nabla \log \mu_t(X_t) + \nabla V(X_t).$$

The gradient flow expression for $\partial_t\mu_t$ then becomes

$$\partial_t\mu_t = \operatorname{div}(\mu_t(\nabla \log \mu_t - \nabla V(X_t))) = \operatorname{div}(\nabla\mu_t) - \operatorname{div}(\mu_t\nabla V(X_t)) = \Delta\mu_t - \operatorname{div}(\mu_t\nabla V(X_t)),$$

and we recover the Fokker–Planck equation for Langevin (2.1).

💡 This is the result of [JKO98] – except they didn't need a notion of gradient to obtain this.

4.3. Rates of convergence. We already know that if the Hessian of a function f is greater than αI , we can obtain a rate of convergence of $e^{-2\alpha t}$ for f . In our case, what is the Hessian of $\operatorname{KL}(\cdot \| \pi)$? For a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, one has

$$\begin{aligned} \partial_t f(x_t) &= \langle \nabla f(x_t), \dot{x}_t \rangle \\ \partial_t^2 f(x_t) &= \langle \nabla^2 f(x_t) \dot{x}_t, \dot{x}_t \rangle + \underbrace{\langle \nabla f(x_t), \ddot{x}_t \rangle}_{=0}, \end{aligned}$$

where the second term is equal to zero since we only examine constant speed geodesics. In the Wasserstein space, a geodesic is given by $\mu_t = [\operatorname{id} + t(T - \operatorname{id})]_{\#}\mu_0$. By analogy, we therefore say that

the Wasserstein Hessian $\nabla_{\mathbb{W}}^2$ satisfies

$$\partial_t^2 \mathcal{F}(\mu_t) := \langle \nabla_{\mathbb{W}}^2 \mathcal{F}(\mu_t) \cdot (T - \operatorname{id}), T - \operatorname{id} \rangle_{\mu_t}.$$

Let us compute this using the fact that $\mathcal{F} = \mathcal{E} + \mathcal{H}$:

- for the potential \mathcal{E} :

$$\begin{aligned}\partial_t \mathcal{E}(\mu_t) &= \langle \nabla_{\mathbb{W}} \mathcal{E}(\mu_t), T - \text{id} \rangle_{L^2(\mu_t)} = \langle \nabla V, T - \text{id} \rangle_{L^2(\mu_t)} \\ \partial_t \Big|_{t=0} \mathcal{E}(\mu_t) &= \mathbb{E}[\langle \nabla V(X_0), T(X_0) - X_0 \rangle] \\ \partial_t^2 \Big|_{t=0} \mathcal{E}(\mu_t) &= \mathbb{E}[\langle \nabla^2 V(X_0) \cdot (T(X_0) - X_0), T(X_0) - X_0 \rangle]\end{aligned}$$

hence $\nabla_{\mathbb{W}}^2 \mathcal{E}(\mu_0) = \nabla^2 V$; and if V is α -strongly convex, then

$$\partial_t^2 \Big|_{t=0} \mathcal{E}(\mu_t) \geq \alpha \mathbb{E} \|T(X_0) - X_0\|^2 = \alpha W_2^2(\mu \| \pi).$$

- for the entropy \mathcal{H} : Denoting T_t the OT map from μ_0 to μ_t , we have for any function h

$$\begin{aligned}\int h \, d\mu_0 &= \mathbb{E}[h(X_0)] \\ &= \mathbb{E}[h(T_t^{-1}(X_t))] \\ &= \int h(T_t^{-1}(y)) \mu_t(y) \, dy \\ &= \int h(x) \mu_t \circ T_t(x) \det(\nabla T_t(x)) \, dx \quad (y = T_t(x))\end{aligned}$$

hence $\mu_0(x) = \mu_t \circ T_t(x) \det(\nabla T_t(x))$. Therefore

$$\begin{aligned}\mathcal{H}(\mu_t) &= \int \mu_t \log \mu_t \\ &= \mathbb{E}[\log \mu_t \circ T_t(X_0)] \\ &= \mathbb{E}[\log(\det(\nabla T_t(X_0))^{-1} \mu_0(X_0))] \\ &= \mathcal{H}(\mu_0) - \mathbb{E} \log \det(\nabla T_t(X_0)).\end{aligned}$$

Moreover, $\nabla T_t = (1-t)\text{id} + t\nabla T$, which is the linear combination of the identity and the Hessian of a convex function, hence $\nabla T_t(X_0)$ is a positive semi-definite matrix. The log det being concave over the space of PSD matrices [BBV04], we get that $t \mapsto \mathcal{H}(\mu_t)$ is convex.

