

# Sampling Methods: From MCMC to Generative Modeling

## Part II: Gradient flows and Langevin Monte Carlo

Anna Korba

CREST, ENSAE, Institut Polytechnique de Paris

# Outline

## Introduction

Few words about this part

Overview

## Optimization over $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

## Optimization over $\mathcal{P}_2(\mathbb{R}^d)$

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

## Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

## Conclusion

# Outline

## Introduction

Few words about this part

Overview

## Optimization over $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

## Optimization over $\mathcal{P}_2(\mathbb{R}^d)$

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

## Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

## Conclusion

## About this part

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 closely related algorithms.

# Outline

## Introduction

Few words about this part

## Overview

### Optimization over $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

### Optimization over $\mathcal{P}_2(\mathbb{R}^d)$

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

## Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

## Conclusion

## (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]

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

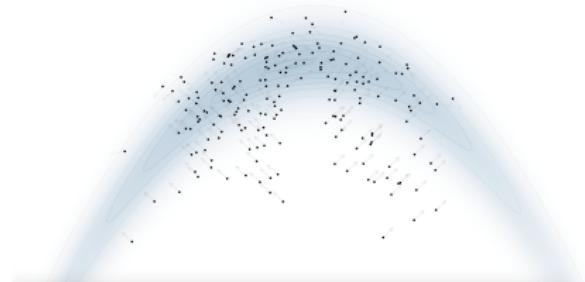


Picture from <https://chi-feng.github.io/mcmc-demo/app.html>.

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

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

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



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:

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

Note that

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

## Sampling as minimization of the KL

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

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

Note that

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

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:

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

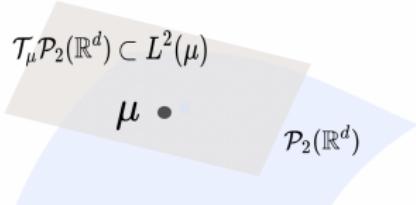
## Sampling as optimization over $\mathcal{P}_2(\mathbb{R}^d)$

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

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

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

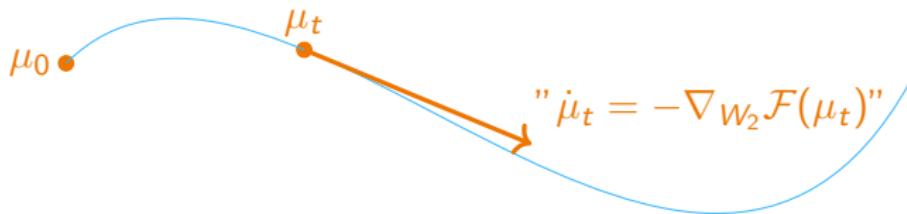



---

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

Starting from some  $\mu_0$ , one can then consider the **Wasserstein gradient flow** of  $\mathcal{F} = \text{KL}(\cdot | \pi)$  over  $\mathcal{P}_2(\mathbb{R}^d)$ , i.e. **path of distributions**  $(\mu_t)_{t \geq 0}$  **decreasing**  $\mathcal{F}$ , to transport  $\mu_0$  to  $\pi$ .

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  $\mathbb{R}^d$

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

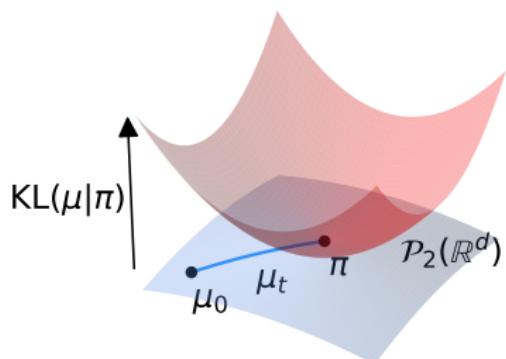
**Discretizing these dynamics**  $(x_t)_{t \geq 0}$  **yields sampling algorithms.**

