# Sampling as optimization of the relative entropy over the space of measures

## Non asymptotic analysis of SVGD and the Forward-Backward scheme


Anna Korba

CREST, ENSAE, Paris, France


Séminaire d'optimisation
Institut Henri Poincaré
September 2020

Joint work with Adil Salim (KAUST), Michael Arbel (Gatsby Unit, UCL), Giulia Luise (CS Department, UCL), Arthur Gretton (Gatsby Unit, UCL).

# Outline

**Problem :** Sample from a target distribution $\pi$ over $\mathbb{R}^d$, whose density w.r.t. Lebesgue is written :

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

where $V : \mathbb{R}^d \to \mathbb{R}$ is the potential function.

**Problem :** Sample from a target distribution $\pi$ over $\mathbb{R}^d$, whose density w.r.t. Lebesgue is written :

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

where $V : \mathbb{R}^d \to \mathbb{R}$ is the potential function.

**Motivation : Bayesian statistics.**

- Let $\mathcal{D} = (x_i, y_i)_{i=1,\dots,N}$ observed data.
- Assume an underlying model parametrized by $\theta$ (e.g. $p(y|x,\theta)$ gaussian)
  $\implies$ Likelihood: $p(\mathcal{D}|\theta) = \prod_{i=1}^{N} p(y_i|\theta, x_i)$
- The parameter $\theta \sim p$ the prior distribution.

Bayes' rule : $p(\theta|\mathcal{D}) = \dfrac{p(\mathcal{D}|\theta)p(\theta)}{Z}$ where $Z = \int_{\mathbb{R}^d} p(\mathcal{D}|\theta)p(\theta)d\theta$.

*How to sample from $\theta \mapsto p(\theta|\mathcal{D})$? (Z unknown).*

# The relative entropy/Kullback-Leibler divergence

For any $\mu, \pi \in \mathcal{P}(\mathbb{R}^d)$, the Kullback-Leibler divergence of $\mu$ w.r.t. $\pi$ is defined by

$$\mathrm{KL}(\mu|\pi) = \int_{\mathbb{R}^d} \log\left(\frac{d\mu}{d\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi$$

and is $+\infty$ otherwise.

We will consider the functional $\mathrm{KL}(\cdot|\pi) : \mathcal{P}(\mathbb{R}^d) \to [0, +\infty]$.

# Sampling as optimization of the KL

The target distribution $\pi$ is solution of :

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{KL}(\mu|\pi) \tag{1}$$

# Sampling as optimization of the KL

The target distribution $\pi$ is solution of :

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{KL}(\mu|\pi) \tag{1}$$

1. **Variants of Langevin Monte Carlo (LMC)**

   [Dalalyan, 2017], [Durmus and Moulines, 2016], [Durmus et al., 2019], [Salim and Richtárik, 2020]

   - generates a Markov chain whose law converges to $\pi$
   - corresponds to a time-discretization of the gradient flow of the KL
   - rates of convergence deteriorates quickly in high dimensions

# Sampling as optimization of the KL

The target distribution $\pi$ is solution of :

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{KL}(\mu|\pi) \tag{1}$$

1. **Variants of Langevin Monte Carlo (LMC)**

   [Dalalyan, 2017], [Durmus and Moulines, 2016], [Durmus et al., 2019], [Salim and Richtárik, 2020]

   - generates a Markov chain whose law converges to $\pi$
   - corresponds to a time-discretization of the gradient flow of the KL
   - rates of convergence deteriorates quickly in high dimensions

2. **Variational Inference (VI):**

   [Alquier and Ridgway, 2017], [Zhang et al., 2018]

   - restrict the search space in (1) to a parametric family
   - tractable in the large scale setting
   - only returns an approximation of $\pi$

# Sampling as optimization of the KL

The target distribution $\pi$ is solution of :

$$\min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \mathrm{KL}(\mu|\pi) \tag{1}$$

1. **Variants of Langevin Monte Carlo (LMC)**

   [Dalalyan, 2017], [Durmus and Moulines, 2016], [Durmus et al., 2019], [Salim and Richtárik, 2020]

   - generates a Markov chain whose law converges to $\pi$
   - corresponds to a time-discretization of the gradient flow of the KL
   - rates of convergence deteriorates quickly in high dimensions

2. **Variational Inference (VI):**

   [Alquier and Ridgway, 2017], [Zhang et al., 2018]

   - restrict the search space in (1) to a parametric family
   - tractable in the large scale setting
   - only returns an approximation of $\pi$

$\Longrightarrow$ Other algorithms from the gradient flow of the KL...

Sampling can be written as an optimization problem on $\mathcal{P}$ :

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

A general strategy to minimize a function is to run the gradient flow dynamics.

$\implies$ **Wasserstein GF find such *continuous* path on the space of distributions (equipped with the Wasserstein geometry).**

Sampling can be written as an optimization problem on $\mathcal{P}$ :

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

A general strategy to minimize a function is to run the gradient flow dynamics.

$\implies$ **Wasserstein GF find such *continuous* path on the space of distributions (equipped with the Wasserstein geometry).**

Different algorithms result from different time-space discretizations.

1. What is the Wasserstein GF of the relative entropy?
2. Tools for non-asymptotic analysis
3. The Wasserstein Proximal Gradient Algorithm

   [Wibisono, 2018][Salim et al., 2020]
4. Stein Variational Gradient Descent

   [Liu and Wang, 2016][Korba et al., 2020]

# Outline

# Setting - The Wasserstein space

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

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

# Setting - The Wasserstein space

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

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

# Setting - The Wasserstein space

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

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

$\mathcal{P}$ is endowed with the Wasserstein-2 distance from **Optimal transport** :

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

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

# $W_2$ geodesics

**Def (pushforward) :** Let $\mu \in \mathcal{P}$, $T : \mathbb{R}^d \to \mathbb{R}^d$. The pushforward measure $T_\# \mu$ is characterized by:

- $\forall$ B meas. set, $T_\# \mu(B) = \mu(T^{-1}(B))$
- $x \sim \mu$, $T(x) \sim T_\# \mu$

# $W_2$ geodesics

**Def (pushforward) :** Let $\mu \in \mathcal{P}$, $T : \mathbb{R}^d \to \mathbb{R}^d$. The pushforward measure $T_{\#}\mu$ is characterized by:

- $\forall$ B meas. set, $T_{\#}\mu(B) = \mu(T^{-1}(B))$
- $x \sim \mu$, $T(x) \sim T_{\#}\mu$

**Brenier's theorem :** Let $\mu, \nu \in \mathcal{P}$ s.t. $\mu \ll Leb$. Then $\exists$ $T_\mu^\nu : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $T^\nu_{\mu\#}\mu = \nu$ and :

$$W_2^2(\mu, \nu) = \|I - T_\mu^\nu\|^2_{L_2(\mu)} = \inf_{T \in L_2(\mu)} \int (x - T(x))^2 d\mu(x)$$

Also if $\nu \ll Leb$, then $T_\mu^\nu \circ T_\nu^\mu = I$ $\nu$-a.e. and $T_\nu^\mu \circ T_\mu^\nu = I$ $\mu$-a.e.

# $W_2$ geodesics

**Def (pushforward) :** Let $\mu \in \mathcal{P}$, $T : \mathbb{R}^d \to \mathbb{R}^d$. The pushforward measure $T_{\#}\mu$ is characterized by:

- $\forall$ B meas. set, $T_{\#}\mu(B) = \mu(T^{-1}(B))$
- $x \sim \mu$, $T(x) \sim T_{\#}\mu$

**Brenier's theorem :** Let $\mu, \nu \in \mathcal{P}$ s.t. $\mu \ll Leb$. Then $\exists$ $T_\mu^\nu : \mathbb{R}^d \to \mathbb{R}^d$ s.t. $T_{\mu\#}^{\nu}\mu = \nu$ and :

$$W_2^2(\mu, \nu) = \|I - T_\mu^\nu\|^2_{L_2(\mu)} = \inf_{T \in L_2(\mu)} \int (x - T(x))^2 d\mu(x)$$

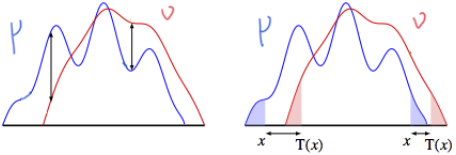
Also if $\nu \ll Leb$, then $T_\mu^\nu \circ T_\nu^\mu = I$ $\nu$-a.e. and $T_\nu^\mu \circ T_\mu^\nu = I$ $\mu$-a.e.