Hence

$$\partial_t^2 \text{KL}(\mu_t \| \pi) = \partial_t^2 \mathcal{E}(\mu_t) + \partial_t^2 \mathcal{H}(\mu_t) \geq \alpha W_2^2(\mu \| \pi) + 0,$$

where we used the convexity of \mathcal{H} . In fact, we have $\partial_t^2|_{t=0} \mathcal{H}(\mu_t) = \int \|\nabla T - \text{id}\|_F^2 \, d\mu_0 \geq 0$. See [Gig12] for more information. We have proved the following theorem:

Theorem 4.5. *If $\pi \propto e^{-V}$ where V is α -strongly convex, then $\text{KL}(\cdot \| \pi)$ is α -strongly convex along Wasserstein geodesics.*

Let us now study the rate of convergence.

$$\begin{aligned}\mathcal{F}(\mu_t) &= \text{KL}(\mu_t \| \pi) \\ \partial_t \mathcal{F}(\mu_t) &= \langle \nabla_{\mathbb{W}} \mathcal{F}(\mu_t), -\nabla_{\mathbb{W}} \mathcal{F}(\mu_t) \rangle_{L^2(\mu_t)} \quad (\text{just the gradient flow}) \\ &= -\|\nabla_{\mathbb{W}} \mathcal{F}(\mu_t)\|_{L^2(\mu_t)}^2 \\ &\leq -2\alpha \mathcal{F}(\mu_t) \quad \text{by (PL)}.\end{aligned}$$

Then by Grönwall's lemma $\mathcal{F}(\mu_t) \leq \mathcal{F}(\mu_0) e^{-2\alpha t}$. In order to show that the KL satisfies the (PL) inequality, we first use a Taylor approximation of the KL between μ_0 and μ_1 :

$$\begin{aligned}f(t) &= f(0) + t f'(0) + \int_0^t \frac{t-s}{2} f''(s) \, ds \\ \text{KL}(\mu_t \| \pi) &\geq \text{KL}(\mu_0 \| \pi) + \langle \nabla_{\mathbb{W}} \text{KL}(\mu_0 \| \pi), t(T - \text{id}) \rangle_{L^2(\mu_0)} + \frac{\alpha}{2} t^2 \underbrace{\|T - \text{id}\|_{L^2(\mu_0)}^2}_{W_2^2(\mu_0, \mu_1)}.\end{aligned} \quad (4.1)$$

By taking $t = 1$, $\mu_0 := \mu$ and $\mu_1 := \arg \min_{\mu} \mathcal{F}(\mu) = \pi$, this inequality becomes

$$\begin{aligned} \mathcal{F}(\mu) &\leq -\langle \nabla_{\mathbb{W}} \text{KL}(\mu \parallel \pi), T - \text{id} \rangle_{L^2(\mu)} - \frac{\alpha}{2} \int \|T - \text{id}\|^2 d\mu \\ &\leq \frac{\delta}{2} \|\nabla_{\mathbb{W}} \text{KL}(\mu \parallel \pi)\|_{L^2(\mu)}^2 + \frac{1}{2\delta} \|T - \text{id}\|_{L^2(\mu)}^2 - \frac{\alpha}{2} \|T - \text{id}\|_{L^2(\mu)}^2 \quad (\text{Young's inequality}) \\ &= \frac{1}{2\alpha} \|\nabla_{\mathbb{W}} \text{KL}(\mu \parallel \pi)\|_{L^2(\mu)}^2 \quad \text{setting } \delta = \frac{1}{\alpha}, \end{aligned}$$

hence (PL). $\textcircled{?}$ Now, what can be said about the convergence of $W_2^2(\mu_t, \pi)$?

$$\begin{aligned} \partial_t W_2^2(\mu_t, \pi) &= \partial_t \|T_{\mu_0 \rightarrow \pi} - \text{id}\|_{L^2(\mu_t)}^2 \\ &= \partial_t \mathbb{E} \|X_t - X_\infty\|^2 \quad \text{with } X_\infty \sim \pi, (X_t, X_\infty) \text{ optimally coupled} \\ &= 2\mathbb{E} \langle X_t - X_\infty, \dot{X}_t \rangle \\ &= -2\mathbb{E} \langle X_t - X_\infty, \nabla_{\mathbb{W}} \mathcal{F}(\mu_t)(X_t) \rangle_{\mu_t} \\ &\leq -2 \left(\underbrace{\text{KL}(\mu_t \parallel \pi)}_{\geq 0} + \frac{\alpha}{2} \underbrace{\int \|T_{\mu_t \rightarrow \pi} - \text{id}\|^2 d\mu_t}_{=W_2^2(\mu_t, \pi)} \right) \quad \text{where we used (4.1) but between } \mu_t \text{ and } \pi \\ &\leq -\alpha W_2^2(\mu_t, \pi), \end{aligned}$$

and by Grönwall's lemma, we get that $W_2^2(\mu_t, \pi) \leq W_2^2(\mu_0, \pi)e^{-\alpha t}$.

Remark 4.6 (Functional inequalities). Strong convexity (SC) implies *quadratic growth* (QG):

$$\mathcal{F}(\mu) \geq \frac{\alpha}{2} W_2^2(\mu, \pi) \quad \text{for all } \mu. \quad (\text{QG})$$

For Langevin, this becomes Talagrand's inequality

$$\text{KL}(\mu \parallel \pi) \geq \frac{\alpha}{2} W_2^2(\mu, \pi) \quad \text{for all } \mu. \quad (T_2)$$

We saw earlier (2.2) that

$$\chi^2(\mu \parallel \pi) \geq \text{KL}(\mu \parallel \pi);$$

we also have *Pinsker's inequality*:

$$\text{KL}(\mu \parallel \pi) \geq 2 \|\mu - \pi\|_{\text{TV}}^2. \quad (\text{Pinsker})$$

