Wasserstein gradient flows and applications to sampling and machine learning

Anna Korba

CREST, ENSAE, Institut Polytechnique de Paris

Outline

Self-introduction

- Graduated from ENSAE (Stats school) and ENS Cachan (MVA) in 2015
- PhD in Machine Learning at Telecom Paris Tech (2015-2018)
- Postdoc at University College London, Gatbsy Computational Neuroscience Unit (2018-2020)
- Assistant Prof at ENSAE since 2020
- Active research in ML community, with a theoretical flavor. Attend & publish regularly in NeurIPS (Advances in Neural Information Processing Systems) & ICML (International Conference of Machine Learning), Aistats (Artificial Intelligence and Statistics), ICLR (International Conference of Learning Representations).

About this course

We view the Sampling problem as an Optimization problem over the space of probability distributions.

Objective

- Leverage the powerful geometry of optimal transport on the space of probability distributions and in particular Wasserstein gradient flows
- Exploit the analogy between Euclidean gradient flows and Wasserstein gradient flows to design and analyze sampling algorithms
- Typically, this is done by discretizing PDEs and lead to the simulation of interacting particle systems.

Structure of this course

1. Motivation for Sampling, Sampling as Optimization and high-level presentation of the ideas
2. Review of Euclidean Gradient Flows (GF) on $\mathbb{R}^d$ and their properties, rates of convergence for discretized GF (=optimization algorithms)
3. Introduction of Wasserstein Gradient Flows (WGF) and analogies with $\mathbb{R}^d$
4. Illustrations with sampling algorithms as discretizations of Wasserstein GF: rates on Langevin Monte Carlo and Stein Variational Gradient Descent, quick tour of closely related algorithms.
5. (depending on time) WGF of other functionals, other applications, e.g. optimization of shallow neural networks and/or generative modelling, gradient flows with moving targets , Wasserstein-Fisher-Rao.

What is sampling?

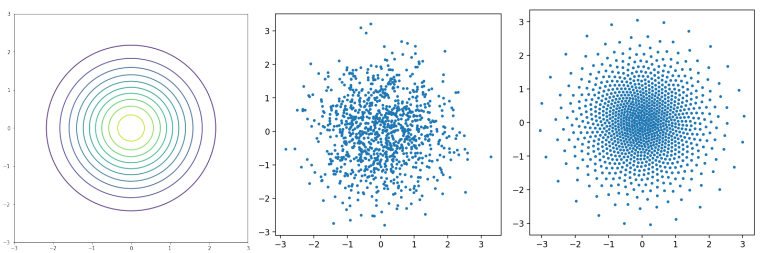
Suppose you are interested in approximating some target probability distribution on $\mathbb{R}^d$, denoted π, and you have access only to partial information on it, e.g.:

1. its unnormalized density (as in Bayesian inference)
2. a discrete approximation $\frac{1}{m}\sum_{k=1}^{m}\delta_{x_i} \approx \pi$ (e.g. i.i.d. samples, iterates of MCMC algorithms...)

Example: approximate $\pi \in \mathcal{P}(\mathbb{R}^d)$ by a finite set of n points $x_1, \dots, x_n$, e.g. to compute functionals $\int_{\mathbb{R}^d} f(x)d\pi(x)$.

The quality of the set can be measured by the integral error:

$$\left| \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int_{\mathbb{R}^d} f(x)d\pi(x) \right|.$$



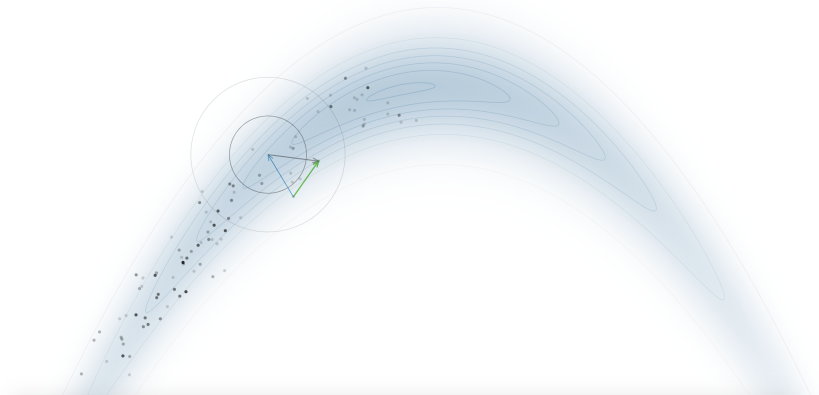
(a) a Gaussian density (b) i.i.d. samples. (c) Particle scheme (SVGD).

(Some, Non parametric) Sampling methods

(1) Markov Chain Monte Carlo (MCMC) methods: generate a Markov chain in $\mathbb{R}^d$ whose law converges to $\pi \propto \exp(-V)$

Example: Langevin Monte Carlo (LMC) [Roberts and Tweedie (1996)]

$$x_{m+1} = x_m - \gamma \nabla V(x_m) + \sqrt{2\gamma}\eta_m, \quad \eta_m \sim \mathcal{N}(0, \mathrm{Id}).$$



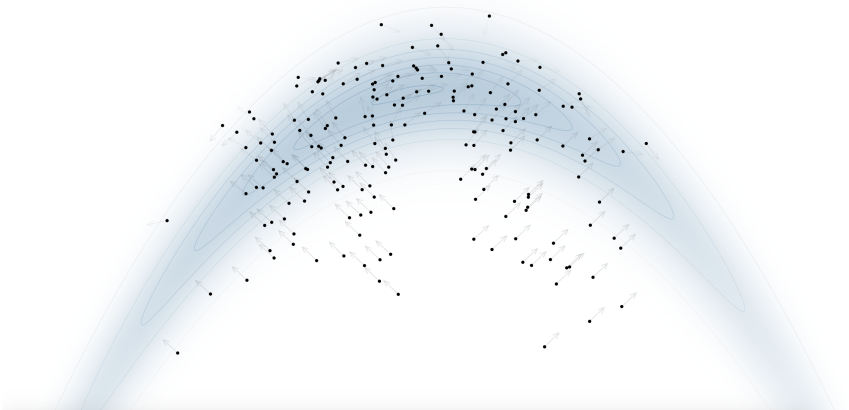
Picture from `https://chi-feng.github.io/mcmc-demo/app.html`.

(2) Interacting particle systems, whose empirical measure at stationarity approximates $\pi \propto \exp(-V)$

Example: Stein Variational Gradient Descent (SVGD) [Liu and Wang (2016)]

$$x_{m+1}^i = x_m^i - \frac{\gamma}{N} \sum_{j=1}^{N} \nabla V(x_m^j) k(x_m^i, x_m^j) - \nabla_2 k(x_m^i, x_m^j), \quad i = 1, \dots, N.$$

where $k(x, y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a smooth kernel (e.g. $k(x, y) = \exp\left(-\|x - y\|^2\right)$).



Picture from `https://chi-feng.github.io/mcmc-demo/app.html`.

Sampling as optimization over $\mathcal{P}_2(\mathbb{R}^d)$

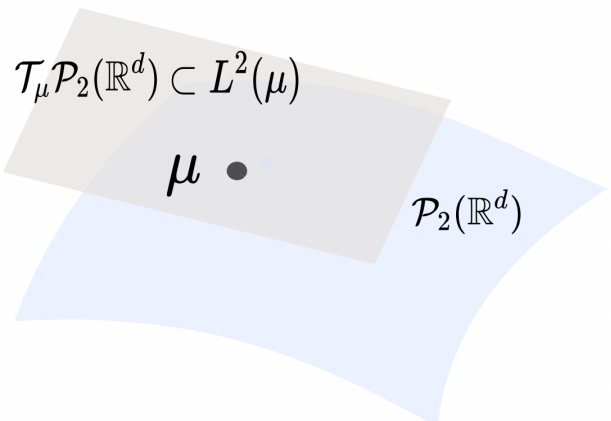
Assume $\pi \in \mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d), \int_{\mathbb{R}^d} \|x\|^2 d\mu(x) < \infty\}$.

Sampling can be recast as optimization over $\mathcal{P}_2(\mathbb{R}^d)$:

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu), \quad \mathcal{F}(\mu) := \mathrm{KL}(\mu|\pi),$$

where $\mathrm{KL}(\mu|\pi) = \int \log(\mu/\pi) d\mu$ if $\mu \ll \pi$.

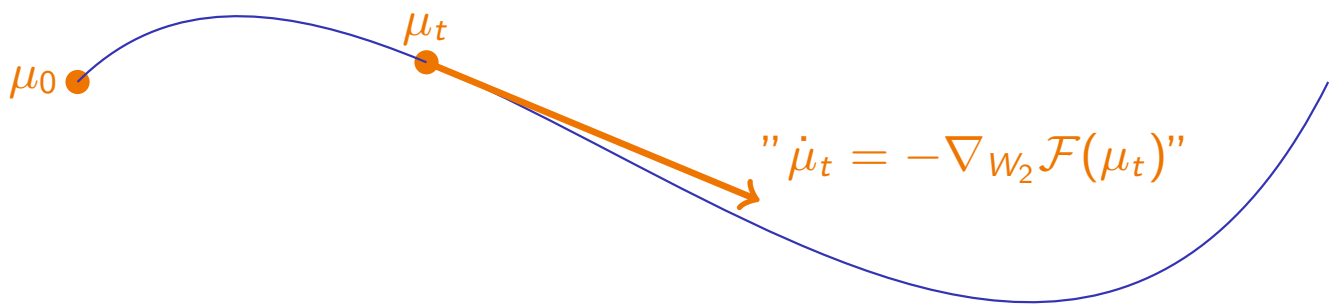
Equipped with the Wasserstein-2 (W_2) distance from optimal transport[1], the metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ has a convenient **Riemannian structure** [Otto (2001)].



[1] $W_2^2(\mu, \nu) = \inf_{s \text{ coupling of } \mu, \nu} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, ds(x, y)$.

Starting from some μ_0, one can then consider the **Wasserstein gradient flow** of $\mathcal{F} = \mathrm{KL}(\cdot|\pi)$ over $\mathcal{P}_2(\mathbb{R}^d)$, **i.e. path of distributions** $(\mu_t)_{t\geq 0}$ **decreasing** $\mathcal{F}$, to transport μ_0 to π.

We will see that these paths $(\mu_t)_{t\geq 0}$ obey PDE (Partial Differential Equations)



which themselves rule the dynamics of particles $(x_t)_{t\geq 0}$ in $\mathbb{R}^d$

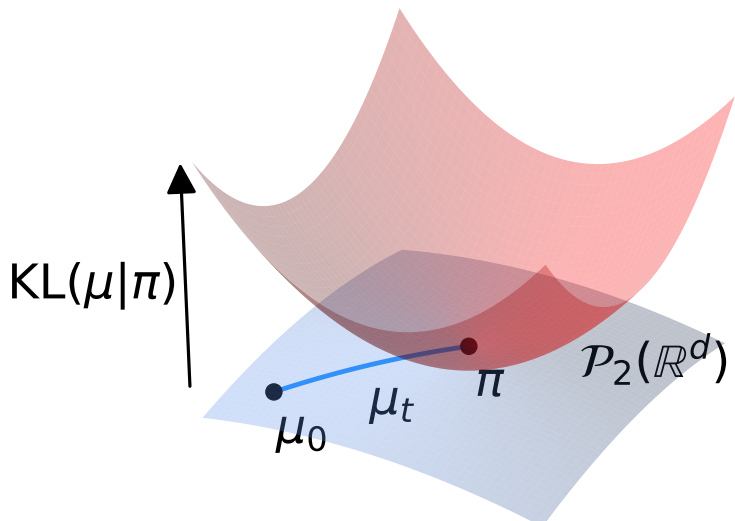
$$dx_t = v(x_t, \mu_t)dt + \sigma(x_t, \mu_t)db_t, \quad x_t \sim \mu_t, \quad (b_t)_{t\geq 0} \text{ Brownian motion.}$$

Discretizing these dynamics $(x_t)_{t\geq 0}$ **yields sampling algorithms.**

Recall that $\pi(x) \propto \exp(-V(x))$.

We will see that in the Wasserstein geometry, the $\mathrm{KL}(\cdot|\pi)$ objective inherits convexity properties of V, i.e.:

- if V is convex, π is "log-concave" and "sampling is easy", i.e. the law of particles will converge in practice to π

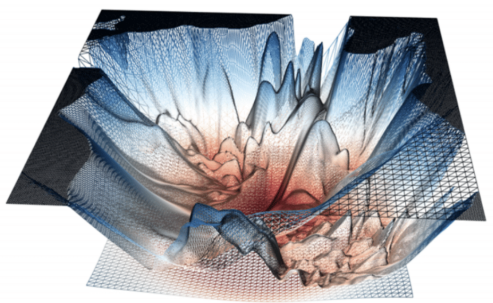


When π is log-concave, $\mathrm{KL}(\cdot|\pi) : \mathcal{P}_2(\mathbb{R}^d) \to \bar{\mathbb{R}}^+$ is (geodesically) convex as represented here.

Recall that $\pi(x) \propto \exp(-V(x))$.

We will see that in the Wasserstein geometry, the $\mathrm{KL}(\cdot|\pi)$ objective inherits convexity properties of V, i.e.:

- if V is nonconvex, π is "non log-concave" and "sampling is hard", i.e. the law of particles may not converge to π



A highly nonconvex loss surface, as is common in deep neural nets. From `https://www.telesens.co/2019/01/16/neural-network-loss-visualization`.

WGFs/Sampling as optimization: context in the stats/ML field

Since the seminal paper of [Jordan et al. (1998)], it is known that the distributions $(\mu_t)_{t\geq 0}$ of Langevin dynamics in $\mathbb{R}^d$

$$dx_t = -\nabla V(x_t)dt + \sqrt{2}db_t,$$

where $(b_t)_{t\geq 0}$ is the Brownian motion in $\mathbb{R}^d$, follow a Wasserstein Gradient Flow (WGF) of the Kullback-Leibler divergence.

This optimization point of view has been used to derive rates of convergence for variants of the Langevin Monte Carlo algorithm (which can be seen as "noisy gradient descent") [Wibisono (2018); Durmus et al. (2019); Bernton (2018)]. Also SVGD was introduced same time [Liu and Wang (2016)].

At the same time, [Chizat and Bach (2018); Mei et al. (2018); Rotskoff and Vanden-Eijnden (2018)] studied the optimization of shallow neural networks through a similar point of view, seeing gradient descent on the weights as an interacting particle system (optimizing a different functional).

Nowadays, WGFs have started to be more part of the folklore and have also been used to model and study generative modeling [Yi et al. (2023); Franceschi et al. (2023)].

Outline

Motivation for Sampling (1): Bayesian inference

Goal of Bayesian inference: learn the best distribution over a parameter x to fit observed data.

Motivation for Sampling (1): Bayesian inference

Goal of Bayesian inference: learn the best distribution over a parameter x to fit observed data.

(1) Let $\mathcal{D} = (w_i, y_i)_{i=1}^p$ a dataset of i.i.d. examples with features w, label y.

(2) Assume an underlying model parametrized by $x \in \mathbb{R}^d$, e.g.:

$$y = g(w, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{Id}).$$

Motivation for Sampling (1): Bayesian inference

Goal of Bayesian inference: learn the best distribution over a parameter x to fit observed data.

(1) Let $\mathcal{D} = (w_i, y_i)_{i=1}^p$ a dataset of i.i.d. examples with features w, label y.

(2) Assume an underlying model parametrized by $x \in \mathbb{R}^d$, e.g.:

$$y = g(w, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathrm{Id}).$$

Step 1. Compute the Likelihood:

$$p(\mathcal{D}|x) \overset{(1)}{\propto} \prod_{i=1}^{p} p(y_i|x, w_i) \overset{(2)}{\propto} \exp\left(-\frac{1}{2} \sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2 \right).$$

Step 2. Choose a prior distribution (initial guess) on the parameter:

$$x \sim p_0, \quad \text{e.g. } p_0(x) \propto \exp\left(-\frac{\|x\|^2}{2}\right).$$

Step 2. Choose a prior distribution (initial guess) on the parameter:

$$x \sim p_0, \quad \text{e.g. } p_0(x) \propto \exp\left(-\frac{\|x\|^2}{2}\right).$$

Step 3. Bayes' rule yields the formula for the posterior distribution over the parameter x:

$$p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p_0(x)}{Z} \quad \text{where} \quad Z = \int_{\mathbb{R}^d} p(\mathcal{D}|x)p_0(x)dx$$

is called the normalization constant and is **intractable**.