## $W_2$ geodesics?

$\rho(0) = \mu, \rho(1) = \nu.$

$\rho(t) = ((1-t)I + tT_\mu^\nu)_{\#}\mu$

$\neq \rho(t) = \underbrace{(1-t)\mu + t\nu}_{\text{mixture}}$



# What is the (Wasserstein) gradient flow of the relative entropy?

The **Wasserstein gradient flow of the functional** $\mathrm{KL}(\cdot|\pi)$ is a curve $\mu : [0,\infty] \to \mathcal{P},\ t \mapsto \mu_t$ that satisfies:

$$\boxed{\frac{\partial \mu_t}{\partial t} = \text{``} - \nabla_{W_2} \mathrm{KL}(\mu_t|\pi)\text{''}}$$

# A dual point of view

Consider the gradient flow

$$x'(t) = -\nabla V(x(t))$$

for $V : \mathbb{R}^d \to \mathbb{R}$ smooth and assume $x(0)$ random with density $\mu_0$. What is the dynamics of the density $\mu_t$ of $x(t)$ ?

[1] $\mathcal{C}^\infty$ function from $\mathbb{R}^d$ to $\mathbb{R}$ with compact support.

# A dual point of view

Consider the gradient flow

$$x'(t) = -\nabla V(x(t))$$

for $V : \mathbb{R}^d \to \mathbb{R}$ smooth and assume $x(0)$ random with density $\mu_0$. What is the dynamics of the density $\mu_t$ of $x(t)$ ?

Let $\phi : \mathbb{R}^d \to \mathbb{R}$ a test function[1].

$$\frac{d}{dt}\mathbb{E}(\phi(x(t))) = \int \phi(x)\frac{\partial \mu_t}{\partial t}(x)dx.$$

and

$$\frac{d}{dt}\mathbb{E}(\phi(x(t))) = -\int \langle \nabla \phi, \nabla V\rangle \mu_t(x)dx = \int \phi(x) div(\mu_t \nabla V)(x)dx,$$

Therefore,

$$\frac{\partial \mu_t}{\partial t} = div(\mu_t \nabla V).$$

[1] $\mathcal{C}^\infty$ function from $\mathbb{R}^d$ to $\mathbb{R}$ with compact support.

# Continuity equations

Let $T > 0$. Consider a family $\mu : [0, T] \to \mathcal{P}, t \mapsto \mu_t$. It satisfies a continuity equation if there exists $(V_t)_{t\in[0,T]}$ such that $V_t \in L^2(\mu_t)$ and distributionnally:

$$\boxed{\frac{\partial \mu_t}{\partial t} + div(\mu_t V_t) = 0}$$

*Density $\mu_t$ of particles $x_t \in \mathbb{R}^d$ driven by a vector field $V_t$:*

$$\frac{dx_t}{dt} = V_t(x_t)$$

**Riemannian interpretation** [Otto, 2001] : tangent space of $\mathcal{P}$ at $\mu_t$
$\mathcal{T}_{\mu_t}\mathcal{P} \subset L^2(\mu_t) = \{f : \mathbb{R}^d \to \mathbb{R}^d, \int \|f(x)\|^2 d\mu_t(x) < \infty\}$.
$L^2(\mu_t)$ is a Hilbert space equipped with $\langle \cdot, \cdot \rangle_{\mu_t}$ and $\| \cdot \|_{\mu_t}$.

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

Let $\mathcal{F} : \mathcal{P} \to \mathbb{R} \cup \{+\infty\}$ a regular functional.

The differential of $\mu \mapsto \mathcal{F}(\mu)$ evaluated at $\mu \in \mathcal{P}$ is the unique function $\frac{\partial \mathcal{F}(\mu)}{\partial \mu} : \mathbb{R}^d \to \mathbb{R}$ s. t. for any $\mu, \mu' \in \mathcal{P}$, $\mu' - \mu \in \mathcal{P}$:

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon}(\mathcal{F}(\mu + \epsilon(\mu' - \mu)) - \mathcal{F}(\mu)) = \int_{\mathbb{R}^d} \frac{\partial \mathcal{F}(\mu)}{\partial \mu}(x)(d\mu' - d\mu)(x).$$

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

Let $\mathcal{F} : \mathcal{P} \to \mathbb{R} \cup \{+\infty\}$ a regular functional.

The differential of $\mu \mapsto \mathcal{F}(\mu)$ evaluated at $\mu \in \mathcal{P}$ is the unique function $\frac{\partial \mathcal{F}(\mu)}{\partial \mu} : \mathbb{R}^d \to \mathbb{R}$ s. t. for any $\mu, \mu' \in \mathcal{P}$, $\mu' - \mu \in \mathcal{P}$:

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon}(\mathcal{F}(\mu + \epsilon(\mu' - \mu)) - \mathcal{F}(\mu)) = \int_{\mathbb{R}^d} \frac{\partial \mathcal{F}(\mu)}{\partial \mu}(x)(d\mu' - d\mu)(x).$$

Then $\mu : [0, \infty] \to \mathcal{P}, t \mapsto \mu_t$ satisfies a Wasserstein gradient flow of $\mathcal{F}$ if distributionnally:

$$\boxed{\frac{\partial \mu_t}{\partial t} = div \left( \mu_t \nabla \frac{\partial \mathcal{F}(\mu_t)}{\partial \mu_t} \right),}$$

where $\nabla_W \mathcal{F}(\mu) := \nabla \frac{\partial \mathcal{F}(\mu)}{\partial \mu} \in L^2(\mu)$ is called the Wasserstein gradient of $\mathcal{F}$.

# Wasserstein gradient flow of the relative entropy

We consider the functional $\mathrm{KL}(\cdot|\pi) : \mathcal{P} \to [0, +\infty]$.

For any $\mu \in \mathcal{P}, \mu \ll \pi$, the differential of $\mathrm{KL}(\cdot|\pi)$ evaluated at $\mu$, $\frac{\partial \mathrm{KL}(\mu|\pi)}{\partial \mu} : \mathbb{R}^d \to \mathbb{R}$ is the function

$$\log\left(\frac{\mu}{\pi}\right) + 1 : \mathbb{R}^d \to \mathbb{R}.$$

Hence, the Wassertein GF of $\mathrm{KL}(\cdot|\pi)$ is written :

$$\boxed{\frac{\partial \mu_t}{\partial t} - div(\mu_t \underbrace{\nabla \frac{\partial \mathrm{KL}(\mu_t|\pi)}{\partial \mu}}_{\nabla \log\left(\frac{\mu_t}{\pi}\right)}) = 0}$$

where $\mu_t$ is a smooth positive density evolving over time.

# The KL as a composite functional

$$\mathrm{KL}(\mu|\pi) = \int \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi, +\infty \text{ else.}$$

# The KL as a composite functional

$$\mathrm{KL}(\mu|\pi) = \int \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi, +\infty \text{ else.}$$

It is written as **a composite functional** :

$$\mathrm{KL}(\mu|\pi) = \underbrace{\int V(x) d\mu(x)}_{\mathcal{E}_V(\mu) \text{ external potential}} + \underbrace{\int \log\left(\frac{\mu}{Leb}\right) d\mu(x)}_{\mathcal{U}(\mu) \text{ negative entropy}} + cte$$

# The KL as a composite functional

$$\mathrm{KL}(\mu|\pi) = \int \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi, +\infty \text{ else.}$$

It is written as **a composite functional** :

$$\mathrm{KL}(\mu|\pi) = \underbrace{\int V(x) d\mu(x)}_{\color{red}{\mathcal{E}_V(\mu) \text{ external potential}}} + \underbrace{\int \log(\frac{\mu}{Leb}) d\mu(x)}_{\color{green}{\mathcal{U}(\mu) \text{ negative entropy}}} + cte$$

$W_2$ gradient flow of the KL is the Fokker-Planck equation:

$$\frac{\partial \mu_t}{\partial t} = div(\mu_t \underbrace{\nabla \log\left(\frac{\mu_t}{\pi}\right)}_{\nabla_W \mathrm{KL}(\mu_t|\pi)}) = div(\mu_t \underbrace{\nabla V}_{\color{red}{\nabla_W \mathcal{E}_V(\mu)}}) + div(\mu_t \underbrace{\nabla \log(\mu_t)}_{\color{green}{\mathcal{U}(\mu)}})$$

# The KL as a composite functional

$$\mathrm{KL}(\mu|\pi) = \int \log\left(\frac{\mu}{\pi}(x)\right) d\mu(x) \text{ if } \mu \ll \pi, +\infty \text{ else.}$$

It is written as **a composite functional** :

$$\mathrm{KL}(\mu|\pi) = \underbrace{\int V(x) d\mu(x)}_{\mathcal{E}_V(\mu) \text{ external potential}} + \underbrace{\int \log\left(\frac{\mu}{Leb}\right) d\mu(x)}_{\mathcal{U}(\mu) \text{ negative entropy}} + cte$$

$W_2$ gradient flow of the KL is the Fokker-Planck equation:

$$\frac{\partial \mu_t}{\partial t} = div(\mu_t \underbrace{\nabla \log\left(\frac{\mu_t}{\pi}\right)}_{\nabla_W \mathrm{KL}(\mu_t|\pi)}) = div(\mu_t \underbrace{\nabla V}_{\nabla_W \mathcal{E}_V(\mu)}) + div(\mu_t \underbrace{\nabla \log(\mu_t)}_{\mathcal{U}(\mu)})$$

It is the continuity equation ($X_t \sim \mu_t$) of the Langevin dynamics :

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

where $(B_t)$ is the brownian motion in $\mathbb{R}^d$.