One also has:

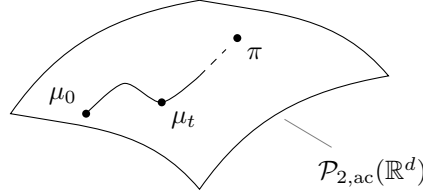
$$\begin{aligned} (\text{SC}) &\implies (\text{LSI}) \xrightarrow{(*)} (T_2) \quad \text{and} \quad (\text{SC}) \implies (\text{PL}) \xrightarrow{(*)} (\text{QG}) \\ &\implies (\text{Poincaré}) \end{aligned}$$

where $(*)$ are equivalences if π is log-concave (see [KNS16] for details), and

$$(\text{LSI}) \iff (\text{PL}) \quad \text{in the Wasserstein space.}$$

Summary.

- (1) $\text{KL}(\cdot \parallel \pi)$ is α -strongly convex along Wasserstein geodesics if V is α -strongly convex, and this implies $W_2^2(\mu_t, \nu_t) \leq W_2^2(\mu_0, \nu_0)e^{-2\alpha t}$ (generalization of what we shown with 2 Langevin processes, can be done as an exercise)
- (2) π satisfies (LSI) with constant $\frac{1}{\alpha} \iff \text{KL}(\pi_t \parallel \pi) \leq \text{KL}(\pi_0 \parallel \pi)e^{-2\alpha t}$
- (3) π satisfies (Poincaré) with constant $\frac{1}{\alpha} \implies \chi^2(\pi_t \parallel \pi) \leq \chi^2(\pi_0 \parallel \pi)e^{-2\alpha t}$



5. APPLICATIONS

💡 This new perspective on sampling opens the possibility of algorithms that consist in deterministically implementing the Wasserstein gradient flow. We review two of them: Stein Variational Gradient Descent (SVGD) and Variational Inference (VI).

Recall that our goal is still to be able to compute any $\int \varphi d\pi$, with the posterior $\pi \propto e^{-V}$. Let's say that we dispose of a discrete measure approximating π , for instance $\pi \approx \sum w_i \delta_{x_i}$; then $\int \varphi d\pi \approx \sum w_i \varphi(x_i)$. In order to trace a curve $t \mapsto \mu_t$ that will asymptotically be close to π , we have 3 possibilities:

- (1) *Solve the Fokker–Planck equation* $\partial_t \mu_t = \text{div}(\mu_t \nabla V) + \Delta \mu_t$ by finite elements methods (deterministic): this is implementable via discretization of \mathbb{R}^d but not scalable as exponential in the dimension d ;
- (2) *Run the Langevin diffusion* $dX_t = -\nabla V(X_t) + \sqrt{2} dB_t$ (random): this is implementable, and actually better than Fokker–Planck since parametrizable;
- (3) *Perform a Wasserstein gradient flow* of $\text{KL}(\cdot \| \pi)$ using $\dot{X}_t = -\nabla_{\text{W}} \text{KL}(\mu_t \| \pi) = -\nabla V(X_t) - \nabla \log \mu_t(X_t)$ (deterministic): this is not readily implementable since we need μ_t from X_t . The idea is then to project $\nabla_{\text{W}} \text{KL}(\mu_t \| \pi)(\cdot)$ onto “tractable” vector fields.

5.1. Stein Variational gradient descent (SVGD). 💡 Introduced in [LW16], doesn't really work or scale but people like it since it is an alternative to MCMC.

Background on kernels and RKHS.

Definition 5.1 (p.s.d. kernel). *We say that $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ is a positive semi-definite (p.s.d.) kernel if for all $x_1, \dots, x_n \in \mathbb{R}^d$, $(k(x_i, x_j))_{1 \leq i, j \leq n}$ is a p.s.d. matrix. This creates a Hilbert space \mathcal{H} called Reproducing Kernel Hilbert Space (RKHS) of functions from \mathbb{R}^d to \mathbb{R}*

$$\mathcal{H} = \left\{ \sum_{j=1}^n \alpha_j k(x_j, \cdot) \mid n \geq 1, \alpha_j \in \mathbb{R}, x_j \in \mathbb{R}^d \right\},$$

equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and an associated norm $\| \cdot \|_{\mathcal{H}}$.

Example 5.2 (Kernels).

- Gaussian kernel $k(x, y) = e^{-\frac{1}{2}\|x-y\|^2}$
- Laplace kernel $k(x, y) = e^{-\|x-y\|}$
- Linear kernel $k(x, y) = \langle x, y \rangle$ that allows us to fall back on what we know.

