Sampling as Optimization of Divergences

Anna Korba

CREST, ENSAE, Institut Polytechnique de Paris

Stochastics and Statistics seminar, MIT

Outline

Introduction
- Motivation for sampling
- Sampling as optimization: context
- Choice of the divergence

Some recent work
- De-regularized MMD
- Mollified χ^2
- Quantization

Main ingredients
- Background on Wasserstein gradient flow
- Geometry of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Conclusion
- Conclusion

Outline

Introduction
- Motivation for sampling
- Sampling as optimization: context
- Choice of the divergence

Some recent work
- De-regularized MMD
- Mollified χ^2
- Quantization

Main ingredients
- Background on Wasserstein gradient flow
- Geometry of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Conclusion
- Conclusion

General sampling problem

Suppose you are interested in some target probability distribution $\pi \in \mathcal{P}(\mathbb{R}^d)$.

The goal of sampling is to generate samples $x_1, \ldots, x_n$ from π, having access only to some partial information, e.g.:

1. its unnormalized density

$$\pi(x) = \frac{\exp(-V(x))}{Z}, \quad Z = \int \exp(-V(x))dx$$

   Example: π is a posterior distribution in Bayesian inference

2. i.i.d. samples

$$y_1, \ldots y_n \sim \pi$$

   Example: $\pi = p_{data}$ for some data of interest (e.g. images)

Motivation for Sampling (1): Bayesian (or Variational) inference

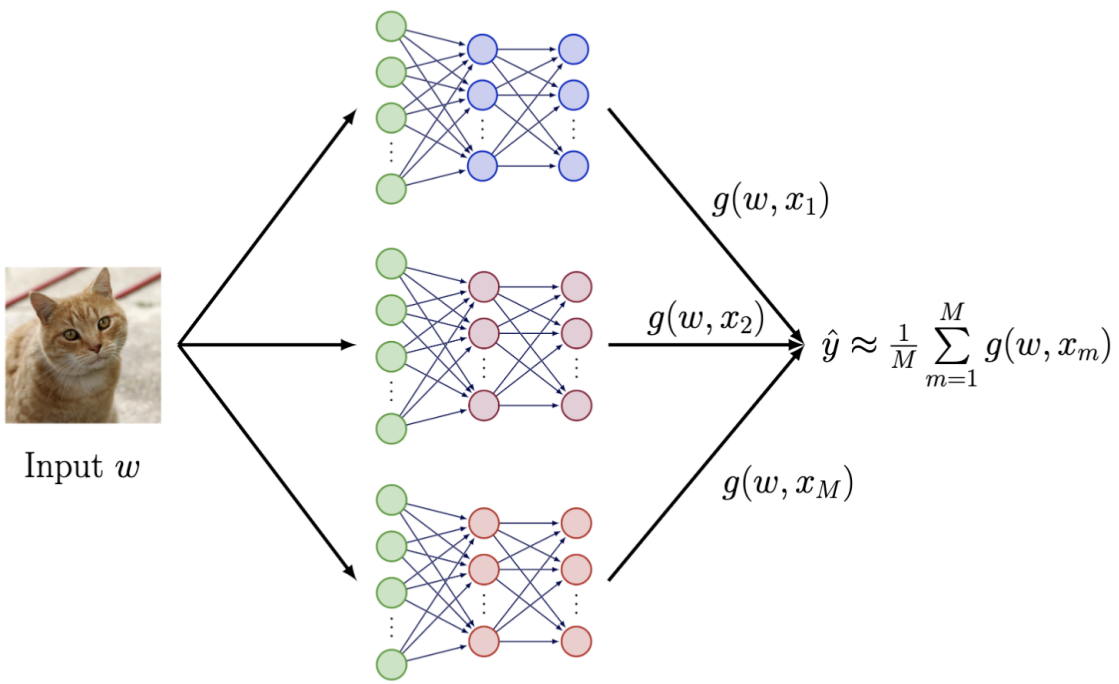
Given labelled data $(w_i, y_i)_{i=1}^m$, we want to sample from the posterior distribution over the parameters of a model $g(\cdot, x)$

$$\pi(x) \propto \exp\left(-V(x)\right), \quad V(x) = \underbrace{\sum_{i=1}^{m} \|y_i - g(w_i, x)\|^2}_{\text{loss on labeled data } (w_i, y_i)_{i=1}^m} + \underbrace{\frac{\|x\|^2}{2}}_{\text{prior reg.}}.$$

Ensemble prediction for a new input w:

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

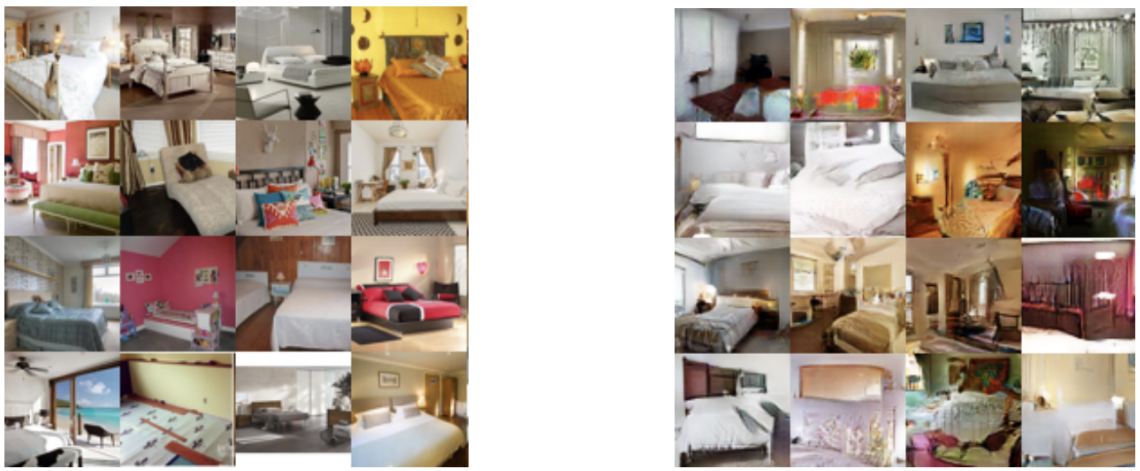
Predictions of models parametrized by $x \in \mathbb{R}^d$ are reweighted by $\pi(x)$.



Motivation for Sampling (2): Generative modeling

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

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



LSUN bedroom samples vs MMD GAN [Li et al. [2017]].

Outline

Introduction
- Motivation for sampling
- Sampling as optimization: context
- Choice of the divergence

Some recent work
- De-regularized MMD
- Mollified χ^2
- Quantization

Main ingredients
- Background on Wasserstein gradient flow
- Geometry of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Conclusion
- Conclusion

The Sampling literature (in a nutshell)

Two different settings:

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

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

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

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

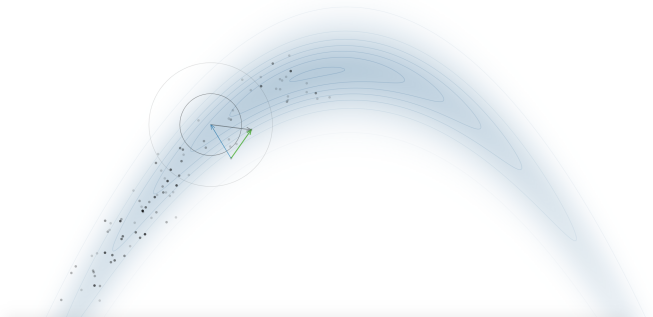
Many of these schemes can be seen through a common framework: optimization over probability distributions (Wasserstein gradient flows are coming) of some **objective functional.**

Langevin Monte Carlo

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

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

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



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

It is a time-discretization of

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

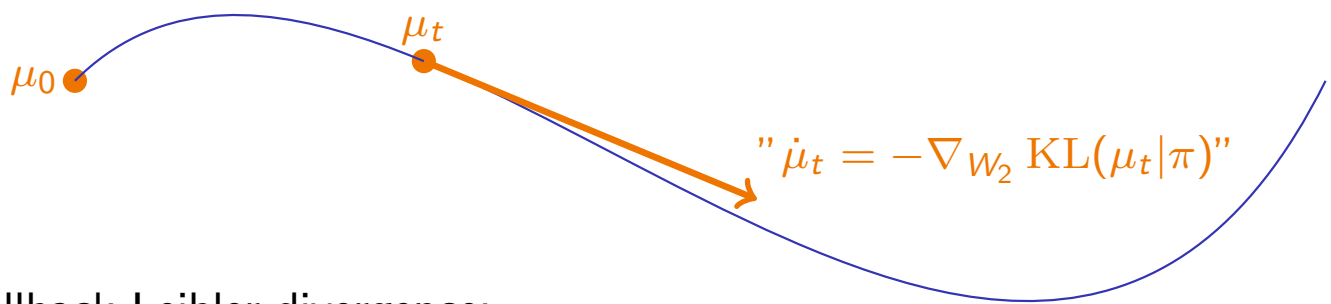
where $(b_t)_{t\geq 0}$ is the Brownian motion in $\mathbb{R}^d$.