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

loss of the model  $g(\cdot, x)$

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

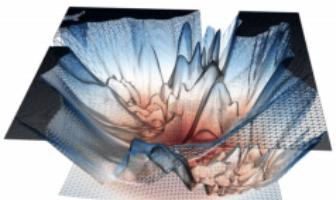
- if  $V$  is convex (e.g.  $g(w, x) = \langle w, x \rangle$  linear),  $\pi$  is "log-concave" and "sampling is easy"



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  $\text{KL}(\cdot | \pi)$  objective inherits convexity properties of  $V$ , i.e.:

- if  $V$  is nonconvex (e.g.  $g(w, x)$  is a neural network),  $\pi$  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]

# Outline

Introduction

Few words about this part

Overview

Optimization over  $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

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

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

Conclusion

# Outline

Introduction

Few words about this part

Overview

Optimization over  $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

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

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

Conclusion

## Gradient

Let  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  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  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  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  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  s.t.  $\nabla V$  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)  $\dot{x}_t = -\nabla V(x_t)$ ,

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

**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  $\lambda$ -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  $\lambda$ -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  $\lambda$ -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  $\lambda$ -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  $\nabla V$ .**

## 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  $\nabla V$ .**

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

Introduction

Few words about this part

Overview

Optimization over  $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

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

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

Conclusion

## 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):

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

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):

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

$$x_{m+1} = \text{prox}_{\gamma G}(x_{m+\frac{1}{2}})$$

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  $x_* \in \arg \min V$  as a fixed point).

## Other time discretizations: splitting schemes

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

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

$$x_{m+1} = \text{prox}_{\gamma G}(x_{m+\frac{1}{2}})$$

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  $x_* \in \arg \min V$  as a fixed point).

**Time discretization  $\Rightarrow$  Optimization algorithm**  
**Discrete Descent/EVI  $\Rightarrow$  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  $\lambda$ , i.e.:

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

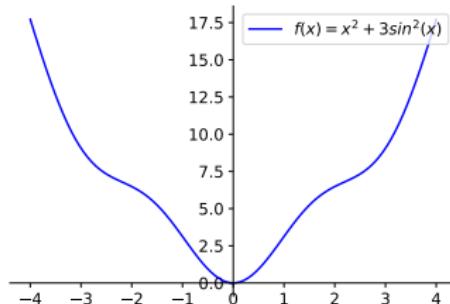
then we can also obtain:

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

## Gradient dominance is more general than convexity

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

- $\lambda$ -Strong convexity  $\Rightarrow$  gradient dominance with the same constant  $\lambda > 0$
- Gradient dominance  $\Rightarrow$  invexity<sup>1</sup>
- Gradient dominance  $\not\Rightarrow$  convexity



---

<sup>1</sup>any local minimum of  $V$  is a global minimum.

## Convex case - Discrete EVI

**Assume  $V$   $\lambda$ -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

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

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

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

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

# Outline

Introduction

Few words about this part

Overview

Optimization over  $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

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

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

Conclusion

# Outline

Introduction

Few words about this part

Overview

Optimization over  $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

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

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

Conclusion

## 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.

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

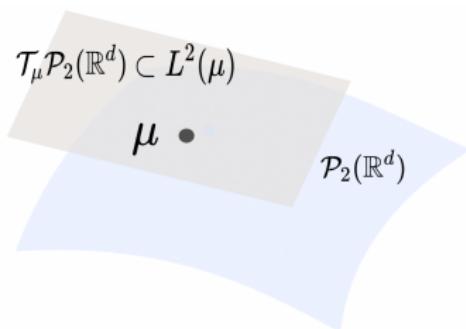
$\mathcal{P}_2(\mathbb{R}^d)$  is endowed with the Wasserstein-2 distance from Optimal transport:  $\forall \mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ ,

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

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

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

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

the space of vector-valued, square-integrable functions w.r.t  $\mu$ .