Step 2. Choose a prior distribution (initial guess) on the parameter:

$$x \sim p_0, \quad \text{e.g. } p_0(x) \propto \exp\left(-\frac{\|x\|^2}{2}\right).$$

Step 3. Bayes' rule yields the formula for the posterior distribution over the parameter x:

$$p(x|\mathcal{D}) = \frac{p(\mathcal{D}|x)p_0(x)}{Z} \quad \text{where} \quad Z = \int_{\mathbb{R}^d} p(\mathcal{D}|x)p_0(x)dx$$

is called the normalization constant and is **intractable**.

Denoting $\pi := p(\cdot|\mathcal{D})$ the posterior on parameters $x \in \mathbb{R}^d$, we have:

$$\pi(x) \propto \exp\left(-V(x)\right), \quad V(x) = \frac{1}{2}\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2 + \frac{\|x\|^2}{2}.$$

i.e. π's density is known "up to a normalization constant".
π is a probability distribution over parameters of a model.

The posterior π is interesting for

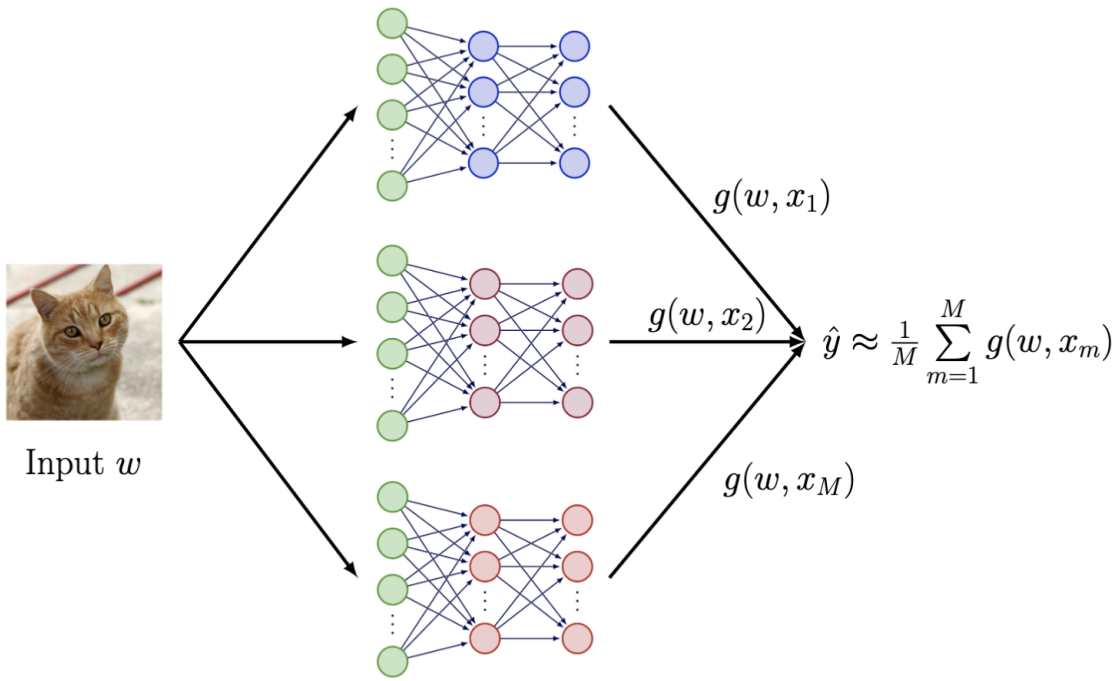
- measuring uncertainty on prediction, for a given input w, through the distribution of $g(w, \cdot)$, $x \sim \pi$.
- we can also output a pointwise prediction for a given input w:

$$\hat{y} = \underbrace{\int_{\mathbb{R}^d} g(w, x) d\pi(x)}_{\text{"Bayesian model averaging"}}$$

  i.e. predictions of models parametrized by $x \in \mathbb{R}^d$ are reweighted by $\pi(x)$.

Here, Sampling methods construct an approximation $\mu_M = \frac{1}{M}\sum_{m=1}^{M} \delta_{x_m}$ of π

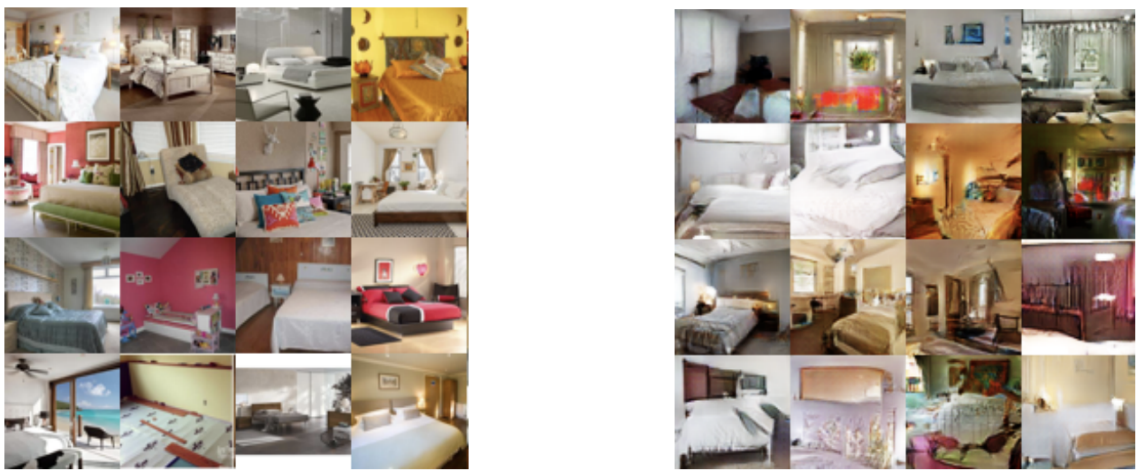
(where $x_1, \ldots, x_m$ are not necessarily samples from π).



Motivation for Sampling (2): Generative modeling

In this setting, we have a collection of samples (data) $x_1, \dots, x_n \sim \pi$.

Goal of Generative Modeling: generate new samples that look like π.



LSUN bedroom samples vs MMD GAN [Li et al. (2017)]. In that context, we do not have any information on the density of π (whereas in Bayesian inference, we knew $\pi \propto e^{-V}$). We only have samples.

The Sampling literature

Two different settings:

- **(1)** the "Bayesian inference" one, where $\pi \propto e^{-V}$
- **(2)** the "Generative Modeling" one, where $x_1, \dots, x_n \sim \pi$

For **(1)**, you may have heard of: Importance Sampling, MCMC algorithms ...

For **(2)**, you may have heard of: Generative Adversarial Networks, Normalizing Flows, Diffusion Models...

There is no clear winner algorithm on the quality of approximation vs computational complexity. Also, these methods are nowadays sometimes used jointly.

We are going to introduce and study these through a common framework: optimization over probability distributions (Wasserstein gradient flows are coming) of some **objective functional**:

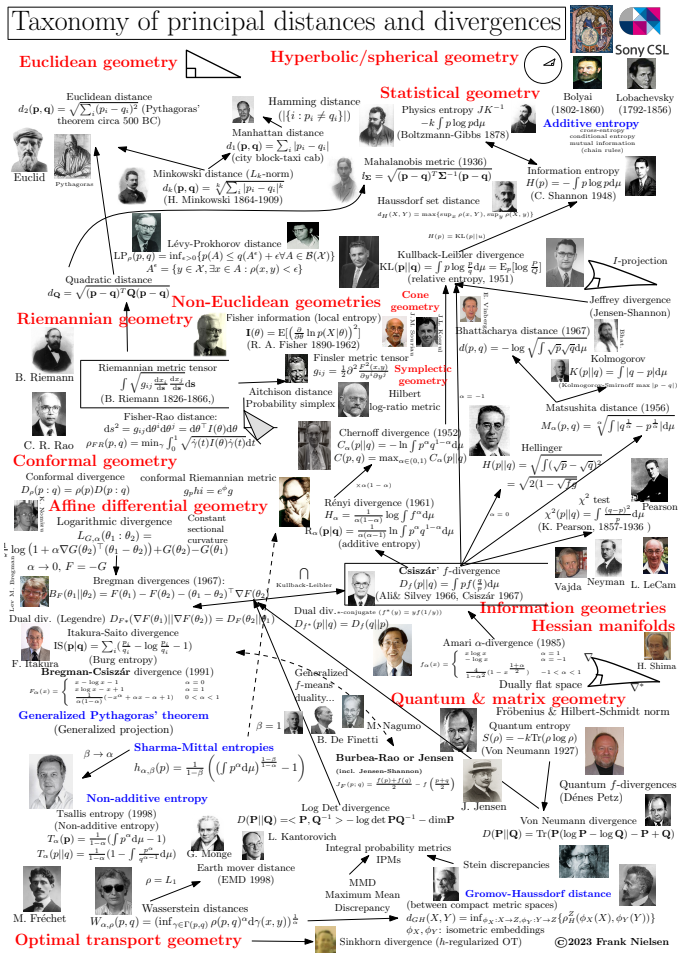
$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{D}(\mu|\pi)$$

where D is a divergence or distance, hence that is minimized for $\mu = \pi$.

Outline

Assume that we have an algorithm that outputs a candidate probability distribution μ, we want to know how close it is from π.

One way is to pick a distance or divergence between probability distributions.



From https://franknielsen.github.io/.

Main families of divergences and distances

Let $\mathcal{P}(\mathbb{R}^d)$ denote the space of probability distributions over $\mathbb{R}^d$.

We will pick D a divergence, i.e. s.t. $\mathrm{D}(\mu||\pi) \geq 0$ for any $\mu \in \mathcal{P}(\mathbb{R}^d)$, $\mathrm{D}(\mu||\pi) \Leftrightarrow \mu = \pi$; or a distance (i.e. satisfies triangle inequality).

Main families of divergences and distances are:

- f-divergences:

$$\int f\left(\frac{\mu}{\pi}\right) d\pi, \quad f \text{ convex, } f(1) = 0$$

  defined for $\mu \ll \pi$ (μ absolutely continuous w.r.t. π)

- integral probability metrics (IPM):

$$\sup_{f \in \mathcal{G}} \left| \int f d\mu - \int f d\pi \right|$$

  for $\mathcal{G}$ a class of functions "rich enough"

- optimal transport (OT) distances

Example (1): the Kullback-Leibler divergence

D could be the (reverse) Kullback-Leibler (KL) divergence:

$$\mathrm{KL}(\mu|\pi) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) & \text{if } \mu \ll \pi \\ +\infty & \text{otherwise.} \end{cases}$$

We recognize a f-divergence $\int f\left(\frac{\mu}{\pi}\right) d\pi$ where $f(x) = x\log(x)$. Taking $f(x) = -\log(x)$ yields the (forward) KL i.e. $\mathrm{KL}(\pi|\mu)$.

Example (1): the Kullback-Leibler divergence

D could be the (reverse) Kullback-Leibler (KL) divergence:

$$\mathrm{KL}(\mu|\pi) = \begin{cases} \int_{\mathbb{R}^d} \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) & \text{if } \mu \ll \pi \\ +\infty & \text{otherwise.} \end{cases}$$

We recognize a f-divergence $\int f\left(\frac{\mu}{\pi}\right) d\pi$ where $f(x) = x\log(x)$. Taking $f(x) = -\log(x)$ yields the (forward) KL i.e. $\mathrm{KL}(\pi|\mu)$.

The (reverse) KL as an objective is convenient when the unnormalized density of π is known since it **does not depend on the normalization constant!**

Indeed writing $\pi(x) = e^{-V(x)}/Z$ we have:

$$\mathrm{KL}(\mu|\pi) = \int_{\mathbb{R}^d} \log\left(\frac{\mu}{e^{-V}}(x)\right) d\mu(x) + \log(Z).$$

But, it is not convenient when μ or π are discrete, because the KL is $+\infty$ unless $supp(\mu) \subset supp(\pi)$.

Example (2): the Maximum Mean Discrepancy

When we have π (or an approximation) as a discrete measure, it is convenient to choose D as an IPM, i.e. integral probability metric (to approximate integrals).

For instance, D could be the MMD (Maximum Mean Discrepancy):

$$\begin{aligned}\mathrm{MMD}^2(\mu,\pi) &= \sup_{f\in \mathrm{H}_k, \|f\|_{\mathcal{H}_k}\leq 1} \left| \int f d\mu - \int f d\pi \right| \\ &= \|m_\mu - m_\pi\|^2_{\mathrm{H}_k}, \quad \text{where } m_\mu = \int k(x,\cdot) d\mu(x) \\ &= \iint_{\mathbb{R}^d} k(x,y) d\mu(x) d\mu(y) \\ &\quad + \iint_{\mathbb{R}^d} k(x,y) d\pi(x) d\pi(y) - 2 \iint_{\mathbb{R}^d} k(x,y) d\mu(x) d\pi(y).\end{aligned}$$

where $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a p.s.d. kernel (e.g. $k(x,y) = e^{-\|x-y\|^2}$) and H_k is the RKHS associated to k:

$$\mathrm{H}_k = \overline{\left\{ \sum_{i=1}^m \alpha_i k(\cdot, x_i);\ m \in \mathbb{N};\ \alpha_1, \dots, \alpha_m \in \mathbb{R};\ x_1, \dots, x_m \in \mathbb{R}^d \right\}}.$$

Example: Take $k(x,y) = e^{-\frac{\|x-y\|^2}{\sigma}}$, $\mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x^i}$, $\pi = \frac{1}{m}\sum_{j=1}^{m} \delta_{y^j}$.

$$\mathrm{MMD}^2(\mu,\pi) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} k(x^i,x^j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} k(y^i,y^j) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} k(x^i,y^j).$$

Wasserstein-p distances

Let $\mathcal{P}_p(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$ with finite p moments, i.e. $\mathcal{P}_p(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d),\ \int \|x\|^p d\mu(x) < \infty\}$.