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:

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

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]].

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

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

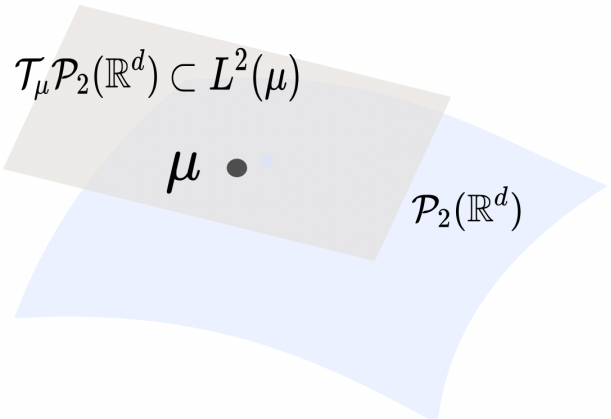
Sampling can be recast as optimization over $\mathcal{P}_2(\mathbb{R}^d)$:

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

Equipped with the Wasserstein-2 (W_2) distance from optimal transport

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

the metric space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ has a convenient **Riemannian structure** [Otto [2001]].

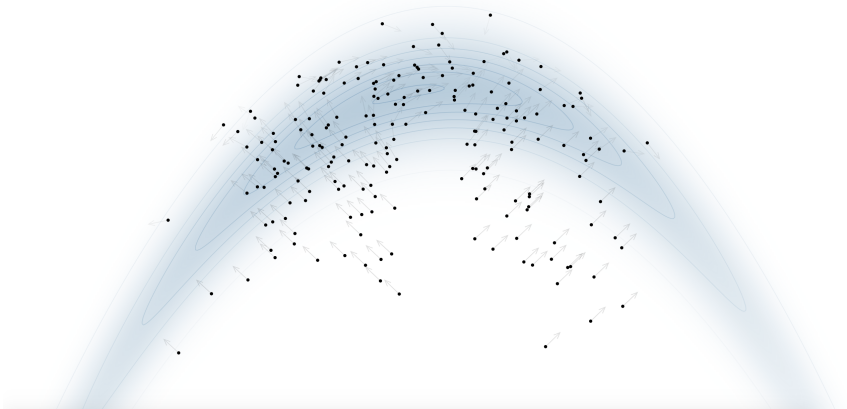


where $L^2(\mu) = \{f : \mathbb{R}^d \to \mathbb{R}^d, \int_{\mathbb{R}^d} \|f(x)\|^2 d\mu(x) < \infty\}$.

Other "KL minimization" schemes

- Stein Variational Gradient Descent [Liu and Wang [2016]], interacting particle scheme, whose empirical measure at stationarity approximates $\pi \propto \exp(-V)$:

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



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

Known to be a gradient flow of KL w.r.t. a kernelized Wasserstein metric [Liu [2017]].

- Variational Inference

$$\min_{\theta} \mathrm{KL}(\mu_\theta | \pi)$$

where μ_θ is parametric (e.g., Gaussian, mixture of Gaussians...)

Sampling as Optimization

To go further: [Ambrosio et al. [2008]], [Santambrogio [2017]], [Chewi et al. [2024]] (Philippe Rigollet, Jonathan Niles-Weed, Sinho Chewi's book).

We can generally view the sampling problem (approximating π) as an optimization one over $\mathcal{P}_2(\mathbb{R}^d)$:

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

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

We will consider several aspects

- can we compute D (or its gradient) in practice given some partial information on π (unnormalized density, or samples)?
- does the optimization algorithm converge ?

Outline

Introduction
- Motivation for sampling
- Sampling as optimization: context
- Choice of the divergence

Some recent work
- De-regularized MMD
- Mollified χ^2
- Quantization

Main ingredients
- Background on Wasserstein gradient flow
- Geometry of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