# Gradient flow of the entropy

The gradient flow of the negative entropy $\mathcal{U}(\mu)$ is the heat equation

$$\frac{\partial \mu_t}{\partial t} = \Delta \mu_t$$

This has an exact solution which is the heat flow $\mu_t = \mu_0 * \mathcal{N}(0, 2tI_d)$.

In space, this is implemented via the addition of Gaussian noise [2]

$$X_t = X_0 + \sqrt{2t}Z \tag{2}$$

where $Z \sim \mathcal{N}(0, I_d)$ and $Z$ independent of $X_0$.

Some time-discretizations of the KL gradient flow...

[2] The true solution of the heat flow is the Brownian motion in space. However, at each time, the solution has the same distribution as (2)

# Unadjusted Langevin Algorithm (ULA)

$$X_{n+1} = X_n - \gamma \nabla V(X_n) + \sqrt{2\gamma}\xi_n \text{ where } \xi_n \sim \mathcal{N}(0, I_d)$$

and $\gamma > 0$ is a step-size.

# Unadjusted Langevin Algorithm (ULA)

$$X_{n+1} = X_n - \gamma \nabla V(X_n) + \sqrt{2\gamma}\xi_n \text{ where } \xi_n \sim \mathcal{N}(0, I_d)$$

and $\gamma > 0$ is a step-size.

**Problem : ULA is biased (has stationary distribution $\pi_\gamma \neq \pi$).**

# Unadjusted Langevin Algorithm (ULA)

$$X_{n+1} = X_n - \gamma \nabla V(X_n) + \sqrt{2\gamma}\xi_n \text{ where } \xi_n \sim \mathcal{N}(0, I_d)$$

and $\gamma > 0$ is a step-size.

**Problem : ULA is biased (has stationary distribution $\pi_\gamma \neq \pi$).**

We can write ULA as the composition :

$$Y_{n+1} = X_n - \gamma \nabla V(X_n) \quad \text{gradient descent/forward method for V}$$

$$X_{n+1} = Y_{n+1} + \sqrt{2\gamma}\xi_n \quad \text{exact solution for the heat flow}$$

$\Longrightarrow$ Forward-Flow discretization

# Unadjusted Langevin Algorithm (ULA)

$$X_{n+1} = X_n - \gamma \nabla V(X_n) + \sqrt{2\gamma}\xi_n \text{ where } \xi_n \sim \mathcal{N}(0, I_d)$$

and $\gamma > 0$ is a step-size.

**Problem : ULA is biased (has stationary distribution $\pi_\gamma \neq \pi$).**

We can write ULA as the composition :

$$Y_{n+1} = X_n - \gamma \nabla V(X_n) \quad \text{gradient descent/forward method for V}$$

$$X_{n+1} = Y_{n+1} + \sqrt{2\gamma}\xi_n \quad \text{exact solution for the heat flow}$$

$\Longrightarrow$ Forward-Flow discretization

In the space of measures $\mathcal{P}$:

$$\nu_{n+1} = (I - \gamma \nabla V)_{\#}\mu_n \quad \text{gradient descent for } \mathcal{E}_V$$

$$\mu_{n+1} = \mathcal{N}(0, 2\gamma I) * \nu_{n+1} \quad \text{exact gradient flow for } \mathcal{U}$$

**This Forward-flow discretization is biased** [Wibisono, 2018]**.**

# Other (unbiased) time discretizations

1. Forward method :

$$\mu_{n+1} = exp_{\mu_n}(-\gamma \nabla_{W_2} \mathrm{KL}(\mu_n|\pi)) = \left(I - \gamma \nabla \log\left(\frac{\mu_n}{\pi}\right)\right)_{\#} \mu_n$$

 where $exp_\mu : L^2(\mu) \to \mathcal{P}, \phi \mapsto (I + \phi)_{\#}\mu$,
 and which corresponds in $\mathbb{R}^d$ to:

$$X_{n+1} = X_n - \gamma \nabla \log\left(\frac{\mu_n}{\pi}\right)(X_n) \sim \mu_{n+1}$$

2. Forward-Backward method :

$$\nu_{n+1} = (I - \gamma \nabla V)_{\#}\mu_n$$
$$\mu_{n+1} = JKO_{\gamma\mathcal{U}}(\nu_{n+1})$$

$$\text{where } JKO_{\gamma\mathcal{U}}(\nu_{n+1}) = \underset{\mu \in \mathcal{P}}{\operatorname{argmin}} \, \mathcal{U}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \nu_{n+1}).$$

 It is unbiased because the backward method is the adjoint of the forward method, so the minimizer is conserved.

# Outline

# Euclidean Gradient Flows

Let $V : \mathbb{R}^d \to \mathbb{R}$ smooth. The (Euclidean) Gradient Flow (GF) of $V$ is given by the solution to

$$\boxed{x'(t) = -\nabla V(x(t))}$$

Continuous time version of gradient descent:

$$\frac{x_{n+1} - x_n}{\gamma} = -\nabla V(x_n)$$

The GF tends to minimize $V$. Let $x^*$ a minimizer of $V$.

# Lyapunov functions for the GF

1. Denote $\mathcal{L}(t) = V(x(t)) - V(x^*)$.

$$\mathcal{L}'(t) = \langle x'(t), \nabla V(x(t)) \rangle = -\|\nabla V(x(t))\|^2 \leq 0,$$

therefore $V(x(t)) \searrow$. Moreover,

$$\frac{1}{T} \int_0^T \|\nabla V(x(t))\|^2 dt \leq \frac{V(x(0)) - V(x^*)}{T}.$$

# Lyapunov functions for the GF

1. Denote $\mathcal{L}(t) = V(x(t)) - V(x^*)$.

$$\mathcal{L}'(t) = \langle x'(t), \nabla V(x(t)) \rangle = -\|\nabla V(x(t))\|^2 \leq 0,$$

therefore $V(x(t)) \searrow$. Moreover,

$$\frac{1}{T} \int_0^T \|\nabla V(x(t))\|^2 dt \leq \frac{V(x(0)) - V(x^*)}{T}.$$

2. Denote $\mathcal{L}_c(t) = \|x(t) - x^*\|^2$. Assume $V$ convex.

$$\mathcal{L}_c'(t) = 2\langle x(t) - x^*, -\nabla V(x(t)) \rangle \leq -2(V(x(t)) - V(x^*)) \leq 0,$$

therefore $\|x(t) - x^*\|^2 \searrow$. Moreover,

$$V(x(T)) - V(x^*) \leq \frac{1}{T} \int_0^T (V(x(t)) - V(x^*))dt \leq \frac{\|x(0) - x^*\|^2}{2T}.$$

# Lyapunov functions for the Wasserstein GF

Let $\mathcal{F} : \mathcal{P} \to \mathbb{R} \cup \{+\infty\}$ a regular functional and $\pi$ a minimizer of $\mathcal{F}$. The Wasserstein GF tends to minimize $\mathcal{F}$:

$$\frac{\partial \mu_t}{\partial t} = div(\mu_t \underbrace{\nabla_W \mathcal{F}(\mu_t)}_{v_t})$$

# Lyapunov functions for the Wasserstein GF

Let $\mathcal{F} : \mathcal{P} \to \mathbb{R} \cup \{+\infty\}$ a regular functional and $\pi$ a minimizer of $\mathcal{F}$. The Wasserstein GF tends to minimize $\mathcal{F}$:

$$\frac{\partial \mu_t}{\partial t} = div(\mu_t \underbrace{\nabla_W \mathcal{F}(\mu_t)}_{V_t})$$

Denote $\mathcal{L}(t) = \mathcal{F}(\mu_t) - \mathcal{F}(\pi)$.

$$\mathcal{L}'(t) = \langle V_t, \nabla_W \mathcal{F}(\mu_t) \rangle_{\mu_t} = -\|\nabla_W \mathcal{F}(\mu_t)\|^2_{\mu_t} \leq 0,$$

therefore $\mathcal{F}(\mu_t) \searrow$. Moreover,

$$\frac{1}{T} \int_0^T \|\nabla_W \mathcal{F}(\mu_t)\|^2_{\mu_t} dt \leq \frac{\mathcal{F}(\mu_0) - \mathcal{F}(\pi)}{T}.$$

# Lyapunov functions for the Wasserstein GF

Denote $\mathcal{L}_c(t) = W_2^2(\mu_t, \pi)$[3]. Assume $\mathcal{F}$ geodesically convex.