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

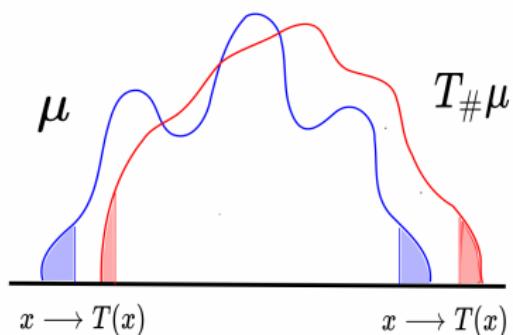
$$\langle f, g \rangle_\mu = \int_{\mathbb{R}^d} \langle f(x), g(x) \rangle_{\mathbb{R}^d} d\mu(x).$$

## Pushforward measure

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

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

$$X \sim \mu \implies T(X) \sim T_\# \mu.$$



**Remark:**  $\text{Id}_\# \mu = \mu$  where  $\text{Id}$  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.  $\mu \ll \text{Leb}$ . Then, there exists a unique  $T_\mu^\nu : \mathbb{R}^d \rightarrow \mathbb{R}^d$  such that

1.  $T_{\mu\#}^\nu \mu = \nu$

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

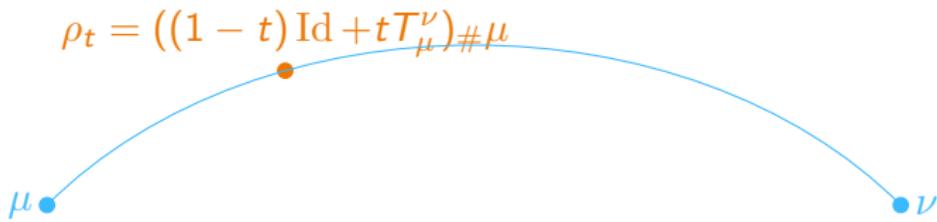
and  $T_\mu^\nu$  is called **the Optimal Transport map** between  $\mu$  and  $\nu$ .

## Wasserstein geodesics between $\mu, \nu$ ?

The path

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

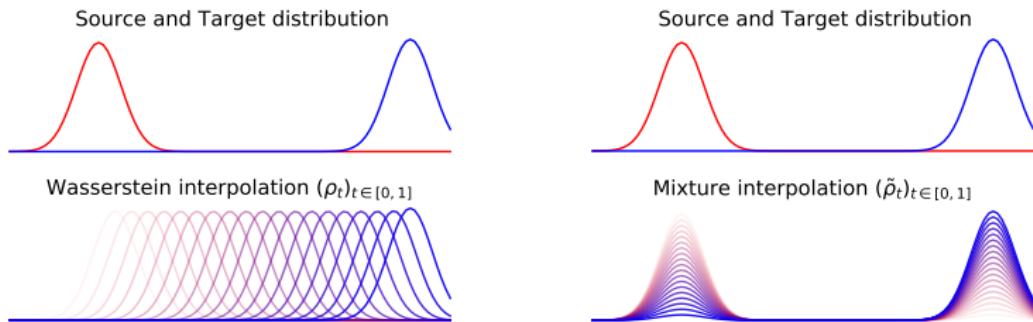
is the Wasserstein geodesic between  $\rho_0 = \mu$  and  $\rho_1 = \nu$ .



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

Figure made with <https://pythonot.github.io/>.

## Convexity along Wasserstein geodesics

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

$\mathcal{F}$   $\lambda$ -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  $\mu$  and  $\nu$ .

## Examples of geo. convex functionals

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

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

## Examples of geo. convex functionals

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

**Proof:** write  $\mathcal{F}(\rho_t)$  along a geodesic  $\rho_t = ((1-t)\text{Id} + tT_\mu^\nu)_\#\mu$  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  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  convex.

**Proof:** write  $\mathcal{F}(\rho_t)$  along a geodesic  $\rho_t = ((1-t)\text{Id} + tT_\mu^\nu)_\#\mu$  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  $\mathcal{F}(\mu) = \text{KL}(\mu|\pi)$ , where  $\pi \propto \exp(-V)$ ,  $V$  convex.