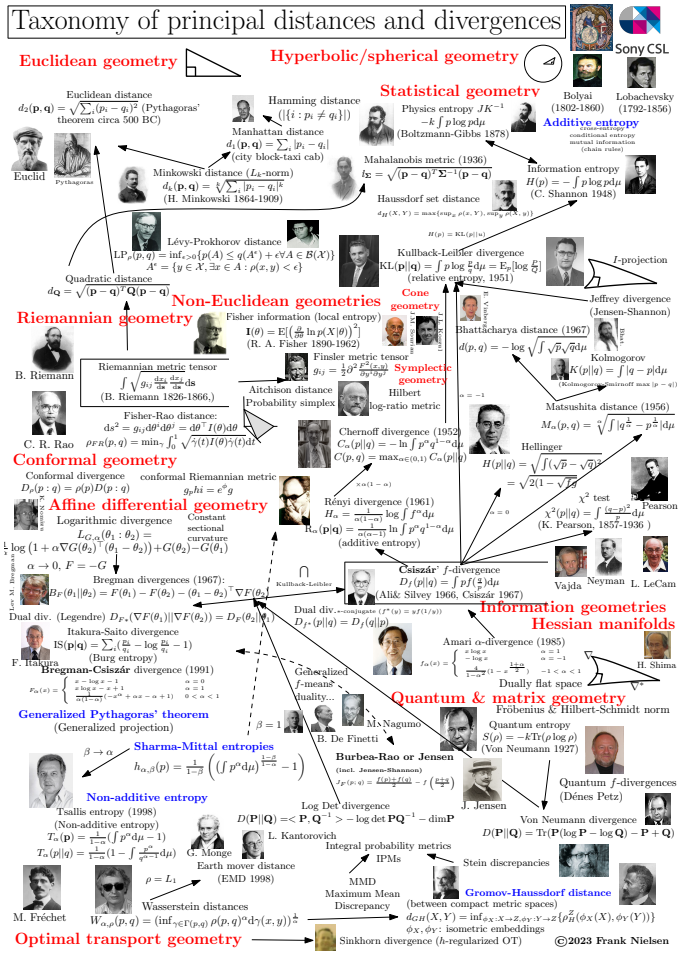
Conclusion
- Conclusion

Which distance or divergence?

Recall we want to solve

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

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



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

Main families of divergences and distances

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

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

Main families of divergences and distances are:

- f-divergences:

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

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

- integral probability metrics (IPM):

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

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

- optimal transport (OT) distances, Sinkhorn divergences

The Kullback-Leibler divergence

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

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

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

The Kullback-Leibler divergence

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

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

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

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

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

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

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

The Maximum Mean Discrepancy

The MMD (Maximum Mean Discrepancy) is defined as:

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

where $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a p.s.d. kernel (e.g. $k(x,y) = e^{-\|x-y\|^2}$) and H_k is the RKHS associated to k $\left(\mathrm{H}_k = \overline{\left\{ \sum_{i=1}^m \alpha_i k(\cdot, x_i);\ \ (\alpha_i)_{i=1}^n \in \mathbb{R}^m;\ (x_i) \in (\mathbb{R}^d)^m \right\}}.\right)$

Can be computed for discrete measures !

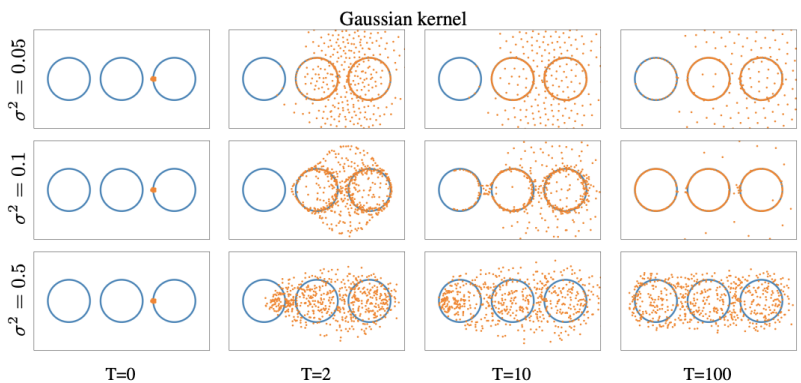
$$\mathrm{MMD}^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x^i}, \frac{1}{m}\sum_{j=1}^m \delta_{y^j}\right) = \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x^i,x^j) + \frac{1}{m^2}\sum_{i=1}^m\sum_{j=1}^m k(y^i,y^j) - \frac{2}{nm}\sum_{i=1}^n\sum_{j=1}^m k(x^i,y^j).$$

Example of the MMD

Let $y_1, \ldots, y_m \sim \pi$, and recall that our goal is to minimize $\mathrm{D}(\mu, \pi)$. We consider the MMD here:

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

What happens if we optimize MMD, through gradient descent on $x_1, \ldots, x_n$?



Picture from [Hertrich et al. [2023]], but known since [Arbel et al. [2019]].

Why is it so bad?