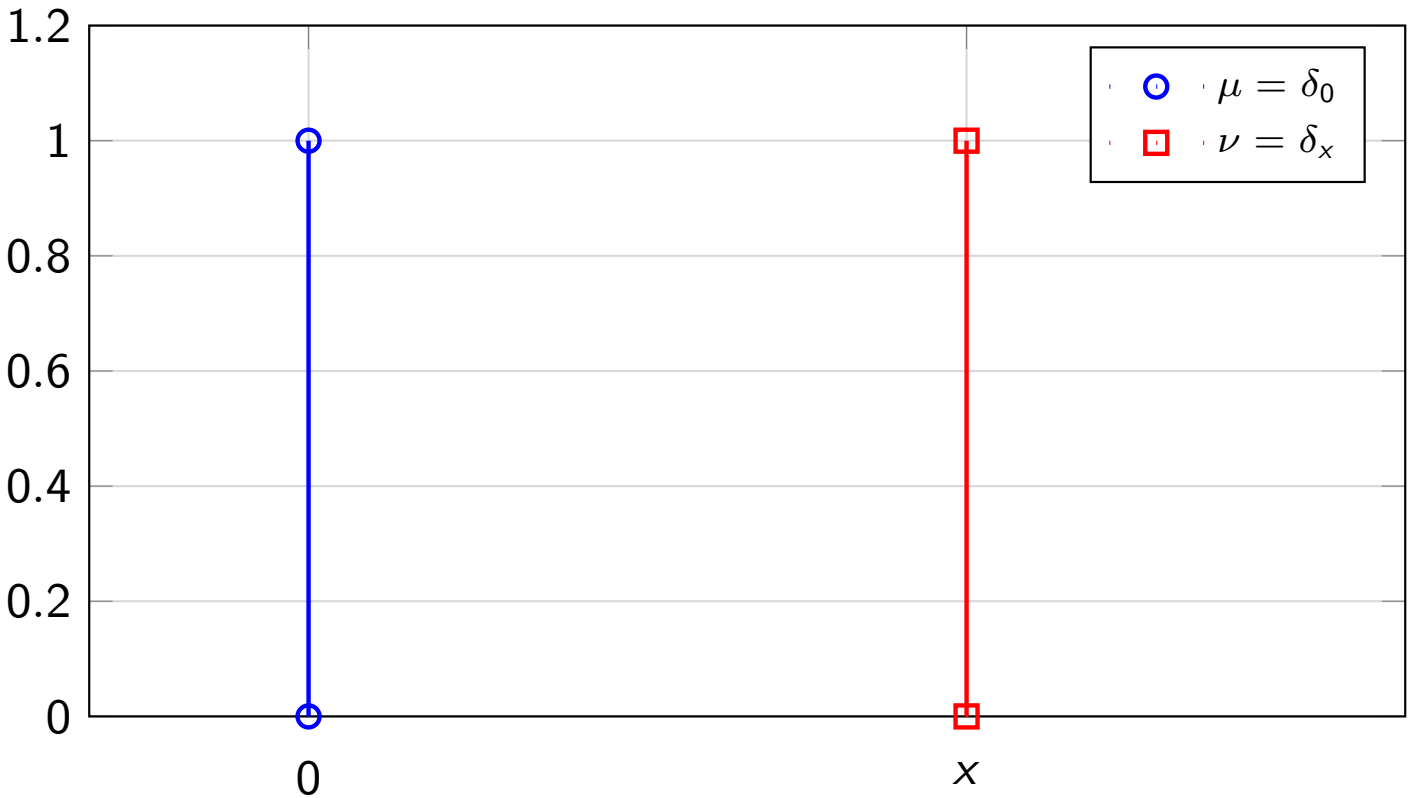
The Wasserstein-p distance from Optimal transport is defined as :

$$\forall \mu, \nu \in \mathcal{P}_p(\mathbb{R}^d), \quad W_p^p(\mu, \nu) = \inf_{s \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^p \, ds(x, y),$$

where $\Gamma(\mu, \nu)$ is the set of possible couplings between μ and ν (probabilities on $\mathbb{R}^d \times \mathbb{R}^d$ with first and second marginal equal to μ and ν). Most popular ones are:

- The W_2 (in many ways analog to an "euclidean distance" but on $\mathcal{P}_2(\mathbb{R}^d)$)
- The W_1, which interestingly can be written as an IPM:

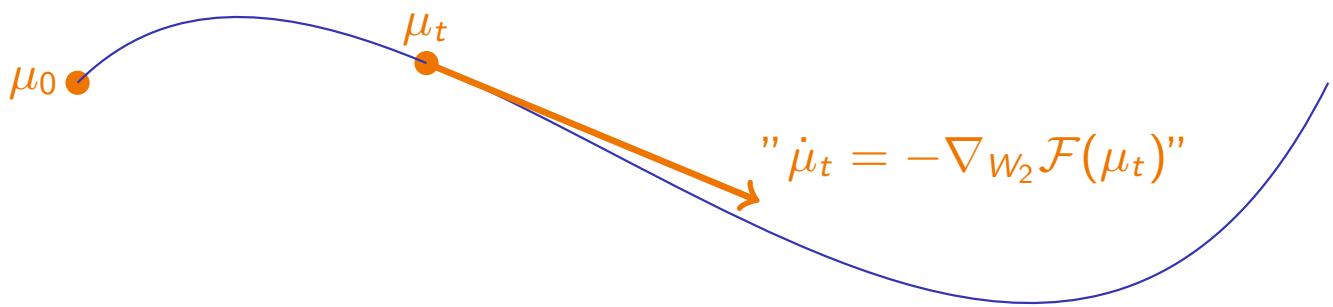
$$W_1(\mu, \nu) = \inf_{f:\mathbb{R}^d \to \mathbb{R}, f \text{ is 1-Lipschitz}} \left| \int f d\mu - \int f d\nu \right|$$



We have $\mathrm{KL}(\mu|\nu) = \mathrm{KL}(\nu|\mu) = +\infty$, and $W_2(\mu, \nu) = |x|$.

Starting from some μ_0, one can then consider the **Wasserstein gradient flow** of $\mathcal{F} = \mathrm{KL}(\cdot|\pi)$ over $\mathcal{P}_2(\mathbb{R}^d)$, **i.e. path of distributions** $(\mu_t)_{t\geq 0}$ **decreasing** $\mathcal{F}$, to transport μ_0 to π.

We will see that these paths $(\mu_t)_{t\geq 0}$ obey PDE (Partial Differential Equations)



which themselves rule the dynamics of particles $(x_t)_{t\geq 0}$ in $\mathbb{R}^d$

$$dx_t = v(x_t, \mu_t)dt + \sigma(x_t, \mu_t)db_t, \quad x_t \sim \mu_t, \quad (b_t)_{t\geq 0} \text{ Brownian motion.}$$

Discretizing these dynamics $(x_t)_{t\geq 0}$ yields sampling algorithms.

How to define the gradient? how to define the evolution of μ_t? then how to discretize this in space and time? how to analyze them? Let's review gradient flows in $\mathbb{R}^d$ first.

Outline

Gradient

Let $V : \mathbb{R}^d \to \mathbb{R}$ differentiable. What is the gradient of V?

Definition: If a Taylor expansion of V yields:

$$V(x + \varepsilon h) = V(x) + \varepsilon \langle g_x, h \rangle + o(\varepsilon),$$

where $\langle \cdot, \cdot \rangle$ is some inner product, then g_x is the gradient of V at x under the inner product $\langle \cdot, \cdot \rangle$.

Gradient

Let $V : \mathbb{R}^d \to \mathbb{R}$ differentiable. What is the gradient of V?

Definition: If a Taylor expansion of V yields:

$$V(x + \varepsilon h) = V(x) + \varepsilon \langle g_x, h \rangle + o(\varepsilon),$$

where $\langle \cdot, \cdot \rangle$ is some inner product, then g_x is the gradient of V at x under the inner product $\langle \cdot, \cdot \rangle$.

- If $\langle \cdot, \cdot \rangle_{\mathbb{R}^d}$ is the Euclidean inner product then $g_x = \nabla V(x)$.
- If $\langle \cdot, \cdot \rangle_P$ is the inner product induced by a positive definite matrix P (i.e. $\langle x, y \rangle_P = \langle Px, y \rangle_{\mathbb{R}^d}$) then $g_x = P^{-1} \nabla V(x)$.

Euclidean Gradient Flow

Problem:

$$\min_{x \in \mathbb{R}^d} V(x),$$

where $V : \mathbb{R}^d \to \mathbb{R}$ s.t. ∇V is L-Lipschitz (V is L-smooth).

Using Cauchy-Lipschitz, consider

$$\dot{x}_t = -\nabla V(x_t), \quad t \geq 0,$$

where we denote $x_t = x(t)$, $\dot{x}_t = \frac{dx_t}{dt}$.

Gradient flow of V = the solution of this Ordinary Differential Equation (ODE) for any initial data $x(0)$.

Descent property of gradient flows

Using (1) the chain rule and (2) $\dot{x}_t = -\nabla V(x_t)$,

$$\frac{dV(x_t)}{dt} \overset{(1)}{=} \langle \dot{x}_t, \nabla V(x_t) \rangle \overset{(2)}{=} -\|\nabla V(x_t)\|^2 \leq 0.$$

The gradient flow decreases the objective function.

This is a fundamental property of the gradient flow [De Giorgi et al. (1980); De Giorgi (1993)].

Particular case: V convex

Let $\lambda \geq 0$. V is λ-strongly convex if
$\forall x, y \in \mathbb{R}^d, t \in [0,1]$,

$$V((1-t)x + ty) \leq (1-t)V(x) + tV(y) - \frac{\lambda t(1-t)}{2}\|x-y\|^2.$$

0-strong convexity is simply convexity.

Particular case: V convex

Let $\lambda \geq 0$. V is λ-strongly convex if
$\forall x, y \in \mathbb{R}^d, t \in [0,1]$,

$$V((1-t)x + ty) \leq (1-t)V(x) + tV(y) - \frac{\lambda t(1-t)}{2}\|x-y\|^2.$$

0-strong convexity is simply convexity.
Since V smooth, this is equivalent to

$$\forall y \in \mathbb{R}^d, V(x) + \langle \nabla V(x), y - x\rangle + \frac{\lambda}{2}\|y-x\|^2 \leq V(y).$$

Evolution Variational Inequality (EVI)

Assume V is λ-strongly convex. Then, the gradient flow satisfies the following variational inequality: for every $y \in \mathbb{R}^d$,

$$\frac{d}{dt}\|x_t - y\|^2 \leq -2(V(x_t) - V(y)) - \lambda\|x_t - y\|^2.$$

Evolution Variational Inequality (EVI)

Assume V is λ-strongly convex. Then, the gradient flow satisfies the following variational inequality: for every $y \in \mathbb{R}^d$,

$$\frac{d}{dt}\|x_t - y\|^2 \leq -2(V(x_t) - V(y)) - \lambda\|x_t - y\|^2.$$

Proof: Using the chain rule and convexity,

$$\begin{aligned}\frac{d}{dt}\|x_t - y\|^2 &= 2\langle \dot{x}_t, x_t - y\rangle \\ &= -2\langle \nabla V(x_t), x_t - y\rangle \\ &\leq -2(V(x_t) - V(y)) - \lambda\|x_t - y\|^2.\end{aligned}$$

Taking $y = x^*$ the global optimum of V, using Gronwall lemma, we obtain exponential (linear) convergence of $\|x_t - x^*\|^2$.

Outline

Time discretizations of the gradient flow

Let $\gamma > 0$ a step-size.

- Gradient descent algorithm:

$$x_{m+1} = x_m - \gamma \nabla V(x_m),$$

  i.e. **Forward** Euler (**explicit**):

$$\frac{x_{m+1} - x_m}{\gamma} = -\nabla V(x_m).$$

- Proximal point algorithm:

$$x_{m+1} = \mathrm{prox}_{\gamma V}(x_m) := \operatorname*{arg\,min}_{y \in \mathbb{R}^d} \gamma V(y) + \frac{1}{2}\|x_m - y\|^2$$

  i.e. **Backward** Euler (**implicit**):

$$\frac{x_{m+1} - x_m}{\gamma} = -\nabla V(x_{m+1}).$$

These time discretizations are unbiased (i.e. they preserve $x_\star \in \arg\min V$ as a fixed point).

Other time discretizations: splitting schemes

- Proximal gradient algorithm ($V = F + G$, G convex):

$$x_{m+\frac{1}{2}} = x_m - \gamma \nabla F(x_m)$$
$$x_{m+1} = \mathrm{prox}_{\gamma G}(x_{m+\frac{1}{2}})$$

  i.e. Forward Backward Euler (explicit implicit):

$$\frac{x_{m+1} - x_m}{\gamma} = -\nabla F(x_m) - \nabla G(x_{m+1}).$$

This time discretizations is also unbiased (i.e. it preserves $x_\star \in \arg\min V$ as a fixed point).

Exercise.

Other time discretizations: splitting schemes

- Proximal gradient algorithm ($V = F + G$, G convex):

$$x_{m+\frac{1}{2}} = x_m - \gamma \nabla F(x_m)$$
$$x_{m+1} = \mathrm{prox}_{\gamma G}(x_{m+\frac{1}{2}})$$

  i.e. Forward Backward Euler (explicit implicit):

$$\frac{x_{m+1} - x_m}{\gamma} = -\nabla F(x_m) - \nabla G(x_{m+1}).$$

This time discretizations is also unbiased (i.e. it preserves $x_\star \in \arg\min V$ as a fixed point).

Exercise.

Time discretization of a flow $\Rightarrow$ Optimization algorithm

Descent lemma

The time discretizations of the gradient flow decrease the objective function:

$$\frac{V(x_{m+1}) - V(x_m)}{\gamma} \leq -\frac{1}{2}\|\nabla V(\hat{x}_m)\|^2.$$

- For Forward Euler (i.e. gradient descent), $\hat{x}_m = x_m$ and $\gamma \leq 1/L$ (we need smoothness of V),
- For Backward Euler $\hat{x}_m = x_{m+1}$ (we don't need smoothness of V)

Exercise.

It is known that gradient descent converges at $1/M$ rate when V is convex, and faster if V is λ-strongly convex. But we can actually ask a bit less than convexity (see next slide).

Example: Convergence rate for averaged gradient

Descent lemma is useful to obtain convergence properties of the time discretizations, viewed as optimization algorithms.

Let $x_\star \in \arg\min V$.

Summing from $m = 0$ to $m = M - 1$ the Descent lemma:

$$\frac{(V(x_M) - V(x_\star)) - (V(x_0) - V(x_\star))}{\gamma} \leq -\frac{1}{2} \sum_{m=0}^{M-1} \|\nabla V(x_m)\|^2,$$

therefore

$$\frac{1}{M} \sum_{m=0}^{M-1} \|\nabla V(x_m)\|^2 \leq \frac{2(V(x_0) - V(x_\star))}{\gamma M}.$$

Nonconvex rates for gradient descent

Generally, nonconvex rates can be obtained using Descent lemma:

1. we first obtain

$$\frac{1}{M}\sum_{m=0}^{M-1}\|\nabla V(x_m)\|^2 \leq \frac{2(V(x_0)-V(x_\star))}{\gamma M}.$$

Nonconvex rates for gradient descent

Generally, nonconvex rates can be obtained using Descent lemma:

1. we first obtain

$$\frac{1}{M}\sum_{m=0}^{M-1}\|\nabla V(x_m)\|^2 \leq \frac{2(V(x_0)-V(x_\star))}{\gamma M}.$$

2. If V satisfies a Gradient dominance condition (a.k.a. Polyak-Łojasiewicz) with λ, i.e.:

$$\forall x \in \mathbb{R}^d, \quad V(x)-V(x_\star) \leq \frac{1}{2\lambda}\|\nabla V(x)\|^2,$$

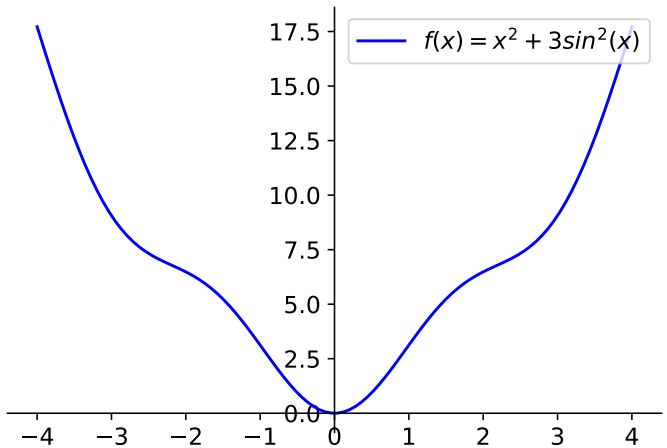
then we can also obtain:

$$V(x_M)-V(x_\star) \leq (1-\gamma\lambda)^M(V(x_0)-V(x_\star)).$$

Gradient dominance is more general than convexity

$$\forall x \in \mathbb{R}^d, \quad V(x) - V_\star \leq \frac{1}{2\lambda} \|\nabla V(x)\|^2.$$

- λ-Strong convexity $\Rightarrow$ gradient dominance with the same constant $\lambda > 0$
- Gradient dominance $\Rightarrow$ invexity[1]
- Gradient dominance $\not\Rightarrow$ convexity



[1] any local minimum of V is a global minimum.

Convex case - Discrete EVI

Assume V λ-strongly convex. Then, the time discretizations of the gradient flow satisfy a discrete variational inequality: for every $y \in \mathbb{R}^d$,

$$\frac{\|x_{m+1} - y\|^2 - \|x_m - y\|^2}{\gamma} \leq -2(V(x_{m+1}) - V(y)) - \lambda \|\hat{x}_m - y\|^2.$$

- For Forward Euler (i.e. gradient descent), $\hat{x}_m = x_m$ and $\gamma \leq 1/L$,
- For Backward Euler $\hat{x}_m = x_{m+1}$.

