Sampling as First-Order Optimization over a space of probability measures

Anna Korba[1], Adil Salim[2]

[1]CREST, ENSAE, Institut Polytechnique de Paris
[2]Microsoft Research, Redmond

International Conference of Machine Learning 2022

Outline

Outline

Scope of this tutorial regarding Sampling

Generally, sampling refers to the problem of generating new samples from a distribution π, given some information on π, e.g.:

1. **π's density is known up to a normalization constant (e.g. as in Bayesian inference)**
2. some samples of π are known (e.g. images as in generative modelling).

We will focus on the first setting and non parametric methods, which includes algorithms such as Langevin Monte Carlo or Stein Variational Gradient Descent.

We will not cover parametric methods i.e. Variational Inference.

We will not cover the second setting and methods such as Generative Adversarial Networks, Score-based Generative modelling...

About this tutorial

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

Structure of this tutorial

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 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 other algorithms.

Disclaimer

We do not claim generality and/or optimality of the results in this talk.

In particular,

- We will not work under minimal assumptions (see [Ambrosio et al., 2008] for that)
- We will not provide the best known convergence rates
- We will not study the dimension dependence of the algorithms (important, but does not fit in our story line)
- We will not cover *all* the literature on this topic (Sorry!)[1]

We focus on the underlying geometry of the problems and some examples.

[1] If you feel we should have included something, please send us an email!

Outline

Motivation for Sampling: Bayesian inference

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

Motivation for Sampling: 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: 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:

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

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".

The posterior π is interesting for

- measuring uncertainty on prediction through the distribution of $g(w,\cdot)$, $x \sim \pi$.
- prediction for a new input w:

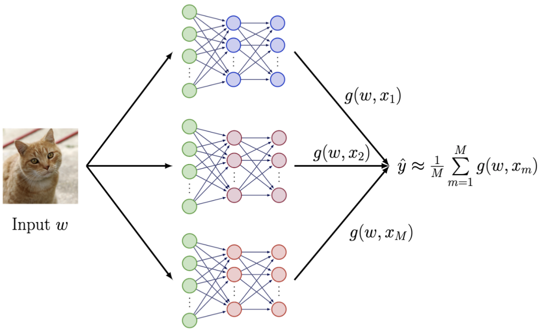
$$\hat{y} = \underbrace{\int_{\mathbb{R}^d} g(w,x)d\pi(x)}_{\text{"Bayesian model averaging"}}$$

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

In this talk, Sampling

=

construct an approximation $\mu_M = \frac{1}{M}\sum_{m=1}^{M} \delta_{x_m}$ of π.

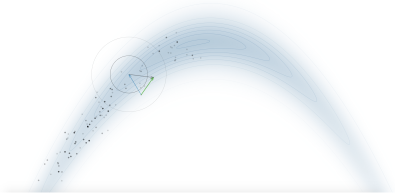


(Some, Non parametric) Sampling methods

(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]

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

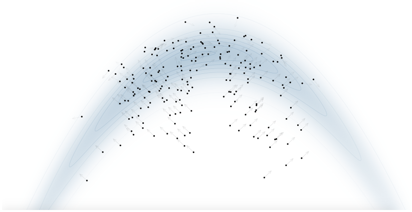


Picture from `https://chi-feng.github.io/mcmc-demo/app.html`.

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

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

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



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

Sampling as minimization of the KL

The Kullback-Leibler (KL) divergence between $\mu, \pi \in \mathcal{P}(\mathbb{R}^d)$ is:

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

Note that

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

Sampling as minimization of the KL

The Kullback-Leibler (KL) divergence between $\mu, \pi \in \mathcal{P}(\mathbb{R}^d)$ is:

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

Note that

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

The KL as an objective is convenient since it **does not depend on the normalization constant Z!**