Euclidean Gradient Flow

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

where $V : \mathbb{R}^d \to \mathbb{R}$ s.t. ∇V is L-Lipschitz (V is L-smooth).
Using Cauchy-Lipschitz, consider (where we denote $x_t = x(t)$, $\dot{x}_t = \frac{dx_t}{dt}$):

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

Gradient flow of V = solution of this Ordinary Differential Equation (ODE) for some $x(0)$.
Let $\gamma > 0$ a step-size. **Optimization algorithms = time-discretizations of the GF**:

- Gradient descent algorithm: $x_{m+1} = x_m - \gamma \nabla V(x_m)$, i.e. Forward Euler (explicit): $\frac{x_{m+1} - x_m}{\gamma} = -\nabla V(x_m)$.
- Proximal point algorithm (V convex): $x_{m+1} = \mathrm{prox}_{\gamma V}(x_m) := \arg\min_{y \in \mathbb{R}^d} \gamma V(y) + \frac{1}{2}\|x_m - y\|^2$ i.e. Backward Euler (implicit): $\frac{x_{m+1} - x_m}{\gamma} = -\nabla V(x_{m+1})$.

When does gradient descent work well?

Assume V is differentiable.

- $V : \mathbb{R}^d \to \mathbb{R}$ is L-smooth if for all $x, y \in \mathbb{R}^d$,

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

$$\text{Then for } \gamma \leq 1/L\text{: } \frac{V(x_{m+1}) - V(x_m)}{\gamma} \leq -\gamma \left(1 - \frac{\gamma L}{2}\right) \|\nabla V(x_m)\|^2.$$

- If $V : \mathbb{R}^d \to \mathbb{R}$ is λ-strongly convex, then for all $x, y \in \mathbb{R}^d$,

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

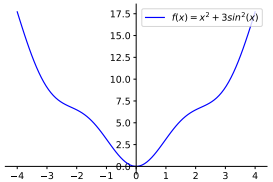
$$\text{Then: } V(x_m) - V_\star \leq (1 - \gamma\lambda)^m (V(x_0) - V_\star).$$

If V twice differentiable, this amounts to

$$\lambda \|v\|_2^2 \leq v^T \nabla^2 f(x) v \leq M \|v\|_2^2 \quad \forall x, v \in \mathbb{R}^d.$$

Generalizations

- Polyak Lojiasiewcz inequality: Strong convexity can be relaxed to: $\forall x \in \mathbb{R}^d, \quad V(x) - V_\star \leq \frac{1}{2\lambda}\|\nabla V(x)\|^2$.



- **Mirror descent** Notice that gradient descent can be written

$$x_{m+1} = \arg\min_{x \in \mathbb{R}^d} \langle \nabla V(x_m), x - x_m \rangle + \frac{1}{2\gamma}\|x - x_m\|^2.$$

If the squared Euclidean distance is replaced by a Bregman divergence,

$$B_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(x), x - y \rangle, \text{ for } \phi \text{ strictly convex}$$

The scheme becomes:

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

Then one can obtain similar rates, considering relative notions of smoothness and convexity [Lu et al. [2018]].

(Some) questions

Recall that in the context of sampling, we mainly used KL when π's density was known up to a normalization constant, while we used MMD when samples were available.

1. when samples of π are available: are there losses that enjoys a better behavior than the MMD?
2. when the unnormalized density is available: are there good alternatives to the KL ?
3. what error can we achieve with a finite number of particles?
4. If time: why some gradient flows converge better than others?

Outline

Introduction
- Motivation for sampling
- Sampling as optimization: context
- Choice of the divergence

Some recent work
- De-regularized MMD
- Mollified χ^2
- Quantization

Main ingredients
- Background on Wasserstein gradient flow
- Geometry of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Conclusion
- Conclusion

Discrete π, and Variational formula of f-divergences

Assume we have sample access to π (e.g. i.i.d. samples from π).

Remember that MMD is convenient as an optimization objective but its gradient descent converges poorly, and KL is not well-suited for a discrete π.

Can we design a better IPM (Integral Probability Metric) **than MMD**?

Discrete π, and Variational formula of f-divergences

Assume we have sample access to π (e.g. i.i.d. samples from π).

Remember that MMD is convenient as an optimization objective but its gradient descent converges poorly, and KL is not well-suited for a discrete π.

Can we design a better IPM (Integral Probability Metric) **than MMD**?