Example: Convergence rate for averaged iterate with $\lambda = 0$

Summing discrete EVI from $m = 0$ to $m = M - 1$:

$$\frac{\|x_M - y\|^2 - \|x_0 - y\|^2}{\gamma} \leq -2 \sum_{m=1}^{M} (V(x_m) - V(y)),$$

therefore

$$\frac{1}{M} \sum_{m=1}^{M} V(x_m) - V(y) \leq \frac{\|x_0 - y\|^2}{2\gamma M}.$$

Finally, using convexity

$$V(\bar{x}_M) - V(y) \leq \frac{\|x_0 - y\|^2}{2\gamma M},$$

where $\bar{x}_M = \frac{1}{M} \sum_{m=1}^{M} x_m$. We can take $y \in \arg\min V$.

Convex rates for gradient descent

Generally, convex rates can be obtained using discrete EVI + Descent lemma:

1. for $\lambda \geq 0$ we can obtain

$$V(\bar{x}_M) - V(x_\star) \leq \frac{\|x_0 - x_\star\|^2}{2\gamma M}, \text{ where } \bar{x}_M = \frac{1}{M}\sum_{m=1}^{M} x_m$$

$$V(x_M) - V(x_\star) \leq \frac{\|x_0 - x_\star\|^2}{2\gamma M},$$

2. and, if $\lambda > 0$,

$$\|x_M - x_\star\|^2 \leq (1 - \gamma\lambda)^M \|x_0 - x_\star\|^2.$$

Highly recommended reference: [Garrigos and Gower (2023)].

Outline

Definition of the Wasserstein space

Let $\mathcal{P}_2(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$ with finite second moments, i.e.

$$\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\}$$

Definition of the Wasserstein space

Let $\mathcal{P}_2(\mathbb{R}^d)$ the space of probability measures on $\mathbb{R}^d$ with finite second moments, i.e.

$$\mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\}$$

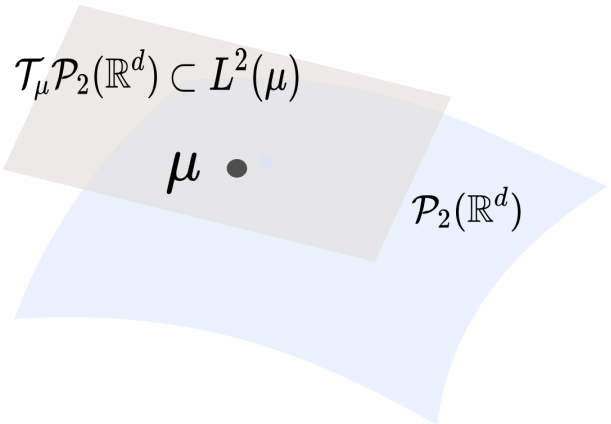
$\mathcal{P}_2(\mathbb{R}^d)$ is endowed with the Wasserstein-2 distance from Optimal transport: $\forall \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W_2^2(\mu, \nu) = \inf_{s \in \Gamma(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, ds(x, y),$$

where $\Gamma(\mu, \nu)$ is the set of possible couplings between μ and ν.

The metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is called **the Wasserstein space**.

Riemannian structure of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ and L^2 spaces



Denote by

$$L^2(\mu) = \{f : \mathbb{R}^d \to \mathbb{R}^d, \int_{\mathbb{R}^d} \|f(x)\|^2 d\mu(x) < \infty\}$$

the space of vector-valued, square-integrable functions w.r.t μ.

It is a Hilbert space of functions equipped with the inner product

$$\langle f, g\rangle_\mu = \int_{\mathbb{R}^d} \langle f(x), g(x)\rangle_{\mathbb{R}^d} d\mu(x).$$

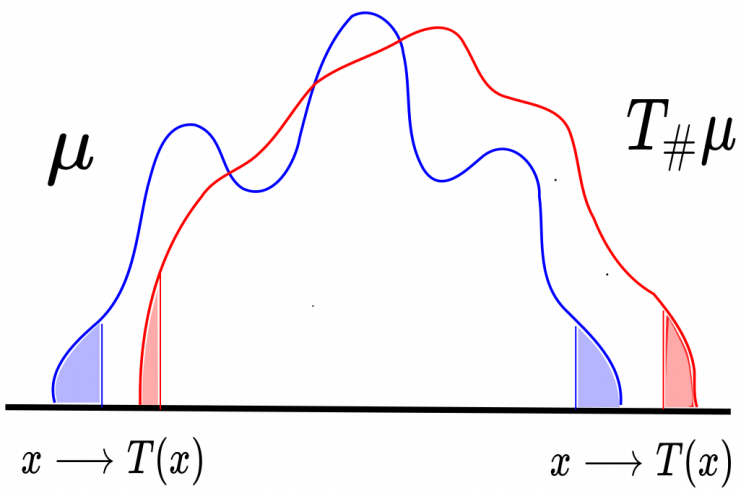
The Riemannian structure of $\mathcal{P}_2(\mathbb{R}^d), W_2)$ was introduced in [Otto (2001)]. Formally, the tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ at μ is defined as $\mathcal{T}_\mu\mathcal{P}_2(\mathbb{R}^d) = \overline{\{\nabla\psi,\ \psi \in \mathcal{C}_c^\infty(\mathbb{R}^d)\}} \subset L^2(\mu)$ [Ambrosio et al. (2008)](Definition 8.4.).

Pushforward measure

Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, $T : \mathbb{R}^d \to \mathbb{R}^d$ a measurable map.

The pushforward measure $T_{\#}\mu$ is characterized by:

$$X \sim \mu \Longrightarrow T(X) \sim T_{\#}\mu.$$



Remark: $\mathrm{Id}_{\#}\,\mu = \mu$ where Id denotes the identity map.

Moving on $\mathcal{P}_2(\mathbb{R}^d)$ through L^2 maps

Note that if $T \in L^2(\mu)$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, then $T_{\#}\mu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\int \|y\|^2 d(T_{\#}\mu)(y) = \int \|T(x)\|^2 d\mu(x) < \infty,$$

since $T \in L^2(\mu)$.

Moving on $\mathcal{P}_2(\mathbb{R}^d)$ through L^2 maps

Note that if $T \in L^2(\mu)$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, then $T_\# \mu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\int \|y\|^2 d(T_\# \mu)(y) = \int \|T(x)\|^2 d\mu(x) < \infty,$$

since $T \in L^2(\mu)$.

Brenier's theorem [Brenier (1991)] : Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ s.t. $\mu \ll$ Leb. Then, there exists a unique $T_\mu^\nu : \mathbb{R}^d \to \mathbb{R}^d$ such that

1. $T_{\mu\#}^\nu \mu = \nu$
2. $W_2^2(\mu, \nu) = \| \mathrm{Id} - T_\mu^\nu \|_\mu^2 \overset{\text{def.}}{=} \int \|x - T_\mu^\nu(x)\|^2 d\mu(x)$.

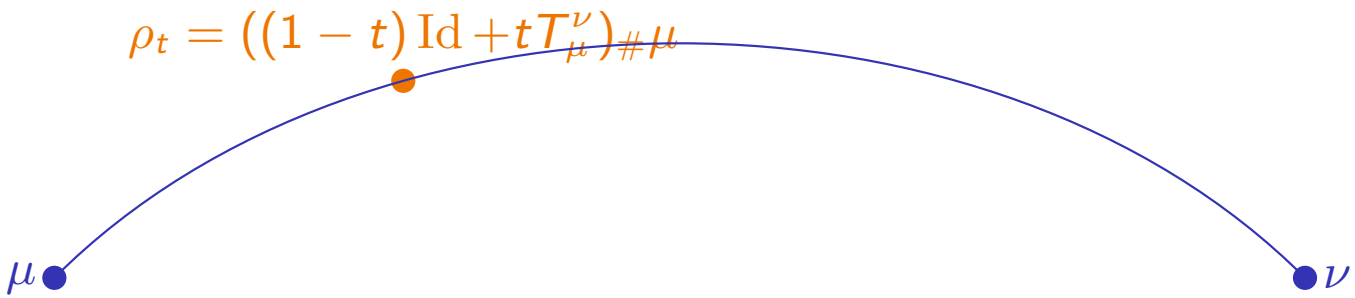
and T_μ^ν is called the Optimal Transport map between μ and ν.

Wasserstein geodesics between μ, ν?

The path

$$\rho_t = ((1-t)\,\mathrm{Id} + tT_\mu^\nu)_\# \mu, \quad t \in [0,1]$$

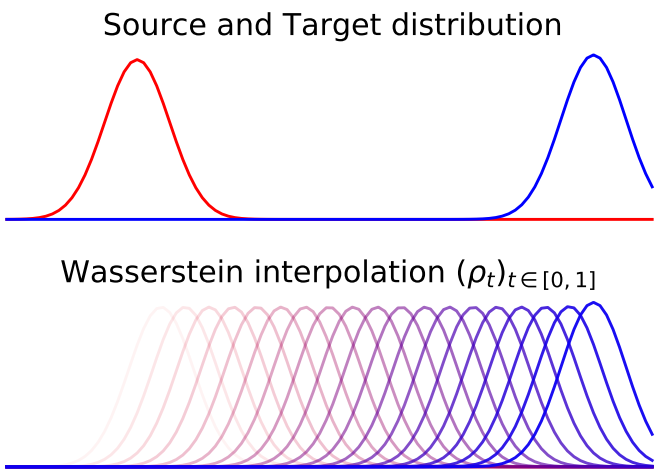
is the Wasserstein geodesic between $\rho_0 = \mu$ and $\rho_1 = \nu$.

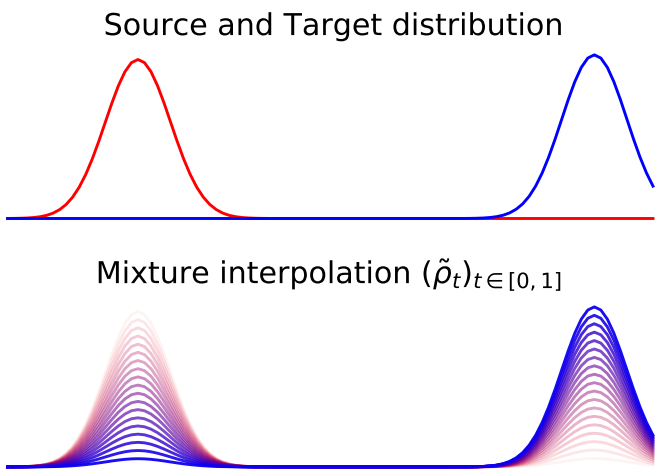


It differs completely from the (mixture) path

$$\tilde{\rho}_t = (1-t)\mu + t\nu$$

which also interpolates between $\tilde{\rho}_0 = \rho_0 = \mu, \tilde{\rho}_1 = \rho_1 = \nu$.





If μ is supported on a set of particles $x^1, \dots, x^N$,
these particles would be **pushed continuously through** ρ_t,
while they would be **teleported to other locations through** $\tilde{\rho}_t$.

Figure made with `https://pythonot.github.io/`.

Convexity along Wasserstein geodesics

Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$.

$\mathcal{F}$ λ-strongly geo. convex with $\lambda \geq 0$, if for any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\mu) + t\mathcal{F}(\nu) - \frac{\lambda t(1-t)}{2} W_2^2(\mu, \nu),$$

where $(\rho_t)_{t\in[0,1]}$ is a Wasserstein-2 geodesic between μ and ν.

Introduced by [McCann (1997)]. If $\lambda > -\infty$, $\mathcal{F}$ is said semi-convex. If $\lambda \geq 0$, $\mathcal{F}$ is said to be displacement (or geodesically) convex.

Examples of geo. convex functionals

1. Potential energy $\mathcal{F}(\mu) = \int V(x) d\mu(x)$ with $V : \mathbb{R}^d \to \mathbb{R}$ convex.

Proof: write $\mathcal{F}(\rho_t)$ along a geodesic $\rho_t = ((1-t)\,\mathrm{Id} + tT_\mu^\nu)_\# \mu$ and use V convex.

Examples of geo. convex functionals

1. Potential energy $\mathcal{F}(\mu) = \int V(x) d\mu(x)$ with $V : \mathbb{R}^d \to \mathbb{R}$ convex.

Proof: write $\mathcal{F}(\rho_t)$ along a geodesic $\rho_t = ((1-t)\,\mathrm{Id} + tT_\mu^\nu)_\# \mu$ and use V convex.

2. Negative entropy (**non trivial**) $\mathcal{F}(\mu) = \int \log(\mu(x)) d\mu(x)$.

Examples of geo. convex functionals

1. Potential energy $\mathcal{F}(\mu) = \int V(x) d\mu(x)$ with $V : \mathbb{R}^d \to \mathbb{R}$ convex.

Proof: write $\mathcal{F}(\rho_t)$ along a geodesic $\rho_t = ((1-t)\,\mathrm{Id} + tT_\mu^\nu)_\# \mu$ and use V convex.

2. Negative entropy (**non trivial**) $\mathcal{F}(\mu) = \int \log(\mu(x)) d\mu(x)$.

3. KL w.r.t. log concave distribution $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$, where $\pi \propto \exp(-V)$, V convex.

Proof:

$$\begin{aligned}\mathrm{KL}(\mu|\pi) &= \int \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) \\ &= \underbrace{\int V(x) d\mu(x)}_{\text{Potential}} + \underbrace{\int \log(\mu(x)) d\mu(x)}_{\text{(Neg.) Entropy}} + C.\end{aligned}$$

Outline

Gradient flows on probability distributions?

Recall that we want to approximate a distribution $\pi \propto e^{-V}$ by solving

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu), \quad \mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi).$$

We have reviewed Euclidean GF of $V : \mathbb{R}^d \to \mathbb{R}$:

$$\dot{x}_t = -\nabla V(x_t), \quad x_t \in \mathbb{R}^d.$$

In an analog manner, what is the gradient flow of $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$? i.e. something of the form

$$"\dot{\mu}_t = -\nabla_{W_2} \mathcal{F}(\mu_t)", \quad \mu_t \in \mathcal{P}_2(\mathbb{R}^d).$$

We need to define both sides of the equality.

LHS: Velocity field

Let $(\mu_t)_{t\geq 0} \in (\mathcal{P}_2(\mathbb{R}^d))^{\mathbb{R}^+}$. What is the time derivative of $(\mu_t)_{t\geq 0}$?

Definition: If there exists $(v_t)_{t\geq 0} \in (L^2(\mu_t))_{t\geq 0}$ such that,

$$\frac{d}{dt}\int \varphi d\mu_t = \langle \nabla\varphi, v_t\rangle_{\mu_t}$$

for every test function $\varphi \in C^\infty(\mathbb{R}^d)$ (smooth functions with compact support), then $(v_t)_{t\geq 0}$ is a velocity field of $(\mu_t)_{t\geq 0}$.

The velocity field rules the dynamics of $(\mu_t)_{t\geq 0}$.

Continuity Equation

Equivalently, a velocity field $(v_t)_{t\geq 0}$ of $(\mu_t)_{t\geq 0}$ satisfies the PDE:

$$\frac{\partial \mu_t}{\partial t} + \boldsymbol{\nabla} \cdot (\mu_t v_t) = 0, \quad t \geq 0.$$

where $\boldsymbol{\nabla} \cdot A(x) = \sum_{i=1}^{d} \frac{\partial A_i(x)}{\partial x_i}$ for $A(x) = (A_1(x), \ldots, A_d(x))$, $A : \mathbb{R}^d \to \mathbb{R}^d$.

Continuity Equation

Equivalently, a velocity field $(v_t)_{t\geq 0}$ of $(\mu_t)_{t\geq 0}$ satisfies the PDE:

$$\frac{\partial \mu_t}{\partial t} + \nabla \cdot (\mu_t v_t) = 0, \quad t \geq 0.$$

where $\nabla \cdot A(x) = \sum_{i=1}^{d} \frac{\partial A_i(x)}{\partial x_i}$ for $A(x) = (A_1(x), \ldots, A_d(x))$, $A : \mathbb{R}^d \to \mathbb{R}^d$.

Proof: If $\mu_t(\cdot)$ density of μ_t, for every test function $\varphi \in C^\infty(\mathbb{R}^d)$,

$$(1): \quad \frac{d}{dt} \int \varphi(x)\mu_t(x)dx = \int \varphi(x) \frac{\partial \mu_t}{\partial t}(x)dx$$

$$(2): \quad \frac{d}{dt} \int \varphi(x)\mu_t(x)dx \overset{\text{def.}}{=} \int \langle \nabla \varphi(x), v_t(x) \rangle_{\mathbb{R}^d} \mu_t(x)dx$$

$$\overset{\text{i.b.p.}}{=} - \int \varphi(x) \nabla \cdot (v_t(x)\mu_t(x))dx.$$

This equation describes the dynamics of $(\mu_t)_{t\geq 0}$.