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

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

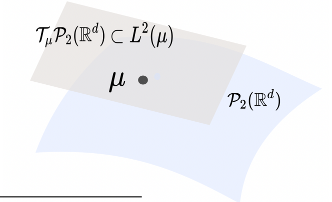
Sampling as optimization over $\mathcal{P}_2(\mathbb{R}^d)$

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 $\mathcal{P}_2(\mathbb{R}^d)$:

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

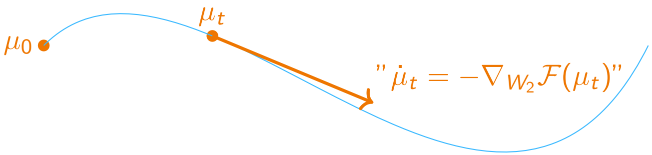
Equipped with the Wasserstein-2 (W_2) distance from optimal transport[1], the metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ has a convenient **Riemannian structure** [Otto and Villani, 2000].



[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 $\mathbb{R}^d$

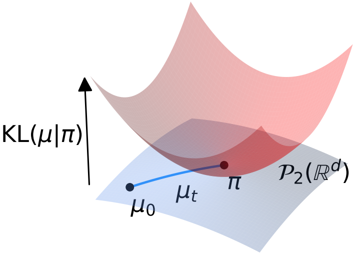
$dx_t = v(x_t, \mu_t)dt + \sigma(x_t, \mu_t)db_t, \quad x_t \sim \mu_t, \; (b_t)_{t\geq 0}$ Brownian motion.

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

Recall that $\pi(x) \propto \exp(-V(x))$, $V(x) = \underbrace{\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2}_{\text{loss of the model } g(\cdot,x)} + \frac{\|x\|^2}{2}$.

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

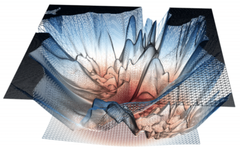
- if V is convex (e.g. $g(w,x) = \langle w, x \rangle$ linear), π is "log-concave" and "sampling is easy"



Recall that $\pi(x) \propto \exp(-V(x))$, $V(x) = \underbrace{\sum_{i=1}^{p} \|y_i - g(w_i, x)\|^2}_{\text{loss of the model } g(\cdot, x)} + \frac{\|x\|^2}{2}$.

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 (e.g. $g(w, x)$ is a neural network), π is "non log-concave" and "sampling is hard"



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

Sampling as optimization: how it started

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 of the Kullback-Leibler divergence.

Recently, this optimization point of view has been used to derive rates of convergence for variants of the Langevin Monte Carlo algorithm:

- [Wibisono, 2018]
- [Durmus et al., 2019]
- [Bernton, 2018]

Recent synergies between Sampling and PDE

- Simons institute program *"Geometric Methods in Optimization and Sampling"*[1], Fall 2021. Co-organized by Philippe Rigollet, Katy Craig, Simone di Marino and Ashia Wilson.

- Book to appear by Sinho Chewi.

[1] https://simons.berkeley.edu/workshops/gmos2021-bc

Outline

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

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

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,

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

The EVI is fundamental

Rewrite the EVI as

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

This inequality characterizes the gradient flow when V is convex. Note that it does not use ∇V.

The EVI is fundamental

Rewrite the EVI as

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

This inequality characterizes the gradient flow when V is convex. Note that it does not use ∇V.

Indeed, any curve $(x_t)_{t\geq 0}$ satisfying this inequality also satisfies

$$2\langle \dot{x}_t, x_t - y\rangle \leq -2(V(x_t) - V(y)), \quad \forall y \in \mathbb{R}^d,$$

which implies $\dot{x}_t = -\nabla V(x_t)$ using convexity.

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 (V convex):

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

  i.e. Backward Euler (implicit):

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

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):

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

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):

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

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

Time discretization ⇒ Optimization algorithm
Discrete Descent/EVI ⇒ Convergence rates

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 $\gamma \leq 1/L$,
- For Backward Euler $\hat{x}_m = x_{m+1}$.