Recall that f-divergences write $D(\mu|\pi) = \int f\left(\frac{\mu}{\pi}\right) d\pi$, f convex, $f(1) = 0$. They admit a variational form [Nguyen et al. [2010]]:

$$D(\mu|\pi) = \sup_{h:\mathbb{R}^d \to \mathbb{R}} \int h d\mu - \int f^\star(h) d\pi$$

where $f^\star(y) = \sup_x \langle x, y\rangle - f(x)$ is the convex conjugate of f and h measurable.

Examples:

- $\mathrm{KL}(\mu|\pi)$: $f(x) = x\log(x) - x + 1$, $f^\star(y) = e^y - 1$
- $\chi^2(\mu|\pi)$: $f(x) = (x-1)^2$, $f^\star(y) = y + \frac{1}{4}y^2$

De-Regularized MMD

In [Chen et al. [2024]] we propose to interpolate between MMD and χ^2

$$\mathrm{DMMD}(\mu||\pi) = (1+\lambda)\Big\{ \max_{h\in \mathrm{H}_k} \int h d\mu - \int (h + \frac{1}{4}h^2) d\pi - \frac{1}{4}\lambda \|h\|^2_{\mathrm{H}_k} \Big\} \qquad (1)$$

- **It is a divergence for any λ, recovers χ^2 for $\lambda = 0$ and MMD for $\lambda = +\infty$.**
- DMMD **and its gradient can be written in closed-form**

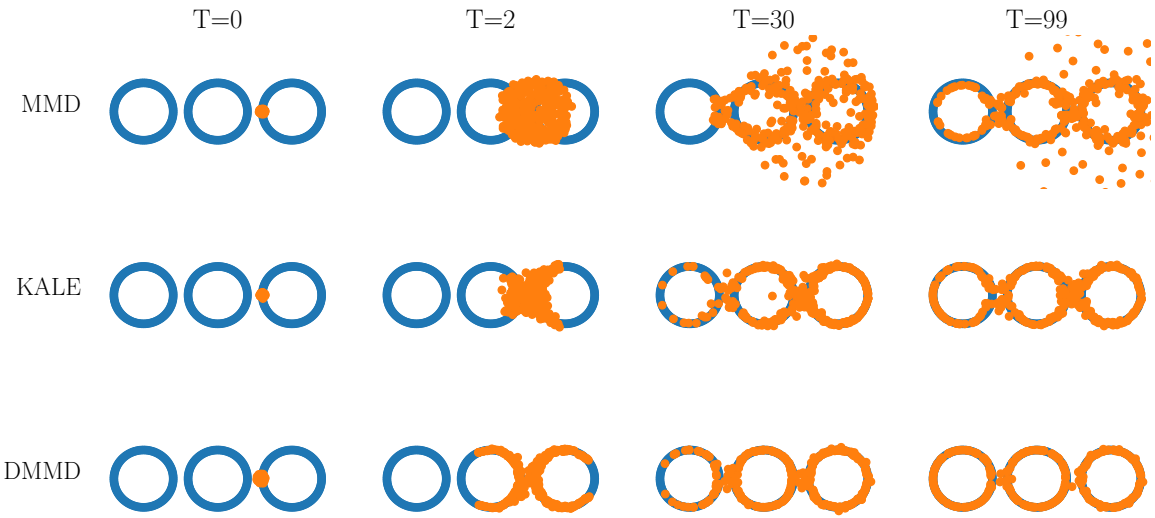
$$\mathrm{DMMD}(\mu||\pi) = (1+\lambda) \left\| (\Sigma_\pi + \lambda \,\mathrm{Id})^{-\frac{1}{2}} (m_\mu - m_\pi) \right\|^2_{\mathrm{H}_k} = \|\nabla h_{\mu,k}\|^2_{\mathrm{H}_k},$$

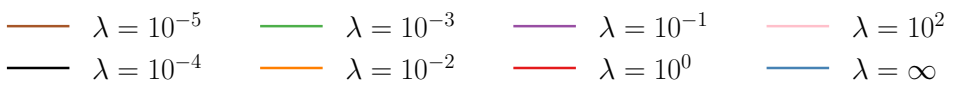
$$\nabla \,\mathrm{DMMD}(\mu||\pi) = \nabla h_{\mu,\pi}$$