A functional $\mathcal{F}$ is geodesically convex if it is convex along $W_2$ geodesics, i.e. if for any $t \in [0,1]$:

$$\mathcal{F}(\rho(t)) \leq (1-t)\mathcal{F}(\rho(0)) + t\mathcal{F}(\rho(1))$$

where $\rho(t) = ((1-t)I + tT_{\rho(0)}^{\rho(1)})_{\#}\rho(0)$

Then

$$\mathcal{L}_c'(t) = 2\langle I - T_{\mu_t}^{\pi}, \underbrace{-\nabla_W \mathcal{F}(\mu_t)}_{v_t}\rangle_{\mu_t} \leq -2(\mathcal{F}(\mu_t) - \mathcal{F}(\mu^*)) \leq 0,$$

therefore $W_2^2(\mu_t, \pi) \searrow$. Moreover,

$$\mathcal{F}(\mu_t) - \mathcal{F}(\mu^*) \leq \frac{1}{T}\int_0^T \mathcal{F}(\mu_t) - \mathcal{F}(\pi) dt \leq \frac{W_2^2(\mu_0, \pi)}{2T}.$$

[3] $= \|I - T_{\mu_t}^{\pi}\|_{\mu}^2$

# Our approach

**Similarly to the transition**

***Euclidean gradient flow*** $\longrightarrow$ ***gradient descent***,

**we use**

***Wasserstein gradient flow point*** $\longrightarrow$ ***Wass Prox Grad, SVGD***

**(*discretized Wasserstein gradient flows*).**

If convexity is involved, we use the Lyapunov function $\mathcal{L}_c$, otherwise we use $\mathcal{L}$.

# Outline

# Algorithm - Forward Backward discretization

$$\mathrm{KL}(\mu|\pi) = \mathcal{E}_V(\mu) + \mathcal{U}(\mu) + cte$$

$\Longrightarrow$ We propose to analyze [Wibisono, 2018] :

$$\nu_{n+1} = (I - \gamma \nabla V)_{\#}\mu_n$$
$$\mu_{n+1} = JKO_{\gamma\mathcal{U}}(\nu_{n+1})$$

where $JKO_{\mathcal{U}}(\nu_{n+1}) = \underset{\mu \in \mathcal{P}}{\mathrm{argmin}}\, \mathcal{U}(\mu) + \frac{1}{2\gamma} W_2^2(\mu, \nu_{n+1})$.

Tools for the proof :

- Identification of OT maps
- use geodesic convexity

# Identification of the optimal transport maps

From $\mu_n$ to $\nu_{n+1} = (I - \gamma \nabla V)_{\#}\mu_n$ :

**Assumption :** $V$ is $L$-smooth i.e. $\forall\, (x, y) \in \mathcal{X}$,

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

**Then :** If $\mu_0 \ll Leb$ and $\gamma < 1/L$, the OT map from $\mu_n$ to $\nu_{n+1}$ corresponds to :

$$T_{\mu_n}^{\nu_{n+1}} = (I - \gamma \nabla V)$$

and $\nu_{n+1} \ll Leb$.

# Identification of the optimal transport maps

From $\nu_{n+1}$ to $\mu_{n+1} \in JKO_{\gamma\mathcal{U}}(\nu_{n+1})$ :

There exists a strong Fréchet subgradient at $\nu_{n+1}$ denoted $\nabla_W \mathcal{U}(\mu_{n+1})$, such that the OT map from $\nu_{n+1}$ to $\mu_{n+1}$ corresponds to :

$$T_{\mu_{n+1}}^{\nu_{n+1}} = I + \gamma \nabla_W \mathcal{U}(\mu_{n+1})$$

and $\mu_{n+1} \ll Leb$ [Ambrosio et al., 2008].

By Brenier's theorem ($T_{\mu_{n+1}}^{\nu_{n+1}} \circ T_{\nu_{n+1}}^{\mu_{n+1}} = I$) this also means

$$\mu_{n+1} = (I - \gamma \nabla_W \mathcal{U}(\mu_{n+1}) \circ T_{\nu_{n+1}}^{\mu_{n+1}})_{\#} \nu_{n+1}.$$

# Generalized geodesic convexity of $\mathcal{U}$

**Key fact :** $\mathcal{U}$ is convex along *generalized geodesics* defined by $W_2$, i.e. for any $\mu, \pi, \nu \in \mathcal{P}$ with $\nu \ll Leb$, $t \in [0, 1]$ :

$$\mathcal{U}((tT_\nu^\pi + (1-t)T_\nu^\mu)_\# \nu) \leq t\mathcal{U}(\pi) + (1-t)\mathcal{U}(\mu)$$

where $T_\nu^\pi$ and $T_\nu^\mu$ are the OT maps from $\nu$ to $\pi$ and from $\nu$ to $\mu$.

# Generalized geodesic convexity of $\mathcal{U}$

**Key fact :** $\mathcal{U}$ is convex along *generalized geodesics* defined by $W_2$, i.e. for any $\mu, \pi, \nu \in \mathcal{P}$ with $\nu \ll Leb$, $t \in [0, 1]$ :

$$\mathcal{U}((tT_\nu^\pi + (1-t)T_\nu^\mu)_\# \nu) \leq t\mathcal{U}(\pi) + (1-t)\mathcal{U}(\mu)$$

where $T_\nu^\pi$ and $T_\nu^\mu$ are the OT maps from $\nu$ to $\pi$ and from $\nu$ to $\mu$.

This enables us to prove a **descent lemma** for $V$ being $L$-smooth and $\gamma < 1/L$:

$$\mathrm{KL}(\mu_{n+1}|\pi) \leq \mathrm{KL}(\mu_n|\pi) - \gamma \left(1 - \frac{L\gamma}{2}\right) \|\nabla V + \nabla_W \mathcal{U}(\mu_{n+1}) \circ X_{n+1}\|^2_{L_2(\mu_n)},$$

where $X_{n+1} = T_{\nu_{n+1}}^{\mu_{n+1}} \circ (I - \gamma \nabla V)$.

# Rates of convergence in the convex case

**Assumptions :** $V$ is $\lambda$-strongly convex, i.e. $\forall\,(x,y)\in\mathcal{X}$,

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

# Rates of convergence in the convex case

**Assumptions :** $V$ is $\lambda$-strongly convex, i.e. $\forall\,(x,y)\in\mathcal{X}$,

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

**Results :** Assume the step size $\gamma<1/L$ and $\mu_0\ll Leb$. Then for all $n\geq 0$

$$W_2^2(\mu_{n+1},\pi)\leq(1-\gamma\lambda)W_2^2(\mu_n,\pi)-2\gamma\,\mathrm{KL}(\mu_{n+1}|\pi).$$

which implies:

1. $\mathrm{KL}(\mu_{n+1}|\pi)\leq\frac{W_2^2(\mu_0,\pi)}{2\gamma n}$ in the convex case ($\lambda=0$)
2. $W_2^2(\mu_n,\pi)\leq(1-\gamma\lambda)^n W_2^2(\mu_0,\pi)$ when $\lambda>0$