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.:

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

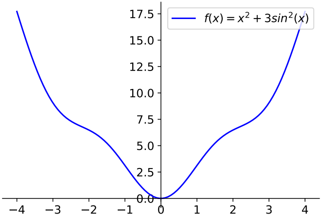
then we can also obtain:

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

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 $\gamma \leq 1/M$,
- For Backward Euler $\hat{x}_m = x_{m+1}$.

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 $\lambda > 0$,

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

Outline

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.

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

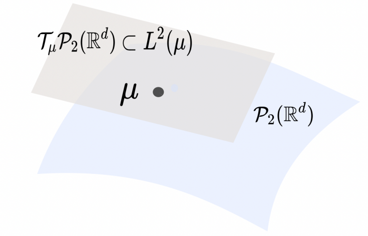
$\mathcal{P}_2(\mathbb{R}^d)$ is endowed with the Wasserstein-2 distance from Optimal transport: $\forall \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

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

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

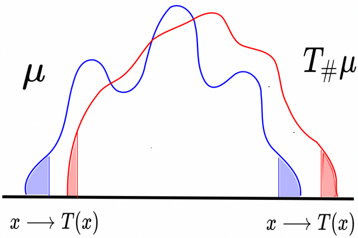
$$\langle f, g \rangle_\mu = \int_{\mathbb{R}^d} \langle f(x), g(x) \rangle_{\mathbb{R}^d} d\mu(x).$$

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

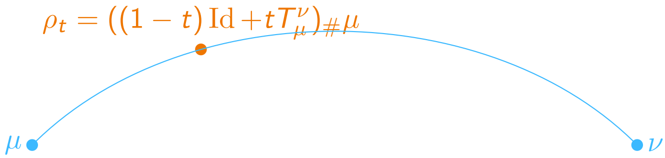
and T_μ^ν is called the Optimal Transport map between μ and ν.

Wasserstein geodesics between μ, ν?

The path

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

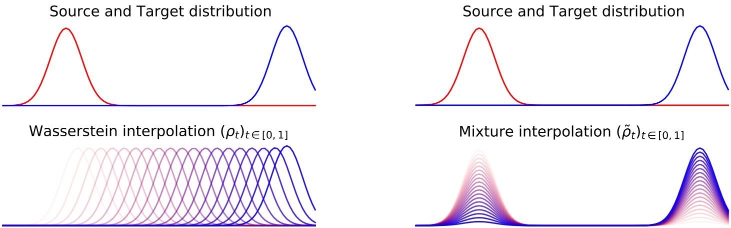
is the Wasserstein geodesic between $\rho_0 = \mu$ and $\rho_1 = \nu$.



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 $x^1, \ldots, x^N$,
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.

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 $\rho_t = ((1-t)\operatorname{Id} + tT_\mu^\nu)_\# \mu$ 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:

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

Outline

Gradient flows on probability distributions?

Recall that we want to approximate a distribution π 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

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

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 $\varphi : \mathbb{R}^d \to \mathbb{R}$ (e.g., $C^\infty(\mathbb{R}^d)$ 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:

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

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

Continuity Equation

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

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

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

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

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

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

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

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 $\mu + \varepsilon\xi \in \mathcal{P}_2(\mathbb{R}^d)$ for a perturbation ξ.

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 $\mu_t = (\mathrm{Id} + th)_{\#}\mu$.
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

$$\mathcal{F}(\mu) = \int \log(\mu(x)) d\mu(x)^1 \longrightarrow \log(\mu) + 1^2 \longrightarrow \nabla \log \mu.$$

[1] The Negative entropy $\mathcal{F}(\mu) = +\infty$ if μ does not have a density.
[2] $(y \log y)' = \log y + 1$

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 [Ambrosio et al., 2008, Th. 10.4.13] 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 will ensure that a Descent property holds.