**Proof:**

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

# Outline

Introduction

Few words about this part

Overview

Optimization over  $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

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

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

Conclusion

## Gradient flows on probability distributions?

Recall that we want to approximate a distribution  $\pi$  by solving

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

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

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

In an analog manner, what is the gradient flow of  $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \rightarrow (-\infty, +\infty]$ ? 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  $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}$  (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  $A(x) = (A_1(x), \dots, A_d(x))$ ,  $A : \mathbb{R}^d \rightarrow \mathbb{R}^d$ .

## 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  $A(x) = (A_1(x), \dots, A_d(x))$ ,  $A : \mathbb{R}^d \rightarrow \mathbb{R}^d$ .

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

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

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

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

## RHS: Wasserstein gradient

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

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

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

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

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

## First Variation

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

**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  $\xi$ .

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

# Wasserstein gradient = Gradient of First Variation

Typically<sup>1</sup>,

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

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

---

<sup>1</sup>see [Ambrosio et al., 2008, Th. 10.4.13] for precise statement.

# Wasserstein gradient = Gradient of First Variation

Typically<sup>1</sup>,

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

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

**Proof:** Let  $\mu_t = (\text{Id} + th)_\# \mu$ .

First, expand  $\mu_\varepsilon$  around  $\mu$  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.

---

<sup>1</sup>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  $\mu$ )

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

2. Negative entropy

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

---

<sup>1</sup>The Negative entropy  $\mathcal{F}(\mu) = +\infty$  if  $\mu$  does not have a density.

<sup>2</sup> $(y \log y)' = \log y + 1$

## Wasserstein gradient of KL

More generally, let

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

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

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

By additivity, the Wasserstein gradient of KL is given by<sup>1</sup>

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

---

<sup>1</sup>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:**

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

Recall that in  $\mathbb{R}^d$ , a chain rule for  $V : \mathbb{R}^d \rightarrow \mathbb{R}$  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  $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \rightarrow (-\infty, +\infty]$  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,

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

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

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

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

Equivalently:

$$\frac{\partial \mu_t}{\partial t} = \nabla \cdot (\mu_t \nabla_{W_2}\mathcal{F}(\mu_t)), \quad (2)$$

**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  $V : \mathbb{R}^d \rightarrow \mathbb{R}$ . Consider the dynamics:

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

Let  $\mu_t$  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  $V : \mathbb{R}^d \rightarrow \mathbb{R}$ . Consider the dynamics:

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

Let  $\mu_t$  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),

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

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

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

## Example I $\implies$ WGF of Potential energy

- We have just seen that:

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



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

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

---

<sup>1</sup>The randomness only comes from  $x_0 \sim \mu_0$ .

## Example II $\implies$ 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  $\mu_t$  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  $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \rightarrow (-\infty, +\infty]$ . The dynamics

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

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

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

<sup>1</sup>The randomness only comes from  $x_0 \sim \mu_0$ .

## Example III - Brownian motion

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

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

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

---

<sup>1</sup>Using  $\Delta = \nabla \cdot \nabla$  (Divergence of Gradient = Laplacian).

## Example III - Brownian motion

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

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

Let  $\mu_t$  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<sup>1</sup>

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

---

<sup>1</sup>Using  $\Delta = \nabla \cdot \nabla$  (Divergence of Gradient = Laplacian).

## Example III $\implies$ WGF of (Neg.) Entropy

- We have just seen that:

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



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

- 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

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

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

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

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  $\mu_t$  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  $\mu_t$  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**  $\mu_t$ .

**Proof:** Combine Example I and III.

In this case, the Continuity Equation is the [Fokker-Planck equation](#).

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

## Example IV $\implies$ WGF of the KL

- We have just seen that:

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



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

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

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

## Example IV $\implies$ WGF of the KL

- We have just seen that:

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

$$\downarrow$$

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

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

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

**Remark:** Another realization is given by

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

# Outline

Introduction

Few words about this part

Overview

Optimization over  $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

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

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

Conclusion

## 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

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

This is a fundamental property of the Wasserstein gradient flow [Ambrosio et al., 2008, Chap 11].