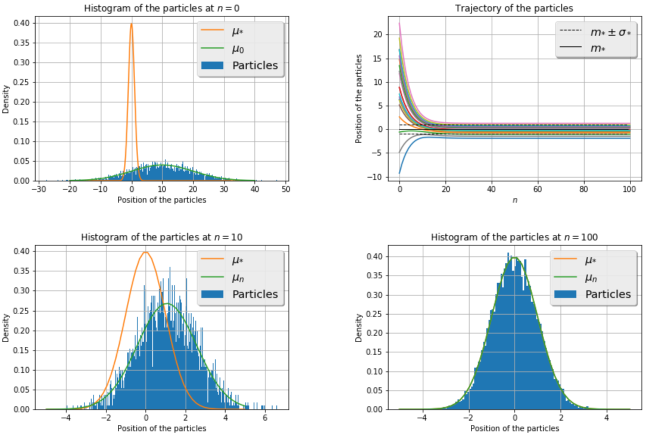
$\Longrightarrow$ same rates than proximal gradient in the euclidean setting!
$\Longrightarrow$ faster than LMC ($1/\sqrt{n}$ for $\lambda=0$ and $1/n$ for $\lambda>0$)

# Implementation of the JKO of the negative entropy

- some subroutines exist to compute the JKO [Santambrogio, 2017], or the JKO w.r.t. the entropy-regularized $W_2$ [Peyré, 2015]
- it is possible to compute the JKO in closed form in the gaussian case (i.e. for $\pi, \mu_0$ gaussians) [Wibisono, 2018].

# Experiments (d=1)

- $\pi = \mu^* = \mathcal{N}(0,1)$ (hence $V(x) = 0.5x^2$ and $\lambda = 1$); $\mu_0 = \mathcal{N}(10,100)$
- we use the closed-form particle implementation for the FB scheme [Wibisono, 2018]



# Linear rate (d=1000)

multi dimensional extension : $V(x) = 0.5\|x\|^2$, target $\mu^{*\otimes d}$ and initial distribution $\mu_0^{\otimes d}$

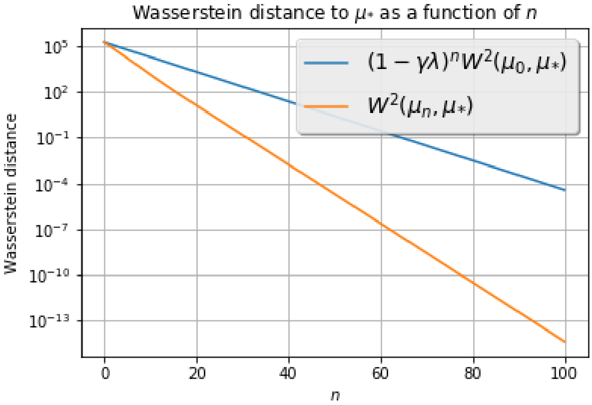



Figure: Linear convergence of $\mu_n$ to $\pi$ in dimension $d = 1000$.

# Contributions and openings

- FB scheme is faster in nb of iterations compared to the Langevin MC algorithm (converges at rate $\mathcal{O}(1/\sqrt{n})$) at the cost of a higher iteration complexity.

- Our proof works for any functional $\mathcal{U}$ that is **convex along generalized geodesics**, and that works for higher order entropies, but also for

$$\text{potential energies } \mathcal{U}(\mu) = \int V(x)\mu(x)dx$$

  for $V$ convex, or

$$\text{interaction energies } \mathcal{U}(\mu) = \int W(x,y)\mu(x)\mu(y)dxdy$$

  for $W$ convex.

- The JKO of entropy deserves more investigation.

# Outline

# Wasserstein Gradient descent for the KL

Let $\mu_0 \in \mathcal{P}$. Gradient descent on $(\mathcal{P}, W_2)$ is written:

$$\mu_{n+1} = \left(I - \gamma \nabla \frac{\partial KL(\mu_n|\pi)}{\partial \mu}\right)_{\#} \mu_n$$

where $\gamma > 0$ is a step-size.

(Particle version) i.e. given $X_0 \sim \mu_0$,

$$X_{n+1} = X_n - \gamma \nabla \frac{\partial \,\mathrm{KL}(\mu_n|\pi)}{\partial \mu}(X_n) \sim \mu_{n+1}$$

# Wasserstein Gradient descent for the KL

Let $\mu_0 \in \mathcal{P}$. Gradient descent on $(\mathcal{P}, W_2)$ is written:

$$\mu_{n+1} = \left( I - \gamma \nabla \frac{\partial KL(\mu_n|\pi)}{\partial \mu} \right)_{\#} \mu_n$$

where $\gamma > 0$ is a step-size.

(Particle version) i.e. given $X_0 \sim \mu_0$,

$$X_{n+1} = X_n - \gamma \nabla \frac{\partial \, \mathrm{KL}(\mu_n|\pi)}{\partial \mu}(X_n) \sim \mu_{n+1}$$

Problem: the $W_2$ gradient of $\mathrm{KL}(\cdot|\pi)$ at $\mu_n$ is the function

$\nabla \log(\frac{\mu_n}{\pi})$. While $\nabla \log \pi$ is known, we do not know what $\mu_n$ is at each $n$, we only have $X_{n+1}$
$\implies \nabla \log \mu_n$ has to be estimated from samples.

# Stein Variational Gradient Descent [Liu and Wang, 2016]

- Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a positive, semi-definite kernel
$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}, \quad \phi : \mathcal{X} \to \mathcal{H}$$
- $\mathcal{H}$ its RKHS : $\overline{\left\{f : \mathbb{R}^d \to \mathbb{R}, f(.) = \sum_{i=1}^n a_i k(x_i, .)\right\}}^{\otimes d}$
Hilbert space of functions equipped with $\langle \cdot, \cdot \rangle_{\mathcal{H}}, \| \cdot \|_{\mathcal{H}}$.
we assume : $\forall \mu$, $\int_{\mathbb{R}^d} k(x, x) d\mu(x) < \infty \implies \mathcal{H} \subset L^2(\mu)$.

# Stein Variational Gradient Descent [Liu and Wang, 2016]

- Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ a positive, semi-definite kernel

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}, \quad \phi : \mathcal{X} \to \mathcal{H}$$

- $\mathcal{H}$ its RKHS : $\overline{\left\{ f : \mathbb{R}^d \to \mathbb{R}, f(.) = \sum_{i=1}^n a_i k(x_i, .) \right\}}^{\otimes d}$
  Hilbert space of functions equipped with $\langle \cdot, \cdot \rangle_{\mathcal{H}}, \| \cdot \|_{\mathcal{H}}$.
  we assume : $\forall \mu, \int_{\mathbb{R}^d} k(x, x) d\mu(x) < \infty \implies \mathcal{H} \subset L^2(\mu)$.

  Define the **kernel integral operator** $S_\mu : L^2(\mu) \to \mathcal{H}$ :

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

  and denote $P_\mu = \iota_{\mathcal{H} \to L^2(\mu)} \circ S_\mu$.

**SVGD trick:** applying this operator to the $W_2$ gradient of $KL(\cdot|\pi)$ leads to (if $\lim_{\|x\| \to \infty} k(x, \cdot)\pi(x) \to 0$)

$$P_\mu \nabla \log \left( \frac{\mu}{\pi} \right) (\cdot) = - \int [\nabla \log \pi(x) k(x, \cdot) + \nabla_x k(x, \cdot)] d\mu(x),$$

# Stein Variational Gradient Descent (SVGD)

**Algorithm :** Starting from $N$ i.i.d. samples $(X_0^i)_{i=1,\dots,N} \sim \mu_0$, SVGD algorithm updates the $N$ particles as follows :

$$X_{n+1}^i = X_n^i - \gamma \underbrace{\left[ \frac{1}{N} \sum_{j=1}^{N} k(X_n^i, X_n^j) \nabla_{X_n^j} \log \pi(X_n^j) + \nabla_{X_n^j} k(X_n^j, X_n^i) \right]}_{P_{\hat{\mu}_n} \nabla \log\left(\frac{\hat{\mu}_n}{\pi}\right)(X_n^i)}$$

where $\hat{\mu}_n = \frac{1}{N} \sum_{j=1}^{N} \delta_{X_n^j}$.

- "non parametric" VI, only depends on the choice of some kernel $k$
- uses a set of interacting particles to approximate $\pi$: `https://chi-feng.github.io/mcmc-demo/app.html?algorithm=HamiltonianMC&target=banana`

# SVGD in the ML literature

- **Empirical performance** demonstrated in various tasks such as:
  - Bayesian inference [Liu and Wang, 2016, Feng et al., 2017, Liu and Zhu, 2018, Detommaso et al., 2018]
  - learning deep probabilistic models [Wang and Liu, 2016, Pu et al., 2017]
  - reinforcement learning [Liu et al., 2017]
- **Theoretical guarantees** : known to converge asymptotically to $\pi$ [Lu et al., 2019] when $V$ grows at most polynomially (in continuous time, infinite number of particles), but no rates of convergence.