If we look again at the definition of velocity field, we can see it as a chain rule:

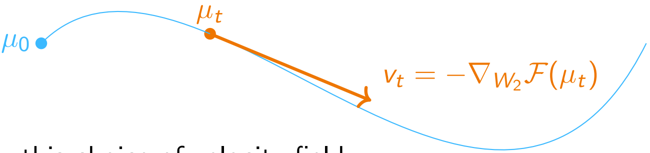
$$\underbrace{\frac{d}{dt}\int \varphi d\mu_t}_{=\mathcal{F}(\mu_t)} = \langle \underbrace{\nabla\varphi}_{=\nabla_{W_2}\mathcal{F}(\mu_t)}, v_t\rangle_{\mu_t}, \text{ for } \mathcal{F}(\mu) = \int \varphi d\mu.$$

Recall that in $\mathbb{R}^d$, a chain rule for $V : \mathbb{R}^d \to \mathbb{R}$ was written $\frac{dV(x_t)}{dt} = \langle \nabla V(x_t), \dot{x}_t\rangle_{\mathbb{R}^d}$.

More generally, we have the following chain rule for any $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to (-\infty, +\infty]$ regular enough and $(v_t)_{t\geq 0}$ velocity field of $(\mu_t)_{t\geq 0}$:

$$\frac{d}{dt}\mathcal{F}(\mu_t) = \langle \nabla_{W_2}\mathcal{F}(\mu_t), v_t\rangle_{\mu_t}.$$

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,

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

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

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

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

Equivalently:

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

Time for Q&A

We now have a break of 5-10 min for questions.

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 $\mathcal{F}$.

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),

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

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

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

Example I $\Longrightarrow$ 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 $\Longrightarrow$ 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 $\Longrightarrow$ 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 <u>other realizations of the Wasser</u>stein 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 $\Delta = \boldsymbol{\nabla} \cdot \nabla$ (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
$\dot{x}_t = -\nabla \log\left(\frac{\mu_t}{\pi}\right)(x_t),\ x_t \sim \mu_t.$

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 [Ambrosio et al., 2008, Chap 11].

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 $\nabla_{W_2} \mathcal{F}$.

Analysis and Design of 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 write **conjectures**:
a Sampling algorithm that is a discretization of the Wasserstein GF of the KL should satisfy a Descent lemma and/or a discrete EVI.

Furthermore, we can **design** Sampling algorithms by discretizing Wasserstein GF.

Outline

Outline

Sampling as Optimization

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

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

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

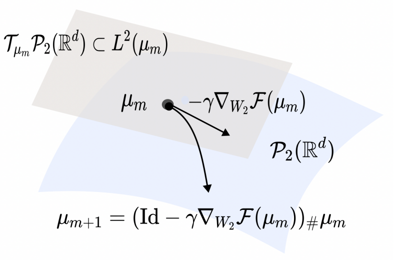
$$\mathcal{F}(\mu) - \underbrace{\mathcal{F}(\pi)}_{\textit{Constant}} = \mathrm{KL}(\mu|\pi).$$

Time discretizations of the Wasserstein GF

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

- Wasserstein gradient descent or Forward Euler (explicit):

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



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):

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

i.e. Backward Euler (implicit) [SKL20].

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

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

  i.e. Backward Euler (implicit) [SKL20].

- Splitting scheme [SKL20] ($\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2$, $\mathcal{F}_2$ 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

Langevin Monte Carlo (LMC) to sample from $\pi \propto \exp(-V)$:

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

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

Intuition: Discretization of Langevin diffusion

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

Can be used for analysis of Langevin
[Durmus and Moulines, 2017, Dalalyan, 2017].

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)

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

$$\mu_{m+1} = \text{flow}_{\gamma,\text{Entropy}}(\mu_{m+\frac{1}{2}})$$

Remark: this splitting scheme is biased.

Consequence: Descent lemma

LMC *almost* decreases the KL [Vempala and Wibisono, 2019], [BCE+22]:

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

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

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

1. we first obtain

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

Nonconvex rates for Langevin Monte Carlo

Nonconvex rates can be obtained using Descent lemma, noting that

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

1. we first obtain

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

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

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

then [Vempala and Wibisono, 2019],

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

Gradient dominance

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

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

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

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).