The idea is then to look for $\arg \min_{g \in \mathcal{H}} \text{KL}((I + \varepsilon g)_{\#} \mu_t \| \pi)$. The projection map we are going to use is $\mathcal{K}_{\mu} : v(\cdot) \mapsto \int K(\cdot, y) v(y) d\mu(y)$, where K has vectorial values but where we assume $K(x, y) = k(x, y)I_d$. If $k(x, y) = \delta_{x=y}$, then $\mathcal{K}_{\mu} v = v$. This is interesting for us since we can compute

$$\begin{aligned} \mathcal{K}_{\mu} \nabla_{\text{W}} \text{KL}(\mu \| \pi) &= \int k(x, y) \nabla \log \frac{\mu}{\pi}(y) d\mu(y) \\ &= \int k(x, y) \frac{\nabla \mu}{\mu}(y) d\mu(y) + \int k(x, y) \nabla V(y) d\mu(y) \\ &= - \int \nabla_y k(x, y) d\mu(y) + \int k(x, y) \nabla V(y) d\mu(y). \quad (\text{integration by parts}) \end{aligned}$$

The new dynamics are:

$$\dot{X}_t = -\mathcal{K}_{\mu_t} \nabla_{\text{W}} \text{KL}(\mu_t \| \pi)(X_t)$$

$$= - \int k(X_t, y) \nabla V(y) d\mu_t(y) + \int \nabla_y k(X_t, y) d\mu_t(y),$$

which is now linear in μ_t !

Remark 5.3 (Discrete case). If we initialize at $\mu_0 = \frac{1}{n} \sum_{j=1}^n \delta_{x_0^{(j)}}$, a mixture of n masses, we will end up at time t with a mixture of n masses $\mu_t = \frac{1}{n} \sum_{j=1}^n \delta_{x_t^{(j)}}$ and the particles are going to evolve in a coupled way:

$$\begin{aligned} \dot{x}_t^{(i)} &= - \int k(x_t^{(i)}, y) \nabla V(y) d\mu_t(y) + \int \nabla_y k(x_t^{(i)}, y) d\mu_t(y) \\ &= - \frac{1}{n} \sum_{j=1}^n k(x_t^{(i)}, x_t^{(j)}) \nabla V(x_t^{(j)}) + \int \nabla_y k(x_t^{(i)}, x_t^{(j)}) d\mu_t(y). \end{aligned}$$

Hence particles that are close from each other in the sense of k will interact.

Remark 5.4. If $k(x, y) = k(x - y)$,

$$\int \nabla_y k(x, y) d\mu(y) = \int \nabla k(x - y) d\mu(y) = \nabla_{\mathbb{W}} \mathcal{W}(\mu),$$

where $\mathcal{W}(\mu) = \int k(x - y) d\mu(x) d\mu(y)$: the second term of the dynamics is an interaction potential.

Remark 5.5. \mathcal{H} does not contain L^2 . How is it that it works then? We modded out some directions, how could we have $\mu_t \rightarrow \pi$? Actually, is kind of a miracle. We would like to have

$$\forall x, \quad \int k(x, y) \nabla \log \frac{\mu}{\pi}(y) d\mu(y) = 0 \implies \mu = \pi;$$

integrating this equality gives

$$S_k(\mu \parallel \pi) := \iint k(x, y) \left\langle \nabla \log \frac{\mu}{\pi}(y), \nabla \log \frac{\mu}{\pi}(x) \right\rangle d\mu(x) d\mu(y) = 0,$$

where S_k is the *kernelized Stein discrepancy*. It is actually a kernelized version of the Fisher Information (FI). Can we then obtain $\nabla \log \frac{\mu}{\pi} = 0$?

Definition 5.6 (ISPD kernel). A kernel k is called integrally strictly positive definite (ISPD) if

$$\forall g \neq 0 \in L^2, \quad \iint k(x, y) g(x) g(y) dx dy > 0.$$

Remark 5.7. The Gaussian and Laplace kernels are ISPD.

Lemma 5.8. Let k be ISPD. Then $S_k(\mu \parallel \pi) \geq 0$ with equality if and only if $\mu = \pi$.

We now have that $\mathcal{K}_\mu \nabla_{\mathbb{W}} \text{KL}(\mu \parallel \pi) = 0 \implies \pi = \mu$ a.e..

But what is the rate of convergence? This is still very unclear, and there are lots of open questions. We will need something that looks like a (PL) inequality:

Definition 5.9 (Stein log-Sobolev inequality). The KL is said to satisfy the Stein log-Sobolev inequality if

$$2\alpha \text{KL}(\mu \parallel \pi) \leq \int \|\mathcal{K}_\mu \nabla_{\mathbb{W}} \text{KL}(\mu \parallel \pi)\|^2.$$

💡 This can be compared with the (PL) over the new space.

[DNS19] defines the Stein geometry:

$$\begin{aligned} d(\mu_0, \mu_1) &= \inf \int_0^1 \|v_t\| d\mu_t && \text{s.t. (CE) holds (before)} \\ d(\mu_0, \mu_1) &= \inf \int_0^1 \|\mathcal{K}_{\mu_t} v_t\| d\mu_t && \text{s.t. (CE) holds (now)} \end{aligned}$$

They also claim that the Stein LSI does not hold; but can we find simple families of μ and π such that it does not hold?