RHS: Wasserstein gradient

Let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$. What is the "gradient" of $\mathcal{F}$ at μ?

Definition: Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Consider a perturbation on the Wasserstein space $(\mathrm{Id} + \varepsilon h)_{\#}\mu$ for $h \in L^2(\mu)$.

If a Taylor expansion of $\mathcal{F}$ yields:

$$\mathcal{F}((\mathrm{Id} + \varepsilon h)_{\#}\mu) = \mathcal{F}(\mu) + \varepsilon \langle \nabla_{W_2} \mathcal{F}(\mu), h \rangle_{\mu} + o(\varepsilon),$$

then $\nabla_{W_2} \mathcal{F}(\mu) \in L^2(\mu)$ is the Wasserstein gradient of $\mathcal{F}$ at μ.

First Variation

In comparison, what is the First Variation of $\mathcal{F}$ at μ?

Definition: Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Consider a linear perturbation $\mu + \varepsilon\xi$, $\xi = \nu - \mu$ for $\nu \in \mathcal{P}_2(\mathbb{R}^d)$.

If a Taylor expansion of $\mathcal{F}$ yields:

$$\mathcal{F}(\mu + \varepsilon\xi) = \mathcal{F}(\mu) + \varepsilon \int \mathcal{F}'(\mu)(x) d\xi(x) + o(\varepsilon),$$

then $\mathcal{F}'(\mu) : \mathbb{R}^d \to \mathbb{R}$ is the First Variation of $\mathcal{F}$ at μ.

Wasserstein gradient = Gradient of First Variation

Typically[1],

$$\nabla_{W_2}\mathcal{F}(\mu) = \nabla\mathcal{F}'(\mu).$$

$\nabla_{W_2}\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}^d$, $\mathcal{F}'(\mu) : \mathbb{R}^d \to \mathbb{R}$.

[1] see [Ambrosio et al. (2008)] (Th. 10.4.13) for precise statement.

Wasserstein gradient = Gradient of First Variation

Typically[1],

$$\nabla_{W_2}\mathcal{F}(\mu) = \nabla\mathcal{F}'(\mu).$$

$\nabla_{W_2}\mathcal{F}(\mu) : \mathbb{R}^d \to \mathbb{R}^d$, $\mathcal{F}'(\mu) : \mathbb{R}^d \to \mathbb{R}$.

Proof: Let $\mu_\epsilon = (\mathrm{Id} + \epsilon h)_\# \mu$.
First, expand μ_ε around μ using the continuity equation of $(\mu_t)_{t\geq 0}$:

$$\mu_\varepsilon = \mu + \varepsilon(\underbrace{-\nabla\cdot(\mu h)}_{=\xi}) + o(\varepsilon).$$

Then, expand $\mathcal{F}(\mu + \varepsilon\xi)$ using the definition of First Variation, and use an i.b.p. to identify the Wasserstein gradient.

[1] see [Ambrosio et al. (2008)] (Th. 10.4.13) for precise statement.

Examples of Wasserstein gradients

Below: $\mathcal{F}(\mu) \longrightarrow \mathcal{F}'(\mu) \longrightarrow \nabla\mathcal{F}'(\mu)$

1. Potential energy (linear function of μ)

$$\mathcal{F}(\mu) = \int V(x) d\mu(x) \longrightarrow V \longrightarrow \nabla V$$

2. Negative entropy (for μ absolutely continuous), using $(y \log y)' = \log y + 1$

$$\mathcal{F}(\mu) = \int \log(\mu(x)) d\mu(x) \longrightarrow \log(\mu) + 1 \longrightarrow \nabla \log \mu.$$

(Important) remark and references. For ease of presentation, I have considered cases where $\mathcal{F}((\mathrm{Id} + \varepsilon h)_{\#}\mu) = \mathcal{F}(\mu) + \varepsilon \langle \nabla_{W_2}\mathcal{F}(\mu), h\rangle_\mu + o(\varepsilon)$ and $\nabla_{W_2}\mathcal{F}(\mu) = \nabla\mathcal{F}'(\mu)$. Originally, [Ambrosio et al. (2008)] introduced Wasserstein subgradients (Def 10.1.1) where the latter equality corresponds to an inequality. They give conditions under which this set is non empty (Th 10.4.13) and is the singleton $\{\nabla\mathcal{F}'(\mu)\}$ (Lemma 10.4.1). Recently [Bonnet (2019); Lanzetti et al. (2022)] formalized the case of the equality and refers to "Wasserstein differentiable" functionals. Includes smooth potential/interaction energies as well as W_p distances for absolutely continuous measures.

Wasserstein gradient of KL

More generally, let

$$\mathcal{F}(\mu) = \underbrace{\int V(x) d\mu(x)}_{\textit{Potential}} + \underbrace{\int \log(\mu(x)) d\mu(x)}_{\text{(Neg.) Entropy}}.$$

Then, for $\pi \propto \exp(-V)$,

$$\mathrm{KL}(\mu|\pi) = \mathcal{F}(\mu) - \underbrace{\mathcal{F}(\pi)}_{\textit{Constant}}.$$

By additivity, the Wasserstein gradient of KL is given by[1]

$$\nabla_{W_2} \mathcal{F}(\mu) = \nabla \mathcal{F}'(\mu) = \nabla V + \nabla \log(\mu) = \nabla \log\left(\frac{\mu}{\pi}\right).$$

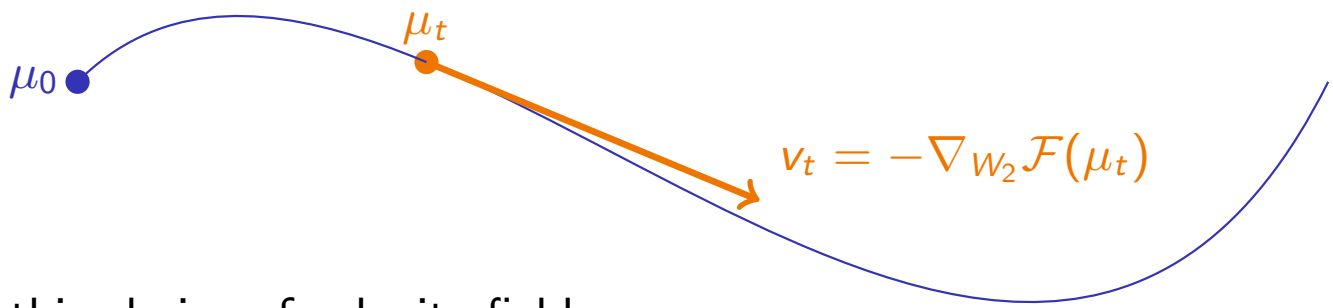
[1] See [?]Th. 10.4.13]ambrosio2008gradient for precise statement.

Velocity field = negative Wasserstein gradient

Recall that we wanted to define the equation

$$"\dot{\mu}_t = -\nabla_{W_2}\mathcal{F}(\mu_t)".$$

We consider the direction $v_t = -\nabla_{W_2}\mathcal{F}(\mu_t)$ at each time to decrease $\mathcal{F}$:



since for this choice of velocity field,

$$\frac{d\mathcal{F}(\mu_t)}{dt} = -\|\nabla_{W_2}\mathcal{F}(\mu_t)\|^2_{\mu_t} \leq 0.$$

Exercise.

Wasserstein gradient flows (WGF) [Ambrosio et al. (2008)]

The Wasserstein GF of $\mathcal{F}$ is ruled by:

$$v_t = -\nabla_{W_2}\mathcal{F}(\mu_t) \tag{1}$$

Equivalently:

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \nabla_{W_2}\mathcal{F}(\mu_t)\right), \tag{2}$$

Problem: How to construct such a flow on $\mathcal{P}_2(\mathbb{R}^d)$?

In the following, we will see some examples of dynamics $(x_t)_{t\geq 0} \in \mathbb{R}^d$ whose law $(\mu_t)_{t\geq 0}$ obeys (2). We will call such dynamics over $\mathbb{R}^d$ **realizations** of the WGF of $\mathcal{F}$.

Outline

Example I - Constant vector field

Let $x_0 \sim \mu_0$ and $V : \mathbb{R}^d \to \mathbb{R}$. Consider the dynamics:

$$\dot{x_t} = -\nabla V(x_t), \quad x_t \in \mathbb{R}^d. \tag{3}$$

Let μ_t be the law of x_t at each time $t \geq 0$. **Then, $v_t = -\nabla V$ is a velocity field of $(\mu_t)_{t \geq 0}$.**

Example I - Constant vector field

Let $x_0 \sim \mu_0$ and $V : \mathbb{R}^d \to \mathbb{R}$. Consider the dynamics:

$$\dot{x_t} = -\nabla V(x_t), \quad x_t \in \mathbb{R}^d. \tag{3}$$

Let μ_t be the law of x_t at each time $t \geq 0$. **Then, $v_t = -\nabla V$ is a velocity field of $(\mu_t)_{t\geq 0}$.**

Proof: Let $t \geq 0$. Using the chain rule and (3),

$$\frac{d}{dt}\varphi(x_t) = \langle \nabla\varphi(x_t), \dot{x_t}\rangle_{\mathbb{R}^d} = \langle \nabla\varphi(x_t), -\nabla V(x_t)\rangle_{\mathbb{R}^d}.$$

$$\begin{aligned}\frac{d}{dt}\int \varphi d\mu_t = \frac{d}{dt}\,\mathbb{E}\left[\varphi(x_t)\right] &= \mathbb{E}\left[\frac{d}{dt}\varphi(x_t)\right] \\ &= \mathbb{E}\left[\langle \nabla\varphi(x_t), -\nabla V(x_t)\rangle_{\mathbb{R}^d}\right] = \langle \nabla\varphi, -\nabla V\rangle_{\mu_t}.\end{aligned}$$

Therefore we can identify $v_t = -\nabla V$.

Example I : WGF of Potential energy

- We have just seen that:

$$\dot{x}_t = -\nabla V(x_t), \quad x_t \in \mathbb{R}^d, \quad x_t \sim \mu_t, \tag{4}$$

$$\Downarrow$$

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot (\mu_t \nabla V). \tag{5}$$

- In other words, $v_t = -\nabla V = -\nabla_{W_2} \mathcal{F}(\mu_t)$ where $\mathcal{F}(\mu) = \int V d\mu$ is a Potential energy.

Hence (4) realizes the WGF of the Potential energy $\mathcal{F}$ (5).

Example II : WGF of generic $\mathcal{F}$

More generally, let $x_0 \sim \mu_0$ and consider the dynamics:

$$\dot{x}_t = v_t(x_t).$$

Let μ_t be the law of x_t at each time $t \geq 0$. **Then, $(v_t)_{t\geq 0}$ is a velocity field of $(\mu_t)_{t\geq 0}$.**

[1] The randomness only comes from $x_0 \sim \mu_0$.

Example II : WGF of generic $\mathcal{F}$

More generally, let $x_0 \sim \mu_0$ and consider the dynamics:

$$\dot{x}_t = v_t(x_t).$$

Let μ_t be the law of x_t at each time $t \geq 0$. **Then, $(v_t)_{t\geq 0}$ is a velocity field of $(\mu_t)_{t\geq 0}$.**

In particular, let $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$. The dynamics

$$\dot{x}_t = -\nabla_{W_2}\mathcal{F}(\mu_t)(x_t), \quad x_t \in \mathbb{R}^d, \quad x_t \sim \mu_t, \tag{6}$$

realizes the Wasserstein GF of $\mathcal{F}$.

Note that $(x_t)_{t\geq 0}$ follows a deterministic dynamics[1]. There may be other realizations of the Wasserstein GF!

[1] The randomness only comes from $x_0 \sim \mu_0$.

Example III : Brownian motion

Let $x_0 \sim \mu_0$ independent of $b_t \sim \mathcal{N}(0, t\,\mathrm{Id})$ the Brownian motion, and consider the dynamics

$$x_t = x_0 + \sqrt{2} b_t.$$

Let μ_t be the law of x_t at each time $t \geq 0$. **Then,** $v_t = -\nabla \log(\mu_t)$ **is a velocity field of** $(\mu_t)_{t \geq 0}$.

[1] Using $\Delta = \nabla \cdot \nabla$ (Divergence of Gradient = Laplacian).

Example III : Brownian motion

Let $x_0 \sim \mu_0$ independent of $b_t \sim \mathcal{N}(0, t\,\mathrm{Id})$ the Brownian motion, and consider the dynamics

$$x_t = x_0 + \sqrt{2} b_t.$$

Let μ_t be the law of x_t at each time $t \geq 0$. **Then, $v_t = -\nabla \log(\mu_t)$ is a velocity field of $(\mu_t)_{t\geq 0}$.**

Proof: Differentiate $\varphi(x_t)$ using Itô formula, take the expectation and identify the velocity field from its definition.

In this case, the Continuity Equation is the Heat equation[1]

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\underbrace{\mu_t \nabla \log(\mu_t)}_{=\mu_t . \nabla \mu_t / \mu_t} \right) = \Delta \mu_t.$$

[1] Using $\Delta = \boldsymbol{\nabla} \cdot \nabla$ (Divergence of Gradient = Laplacian).

Example III $\Longrightarrow$ WGF of (Neg.) Entropy

- We have just seen that:

$$x_t = x_0 + \sqrt{2}b_t, \quad b_t \sim \mathcal{N}(0, t\,\mathrm{Id}), \quad x_t \in \mathbb{R}^d, \quad x_t \sim \mu_t, \tag{7}$$

$$\Downarrow$$

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot (\mu_t \nabla \log(\mu_t)) = \Delta \mu_t. \tag{8}$$

- In other words, $v_t = -\nabla \log(\mu_t) = -\nabla_{W_2} \mathcal{F}(\mu_t)$ where $\mathcal{F}(\mu) = \int \log(\mu(x)) d\mu(x)$ is the Negative entropy.

Hence (7) realizes the WGF of the Negative entropy $\mathcal{F}$ (8).

Other realizations of WGF of (Neg.) Entropy

Remark: While we have just seen that

$$x_t = x_0 + \sqrt{2} b_t, \quad b_t \sim \mathcal{N}(0, t\,\mathrm{Id})$$

realizes the WGF of the Negative entropy, it is also the case of

$$x_t = x_0 + \sqrt{2t}\eta, \quad \eta \sim \mathcal{N}(0, \mathrm{Id}). \tag{9}$$

Indeed, the latter satisfies

$$\dot{x}_t = -\nabla \log(\mu_t)(x_t),$$

which has the same velocity field $v_t = -\nabla \log(\mu_t)$.

All these processes have the same distribution μ_t realizing the WGF of the Negative entropy.

Example IV - Langevin diffusion

More generally, let $x_0 \sim \mu_0$, and consider the dynamics (Langevin diffusion)

$$dx_t = -\nabla V(x_t)dt + \sqrt{2}db_t,$$

where $(b_t)_{t\geq 0}$ is the Brownian motion. Let μ_t be the law of x_t at each time $t \geq 0$. **Then,** $v_t = -\nabla V + \nabla \log(\mu_t) = -\nabla \log\left(\frac{\mu_t}{\pi}\right)$ **where** $\pi \propto \exp(-V)$**, is a velocity field of** μ_t.

Proof: Combine Example I and III.

In this case, the Continuity Equation is the Fokker-Planck equation.

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \nabla \log\left(\frac{\mu_t}{\pi}\right)\right) = \Delta \mu_t + \boldsymbol{\nabla} \cdot (\mu_t \nabla V).$$

Example IV $\Longrightarrow$ WGF of the KL

- We have just seen that:

$$x_t = -\nabla V(x_t) + \sqrt{2} db_t, \quad x_t \in \mathbb{R}^d, \quad x_t \sim \mu_t, \tag{10}$$

$$\Downarrow$$

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \nabla \log \left(\frac{\mu_t}{\pi} \right) \right) = \Delta \mu_t + \boldsymbol{\nabla} \cdot (\mu_t \nabla V). \tag{11}$$

- In other words, $v_t = -\nabla \log\left(\frac{\mu_t}{\pi}\right) = -\nabla_{W_2} \mathcal{F}(\mu_t)$ where $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$ and $\pi \propto \exp(-V)$.