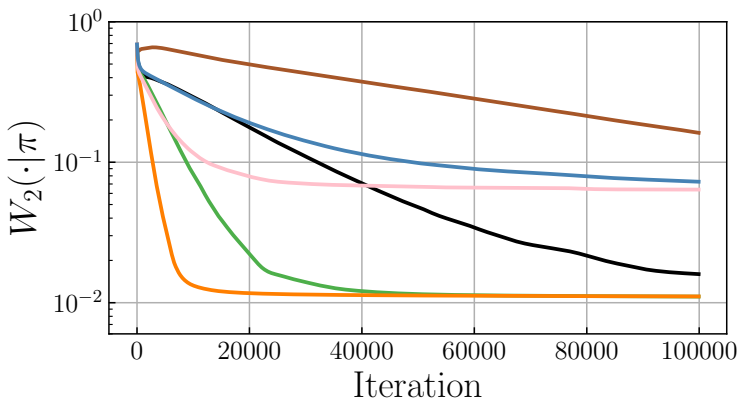
where $\Sigma_\pi = \int k(\cdot, x) \otimes k(\cdot, x) d\pi(x)$, and $h_{\mu,\pi}$ solves (1).

- In particular for μ, π discrete (supported on N, M samples respectively), it writes with kernel Gram matrices over samples of μ, μ^* in complexity $\mathcal{O}(M^3 + NM)$.
- had been proposed for the KL (namely KALE, [Glaser et al. [2021]]) but was not closed-form.

Ring Experiment







Outline

Introduction
- Motivation for sampling
- Sampling as optimization: context
- Choice of the divergence

Some recent work
- De-regularized MMD
- Mollified χ^2
- Quantization

Main ingredients
- Background on Wasserstein gradient flow
- Geometry of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Conclusion
- Conclusion

Mollified χ^2

What if we don't have access to samples of π? (recall that DMMD involves an integral over π) e.g. as in Bayesian inference.

Choose a mollifiers/kernels (Gaussian, Laplace, Riesz-s):

$$k_\epsilon^g(x) := \frac{\exp\left(-\frac{\|x\|_2^2}{2\epsilon^2}\right)}{Z^g(\epsilon)}, \quad k_\epsilon^g(x) := \frac{\exp\left(-\frac{\|x\|_2}{\epsilon}\right)}{Z^l(\epsilon)}, \quad k_\epsilon^s(x) := \frac{1}{(\|x\|_2^2+\epsilon^2)^{s/2} Z^r(s,\epsilon)}$$

In [Li et al. [2022]] we propose the **Mollified chi-square**:

$$\mathcal{E}_\epsilon(\mu) = \iint k_\epsilon(x-y)(\pi(x)\pi(y))^{-1/2}\mu(x)\mu(y)\mathrm{d}x\mathrm{d}y$$

$$= \int \left(k_\epsilon * \frac{\mu}{\sqrt{\pi}}\right)(x)\frac{\mu}{\sqrt{\pi}}(x)\mathrm{d}x \xrightarrow[\varepsilon\to 0]{} \chi^2(\mu|\pi) + 1.$$

It writes as a double integral on μ, allowing to consider μ discrete and π with a density (even unnormalized !). It differs from $\chi^2(k_\epsilon \star \mu|\pi)$ as in [Craig et al. [2022]], whose gradient requires an integration over $\mathbb{R}^d$ (instead of μ).

Sampling/Optimization with constraints

When $\mu = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$ the previous problem becomes a finite-dimensional one on $\omega_N = \{x_1, \dots, x_N\}$,

$$\mathcal{E}_\epsilon(\mu) E_\epsilon(\omega_N) := \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \phi_\epsilon(x_i - x_j)(\pi(x_i)\pi(x_j))^{-1/2}.$$

We can thus resort to finite-dimensional optimization techniques, also for constrained optimization.

- Sampling with (hard/support) constraints, i.e.