## Evolution Variational Inequality (EVI)

**Assume  $\mathcal{F}$   $\lambda$ -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

Introduction

Few words about this part

Overview

Optimization over  $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

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

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

Conclusion

# Outline

## Introduction

Few words about this part

Overview

## Optimization over $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

## Optimization over $\mathcal{P}_2(\mathbb{R}^d)$

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

## Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

## Conclusion

## Sampling as Optimization

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

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

## Sampling as Optimization

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

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

where

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

satisfies

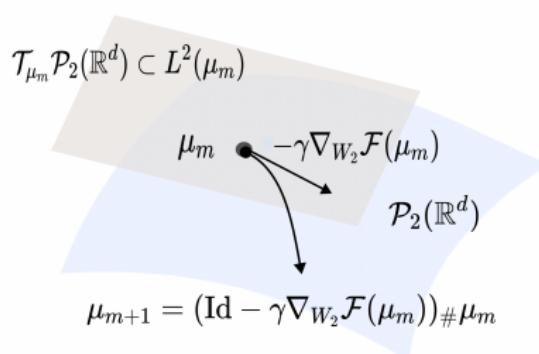
$$\mathcal{F}(\mu) - \underbrace{\mathcal{F}(\pi)}_{\text{Constant}} = \text{KL}(\mu|\pi).$$

## Time discretizations of the Wasserstein GF

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

- Wasserstein gradient descent or Forward Euler (explicit):

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



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

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

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

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

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

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

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

- Splitting scheme [SKL20] ( $\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2$ ,  $\mathcal{F}_2$  geo. convex):

$$\mu_{m+\frac{1}{2}} = (\text{Id} - \gamma \nabla_{W_2} \mathcal{F}_1(\mu_m)) \# \mu_m$$

$$\mu_{m+1} = \text{JKO}_{\gamma \mathcal{F}_2} \left( \mu_{m+\frac{1}{2}} \right)$$

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

# Outline

## Introduction

Few words about this part

Overview

## Optimization over $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

## Optimization over $\mathcal{P}_2(\mathbb{R}^d)$

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

## Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

## Conclusion

## 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

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

Let  $x_m \sim \mu_m$ .

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

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

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

**Remark: this splitting scheme is biased.**

## Consequence: Descent lemma

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

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

where  $\hat{\mu}_m$  "between"  $\mu_m$  and  $\mu_{m+1}$ .

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

## Nonconvex rates for Langevin Monte Carlo

Nonconvex rates can be obtained using Descent lemma, noting that

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

1. we first obtain

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

## Nonconvex rates for Langevin Monte Carlo

Nonconvex rates can be obtained using Descent lemma, noting that

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

1. we first obtain

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

2. If  $\pi$  satisfies Log Sobolev inequality with  $\lambda$ , i.e.:

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

then [Vempala and Wibisono, 2019],

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

## Gradient dominance

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

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

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

## Non log concave $\pi$ 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.  $\pi$  is (1-)strongly log-concave.

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

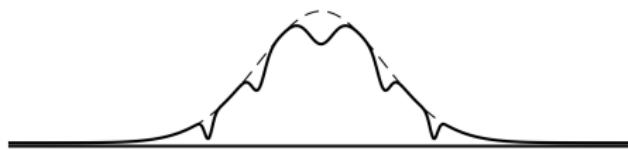


Figure from [Vempala and Wibisono, 2019].