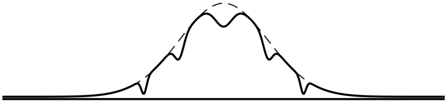


Figure from [Vempala and Wibisono, 2019].

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 $2\gamma Ld$**: 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 $m \geq 0$:

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

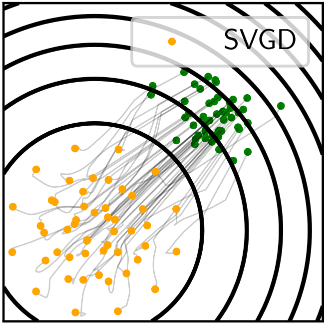
where k is a kernel associated to a **Reproducing Kernel Hilbert Space** H_k.

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 [KAFMA21]. 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,

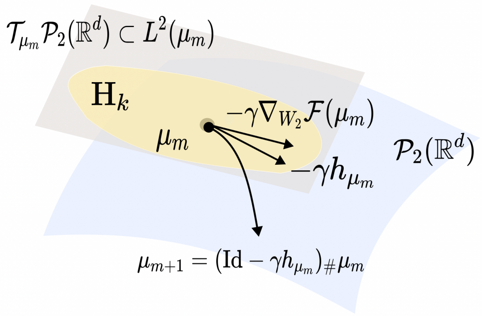
$$h_\mu = P_\mu \nabla \log\left(\frac{\mu}{\pi}\right), \text{ where } P_\mu : L^2(\mu) \to \mathrm{H}_k, f \mapsto \int f(x) k(x, \cdot) d\mu(x).$$

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.



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], [KSA+20, SSR21]:

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

Nonconvex rate for SVGD

Nonconvex rates can be obtained using Descent lemma, noting that

$$\|h_{\mu_m}\|^2_{\mathrm{H}_k} = \left\| P_{\mu_m} \nabla \log \left(\frac{\mu_m}{\pi}\right) \right\|^2_{\mathrm{H}_k} = \mathrm{KSD}^2(\mu_m|\pi).^1$$

We obtain

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

See "A Convergence Theory for SVGD in the Population Limit under Talagrand's Inequality T1" A. Salim, L. Sun, P. Richtárik. ICML 2022. In Session 9 Track 8, Thursday 4:50 PM.

[1] [Liu et al., 2016, Chwialkowski et al., 2016, Gorham and Mackey, 2017].

Outline

Approaches based on the JKO (I)

Recall that the JKO of $\mathcal{F}$ at $\mu_m \in \mathcal{P}_2(\mathbb{R}^d)$ writes

$$\operatorname*{arg\,min}_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu) + \frac{1}{2\gamma} W_2^2(\mu_m, \mu)$$

If $\mathcal{F}$ is the KL

- Blob method considers a regularized KL whose gradient flow can be approximated with particles [Carrillo et al., 2019].
- Restricted Gaussian Oracle [Lee et al., 2021b], [CCSW22] implements in closed-form the JKO of $\mathcal{F}$ if the starting point is a Dirac

Approaches based on the JKO (II)

For a general $\mathcal{F}$ (e.g. the KL), fast methods for computing the JKO are being developed (do not involve discretization of the domain)

- using input-convex neural networks (ICNN) to approximate the transport map [Mokrov et al., 2021, Alvarez-Melis et al., 2021]
- using parametric maps [Fan et al., 2021]
- other approaches based on deep learning [Hwang et al., 2021, Shen et al., 2022]
- change the underlying metric [Peyré, 2015] [Bonet et al., 2021]

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]
- "Proximal" algorithms for non-smooth potentials V (i.e. no gradients of V) [Durmus et al., 2019, Wibisono, 2019], [SKR19, SR20]
- Variance reduction for potentials V written as finite sums [Ding and Li, 2021, Zou et al., 2018, Zou et al., 2019, Dubey et al., 2016, Huang and Becker, 2021], [BCE$^+$22].