This work : non asymptotic analysis of SVGD in the infinite particle regime but discrete time + finite sample approximation.

# Continuous-time dynamics of SVGD

SVGD gradient flow [Liu, 2017]:

$$\frac{\partial \mu_t}{\partial t} + div(\mu_t V_t) = 0, \qquad V_t := -P_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right)$$

# Continuous-time dynamics of SVGD

SVGD gradient flow [Liu, 2017]:

$$\frac{\partial \mu_t}{\partial t} + div(\mu_t V_t) = 0, \qquad V_t := -P_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right)$$

How fast the KL decreases along SVGD dynamics?

$$\begin{aligned}
\frac{dKL(\mu_t|\pi)}{dt} &= \left\langle V_t, \nabla \log\left(\frac{\mu_t}{\pi}\right)\right\rangle_{L^2(\mu_t)} \\
&= -\left\langle \iota_{\mathcal{H}\to L^2(\mu_t)} \circ S_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right), \nabla \log\left(\frac{\mu_t}{\pi}\right)\right\rangle_{L^2(\mu_t)} \\
&= -\left\| S_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right)\right\|^2_{\mathcal{H}} \text{ since } \iota^*_{\mathcal{H}\to L^2(\mu_t)} = S_{\mu_t}.
\end{aligned}$$

# Continuous-time dynamics of SVGD

SVGD gradient flow [Liu, 2017]:

$$\frac{\partial \mu_t}{\partial t} + div(\mu_t V_t) = 0, \qquad V_t := -P_{\mu_t} \nabla \log \left( \frac{\mu_t}{\pi} \right)$$

How fast the KL decreases along SVGD dynamics?

$$\begin{aligned}
\frac{dKL(\mu_t|\pi)}{dt} &= \left\langle V_t, \nabla \log \left( \frac{\mu_t}{\pi} \right) \right\rangle_{L^2(\mu_t)} \\
&= -\left\langle \iota_{\mathcal{H} \to L^2(\mu_t)} \circ S_{\mu_t} \nabla \log \left( \frac{\mu_t}{\pi} \right), \nabla \log \left( \frac{\mu_t}{\pi} \right) \right\rangle_{L^2(\mu_t)} \\
&= -\left\| S_{\mu_t} \nabla \log \left( \frac{\mu_t}{\pi} \right) \right\|^2_{\mathcal{H}} \text{ since } \iota^*_{\mathcal{H} \to L^2(\mu_t)} = S_{\mu_t}.
\end{aligned}$$

On the r.h.s. we have the **Kernel Stein discrepancy** [Chwialkowski et al., 2016] or **Stein Fisher information** at $\mu_t$.

# Continuous-time dynamics of SVGD

SVGD gradient flow [Liu, 2017]:

$$\frac{\partial \mu_t}{\partial t} + div(\mu_t V_t) = 0, \qquad V_t := -P_{\mu_t} \nabla \log \left( \frac{\mu_t}{\pi} \right)$$

How fast the KL decreases along SVGD dynamics?

$$\begin{aligned}
\frac{dKL(\mu_t|\pi)}{dt} &= \left\langle V_t, \nabla \log \left( \frac{\mu_t}{\pi} \right) \right\rangle_{L^2(\mu_t)} \\
&= - \left\langle \iota_{\mathcal{H} \to L^2(\mu_t)} \circ S_{\mu_t} \nabla \log \left( \frac{\mu_t}{\pi} \right), \nabla \log \left( \frac{\mu_t}{\pi} \right) \right\rangle_{L^2(\mu_t)} \\
&= - \left\| S_{\mu_t} \nabla \log \left( \frac{\mu_t}{\pi} \right) \right\|^2_{\mathcal{H}} \text{ since } \iota^*_{\mathcal{H} \to L^2(\mu_t)} = S_{\mu_t}.
\end{aligned}$$

On the r.h.s. we have the **Kernel Stein discrepancy** [Chwialkowski et al., 2016] or **Stein Fisher information** at $\mu_t$.

Along the WGF of the KL we would have obtained the relative Fisher information $\|\nabla \log \left( \frac{\mu_t}{\pi} \right) \|^2_{L^2(\mu_t)}$.

# Discrete time -A descent lemma for SVGD?

In optimization, descent lemmas can be obtained under a boundedness condition on the Hessian matrix.

Gradient descent for $V : \mathbb{R}^d \to \mathbb{R}$ a $C^2(\mathbb{R}^d)$ s.t. $\|H_V(x)\| \leq L$ for any $x$.

$$x_{n+1} = x_n - \gamma \nabla V(x_n).$$

Denote $x(t) = x_n - t\nabla V(x_n)$ and $\varphi(t) = V(x(t))$. Using Taylor expansion :

$$\varphi(\gamma) = \varphi(0) + \gamma\varphi'(0) + \int_0^\gamma (\gamma - t)\varphi''(t)dt.$$

leads to

$$V(x_{n+1}) \leq V(x_n) - \gamma\|\nabla V(x_n)\|^2 + L\int_0^\gamma (\gamma - t)\|\nabla F(x_n)\|^2 dt$$

$$V(x_{n+1}) \leq V(x_n) - \gamma\|\nabla V(x_n)\|^2 + \frac{L\gamma^2}{2}\|\nabla V(x_n)\|^2.$$

Here, the Hessian operator of the KL at $\mu$ is an operator on $T_\mu\mathcal{P} \subset L^2(\mu)$:

$$\langle f, Hess_{KL(.|\pi)}(\mu)f\rangle_{L^2(\mu)} = \mathbb{E}_{X\sim\mu}\left[\langle f(X), H_V(X)f(X)\rangle + \|Jf(X)\|^2_{HS}\right]$$

and yet, this operator is not bounded.

Here, the Hessian operator of the KL at $\mu$ is an operator on $T_\mu \mathcal{P} \subset L^2(\mu)$:

$$\langle f, Hess_{KL(.|\pi)}(\mu) f\rangle_{L^2(\mu)} = \mathbb{E}_{X\sim\mu}\left[\langle f(X), H_V(X) f(X)\rangle + \|Jf(X)\|_{HS}^2\right]$$

and yet, this operator is not bounded.

**In the case of SVGD** one restricts the descent directions $f$ to $\mathcal{H}$. Under several assumptions (boundedness of $k$ and $\nabla k$, of Hessian of $V$ and moments on the trajectory) we could show for $\gamma$ small enough:

$$KL(\mu_{n+1}|\pi) - KL(\mu_n|\pi) \leq -c_\gamma \underbrace{\left\| S_{\mu_n} \nabla \log\left(\frac{\mu_n}{\pi}\right)\right\|_{\mathcal{H}}^2}_{I_{Stein}(\mu_n|\pi)}.$$

# Rates in terms of the Stein Fisher Information

**Consequence :** for $\gamma$ small enough,

$$\min_{k=1,\dots,n} I_{Stein}(\mu_n|\pi) \leq \frac{1}{n}\sum_{k=1}^{n} I_{Stein}(\mu_k|\pi) \leq \frac{KL(\mu_0|\pi)}{c_\gamma n}.$$

This result does not rely on the convexity of $V$, unlike most results on LMC which rely on Log Sobolev inequality or convexity of $V$.

$I_{Stein}(\mu_n|\pi)$ implies weak convergence of $\mu_n$ to $\pi$ if :

- $\pi$ is distantly dissipative[4] (e.g. gaussian mixtures)
- $k$ is translation invariant with a non-vanishing Fourier transform; or $k$ is the IMQ kernel defined by $k(x,y) = (c^2 + \|x-y\|_2^2)^\beta$ for $c > 0$ and $\beta \in [-1, 0]$ (slow decay rate) [Gorham and Mackey, 2017].

---

[4] $\liminf_{r\to\infty} \kappa(r) > 0$ for $\kappa(r) = \inf\{-2\langle \nabla \log \pi(x) - \nabla \log \pi(y), x-y\rangle / \|x-y\|_2^2 ; \|x-y\|_2^2 = r\}$

# Finite number of particles regime

Recall that the practical SVGD implementation is :

$$X_{n+1}^i = X_n^i - \gamma P_{\hat{\mu}_n} \nabla \log\left(\frac{\hat{\mu}_n}{\pi}\right)(X_n^i), \qquad \hat{\mu}_n = \frac{1}{N}\sum_{j=1}^{N} \delta_{X_n^j}.$$

where $\hat{\mu}_n$ denotes the empirical distribution of the interacting particles.