Hence (10) **realizes the WGF of the KL divergence** $\mathcal{F}$ (11).

Example IV $\Longrightarrow$ WGF of the KL

- We have just seen that:

$$x_t = -\nabla V(x_t) + \sqrt{2} db_t, \quad x_t \in \mathbb{R}^d, \quad x_t \sim \mu_t, \tag{10}$$

$$\Downarrow$$

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \nabla \log \left(\frac{\mu_t}{\pi} \right) \right) = \Delta \mu_t + \boldsymbol{\nabla} \cdot (\mu_t \nabla V). \tag{11}$$

- In other words, $v_t = -\nabla \log\left(\frac{\mu_t}{\pi}\right) = -\nabla_{W_2} \mathcal{F}(\mu_t)$ where $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$ and $\pi \propto \exp(-V)$.

Hence (10) **realizes the WGF of the KL divergence** $\mathcal{F}$ (11).

Remark: Another realization is given by

$$\dot{x}_t = -\nabla \log \left(\frac{\mu_t}{\pi} \right)(x_t), \; x_t \sim \mu_t.$$

Design of (Some) Sampling algorithms

A take home message.

As in Optimization, time discretizations of the Wasserstein GF can be seen as Sampling algorithms ($=$ optimization algorithms in $\mathcal{P}_2(\mathbb{R}^d)$).

This point of view allows to **design** (and **analyze !**) Sampling algorithms by discretizing Wasserstein GF.

Outline

Descent property of Wasserstein gradient flows

The Wasserstein GF decreases the objective function.

Using (1) the chain rule, and (2) $v_t = -\nabla_{W_2}\mathcal{F}(\mu_t)$, we have

$$\frac{d\mathcal{F}(\mu_t)}{dt} \overset{(1)}{=} \langle v_t, \nabla_{W_2}\mathcal{F}(\mu_t)\rangle_{\mu_t} \overset{(2)}{=} -\|\nabla_{W_2}\mathcal{F}(\mu_t)\|^2_{\mu_t} \leq 0.$$

This is a fundamental property of the Wasserstein gradient flow.

Evolution Variational Inequality (EVI)

Assume $\mathcal{F}$ λ-strongly geo. convex. Then, the Wasserstein GF satisfies the following variational inequality: for every $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\frac{d}{dt} W_2^2(\mu_t, \nu) \leq -2(\mathcal{F}(\mu_t) - \mathcal{F}(\nu)) - \lambda W_2^2(\mu_t, \nu).$$

The EVI characterizes the WGF when $\mathcal{F}$ is geo. convex. Note that it does not use $\nabla_{W_2} \mathcal{F}$.

Once again, taking $\nu = \pi$ the global optimum of $\mathcal{F}$, using Gronwall Lemma, we obtain linear convergence of $W_2^2(\mu_t, \pi)$ to zero.

Functional inequalities (Gradient dominance condition)

We already know

$$\frac{d\mathcal{F}(\mu_t)}{dt} \overset{(1)}{=} \langle v_t, \nabla_{W_2}\mathcal{F}(\mu_t)\rangle_{\mu_t} \overset{(2)}{=} -\|\nabla_{W_2}\mathcal{F}(\mu_t)\|^2_{\mu_t}.$$

If we have a **functional inequality**, i.e. for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$

$$\|\nabla_{W_2}\mathcal{F}(\mu_t)\|^2_{\mu} \geq 2\lambda\mathcal{F}(\mu)$$

then Gronwall inequality leads to $\mathcal{F}(\mu_t) \leq e^{-2\lambda t}\mathcal{F}(\mu_0)$.

Functional inequalities (Gradient dominance condition)

We already know

$$\frac{d\mathcal{F}(\mu_t)}{dt} \overset{(1)}{=} \langle v_t, \nabla_{W_2}\mathcal{F}(\mu_t)\rangle_{\mu_t} \overset{(2)}{=} -\|\nabla_{W_2}\mathcal{F}(\mu_t)\|^2_{\mu_t}.$$

If we have a **functional inequality**, i.e. for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$

$$\|\nabla_{W_2}\mathcal{F}(\mu_t)\|^2_{\mu} \geq 2\lambda\mathcal{F}(\mu)$$

then Gronwall inequality leads to $\mathcal{F}(\mu_t) \leq e^{-2\lambda t}\mathcal{F}(\mu_0)$.

Log Sobolev inequality is a gradient dominance condition for KL. [Otto and Villani (2000); Blanchet and Bolte (2018)].

$$\forall \mu \in \mathcal{P}_2(\mathbb{R}^d), \quad \mathrm{KL}(\mu|\pi) \leq \frac{1}{2\lambda}\|\nabla \log(\mu|\pi)\|^2_{L^2(\mu)}.$$

- V is λ-strongly convex $\Rightarrow$ $\pi \propto \exp(-V)$ satisfies Log Sobolev with λ (Bakry–Emery theorem)
- Log Sobolev $\not\Rightarrow$ V convex.

Generalized to the case with interaction energies in [Carrillo et al. (2006)].

Non log concave π satisfying Log Sobolev

Example: Consider a standard Gaussian distribution

$$\pi(x) \propto \exp\left(-\frac{\|x\|^2}{2}\right),$$

i.e. $\pi \propto \exp(-V)$ with V 1-strongly convex, i.e. π is (1-)strongly log-concave.

A small (bounded) perturbation of π is not necessarily log-concave, but still verifies a Log Sobolev inequality (Holley–Stroock perturbation theorem).

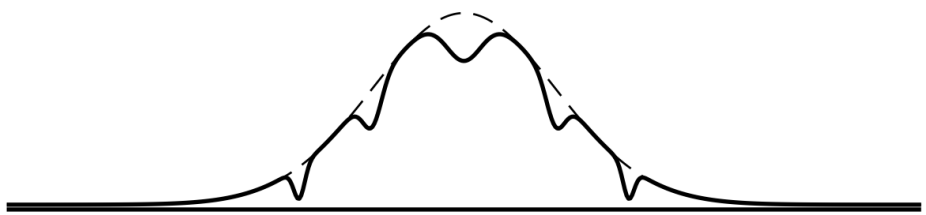


Figure from [Vempala and Wibisono (2019)].

Sampling as Optimization

$$\pi(x) \propto \exp(-V(x)),$$

$$\pi = \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathrm{KL}(\mu|\pi) = \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu),$$

Sampling as Optimization

$$\pi(x) \propto \exp(-V(x)),$$

$$\pi = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min}\ \mathrm{KL}(\mu|\pi) = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\min}\ \mathcal{F}(\mu),$$

where

$$\mathcal{F}(\mu) := \underbrace{\int V(x) d\mu(x)}_{\textit{Potential}} + \underbrace{\int \log(\mu(x)) d\mu(x)}_{\textit{(Neg.)Entropy}}$$

satisfies

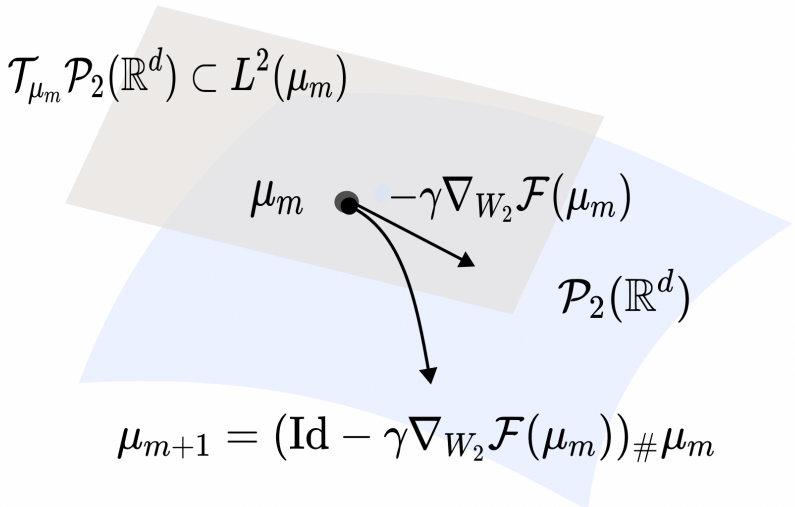
$$\mathcal{F}(\mu) - \underbrace{\mathcal{F}(\pi)}_{\textit{Constant}} = \mathrm{KL}(\mu|\pi).$$

Time discretizations of the Wasserstein GF

Let $\gamma > 0$ a step-size.

- Wasserstein gradient descent or Forward Euler (explicit):

$$\mu_{m+1} = \left(\mathrm{Id} - \gamma \, \nabla_{W_2} \mathcal{F}(\mu_m)\right)_{\#} \mu_m$$



Problem: If $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$, $\nabla_{W_2}\mathcal{F}(\mu_m) = \nabla \log\left(\frac{\mu_m}{\pi}\right)$ requires the knowledge of the density μ_m.

- JKO scheme [Jordan et al. (1998)] ($\mathcal{F}$ geo. convex):

$$\mu_{m+1} \in \mathrm{JKO}_{\gamma\mathcal{F}}(\mu_m) := \operatorname*{arg\,min}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \gamma\mathcal{F}(\mu) + \frac{1}{2} W_2^2(\mu, \mu_m) \right\}.$$

  i.e. Backward Euler (implicit) [Salim et al. (2020)].

- JKO scheme [Jordan et al. (1998)] ($\mathcal{F}$ geo. convex):

$$\mu_{m+1} \in \mathrm{JKO}_{\gamma\mathcal{F}}(\mu_m) := \operatorname*{arg\,min}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \left\{ \gamma\mathcal{F}(\mu) + \frac{1}{2} W_2^2(\mu, \mu_m) \right\}.$$

  i.e. Backward Euler (implicit) [Salim et al. (2020)].

- Splitting scheme [Salim et al. (2020)] ($\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2$, $\mathcal{F}_2$ geo. convex):

$$\mu_{m+\frac{1}{2}} = (\mathrm{Id} - \gamma \nabla_{W_2} \mathcal{F}_1(\mu_m))_{\#}\mu_m$$
$$\mu_{m+1} = \mathrm{JKO}_{\gamma\, \mathcal{F}_2}\left(\mu_{m+\frac{1}{2}}\right)$$

Problem: these (unbiased) schemes are also hard to implement (global optimization subroutine).

Outline

Langevin Monte Carlo

Recall $\pi \propto \exp(-V)$. *Langevin Monte Carlo* (LMC), introduced in [Roberts and Tweedie (1996)], starting from $x_0 \sim \mu_0$, updates the particle position as follows:

$$x_{m+1} = x_m - \gamma \nabla V(x_m) + \sqrt{2\gamma}\eta_m,$$

where $\gamma > 0$ and $(\eta_m)_{m \geq 0}$ i.i.d. standard Gaussian.

It is a discretization of Langevin diffusion

$$dx_t = -\nabla V(x_t)dt + \sqrt{2}db_t.$$

What's happening over the Wasserstein space?

Rewrite LMC as

$$x_{m+\frac{1}{2}} = x_m - \gamma \nabla V(x_m)$$
$$x_{m+1} = x_{m+\frac{1}{2}} + \sqrt{2\gamma}\eta_m.$$

Let $x_m \sim \mu_m$.

LMC can be written as a Forward Flow splitting scheme [Wibisono (2018); Durmus et al. (2019); Bernton (2018)]
($\mathcal{F}$ = Potential + Entropy)

$$\mu_{m+\frac{1}{2}} = (\mathrm{Id} - \gamma \underbrace{\nabla V}_{= \nabla_{W_2} \text{ Potential}})_{\#}\mu_m$$
$$\mu_{m+1} = \mathrm{flow}_{\gamma,\text{Entropy}}(\mu_{m+\frac{1}{2}})$$

Remark: this splitting scheme is biased.

Consequence: Descent lemma

LMC *almost* decreases the KL [Vempala and Wibisono (2019)], [Balasubramanian et al. (2022)]:

$$\frac{\mathcal{F}(\mu_{m+1}) - \mathcal{F}(\mu_m)}{\gamma} \leq -\frac{1}{2}\|\nabla_{W_2}\mathcal{F}(\hat{\mu}_m)\|^2_{\hat{\mu}_m} + 4L^2 d\gamma,$$

where $\hat{\mu}_m$ "between" μ_m and μ_{m+1}.

Error term $4L^2 d\gamma$: LMC is biased, i.e., π is not an invariant distribution.

Nonconvex rates for Langevin Monte Carlo

Nonconvex rates can be obtained using Descent lemma, noting that

$$\|\nabla_{W_2}\mathcal{F}(\mu)\|_\mu^2 = \left\|\nabla \log\left(\frac{\mu}{\pi}\right)\right\|_\mu^2 := \mathrm{FD}(\mu|\pi),$$

where KSD is a divergence [Liu et al. (2016); Chwialkowski et al. (2016)] and can metrize weak convergence in certain cases [Gorham and Mackey (2017)].

1. we first obtain

$$\frac{1}{M}\sum_{m=0}^{M-1}\mathrm{FD}(\hat{\mu}_m|\pi) \le \frac{2\,\mathrm{KL}(\mu_0|\pi)}{\gamma M} + 8L^2 d\gamma.$$

Nonconvex rates for Langevin Monte Carlo

Nonconvex rates can be obtained using Descent lemma, noting that

$$\|\nabla_{W_2}\mathcal{F}(\mu)\|_\mu^2 = \left\|\nabla \log\left(\frac{\mu}{\pi}\right)\right\|_\mu^2 := \mathrm{FD}(\mu|\pi),$$

where KSD is a divergence [Liu et al. (2016); Chwialkowski et al. (2016)] and can metrize weak convergence in certain cases [Gorham and Mackey (2017)].

1. we first obtain

$$\frac{1}{M}\sum_{m=0}^{M-1}\mathrm{FD}(\hat{\mu}_m|\pi) \le \frac{2\,\mathrm{KL}(\mu_0|\pi)}{\gamma M} + 8L^2 d\gamma.$$

2. If π satisfies Log Sobolev inequality with λ, i.e.:

$$\forall \mu \in \mathcal{P}_2(\mathbb{R}^d), \quad \mathrm{KL}(\mu|\pi) \le \frac{1}{2\lambda}\mathrm{FD}(\mu|\pi),$$

then [Vempala and Wibisono (2019)],

$$\mathrm{KL}(\mu_M|\pi) \le \exp(-\gamma M\lambda)\,\mathrm{KL}(\mu_0|\pi) + \frac{8L^2 d\gamma}{\lambda}.$$

Convex case - Discrete EVI

Assume V λ-strongly convex. Then, the Langevin algorithm *almost* satisfies a discrete EVI [Durmus et al. (2019)]; i.e. for every $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\frac{W_2^2(\mu_{m+1}, \nu) - W_2^2(\mu_m, \nu)}{\gamma} \leq -2(\mathcal{F}(\mu_{m+1}) - \mathcal{F}(\nu)) - \lambda W_2^2(\mu_m, \nu) + 2\gamma L d.$$

Convex case - Discrete EVI

Assume V λ-strongly convex. Then, the Langevin algorithm *almost* satisfies a discrete EVI [Durmus et al. (2019)]; i.e. for every $\nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\frac{W_2^2(\mu_{m+1}, \nu) - W_2^2(\mu_m, \nu)}{\gamma} \leq -2(\mathcal{F}(\mu_{m+1}) - \mathcal{F}(\nu)) - \lambda W_2^2(\mu_m, \nu) + 2\gamma L d.$$

Error term $2\gamma L d$: LMC is biased, i.e., π is not an invariant distribution.

Convex rates for Langevin Monte Carlo

Convex rates can be obtained using discrete EVI, noting that $\mathcal{F}(\mu) - \mathcal{F}(\pi) = \mathrm{KL}(\mu|\pi)$,

1. for $\lambda \geq 0$ we can obtain

$$\mathrm{KL}(\bar{\mu}_M|\pi) \leq \frac{W_2^2(\mu_0, \pi)}{2\gamma M} + \gamma L d,$$

where $\bar{\mu}_M = \frac{1}{M} \sum_{m=0}^{M-1} \mu_m$,

2. and, if $\lambda > 0$,

$$W_2^2(\mu_M, \pi) \leq (1 - \gamma\lambda)^M W_2^2(\mu_0, \pi) + \frac{2\gamma L d}{\lambda}.$$

Outline

Stein Variational Gradient Descent (SVGD)

SVGD [Liu and Wang (2016)] to sample from $\pi \propto \exp(-V)$.