Optimization of alternative functionals than the KL

- SVGD can be seen as a gradient flow of the Chi-square divergence [Chewi et al., 2020]
- [KAFMA21] propose to consider the Wasserstein gradient flow the Kernel Stein Discrepancy

Outline

Conclusion

- Sampling can be seen as an optimization problem on a "Wasserstein manifold"
- This point of view enables to leverage its geometry and Wasserstein Gradient Flows (GF)
- 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
- These can be analyzed adapting optimization techniques (e.g. proof of convergence of gradient descent) to the Wasserstein space

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 (last iterates of Langevin Monte Carlo, SVGD stationary particles...)

  See "Accurate Quantization of Measures via Interacting Particle-based Optimization" L. Xu, A. Korba, D. Slepcev. ICML 2022. In Session 3 Track 6, Tuesday 5:40 PM.

Open problems and future directions

Some theoretical questions remain largely open:

- Complexity lower bounds for sampling problems [Lee et al., 2021a, Chewi et al., 2022]
- Convex rates for SVGD/ Stein log Sobolev inequality [Duncan et al., 2019]
- While many works on sampling have mixed first-order optimization and sampling ideas, there may remain some issues regarding implementation or analysis (there is always a balance between both aspects)

... and also practical considerations:

- improving convergence (for π multimodal, high-dimensional)
- improving scaling in the number of particles

Time for Q&A

Questions?

We wish to thank ICML for travel support, and many people for feedback: Pierre-Cyril Aubin-Frankowski, Sebastien Bubeck, Sinho Chewi, Alain Durmus, Eric Moulines, Philippe Rigollet.

References I

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

Alvarez-Melis, D., Schiff, Y., and Mroueh, Y. (2021).
Optimizing functionals on the space of probabilities with input convex neural networks.
arXiv preprint arXiv:2106.00774.

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

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.

Bonet, C., Courty, N., Septier, F., and Drumetz, L. (2021).
Sliced-wasserstein gradient flows.
arXiv preprint arXiv:2110.10972.

References II

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., Craig, K., and Patacchini, F. S. (2019).
A blob method for diffusion.
Calculus of Variations and Partial Differential Equations, 58(2):53.

Chewi, S., Gerber, P., Lu, C., Gouic, T. L., and Rigollet, P. (2022).
The query complexity of sampling from strongly log-concave distributions in one dimension.
Conference on Learning Theory.

Chewi, S., Gouic, T. L., Lu, C., Maunu, T., and Rigollet, P. (2020).
Svgd as a kernelized wasserstein gradient flow of the chi-squared divergence.
In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016).
A kernel test of goodness of fit.
In *International Conference on Machine Learning (ICML)*, pages 2606–2615.

Dalalyan, A. S. (2017).
Theoretical guarantees for approximate sampling from smooth and log-concave densities.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3):651–676.

References III

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

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.

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

References IV

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.

Durmus, A. and Moulines, E. (2017).
Nonasymptotic convergence analysis for the unadjusted Langevin algorithm.
The Annals of Applied Probability, 27(3):1551–1587.

Fan, J., Taghvaei, A., and Chen, Y. (2021).
Variational wasserstein gradient flow.
arXiv preprint arXiv:2112.02424.

Gorham, J. and Mackey, L. (2017).
Measuring sample quality with kernels.
In *International Conference on Machine Learning*, pages 1292–1301. PMLR.

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.

References V

Huang, Z. and Becker, S. (2021).
Stochastic gradient langevin dynamics with variance reduction.
In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Hwang, H. J., Kim, C., Park, M. S., and Son, H. (2021).
The deep minimizing movement scheme.
arXiv preprint arXiv:2109.14851.

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.

Lee, Y. T., Shen, R., and Tian, K. (2021a).
Lower bounds on metropolized sampling methods for well-conditioned distributions.
Advances in Neural Information Processing Systems, 34:18812–18824.