# Finite number of particles regime

Recall that the practical SVGD implementation is :

$$X_{n+1}^i = X_n^i - \gamma P_{\hat{\mu}_n} \nabla \log\left(\frac{\hat{\mu}_n}{\pi}\right)(X_n^i), \qquad \hat{\mu}_n = \frac{1}{N}\sum_{j=1}^{N} \delta_{X_n^j}.$$

where $\hat{\mu}_n$ denotes the empirical distribution of the interacting particles.

**Propagation of chaos result**
Let $n \geq 0$ and $T > 0$. Under boundedness and Lipschitzness assumptions for all $k, \nabla k, V$; for any $0 \leq n \leq \frac{T}{\gamma}$ we have :

$$\mathbb{E}[W_2^2(\mu_n, \hat{\mu}_n)] \leq \frac{1}{2}\left(\frac{1}{\sqrt{N}}\sqrt{var(\mu_0)}e^{LT}\right)(e^{2LT} - 1)$$

where $L$ is a constant depending on $k$ and $\pi$.

# Contributions and openings

- First rates of convergence for SVGD, using techniques from optimal transport and optimization (discrete time - infinite number of particles)
- Propagation of chaos bound (finite number of particles regime)

Open questions :

- Rates in KL? (for $V$ convex)

Open questions :

- Rates in KL? (for $V$ convex)
- Is it possible to obtain a uniform propagation of chaos and a unified convergence bound (decreasing as $n, N \to \infty$)?

Open questions :

- Rates in KL? (for $V$ convex)
- Is it possible to obtain a uniform propagation of chaos and a unified convergence bound (decreasing as $n, N \to \infty$)?
- Properties of the kernel? SVGD dynamics are also relevant for black-box variational inference and Gans [Chu et al., 2020], where the kernel depends on the current distribution.

  $\implies$ in this case the kernel is the neural tangent kernel

  $$k_w(x, y) = \nabla_w f(x, w)^T \nabla_w f(y, w)$$

  (infinite width NN $\approx$ linear models [Jacot et al., 2018])

Open questions :

- Rates in KL? (for $V$ convex)
- Is it possible to obtain a uniform propagation of chaos and a unified convergence bound (decreasing as $n, N \to \infty$)?
- Properties of the kernel? SVGD dynamics are also relevant for black-box variational inference and Gans [Chu et al., 2020], where the kernel depends on the current distribution.

  $\Longrightarrow$ in this case the kernel is the neural tangent kernel

$$k_w(x, y) = \nabla_w f(x, w)^T \nabla_w f(y, w)$$

  (infinite width NN $\approx$ linear models [Jacot et al., 2018])

Thank you!

# References I

Alquier, P. and Ridgway, J. (2017).
Concentration of tempered posteriors and of their variational approximations.
*arXiv preprint arXiv:1706.09293*.

Ambrosio, L., Gigli, N., and Savaré, G. (2008).
*Gradient flows: in metric spaces and in the space of probability measures*.
Springer Science & Business Media.

Chu, C., Minami, K., and Fukumizu, K. (2020).
The equivalence between stein variational gradient descent and black-box variational inference.
*arXiv preprint arXiv:2004.01822*.

# References II


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

Dalalyan, A. S. (2017).
Further and stronger analogy between sampling and optimization: Langevin monte carlo and gradient descent.
*arXiv preprint arXiv:1704.04752*.

Detommaso, G., Cui, T., Marzouk, Y., Spantini, A., and Scheichl, R. (2018).
A stein variational newton method.
In *Advances in Neural Information Processing Systems*, pages 9169–9179.

# References III


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

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

Durmus, A. and Moulines, E. (2016).
Sampling from strongly log-concave distributions with the unadjusted langevin algorithm.
*arXiv preprint arXiv:1605.01559*, 5.

Feng, Y., Wang, D., and Liu, Q. (2017).
Learning to draw samples with amortized stein variational gradient descent.
*arXiv preprint arXiv:1707.06626*.

# References IV


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

Jacot, A., Gabriel, F., and Hongler, C. (2018).
Neural tangent kernel: Convergence and generalization in neural networks.
In *Advances in neural information processing systems*, pages 8571–8580.

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

# References V


Liu, C. and Zhu, J. (2018).
Riemannian stein variational gradient descent for bayesian inference.
In *Thirty-second aaai conference on artificial intelligence*.

Liu, Q. (2017).
Stein variational gradient descent as gradient flow.
In *Advances in neural information processing systems*, pages 3115–3123.

Liu, Q. and Wang, D. (2016).
Stein variational gradient descent: A general purpose bayesian inference algorithm.
In *Advances in neural information processing systems*, pages 2378–2386.

# References VI


Liu, Y., Ramachandran, P., Liu, Q., and Peng, J. (2017).
Stein variational policy gradient.
*arXiv preprint arXiv:1704.02399*.

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

Otto, F. (2001).
The Geometry of Dissipative Evolution Equations: The Porous Medium Equation.
*Communications in Partial Differential Equations*, 26(1-2):101–174.

# References VII


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

Pu, Y., Gan, Z., Henao, R., Li, C., Han, S., and Carin, L. (2017).
Vae learning via stein variational gradient descent.
In *Advances in Neural Information Processing Systems*, pages 4236–4245.

Salim, A., Korba, A., and Luise, G. (2020).
Wasserstein proximal gradient.
*arXiv preprint arXiv:2002.03035*.

# References VIII

Salim, A. and Richtárik, P. (2020).
Primal dual interpretation of the proximal stochastic gradient langevin algorithm.
*arXiv preprint arXiv:2006.09270*.

Santambrogio, F. (2017).
{Euclidean, metric, and Wasserstein} gradient flows: an overview.
*Bulletin of Mathematical Sciences*, 7(1):87–154.

Wang, D. and Liu, Q. (2016).
Learning to draw samples: With application to amortized mle for generative adversarial learning.
*arXiv preprint arXiv:1611.01722*.

# References IX


Wibisono, A. (2018).
Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem.
*arXiv preprint arXiv:1802.08089*.

Zhang, C., Butepage, J., Kjellstrom, H., and Mandt, S. (2018).
Advances in variational inference.
*IEEE transactions on pattern analysis and machine intelligence*.

# Free energies

In particular, if the functional $\mathcal{F}$ is a free energy:

$$\mathcal{F}(\mu) = \underbrace{\int U(\mu(x))\mu(x)dx}_{\text{internal potential } \mathcal{U}} + \underbrace{\int V(x)\mu(x)dx}_{\text{external potential } \mathcal{E}_V} + \underbrace{\int W(x,y)\mu(x)\mu(y)dxdy}_{\text{interaction energy } \mathcal{W}}$$

$$\text{Then : } \frac{\partial \mu_t}{\partial t} = div(\mu_t \nabla (U'(\mu_t) + V + W * \mu_t)).$$

We recover the Euclidean GF if $U \equiv 0$, $W \equiv 0$.

# Some free energies in Machine Learning

The **relative entropy** $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$ can be written:

$$\mathcal{F}(\mu) = \underbrace{\int U(\mu(x))dx}_{\mathcal{U}} + \underbrace{\int V(x)\mu(x)dx}_{\mathcal{E}_V} - C,$$

$U(s) = s\log(s),\ V(x) = -log(\pi(x)),\ C = \mathcal{U}(\pi) + \mathcal{E}_V(\pi)$.

# Some free energies in Machine Learning

The **relative entropy** $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$ can be written:

$$\mathcal{F}(\mu) = \underbrace{\int U(\mu(x))dx}_{\mathcal{U}} + \underbrace{\int V(x)\mu(x)dx}_{\mathcal{E}_V} - C,$$

$U(s) = s\log(s),\ V(x) = -log(\pi(x)),\ C = \mathcal{U}(\pi) + \mathcal{E}_V(\pi)$.

Application : sampling from a posterior distribution $\mu^* \propto \exp(-V)$ in Bayesian inference.