SVGD updates the positions of a set of N particles $x^1, \dots, x^N$, i.e. for any $i = 1, \dots, N$, at each time $m \geq 0$:

$$x_{m+1}^i = x_m^i - \frac{\gamma}{N} \sum_{j=1}^{N} \nabla V(x_m^j) k(x_m^i, x_m^j) - \nabla_2 k(x_m^i, x_m^j),$$

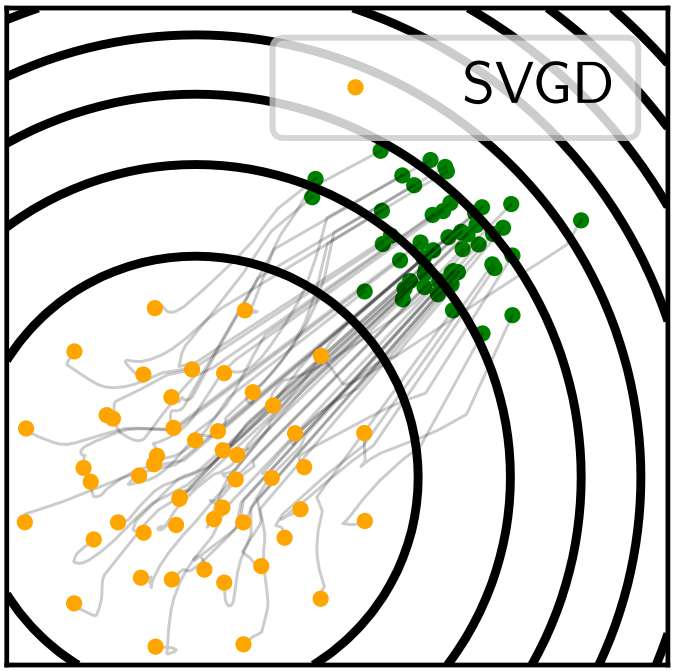
where k is a kernel associated to a **Reproducing Kernel Hilbert Space** H_k.

Reproducing kernel Hilbert Space

- Hilbert space of functions H_k (here, $\mathrm{H}_k \subset L^2(\mu)$ for every μ)
- For every x, $k(x, \cdot) \in \mathrm{H}_k$ ($k(x, \cdot) : \mathbb{R}^d \to \mathbb{R}$)
- Reproducing property: for every $f \in \mathrm{H}_k$, $f(x) = \langle f, k(x, \cdot) \rangle_{\mathrm{H}_k}$.

Example: $k(x, y) = \exp\big(-\|x - y\|^2\big)$.

Two dimensional example



Simulation from [Korba et al. (2021)]. Pytorch code available at `https://github.com/pierreablin/ksddescent`.

What's happening over the Wasserstein space

Let $\mu_m = \frac{1}{N}\sum_{j=1}^{N} \delta_{x_m^j}$. Then,

$$\mu_{m+1} = (\mathrm{Id} - \gamma h_{\mu_m})_{\#}\, \mu_m,$$

where $h_\mu := \int \nabla V(x) k(x,\cdot) - \nabla_1 k(x,\cdot) d\mu(x)$.

Actually,

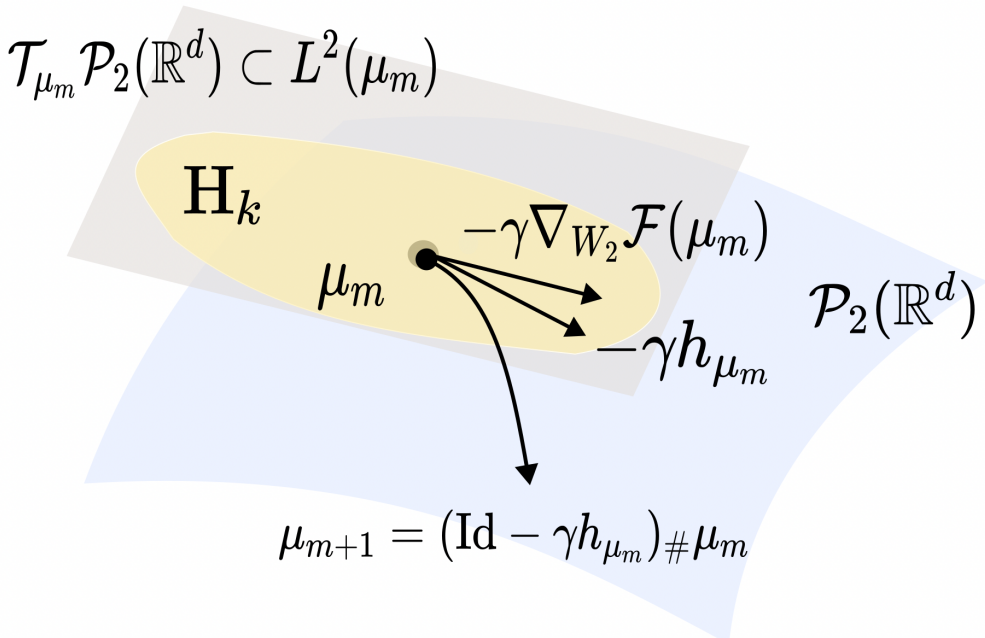
$$h_\mu = P_\mu \nabla \log\left(\frac{\mu}{\pi}\right), \text{ where } P_\mu : L^2(\mu) \to \mathrm{H}_k, f \mapsto \int f(x) k(x,\cdot) d\mu(x).$$

Gradient descent interpretation

A Taylor expansion around μ for $h \in \mathrm{H}_k$, if μ has a density yields [Liu (2017)]:

$$\mathrm{KL}((\mathrm{Id} + \varepsilon h)_{\#}\mu | \pi) = \mathrm{KL}(\mu|\pi) + \varepsilon \langle h_\mu, h \rangle_{\mathrm{H}_k} + o(\varepsilon).$$

Therefore, h_μ plays the role of the Wasserstein gradient in H_k.



It has been proved that h_μ is the Wasserstein gradient of $\mathrm{KL}(\cdot|\pi)$ with respect to a "kernelized" Wasserstein metric [Liu (2017); Duncan et al. (2019)][1].

[1] WGF of KL w.r.t. $W_k^2(\mu,\nu) = \inf_{\mu_t, v_t} \left\{ \int_0^1 \|v_t\|^2_{\mathrm{H}_k} d\,dt : \frac{\partial \mu_t}{\partial t} = \nabla \cdot (\mu_t v_t), \mu_0 = \mu, \mu_1 = \nu \right\}$.

Consequence: Descent lemma

We study

$$\mu_{m+1} = (\mathrm{Id} - \gamma h_{\mu_m})_{\#} \mu_m$$

when μ_m has a density (i.e. "mean field" or "population limit" $=$ SVGD with an infinite number of particles).

In this case, for a bounded k, SVGD decreases the KL [Liu (2017); Gorham et al. (2020)], [Korba et al. (2020); Salim et al. (2021)]:

$$\frac{\mathcal{F}(\mu_{m+1}) - \mathcal{F}(\mu_m)}{\gamma} \leq -\frac{1}{2} \| h_{\mu_m} \|^2_{\mathrm{H}_k}.$$

Nonconvex rates for SVGD

Nonconvex rates can be obtained using Descent lemma, noting that

$$\|h_{\mu_m}\|_{\mathrm{H}_k}^2 = \left\| P_{\mu_m} \nabla \log\left(\frac{\mu_m}{\pi}\right) \right\|_{\mathrm{H}_k}^2 = \mathrm{KSD}^2(\mu_m|\pi).$$

We obtain

$$\mathrm{KSD}^2(\bar{\mu}_M|\pi) \leq \frac{2\,\mathrm{KL}(\mu_0|\pi)}{\gamma M}, \quad \bar{\mu}_M = \frac{1}{M}\sum_{m=0}^{M-1} \mu_m.$$

This holds for $(\mu_m)_{m\in\mathbb{N}}$ with densities (even if $\mathrm{KSD}(\mu|\pi)$ can be defined for discrete measures μ), i.e. the mean-field limit where the number of particles $N \to \infty$.

Many other works have studied the asymptotic behavior for both finite time and number of particles, e.g. in continuous time [Lu et al. (2019); Carrillo and Skrzeczkowski (2023)] and discrete time [Shi and Mackey (2023); Das and Nagaraj (2023)]

Outline

Extensions to other optimization techniques

- Accelerated methods: accelerated LMC [Ma et al. (2019); Dalalyan and Riou-Durand (2020); Shen and Lee (2019)], accelerated particle methods [Liu et al. (2019)]
- "Mirror-descent" like sampling algorithms to sample from a distribution with compact support: Mirror Langevin [Hsieh et al. (2018); Zhang et al. (2020); Ahn and Chewi (2021); Li et al. (2022)], Mirror SVGD [Shi et al. (2021)], recently extended in [Bonet et al. (2024)]
- "Proximal" algorithms for non-smooth potentials V (i.e. no gradients of V) [Durmus et al. (2019); Wibisono (2019)], [Salim et al. (2019); Salim and Richtárik (2020)]
- Variance reduction for potentials V written as finite sums [Ding and Li (2021); Zou et al. (2018, 2019); Dubey et al. (2016)], [Balasubramanian et al. (2022)].

Conclusion

- Sampling can be seen as an optimization problem on a "Wasserstein manifold", and we can consider Wasserstein gradient flows, that decrease a loss (e.g. here the KL)
- Their discretizations (space/time) lead to different algorithms: LMC is a splitting (forward-flow) scheme, SVGD is a gradient descent
- One can design Sampling algorithms by discretizing Wasserstein GF

Some limitations of the framework

- The presented framework does not cover all sampling algorithms, e.g. involving dynamics such as accept/reject steps, birth and death of particles...
- It does not cover neither the analysis for finite number of particles (e.g. iterates of Langevin Monte Carlo, SVGD stationary particles...)
- We did not talk about practical considerations, e.g. improving convergence (for π multimodal, high-dimensional)

Sampling as optimization over probability distributions

Assume that $\pi \in \mathcal{P}_2(\mathbb{R}^d) = \left\{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\right\}$.
The sampling task can be recast as an optimization problem:

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathrm{D}(\mu|\pi) := \mathcal{F}(\mu),$$

where D is a **discrepancy**, for instance:

- a f-divergence: $\int f\left(\frac{\mu}{\pi}\right) d\pi, \quad f$ convex, $f(1) = 0$
- an integral probability metric: $\sup_{f \in \mathcal{G}} \left|\int f d\mu - \int f d\pi\right|$
- an optimal transport distance (e.g. W_1, W_2), or Sinkhorn divergence:

$$S^\epsilon(\mu, \nu) = \mathcal{W}_2^\epsilon(\mu, \nu) - \frac{1}{2}\mathcal{W}_2^\epsilon(\mu, \mu) - \frac{1}{2}\mathcal{W}_2^\epsilon(\nu, \nu)$$

where $\mathcal{W}_2^\epsilon(\mu, \nu) = \inf_{s \in \Gamma(\nu,\mu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, ds(x, y) + \epsilon \, \mathrm{KL}(\pi|\mu \otimes \nu)$.

Starting from an initial distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, one can then consider a **Wasserstein-2**[1] **gradient flow** of $\mathcal{F}$ over $\mathcal{P}_2(\mathbb{R}^d)$ to transport μ_0 to π.

[1] $W_2^2(\nu, \mu) = \inf_{s \in \Gamma(\nu,\mu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, ds(x, y)$, where $\Gamma(\nu, \mu)$= couplings between ν, μ.

Particle system/Gradient descent approximating the WGF

Recall we want to minimize $\mathcal{F}(\mu) = \mathrm{D}(\mu|\pi)$. The family $\mu : [0, \infty] \to \mathcal{P}_2(\mathbb{R}^d), t \mapsto \mu_t$ is a Wasserstein gradient flow of $\mathcal{F}$ if:

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \nabla_{W_2} \mathcal{F}(\mu_t)\right),$$

where $\nabla_{W_2}\mathcal{F}(\mu) := \nabla \frac{\partial \mathcal{F}(\mu)}{\partial \mu} : \mathbb{R}^d \to \mathbb{R}^d$ denotes the Wasserstein gradient of $\mathcal{F}$. It can be implemented by the deterministic process in $\mathbb{R}^d$:

$$\frac{dx_t}{dt} = -\nabla_{W_2}\mathcal{F}(\mu_t)(x_t), \quad \text{where } x_t \sim \mu_t$$

Space/time discretization: Introduce a particle system $x_0^1, \ldots, x_0^n \sim \mu_0$, a step-size γ, and an explicit time discretisation:

$$x_{m+1}^i = x_m^i - \gamma \nabla_{W_2}\mathcal{F}(\hat{\mu}_m)(x_m^i) \quad \text{for } i = 1, \ldots, n, \text{ where } \hat{\mu}_m = \frac{1}{n}\sum_{i=1}^n \delta_{x_m^i}. \tag{12}$$

In particular, if $\mathcal{F}(\mu) = \mathrm{D}(\mu|\pi)$ is well-defined for discrete measures μ, Algorithm (12) simply corresponds to gradient descent of $F : \mathbb{R}^{N\times d} \to \mathbb{R}$, $F(x^1, \ldots, x^n) := \mathcal{F}(\mu^n)$ where $\mu^n = \frac{1}{n}\sum_{i=1}^n \delta_{x^i}$.

Outline

χ^2 flow

Consider the chi-square (CS) divergence, which is a f-divergence:

$$\chi^2(\mu|\pi) := \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\pi} - 1\right)^2 \mathrm{d}\pi \text{ if } \mu \ll \pi; +\infty \text{ else.}$$

- It is not convenient neither when μ, μ^* are discrete
- χ^2-gradient requires the normalizing constant of μ^*: $\nabla \frac{\mu}{\mu^*}$
- However, the GF of χ^2 has interesting properties
  - we have $\chi^2(\mu|\pi) \geq \mathrm{KL}(\mu|\pi)$.
  - KL decreases exp. fast along CS flow/χ^2 decreases exp. fast along KL flow if π satisfies a Poincaré inequality [Matthes et al. (2009)], i.e. for all locally Lipschitz function functions $f : \mathbb{R}^d \to \mathbb{R}$,

$$\int f^2(x)d\pi(x) - \left(\int f(x)d\pi(x)\right)^2 \leq C_P \|\nabla f\|^2_{L^2(\pi)}, \tag{13}$$

    where C_P is the smallest constant for which this holds.
  - χ^2 known to converge polynomially along its WGF under a weak Poincaré inequality [Dolbeault et al. (2007)]

Outline

MMD Gradient flow in practice

Take $\mathcal{F}(\mu) = \mathrm{MMD}^2(\mu, \pi) =$
$\iint k(x,y)d\mu(x)d\mu(y) + \iint k(x,y)d\pi(x)d\pi(y) - 2\iint k(x,y)d\mu(x)d\pi(y)$.

- The first variation and the Wasserstein gradient of $\mathcal{F}$ at μ are

$$\frac{\partial \mathcal{F}(\mu)}{\partial \mu} = \int k(x,\cdot)d\mu(x) - \int k(x,\cdot)d\pi(x),$$

$$\nabla_{W_2}\mathcal{F}(\mu) = \int \nabla_2 k(x,\cdot)d\mu(x) - \int \nabla_2 k(x,\cdot)d\pi(x)$$

- The WGF of the MMD can be implemented via :

$$\frac{dx_t}{dt} = -\nabla_{W_2}\mathcal{F}(\mu_t)(x_t)$$

- in practice we can implement the discrete-time interacting particle system:

$$x^i_{m+1} = x^i_l - \gamma\left(\sum_{j=1}^n \nabla_2 k(x^i_m, x^j_m) - \int \nabla_2 k(x^i_m, y)d\pi(y)\right)$$

which is gradient descent of $(x^1,\dots,x^n) \mapsto \mathrm{MMD}^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x^i}, \pi\right)$

References I

Agapiou, S., Papaspiliopoulos, O., Sanz-Alonso, D., and Stuart, A. M. (2017). Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pages 405–431.