Remark 5.10. Some open questions, some of them being studied by A. Korba, A. Salim [Kor+20]:

- $n \rightarrow \infty$ (number of particles): how close are the discrete and true trajectories? \rightarrow propagation of chaos, the distance between the two is exponential in the time T
- for finite n , what is $\lim_{t \rightarrow \infty} \hat{\mu}_t = \frac{1}{n} \sum_{i=1}^n x_t^{(i)}$? It does have some optimal quantization properties, like following a honeycomb pattern. Can we find bounds on the quantity

$$\left| \frac{1}{n} \sum_{i=1}^n \varphi(x^{(i)}) - \int \varphi d\mu \right|?$$


- Our problem in this course was that we didn't know how to compute $\nabla \log \mu_t = \frac{\nabla \hat{\mu}_t}{\hat{\mu}_t}$; one could build a kernel density estimator with $x_t^{(1)}, \dots, x_t^{(n)}$:

$$\hat{\mu}_t(x) = \frac{1}{nh^d} \sum_{i=1}^n k\left(\frac{x_t^{(i)} - x}{h}\right),$$

then compute $\frac{\nabla \hat{\mu}_t}{\hat{\mu}_t}$, which works well if $\hat{\mu}_t$ is smooth. The dynamics we get is

$$\dot{x}_t^{(i)} = -\nabla V(x_t^{(i)}) - \frac{\sum_{j=1}^n \nabla k\left(\frac{x_t^{(j)} - x}{h}\right)}{h \sum_{j=1}^n k\left(\frac{x_t^{(j)} - x}{h}\right)},$$

which is a competitor for SVGD.

5.2. Variational Inference (VI).  We sketch here the idea of [Lam+22]. The goal: study problems of the form

$$\arg \min_{p \in \mathcal{P}} \text{KL}(p \parallel \pi),$$

where \mathcal{P} is a parametric class, hopefully convenient to compute things like $\mathbb{E}_p[X]$ and $\text{Cov}_p(X)$. Usually, this mainly consists in heuristics, you throw a gradient descent at it and sometimes it works.

Example 5.11 (Some parametric classes \mathcal{P}).

- *Mean-field VI*: $\mathcal{P} \subset \{\text{product distributions}\}$, *i.e.* we do not care about correlation (Figure 6). It is sometimes sufficient!
- *Gaussian VI*: $\mathcal{P} \subset \{\mathcal{N}(m, \Sigma) \mid \Sigma \in S^d, m \in \mathbb{R}^d\}$, where S^d is the set of spd matrices;
- [Lam+22]: $\mathcal{P} \subset \{\text{mixtures of Gaussians}\}$.

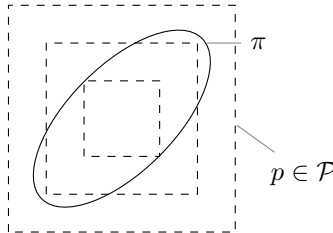


FIGURE 6. Approximating correlated random variables by a product distribution $p \in \mathcal{P}$ (mean-field VI).

In the case where π is a Gaussian distribution (of mean 0 and covariance matrix I):

$$\text{KL}(\mathcal{N}(m, \Sigma) \parallel \mathcal{N}(0, I)) = \frac{1}{2} [\text{Tr} \Sigma - d + \|m\|^2 - \log \det \Sigma].$$

This is convex in (m, Σ) ! Now, for a generic $\pi \propto e^{-V}$:

$$\text{KL}(\mathcal{N}(m, \Sigma) \parallel \pi) = \underbrace{-\frac{1}{2} \log \det \Sigma}_{\approx \text{entropy of } \mathcal{N}(m, \Sigma)} + \underbrace{\int \frac{e^{-\frac{1}{2}(x-m)^\top \Sigma^{-1}(x-m)}}{((2\pi)^d \det \Sigma)^{1/2}} V(x) dx}_{\approx \text{potential energy of } \mathcal{N}(m, \Sigma)} + \text{cst},$$

which is not convex anymore... but the potential energy reads

$$\mathcal{E}_V(\mathcal{N}(m, \Sigma)) = \mathbb{E}_{X \sim \mathcal{N}(m, \Sigma)}[V(X)] = \mathbb{E}_{Z \sim \mathcal{N}(0, I)}[V(m + \Sigma^{1/2} Z)].$$

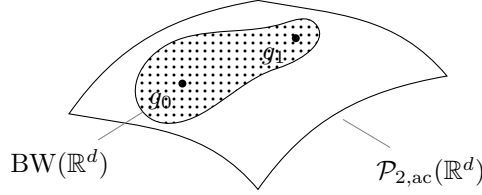
Hence

$$\text{KL}(\mathcal{N}(m, \Sigma) \parallel \pi) = -\log \det \Sigma^{1/2} + \mathbb{E}_{Z \sim \mathcal{N}(0, I)}[V(m + \Sigma^{1/2} Z)],$$

and if V is convex, then $(m, \Sigma^{1/2}) \mapsto \text{KL}(\mathcal{N}(m, \Sigma) \parallel \pi)$ is convex. We could then perform a stochastic GD [ARC16].

💡 What we explore here: does Wasserstein geometry helps (too)?

5.2.1. *Bures-Wasserstein*. The Bures-Wasserstein distance is an extension of the Bures distance between SPD matrices. We denote $\text{BW}(\mathbb{R}^d)$ the set of d -dimensional spd matrices.