# Some free energies in Machine Learning

The **relative entropy** $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$ can be written:

$$\mathcal{F}(\mu) = \underbrace{\int U(\mu(x))dx}_{\mathcal{U}} + \underbrace{\int V(x)\mu(x)dx}_{\mathcal{E}_V} - C,$$

$$U(s) = s\log(s),\ V(x) = -log(\pi(x)),\ C = \mathcal{U}(\pi) + \mathcal{E}_V(\pi).$$

Application : sampling from a posterior distribution
$\mu^* \propto \exp(-V)$ in Bayesian inference.

The **Maximum Mean Discrepancy** $\mathcal{F}(\mu) = \frac{1}{2}MMD^2(\mu, \mu^*)$
also:

$$\mathcal{F}(\mu) = \underbrace{\int V(x)d\mu(x)}_{\mathcal{E}_V} + \underbrace{\frac{1}{2}\int W(x,y)d\mu(x)d\mu(y)}_{\mathcal{W}} + C,$$

$$V(x) = -\int k(x,x')d\mu^*(x'),\ W(x,x') = k(x,x'),\ C = \mathcal{W}(\mu^*).$$

# Some free energies in Machine Learning

The **relative entropy** $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\pi)$ can be written:

$$\mathcal{F}(\mu) = \underbrace{\int U(\mu(x))dx}_{\mathcal{U}} + \underbrace{\int V(x)\mu(x)dx}_{\mathcal{E}_V} - C,$$

$U(s) = s\log(s),\ V(x) = -log(\pi(x)),\ C = \mathcal{U}(\pi) + \mathcal{E}_V(\pi).$

Application : sampling from a posterior distribution $\mu^* \propto \exp(-V)$ in Bayesian inference.

The **Maximum Mean Discrepancy** $\mathcal{F}(\mu) = \frac{1}{2}MMD^2(\mu, \mu^*)$ also:

$$\mathcal{F}(\mu) = \underbrace{\int V(x)d\mu(x)}_{\mathcal{E}_V} + \underbrace{\frac{1}{2}\int W(x,y)d\mu(x)d\mu(y)}_{\mathcal{W}} + C,$$

$V(x) = -\int k(x,x')d\mu^*(x'),\ W(x,x') = k(x,x'),\ C = \mathcal{W}(\mu^*).$

Application : optimizing infinite-width 1 hidden layer NN where $\mu^*$ is the optimal distribution.

# Convergence of continuous-time dynamics

The convergence of the Stein Fisher information to 0 can be slow. **When do we have fast convergence of SVGD dynamics?**

$\pi$ satisfies the Stein log-Sobolev inequality [Duncan et al., 2019] with constant $\lambda > 0$ if for any $\mu$:

$$KL(\mu|\pi) \leq \frac{1}{2\lambda} \left\| S_\mu \nabla \log \left( \frac{\mu}{\pi} \right) \right\|^2_{\mathcal{H}}.$$

# Convergence of continuous-time dynamics

The convergence of the Stein Fisher information to 0 can be slow. **When do we have fast convergence of SVGD dynamics?**

$\pi$ satisfies the Stein log-Sobolev inequality [Duncan et al., 2019] with constant $\lambda > 0$ if for any $\mu$:

$$KL(\mu|\pi) \leq \frac{1}{2\lambda} \left\| S_\mu \nabla \log\left(\frac{\mu}{\pi}\right) \right\|^2_{\mathcal{H}}.$$

If it holds,

$$\frac{dKL(\mu_t|\pi)}{dt} = -\left\| S_{\mu_t} \nabla \log\left(\frac{\mu_t}{\pi}\right) \right\|^2_{\mathcal{H}} \leq -2\lambda KL(\mu_t|\pi)$$

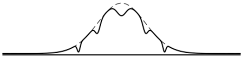
and by integrating :

$$KL(\mu_t|\pi) \leq e^{-2\lambda t} KL(\mu_0|\pi).$$

**"Classic" log-Sobolev inequality** upper bounds the KL by the Fisher divergence :

$$KL(\mu|\pi) \leq \frac{1}{2\lambda} \left\| \nabla \log \left( \frac{\mu}{\pi} \right) \right\|^2_{L^2(\mu)} .$$

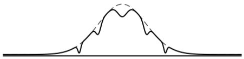
satisfied as soon as $\pi$ is $\lambda$-log concave, but it's more general.



**"Classic" log-Sobolev inequality** upper bounds the KL by the Fisher divergence :

$$KL(\mu|\pi) \leq \frac{1}{2\lambda} \left\| \nabla \log \left( \frac{\mu}{\pi} \right) \right\|^2_{L^2(\mu)} .$$

satisfied as soon as $\pi$ is $\lambda$-log concave, but it's more general.



**When is Stein log-Sobolev satisfied?** not as well known and understood [Duncan et al., 2019], but :

- it fails to hold if $k$ is too regular with respect to $\pi$
- some working examples in dimension 1
- whether it holds in higher dimension is more challenging and subject to further research...

# Rates in terms of the KL objective?

To obtain rates, one may combine a descent lemma (1) of the form

$$KL(\mu_{n+1}|\pi) - KL(\mu_n|\pi) \leq -c_\gamma \left\| S_{\mu_n} \nabla \log\left(\frac{\mu_n}{\pi}\right) \right\|_{\mathcal{H}}^2$$

and the Stein log-Sobolev inequality (2):

$$KL(\mu_{n+1}|\pi) - KL(\mu_n|\pi) \underbrace{\leq}_{(1)} -c_\gamma \left\| S_{\mu_n} \nabla \log\left(\frac{\mu_n}{\pi}\right) \right\|_{\mathcal{H}}^2 \underbrace{\leq}_{(2)} -c_\gamma 2\lambda KL(\mu_n|\pi).$$

Iterating this inequality yields $KL(\mu_n|\pi) \leq (1 - 2c_\gamma\lambda)^n KL(\mu_0|\pi)$.

# Not possible to combine both....

Given that both the kernel and its derivative are bounded, the equation

$$\int \sum_{i=1}^{d} [(\partial_i V(x))^2 k(x,x) \\ - \partial_i V(x)(\partial_i^1 k(x,x) + \partial_i^2 k(x,x)) + \partial_i^1 \partial_i^2 k(x,x)] d\pi(x) < \infty \quad (3)$$

reduces to a property on $V$ which, as far as we can tell, always holds...

# Not possible to combine both....

Given that both the kernel and its derivative are bounded, the equation

$$\int \sum_{i=1}^{d} [(\partial_i V(x))^2 k(x,x) - \partial_i V(x)(\partial_i^1 k(x,x) + \partial_i^2 k(x,x)) + \partial_i^1 \partial_i^2 k(x,x)] d\pi(x) < \infty \quad (3)$$

reduces to a property on $V$ which, as far as we can tell, always holds...

and this implies that Stein LSI does not hold [Duncan et al., 2019].

**Remark :** Equation (3) does not hold for $k$ polynomial of order $\geq 3$ and $\pi$ with exploding $\beta \geq 3$ moments (ex: a student distribution in $\mathcal{P}$ the set of distributions with bounded second moment).

# Experiments

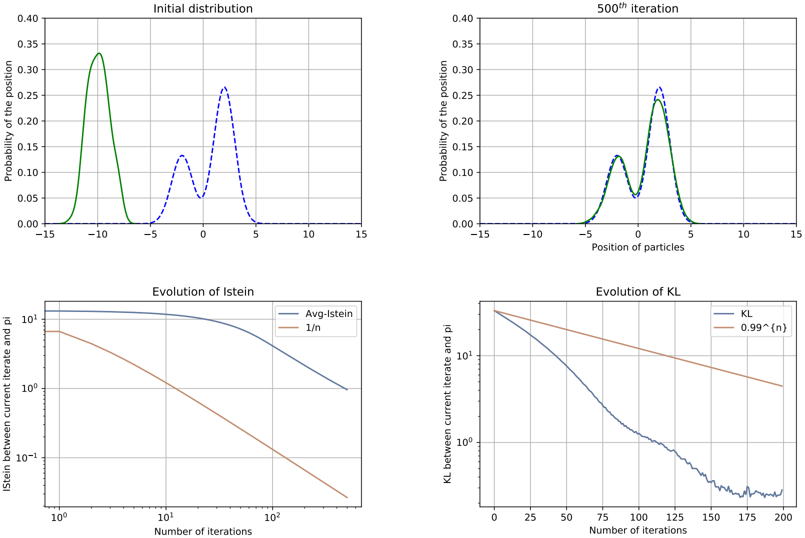



Figure: The particle implementation of the SVGD algorithm illustrates the convergence of $I_{Stein}(\mu_n|\pi)$ and $KL(\mu_n|\pi)$ to 0.