Ahn, K. and Chewi, S. (2021). Efficient constrained sampling via the mirror-langevin algorithm. *Advances in Neural Information Processing Systems*, 34:28405–28418.

Ambrosio, L., Gigli, N., and Savaré, G. (2008). *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media.

Balasubramanian, K., Chewi, S., Erdogdu, M. A., Salim, A., and Zhang, M. (2022). Towards a theory of non-log-concave sampling: First-order stationarity guarantees for langevin monte carlo. *arXiv preprint arXiv:2202.05214*.

Bernton, E. (2018). Langevin Monte Carlo and JKO splitting. In *Conference On Learning Theory (COLT)*, pages 1777–1798.

Blanchet, A. and Bolte, J. (2018). A family of functional inequalities: Łojasiewicz inequalities and displacement convex functions. *Journal of Functional Analysis*, 275(7):1650–1673.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Bonet, C., Uscidda, T., David, A., Aubin-Frankowski, P.-C., and Korba, A. (2024). Mirror and preconditioned gradient descent in wasserstein space. *arXiv preprint arXiv:2406.08938*.

Bonnet, B. (2019). A Pontryagin Maximum Principle in Wasserstein Spaces for Constrained Optimal Control Problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:52.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417.

Carrillo, J. A., McCann, R. J., and Villani, C. (2006). Contractions in the 2-wasserstein length space and thermalization of granular media. *Archive for Rational Mechanics and Analysis*, 179(2):217–263.

Carrillo, J. A. and Skrzeczkowski, J. (2023). Convergence and stability results for the particle system in the stein gradient descent method. *arXiv preprint arXiv:2312.16344*.

Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit. In *International conference on machine learning*.

References II

Dalalyan, A. S. and Riou-Durand, L. (2020). On sampling from a log-concave density using kinetic langevin diffusions. *Bernoulli*, 26(3):1956–1988.

Das, A. and Nagaraj, D. (2023). Provably fast finite particle variants of svgd via virtual particle stochastic approximation. *Advances in Neural Information Processing Systems*, 36:49748–49760.

De Giorgi, E. (1993). New problems on minimizing movements. *Ennio de Giorgi: Selected Papers*, pages 699–713.

De Giorgi, E., Marino, A., and Tosques, M. (1980). Problems of evolution in metric spaces and maximal decreasing curve. *Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur.(8)*, 68(3):180–187.

Ding, Z. and Li, Q. (2021). Langevin monte carlo: random coordinate descent and variance reduction. *J. Mach. Learn. Res.*, 22:205–1.

Dolbeault, J., Gentil, I., Guillin, A., and Wang, F.-Y. (2007). Lq-functional inequalities and weighted porous media equations. *arXiv preprint math/0701037*.

Dubey, K. A., J Reddi, S., Williamson, S. A., Poczos, B., Smola, A. J., and Xing, E. P. (2016). Variance reduction in stochastic gradient langevin dynamics. *Advances in neural information processing systems*, 29.

Duncan, A., Nüsken, N., and Szpruch, L. (2019). On the geometry of stein variational gradient descent. *arXiv preprint arXiv:1912.00894*.

Durmus, A., Majewski, S., and Miasojedow, B. (2019). Analysis of langevin monte carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46.

Franceschi, J.-Y., Gartrell, M., Dos Santos, L., Issenhuth, T., de Bézenac, E., Chen, M., and Rakotomamonjy, A. (2023). Unifying gans and score-based diffusion as generative particle models. *Advances in Neural Information Processing Systems*, 36.

Garrigos, G. and Gower, R. M. (2023). Handbook of convergence theorems for (stochastic) gradient methods. *arXiv preprint arXiv:2301.11235*.

Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1292–1301. JMLR. org.

Gorham, J., Raj, A., and Mackey, L. (2020). Stochastic stein discrepancies. *Advances in Neural Information Processing Systems*, 33:17931–17942.

Hsieh, Y.-P., Kavis, A., Rolland, P., and Cevher, V. (2018). Mirrored Langevin dynamics. In *Advances in Neural Information Processing Systems*, pages 2878–2887.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17.

References III

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. (2021). Kernel stein discrepancy descent. *arXiv preprint arXiv:2105.09994*.

Korba, A., Salim, A., Arbel, M., Luise, G., and Gretton, A. (2020). A non-asymptotic analysis for stein variational gradient descent. *arXiv preprint arXiv:2006.09797*.

Lanzetti, N., Bolognani, S., and Dörfler, F. (2022). First-Order Conditions for Optimization in the Wasserstein Space. *arXiv preprint arXiv:2209.12197*.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., and Póczos, B. (2017). Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30.

Li, R., Tao, M., Vempala, S. S., and Wibisono, A. (2022). The mirror langevin algorithm converges with vanishing bias. In *International Conference on Algorithmic Learning Theory*, pages 718–742. PMLR.

Liu, C., Zhuo, J., Cheng, P., Zhang, R., and Zhu, J. (2019). Understanding and accelerating particle-based variational inference. In *International Conference on Machine Learning*, pages 4082–4092. PMLR.

Liu, Q. (2017). Stein variational gradient descent as gradient flow. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2378–2386.

Lu, J., Lu, Y., and Nolen, J. (2019). Scaling limit of the stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671.

Ma, Y.-A., Chatterji, N., Cheng, X., Flammarion, N., Bartlett, P. L., and Jordan, M. I. (2019). Is there an analog of Nesterov acceleration for MCMC? *arXiv preprint arXiv:1902.00996*.

Matthes, D., McCann, R. J., and Savaré, G. (2009). A family of nonlinear fourth order equations of gradient flow type. *Communications in Partial Differential Equations*, 34(11):1352–1397.

McCann, R. J. (1997). A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179.

Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671.

Otto, F. (2001). The geometry of dissipative evolution equations: the porous medium equation.

References IV

Otto, F. and Villani, C. (2000). Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

Rotskoff, G. M. and Vanden-Eijnden, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *stat*, 1050:22.

Salim, A., Korba, A., and Luise, G. (2020). The Wasserstein proximal gradient algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Salim, A., Kovalev, D., and Richtárik, P. (2019). Stochastic proximal langevin algorithm: Potential splitting and nonasymptotic rates. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6649–6661.

Salim, A. and Richtárik, P. (2020). Primal dual interpretation of the proximal stochastic gradient Langevin algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Salim, A., Sun, L., and Richtárik, P. (2021). Complexity analysis of stein variational gradient descent under talagrand's inequality t1. *arXiv preprint arXiv:2106.03076*.

Shen, R. and Lee, Y. T. (2019). The randomized midpoint method for log-concave sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2100–2111.

Shi, J., Liu, C., and Mackey, L. (2021). Sampling with mirrored stein operators. *arXiv preprint arXiv:2106.12506*.

Shi, J. and Mackey, L. (2023). A finite-particle convergence rate for stein variational gradient descent. *Advances in Neural Information Processing Systems*, 36.

Vempala, S. and Wibisono, A. (2019). Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8092–8104.

Wibisono, A. (2018). Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory (COLT)*, page 2093–3027.

Wibisono, A. (2019). Proximal Langevin algorithm: Rapid convergence under isoperimetry. *arXiv preprint arXiv:1911.01469*.

Yi, M., Zhu, Z., and Liu, S. (2023). Monoflow: Rethinking divergence gans via the perspective of wasserstein gradient flows. In *International Conference on Machine Learning*, pages 39984–40000. PMLR.

Zhang, K. S., Peyré, G., Fadili, J., and Pereyra, M. (2020). Wasserstein control of mirror langevin monte carlo. In *Conference on Learning Theory*, pages 3814–3841. PMLR.

Zou, D., Xu, P., and Gu, Q. (2018). Subsampled stochastic variance-reduced gradient Langevin dynamics. In *International Conference on Uncertainty in Artificial Intelligence*.

Zou, D., Xu, P., and Gu, Q. (2019). Sampling from non-log-concave distributions via variance-reduced gradient Langevin dynamics. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2936–2945.

Reminders (Bayesian setting)

Here we will recall some fundamental methods and principles from Simulation and Monte Carlo (see Nicolas Chopin's course).

We will consider the "Bayesian inference" setting, where the target π has a density that is known to be $\pi \propto e^{-V}$.

Recall that in this setting we are often interested in approximating:

$$\int f(x) d\pi(x) \quad \text{for some } f.$$

Importance Sampling (IS)

Let q be a proposal distribution such that $\mathrm{Supp}(\pi) \subset \mathrm{Supp}(q)$. Define for all $x \in \mathbb{R}^d$

$$w(x) = \frac{\pi(x)}{q(x)}$$

Define the Self-Normalized Importance Sampling (SNIS) estimator of the expectation of f as

$$\int f d\pi \approx \sum_{i=1}^{N} w_N^i f(X_i), \qquad \text{where } w_N^i = \frac{w(X_i)}{\sum_{j=1}^{n} w(X_j)}$$

and $X_1, \ldots, X_N \sim q$.

Importance Sampling (IS)

Let q be a proposal distribution such that $\mathrm{Supp}(\pi) \subset \mathrm{Supp}(q)$. Define for all $x \in \mathbb{R}^d$

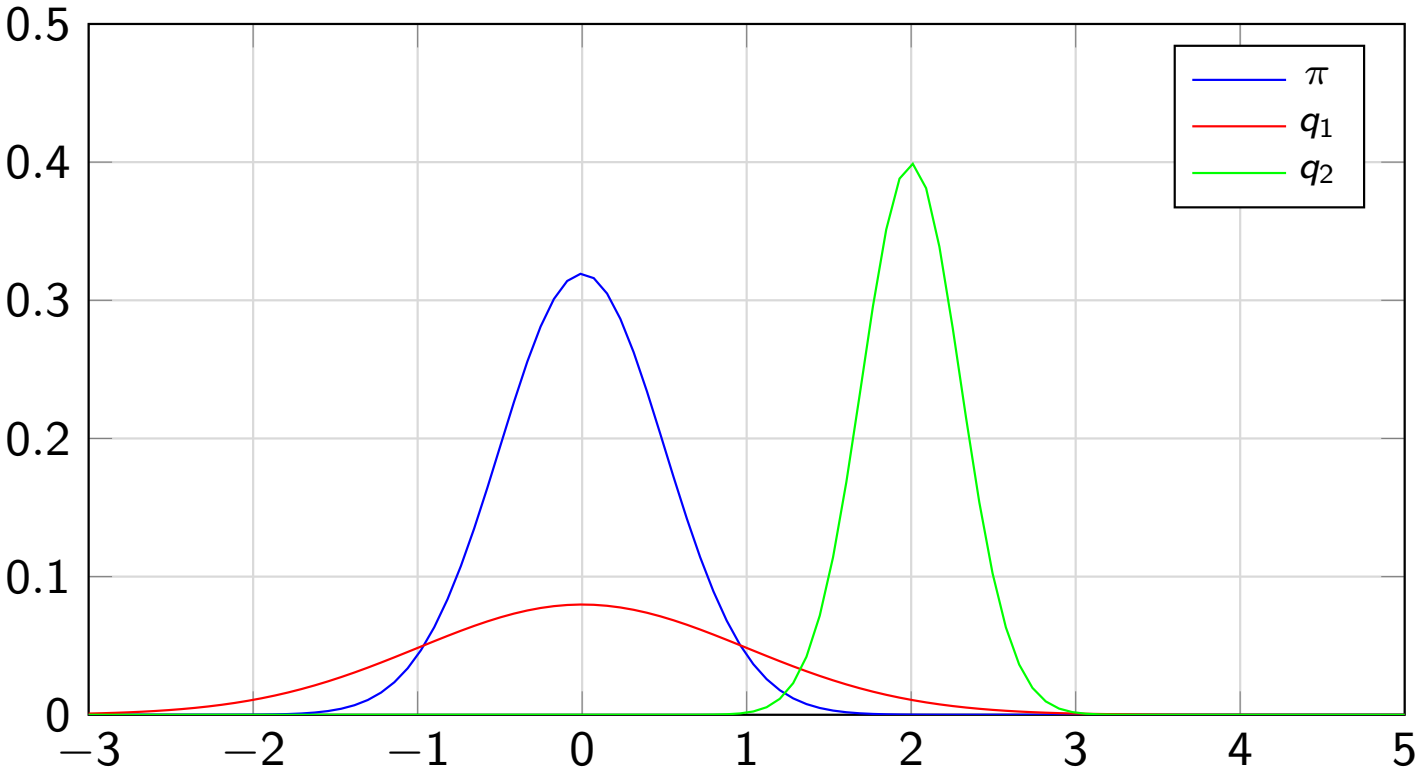
$$w(x) = \frac{\pi(x)}{q(x)}$$

Define the Self-Normalized Importance Sampling (SNIS) estimator of the expectation of f as

$$\int f d\pi \approx \sum_{i=1}^{N} w_N^i f(X_i), \qquad \text{where } w_N^i = \frac{w(X_i)}{\sum_{j=1}^{n} w(X_j)}$$

and $X_1, \ldots, X_N \sim q$.

Remark: For IS to be effective, the proposal q must be close enough to π in χ^2-square distance (see Agapiou et al. (2017, Th1)), which makes IS also notably affected by the curse of dimensionality (e.g., Agapiou et al. (2017, Sec 2.4.1)).



- Designing a good proposal q is critical
- There is a huge literature on Adaptive Importance Sampling

Recall that a Markov kernel $Q(x, dy)$ is an application $\mathbb{R}^d \to \mathcal{P}(\mathbb{R}^d)$.

Let $Q(x, dy)$ a Markov kernel, such that $Q(x, dy) = q(x, y)dy$.

Metropolis-Hastings is a two-step iterative algorithm relying on the proposal Markov kernel Q.

Let x_m be the state at time m.

- **Step 1**: Sample a candidate $y \sim Q(x_m, dy)$
- **Step 2**: The next state is set according to the rule:

$$x_{m+1} = \begin{cases} y & \text{with probability } \mathrm{acc}(x_m, y) \\ x_m & \text{with probability } 1 - \mathrm{acc}(x_m, y) \end{cases}$$

where the acceptance probability is

$$\mathrm{acc}(x_m, y) = \min\left(1, \frac{q(y, x_m)\pi(y)}{q(x_m, y)\pi(x_m)}\right)$$

Examples of Markov kernels

- Gaussian random walk

$$y \sim \mathcal{N}(x, \Sigma)$$

- Langevin proposal (yields "MALA" i.e. Metropolis Adjusted Langevin Algorithm)

$$y \sim \mathcal{N}(x + \nabla \log \pi(x), \mathrm{Id})$$

Recall that if $\pi \propto e^{-V}$, i.e. $\pi = \tilde{\pi}/Z$ where $\tilde{\pi}$ is known and Z unknown, then $\nabla \log(\pi) = \frac{\tilde{\pi}/Z}{\tilde{\pi}/Z} = \nabla V$.

Optimizing the KL with parametric models

Consider the sampling optimization objective:

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{D}(\mu|\pi)$$

But now (in this slide) assume we restrict the search space to a parametric families $\{P_\theta,\ \theta \in \mathbb{R}^p\}$ (ex: Gaussian with diagonal covariance matrices can be parametrized by $\theta = (m, \sigma) \in \mathbb{R}^{2d}$). The problem rewrites as a finite-dimensional optimization problem (i.e. over $\mathbb{R}^p$):

$$\min_{\theta \in \mathbb{R}^p} \mathrm{D}(\mu_\theta|\pi)$$

- Choosing D as the reverse KL, i.e. $\mathrm{D}(\mu_\theta|\pi) = \mathrm{KL}(\mu_\theta|\pi)$ yields Variational Inference [Blei et al. (2017)] which is useful for Bayesian Inference ($\pi \propto e^{-V}$)
- Choosing D as the forward KL, i.e. $\mathrm{D}(\mu_\theta|\pi) = \mathrm{KL}(\pi|\mu_\theta)$ yields Maximum Likelihood, which is useful for fitting a model ($x_1, \ldots, x_n \sim \pi$) since:

$$\min_\theta \mathrm{KL}(\pi|\mu_\theta) = \int \log\left(\frac{\pi}{\mu_\theta}\right) d\pi \Leftrightarrow \min_\theta - \int \log(\mu_\theta(x)) d\pi(x) \approx \sum_{i=1}^n \log(\mu_\theta(x_i))$$