💡 **Insight:** the Wasserstein geodesic between 2 Gaussians stays in the space of Gaussians. For $g_0 \sim \mathcal{N}(m_0, \Sigma_0)$ and $g_1 \sim \mathcal{N}(m_1, \Sigma_1)$, a candidate for an affine transport map is $x \mapsto \Sigma_1^{1/2} \Sigma_0^{-1/2} (x - m_0) + m_1$. It is the gradient of a function but not of a convex one since $\Sigma_1^{1/2} \Sigma_0^{-1/2}$ is not psd... Actually:

$$T_{g_0 \rightarrow g_1}(x) = m_1 + \underbrace{\Sigma_0^{-1/2} (\Sigma_0^{1/2} \Sigma_1 \Sigma_0^{1/2})^{1/2} \Sigma_0^{-1/2}}_{:=S} (x - m_0)$$

has the right mean and covariance. It is the gradient of $x \mapsto m_1 x + \frac{1}{2} \|S^{1/2}(x - m_0)\|^2$, which is a convex function since S is psd. We know the form of the Wasserstein geodesics; they will be of the form $g_t = [(1-t) \text{id} + t T_{g_0 \rightarrow g_1}]_{\#} g_0$. If $X_0 \sim g_0$, then $X_t = (1-t)X_0 + t(m_1 + S(X_0 - m_0)) \sim \mathcal{N}(m_t, \Sigma_t)$. The geodesics stay in $\text{BW}(\mathbb{R}^d)$: we say that $\text{BW}(\mathbb{R}^d)$ is a *geodesically convex subset* of $\mathcal{P}_{2,ac}(\mathbb{R}^d)$. Just rephrasing what we shown above, we know that $\text{KL}(\cdot \parallel \pi)$ is α -strongly convex along BW geodesics $\iff V$ is α -strongly convex.

The tangent vectors to a geodesic are given by

$$\mathcal{T}_g \text{BW}(\mathbb{R}^d) = \{x \mapsto a + S(x - m_g) \mid a \in \mathbb{R}^d, S \text{ symmetric}\},$$

and we can therefore identify $\mathcal{T}_g \text{GW}(\mathbb{R}^d)$ to $\{(a, S) \mid a \in \mathbb{R}^d, S \text{ symmetric}\}$. Our metric on $\mathcal{T}_g \text{BW}(\mathbb{R}^d)$ is then

$$\begin{aligned} \langle (a, S), (a', S') \rangle_g &= \int \langle a + S(x - m_g), a' + S'(x - m_g) \rangle dg \\ &= \langle a, a' \rangle + \mathbb{E}_g[\text{Tr}(SS'(x - m_g)(x - m_g)^\top)] \\ &= \langle a, a' \rangle + \text{Tr}(SS' \Sigma_g) \\ &= \langle a, a' \rangle + \langle S, \Sigma_g S' \rangle. \end{aligned}$$

A geodesic with velocity (a, S) is $g_t = [\text{id} + t(a + S(\cdot - m_0))]_{\#} g_0$, the law of $X_t = X_0 + t(a + S(X_0 - m_0))$, hence $g_t \sim \mathcal{N}(m_0 + ta, (I + tS)\Sigma_0(I + tS))$. This is fully parametric, and we can precisely tell how m_t

and Σ_t are moving:

$$\begin{cases} \dot{m}_0 = a \\ \dot{\Sigma}_0 = S\Sigma_0 + \Sigma_0 S. \end{cases}$$

Now, what is the BW gradient? Considering $g_t \sim \mathcal{N}(m_t, \Sigma_t)$ of velocity (a, S) at 0, we are looking for $(\bar{a}, \bar{S}) = \nabla_{\text{BW}} f(m_0, \Sigma_0) \in \mathcal{T}_{g_0} \text{BW}(\mathbb{R}^d)$, and we know that

$$\partial_t \Big|_{t=0} f(m_t, \Sigma_t) = \langle \nabla_{\text{BW}} f(m_0, \Sigma_0), (a, S) \rangle_{g_0},$$

so we are looking for (\bar{a}, \bar{S}) such that

$$\begin{aligned} \langle \bar{a}, a \rangle + \langle \bar{S}, \Sigma_0 S \rangle &= \langle \nabla_m f(m_0, \Sigma_0), \dot{m}_0 \rangle + \langle \nabla_{\Sigma} f(m_0, \Sigma_0), \dot{\Sigma}_0 \rangle \\ &= \langle \nabla_m f(m_0, \Sigma_0), a \rangle + \langle \nabla_{\Sigma} f(m_0, \Sigma_0), S\Sigma_0 + \Sigma_0 S \rangle \\ &= \langle \nabla_m f(m_0, \Sigma_0), a \rangle + 2\langle \nabla_{\Sigma} f(m_0, \Sigma_0), \Sigma_0 S \rangle \end{aligned}$$

and we identify:

$$\begin{cases} \bar{a} = \nabla_m f(m_0, \Sigma_0) \\ \bar{S} = 2\nabla_{\Sigma} f(m_0, \Sigma_0), \end{cases}$$

which means that $\nabla_{\text{BW}} f(m_0, \Sigma_0) = (\nabla_m f(m_0, \Sigma_0), 2\nabla_{\Sigma} f(m_0, \Sigma_0))$ and this is also what we would have got using Alquier's geometry [ARC16].

BW gradient as a projection of W gradient.