## Convex case - Discrete EVI

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

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

## Convex case - Discrete EVI

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

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

**Error term  $2\gamma Ld$ :** LMC is biased, i.e.,  $\pi$  is not an invariant distribution.

## Convex rates for Langevin Monte Carlo

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

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

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

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

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

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

# Outline

## Introduction

Few words about this part

Overview

## Optimization over $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

## Optimization over $\mathcal{P}_2(\mathbb{R}^d)$

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

## Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

## Conclusion

## 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  $m \geq 0$ :

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

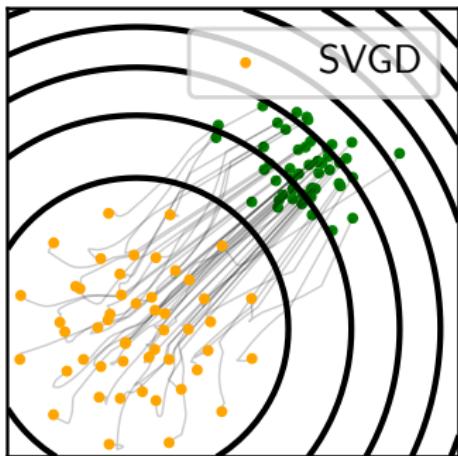
where  $k$  is a kernel associated to a **Reproducing Kernel Hilbert Space**  $H_k$ .

## Reproducing kernel Hilbert Space

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

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

## 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,

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

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

Actually,

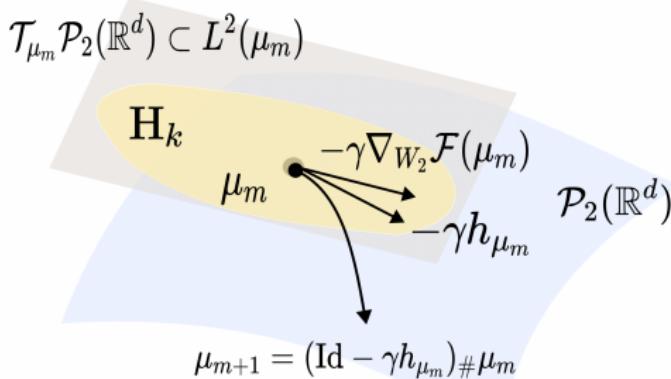
$$h_\mu = P_\mu \nabla \log \left( \frac{\mu}{\pi} \right), \text{ where } P_\mu : L^2(\mu) \rightarrow \mathbf{H}_k, f \mapsto \int f(x)k(x, \cdot) d\mu(x).$$

## Gradient descent interpretation

A Taylor expansion around  $\mu$  for  $h \in H_k$ , if  $\mu$  has a density yields [Liu, 2017]:

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

Therefore,  $h_\mu$  plays the role of the Wasserstein gradient in  $H_k$ .



## Consequence: Descent lemma

We study

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

when  $\mu_m$  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<sup>+</sup>20, SSR21]:

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

## Nonconvex rate for SVGD

Nonconvex rates can be obtained using Descent lemma, noting that

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

We obtain

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

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.

<sup>1</sup>[Liu et al., 2016, Chwialkowski et al., 2016, Gorham and Mackey, 2017].

# Outline

## Introduction

Few words about this part

Overview

## Optimization over $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

## Optimization over $\mathcal{P}_2(\mathbb{R}^d)$

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

## Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

## Conclusion

## Approaches based on the JKO (I)

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

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

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, ?], [BCE<sup>+</sup>22].

# Outline

## Introduction

Few words about this part

Overview

## Optimization over $\mathbb{R}^d$

Euclidean Gradient Flow

Time discretizations of the Euclidean gradient flow

## Optimization over $\mathcal{P}_2(\mathbb{R}^d)$

Geometry of  $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Definition of Wasserstein gradient flows

Properties of Wasserstein gradient flows

## Sampling algorithms

Optimizing the KL

Langevin Monte Carlo

Stein Variational Gradient Descent (SVGD)

Other examples

## Conclusion

## 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...)

## 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  $\pi$  multimodal, high-dimensional)
- improving scaling in the number of particles

# References |

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

## References III

-  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*.
-  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.

## References IV

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

# References V

-  **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.
-  **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.

## References VI

-  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.
-  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.

## References VII

-  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*.
-  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.

## References VIII

-  **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.