Lee, Y. T., Shen, R., and Tian, K. (2021b).
Structured logconcave sampling with a restricted gaussian oracle.
In *Conference on Learning Theory*, pages 2993–3050. PMLR.

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.

References VI

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 (ICML)*, 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.

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.

Mokrov, P., Korotin, A., Li, L., Genevay, A., Solomon, J. M., and Burnaev, E. (2021).
Large-scale wasserstein gradient flows.
Advances in Neural Information Processing Systems, 34:15243–15256.

References VII

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.

Peyré, G. (2015).
Entropic approximation of wasserstein gradient flows.
SIAM Journal on Imaging Sciences, 8(4):2323–2351.

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

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.

Shen, Z., Wang, Z., Kale, S., Ribeiro, A., Karbasi, A., and Hassani, H. (2022).
Self-consistency of the fokker-planck equation.
arXiv preprint arXiv:2206.00860.

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

References VIII

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.

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.

References I

Krishnakumar Balasubramanian, Sinho Chewi, Murat A Erdogdu, Adil Salim, and Matthew Zhang.
Towards a theory of non-log-concave sampling: First-order stationarity guarantees for langevin monte carlo.
arXiv preprint arXiv:2202.05214, 2022.

Yongxin Chen, Sinho Chewi, Adil Salim, and Andre Wibisono.
Improved analysis for a proximal algorithm for sampling.
arXiv preprint arXiv:2202.06386, 2022.

Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin.
Kernel stein discrepancy descent.
arXiv preprint arXiv:2105.09994, 2021.

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 (NeurIPS)*, 2020.

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

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

References II

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

Adil Salim, Lukang Sun, and Peter Richtárik.
Complexity analysis of stein variational gradient descent under talagrand's inequality t1.
arXiv preprint arXiv:2106.03076, 2021.

Forward method for the KL

Problem: $\nabla_{W_2} \mathrm{KL}(\mu_m|\pi) = \nabla \log\left(\frac{\mu_m}{\pi}\right)$ where μ_n is unknown.

While $\nabla \log \pi$ is known, $\nabla \log \mu_n$ has to be estimated from N particles $X_n^1, \dots, X_n^N$, e.g. with[1] :

1. **Kernel Density Estimation (KDE)**:

$$\mu_m(.) \approx \frac{1}{N} \sum_{i=1}^{N} k(X_m^i - .)$$

Then,

$$-\nabla_{W_2} \mathrm{KL}(\mu_m|\pi)(.) \approx -\left(\nabla V(.) + \frac{\sum_{i=1}^{N} \nabla k(. - X_m^i)}{\sum_{i=1}^{N} k(. - X_m^i)}\right)$$

Remark: it is not the W_2 gradient of some functional (see the next slide)

[1] assume a symmetric, translation invariant kernel

2. **Blob Method** [Carrillo et al., 2019]:
Instead of

$$\mathcal{U}(\mu) = \int \log(\mu(x))d\mu(x),$$

consider

$$\mathcal{U}_k(\mu) = \int \log(k \star \mu(x))d\mu(x), \text{ where } k \star \mu(x) = \int k(x-y)d\mu(y).$$

Then,

$$\frac{\partial \mathcal{U}_k(\mu)}{\partial \mu}(.) = k \star \left(\frac{\mu}{k \star \mu}\right) + \log(k \star \mu)$$

$$\implies \nabla_{W_2}\mathcal{U}_k(\mu) = = \nabla k \star \left(\frac{\mu}{k \star \mu}\right) + \underbrace{\nabla \log(k \star \mu)}_{\frac{\nabla k \star \mu}{k \star \mu}}$$

$$\implies \nabla_{W_2} \mathrm{KL}(\mu_m|\pi)(.) \approx -(\nabla V(.) + \sum_{i=1}^{N} \frac{\nabla k(. - X_m^i)}{\sum_{z=1}^{N} k(X_m^i - X_m^z)} + \frac{\sum_{i=1}^{N} \nabla k(. - X_m^i)}{\sum_{i=1}^{N} k(. - X_m^i)}\right)$$

SVGD trick and the kernel integral operator

We assume $\int_{\mathbb{R}^d \times \mathbb{R}^d} k(x,x) d\mu(x) < \infty$ for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.
$\Longrightarrow \mathrm{H}_k \subset L^2(\mu)$.