We consider a curve in BW with tangent vectors $v_t \in \mathcal{T}_{g_t} \text{BW}(\mathbb{R}^d)$. By definition,

$$\partial_t \mathcal{F}(g_t) = \langle \nabla_{\text{BW}} \mathcal{F}(g_t), v_t \rangle_{g_t},$$

but we also have

$$\partial_t \mathcal{F}(g_t) = \langle \nabla_{\text{W}} \mathcal{F}(g_t), v_t \rangle_{g_t},$$

hence $\nabla_{\text{BW}} \mathcal{F}(g_t) = \text{proj}_{\mathcal{T}_{g_t} \text{BW}(\mathbb{R}^d)} \nabla_{\text{W}} \mathcal{F}(g_t)$, and that is convenient:

$$\begin{cases} \nabla_{\text{W}} \text{KL}(\mu \| \pi) = \nabla \log \frac{\mu}{\pi} \\ \nabla_{\text{BW}} \text{KL}(\mu \| \pi) = (\bar{a}, \bar{S}), \end{cases}$$

so we have

$$\begin{aligned} \langle \bar{a}, a \rangle + \langle \bar{S}, \Sigma S \rangle &= \int \langle \nabla \log \frac{\mu}{\pi}(x), S(x - m) \rangle dg(x) \\ &= \langle \int \nabla \log \frac{\mu}{\pi}(x) dg(x), a \rangle + \underbrace{\int \langle \Sigma S \nabla \log \frac{g}{\pi}(x), \Sigma^{-1}(x - m) \rangle dg(x)}_{(*)}. \end{aligned}$$

Since $g \propto e^{-\frac{1}{2}(x-m)^\top \Sigma^{-1}(x-m)}$, we have that $\Sigma^{-1}(x - m) = -\nabla \log g = -\frac{\nabla g}{g}$, hence

$$\begin{aligned} (*) &= - \int \langle \Sigma S \nabla \log \frac{g}{\pi}, \nabla g \rangle dx \\ &= \int \text{div} \left(\Sigma S \nabla \log \frac{g}{\pi} \right) dg \quad (\text{divergence theorem}) \\ &= \langle \Sigma S, \int \nabla^2 \log \frac{g}{\pi} dg \rangle. \end{aligned}$$

We therefore found that

$$\langle \bar{a}, a \rangle + \langle \bar{S}, \Sigma S \rangle = \langle \int \nabla \log \frac{\mu}{\pi}(x) dg(x), a \rangle + \langle \Sigma S, \int \nabla^2 \log \frac{g}{\pi} dg \rangle,$$

and by identification:

$$\begin{cases} \bar{a} = \underbrace{\int \nabla g}_{=\nabla \int g=0} + \int \nabla V dg = \mathbb{E}_g[\nabla V(X)] \\ \bar{S} = \mathbb{E}_g[\nabla^2 V(X)] - \Sigma^{-1}. \end{cases}$$

BW gradient flow. The dynamics are:

$$\begin{cases} \dot{m}_t &= -\mathbb{E}_{g_t}[\nabla V(X_t)] \\ \dot{\Sigma}_t &= -\mathbb{E}_{g_t}[\nabla^2 V(X_t)]\Sigma_t + I_d - \Sigma_t \mathbb{E}_{g_t}[\nabla^2 V(X_t)] + I_d \\ &= 2I_d - (\mathbb{E}_{g_t}[\nabla^2 V(X_t)]\Sigma_t + \Sigma_t \mathbb{E}_{g_t}[\nabla^2 V(X_t)]). \end{cases}$$

Let's work a bit on $\nabla^2 V(X_t)$:

$$\begin{aligned} \int \nabla^2 V dg_t &= - \int \langle \nabla V, \frac{\nabla g_t}{g_t} \rangle dg_t && \text{(divergence theorem)} \\ &= - \int \nabla V \otimes \log g_t dg_t \\ &= \int \nabla V \otimes (\Sigma_t^{-1}(\cdot - m_t)) dg_t \\ \left(\int \nabla^2 V dg_t \right) \Sigma_t &= \int \nabla V \otimes (\cdot - m_t) dg_t \\ \Sigma_t \left(\int \nabla^2 V dg_t \right) &= \int (\cdot - m_t) \otimes \nabla V dg_t. \end{aligned}$$

Hence:

$$\begin{cases} \dot{m}_t = -\mathbb{E}_{g_t}[\nabla V(X_t)] \\ \dot{\Sigma}_t = 2I_d - \mathbb{E}_{g_t}[\nabla V(X_t) \otimes (X_t - m_t) + (X_t - m_t) \otimes \nabla V(X_t)]. \end{cases} \quad (\text{Särkkä})$$

This is Särkkä's heuristic [Sär07]. This is what happens when you constraint the Wasserstein gradient flow to stay in the space of Gaussians.

Rates of convergence. The PL inequality that we had before assumed that the minimum of the KL was 0, which is not the cas anymore here (since we are on the Gaussians only). For V α -strongly convex, we would like to have something like

$$\text{KL}(g_t \parallel \pi) - \min_{g \text{ Gaussian}} \text{KL}(g \parallel \pi) \leq Ce^{-2\alpha t},$$

in order to have that for $g^* = \arg \min_{g \text{ Gaussian}} \text{KL}(g \parallel \pi)$,

$$W_2^2(g_t, g^*) \leq e^{-2\alpha t} W_2^2(g_0, g^*).$$

In order to obtain this, we can for instance integrated the coupled ODE (Särkkä) using Runge-Kutta.

Remark 5.12. For mixtures of Gaussians, we can equip the space of measures with a measure $\mu \in \mathcal{P}_{2,\text{ac}}(\text{BW}(\mathbb{R}^d))$

$$p_\mu(x) = \int d\mu(\theta) p_\theta(x), \quad \text{where } \theta = (m, \Sigma),$$

and everything works (with a different form of divergence), except that we do not have any convergence guarantee.

Remark 5.13. The fact that we stay with the same number of points is annoying. Wasserstein-Fisher-Rao gradient flow allows to play with the weights:

$$\mu_t = \sum_{i=1}^n w_t^{(i)} \delta_{(m_t^{(i)}, \Sigma_t^{(i)})}.$$

See [Lam+22, Section H] for more information.

REFERENCES

- [AG13] Luigi Ambrosio and Nicola Gigli. “A user’s guide to optimal transport”. In: *Modelling and optimisation of flows on networks*. Springer, 2013, pp. 1–155 (cit. on p. 10).
- [ARC16] Pierre Alquier, James Ridgway, and Nicolas Chopin. “On the properties of variational approximations of Gibbs posteriors”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 8374–8414 (cit. on pp. 25, 26).
- [BBV04] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004 (cit. on p. 20).
- [BGL+14] Dominique Bakry, Ivan Gentil, Michel Ledoux, et al. *Analysis and geometry of Markov diffusion operators*. Vol. 103. Springer, 2014 (cit. on pp. 4, 10).
- [Bou22] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. To appear with Cambridge University Press. 2022. URL: <http://www.nicolasboumal.net/book> (cit. on p. 18).
- [Che+20] Sinho Chewi, Thibaut Le Gouic, Chen Lu, Tyler Maunu, Philippe Rigollet, and Austin Stromme. “Exponential ergodicity of mirror-Langevin diffusions”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 19573–19585 (cit. on p. 9).
- [DNS19] Andrew Duncan, Nikolas Nüsken, and Lukasz Szpruch. “On the geometry of Stein variational gradient descent”. In: *arXiv preprint arXiv:1912.00894* (2019) (cit. on p. 23).
- [Do 92] Manfredo Perdigao Do Carmo. *Riemannian geometry*. Vol. 6. Springer, 1992 (cit. on p. 16).
- [Gig12] Nicola Gigli. *Second Order Analysis on $(\mathcal{P}_2(M), W_2)$* . American Mathematical Soc., 2012 (cit. on p. 20).
- [JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. “The variational formulation of the Fokker–Planck equation”. In: *SIAM journal on mathematical analysis* 29.1 (1998), pp. 1–17 (cit. on pp. 1, 19).
- [KNS16] Hamed Karimi, Julie Nutini, and Mark Schmidt. “Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition”. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2016, pp. 795–811 (cit. on pp. 3, 21).
- [Kor+20] Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. “A non-asymptotic analysis for Stein variational gradient descent”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 4672–4682 (cit. on p. 24).
- [Lam+22] Marc Lambert, Sinho Chewi, Francis Bach, Silvère Bonnabel, and Philippe Rigollet. “Variational inference via Wasserstein gradient flows”. In: *arXiv preprint arXiv:2205.15902* (2022) (cit. on pp. 24, 27).
- [LW16] Qiang Liu and Dilin Wang. “Stein variational gradient descent: A general purpose bayesian inference algorithm”. In: *Advances in neural information processing systems* 29 (2016) (cit. on p. 22).
- [OV00] Felix Otto and Cédric Villani. “Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality”. In: *Journal of Functional Analysis* 173.2 (2000), pp. 361–400 (cit. on p. 10).
- [Pet06] Peter Petersen. *Riemannian geometry*. Vol. 171. Springer, 2006 (cit. on p. 16).
- [Roc70] R Tyrrell Rockafellar. *Convex analysis*. Vol. 18. Princeton university press, 1970 (cit. on p. 12).
- [San15] Filippo Santambrogio. “Optimal transport for applied mathematicians”. In: *Birkhäuser, NY* 55.58-63 (2015), p. 94 (cit. on p. 10).
- [Sär07] Simo Särkkä. “On unscented Kalman filtering for state estimation of continuous-time nonlinear systems”. In: *IEEE Transactions on automatic control* 52.9 (2007), pp. 1631–1641 (cit. on p. 27).
- [Sch16] Frederic Schuller. *Lectures on Geometrical Anatomy of Theoretical Physics*. 2016. URL: https://youtube.com/playlist?list=PLPH7f_7ZlzxTi6kS4vCmv4ZKm9u8g5yic (cit. on p. 16).

- [Van16] Ramon Van Handel. “Probability in high dimension apc 550 lecture notes”. In: *Princeton University* (2016) (cit. on pp. 4, 9).
- [Vil03] C Villani. “Topics in optimal transportation: American Mathematical Society”. In: *Graduate studies in mathematics* 58 (2003) (cit. on p. 10).
- [Vil09] Cédric Villani. *Optimal transport: old and new*. Vol. 338. Springer, 2009 (cit. on p. 10).