For instance assume $\|k(x,.)\|^2_{\mathrm{H}_k} = k(x,x) \leq B^2$, then for $f \in \mathrm{H}_k$

$$\|f\|^2_{L^2(\mu)} = \int \|f(x)\|^2 d\mu(x) = \int \langle f, k(x,.)\rangle^2_{\mathrm{H}_k} d\mu(x)$$
$$\leq \|f\|^2_{\mathrm{H}_k} \int k(x,x) d\mu(x) \leq B^2 \|f\|^2_{\mathrm{H}_k}$$

Then, the injection from $\iota : \mathrm{H}_k \to L^2(\mu)$ admits an adjoint $\iota^\star = S_\mu$, where $S_\mu : L^2(\mu) \to \mathrm{H}_k$ is defined by:

$$S_\mu f(\cdot) = \int k(x,.) f(x) d\mu(x), \quad f \in L^2(\mu).$$

SVGD trick and the kernel integral operator

We assume $\int_{\mathbb{R}^d \times \mathbb{R}^d} k(x,x) d\mu(x) < \infty$ for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$.
$\Longrightarrow \mathrm{H}_k \subset L^2(\mu)$.

For instance assume $\|k(x,.)\|^2_{\mathrm{H}_k} = k(x,x) \leq B^2$, then for $f \in \mathrm{H}_k$

$$\|f\|^2_{L^2(\mu)} = \int \|f(x)\|^2 d\mu(x) = \int \langle f, k(x,.)\rangle^2_{\mathrm{H}_k} d\mu(x)$$
$$\leq \|f\|^2_{\mathrm{H}_k} \int k(x,x) d\mu(x) \leq B^2 \|f\|^2_{\mathrm{H}_k}$$

Then, the injection from $\iota : \mathrm{H}_k \to L^2(\mu)$ admits an adjoint $\iota^\star = S_\mu$, where $S_\mu : L^2(\mu) \to \mathrm{H}_k$ is defined by:

$$S_\mu f(\cdot) = \int k(x,.) f(x) d\mu(x), \quad f \in L^2(\mu).$$

We have for any $f, g \in L_2(\mu) \times \mathrm{H}_k$

$\langle f, \iota g\rangle_{L^2(\mu)} = \langle \iota^* f, g\rangle_{\mathrm{H}_k} = \langle S_\mu f, g\rangle_{\mathrm{H}_k}$. We will denote $P_\mu = \iota \circ S_\mu$.

The Descent property is fundamental

Rewrite the descent property as

$$\frac{dV(x_t)}{dt} \leq -\frac{1}{2}\|\nabla V(x_t)\|^2 - \frac{1}{2}\|\dot{x}_t\|^2.$$

This inequality characterizes the gradient flow [De Giorgi et al., 1980, De Giorgi, 1993].

The Descent property is fundamental

Rewrite the descent property as

$$\frac{dV(x_t)}{dt} \leq -\frac{1}{2}\|\nabla V(x_t)\|^2 - \frac{1}{2}\|\dot{x}_t\|^2.$$

This inequality characterizes the gradient flow [De Giorgi et al., 1980, De Giorgi, 1993].

Indeed, any curve $(x_t)_{t\geq 0}$ satisfying this inequality also satisfies

$$\langle \dot{x}_t, \nabla V(x_t)\rangle \leq -\frac{1}{2}\|\nabla V(x_t)\|^2 - \frac{1}{2}\|\dot{x}_t\|^2,$$

which implies

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

using $\langle a, b\rangle \geq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2 \Longrightarrow a = b$.