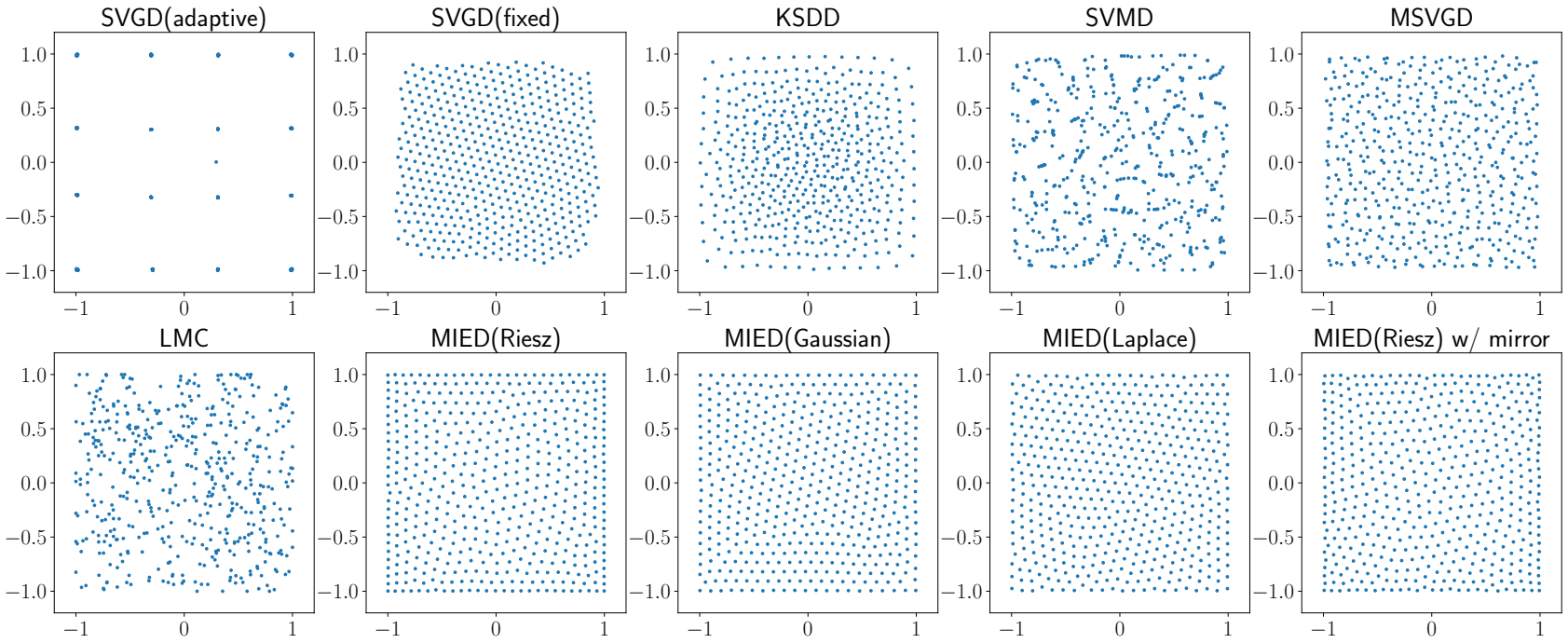
$$\min_{\mu \in \mathcal{P}_2(X)} \mathcal{E}_\epsilon(\mu)$$

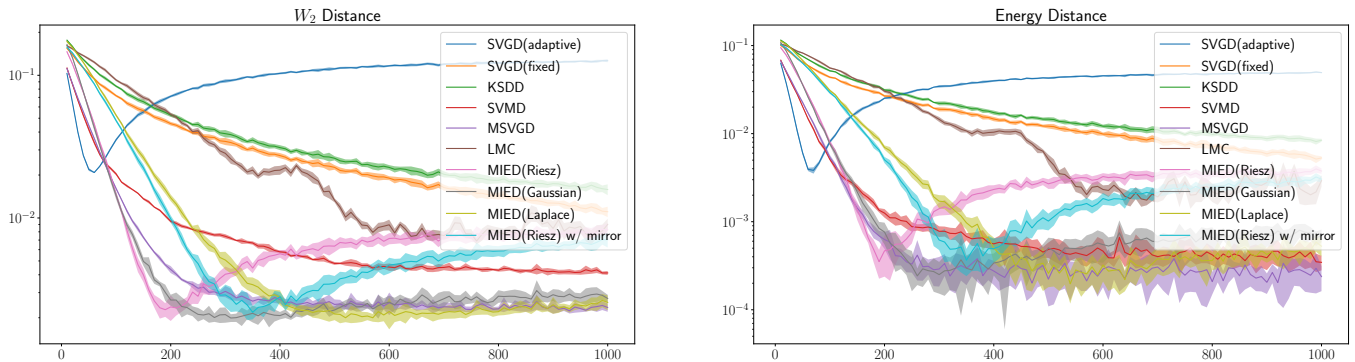
where if we think of x as being parameter of a model and μ the posterior in Bayesian inference, X could encode

  - (1) norm constraints $\|x\|_q \leq C$ (e.g. Bayesian Lasso $q = C = 1$)
  - (2) inequality constraints $X = \left\{x \in \mathbb{R}^d,\ g(x) \leq 0\right\}$ (e.g. fairness constraints)

For (1) we can compare with **"projected/mirror" methods:** Projected LMC [Bubeck et al. [2018]], Mirror LMC [Ahn and Chewi [2021]], Mirror SVGD [Shi et al. [2021]], for (2) we can use dynamic barrier [Gong et al. [2021]].

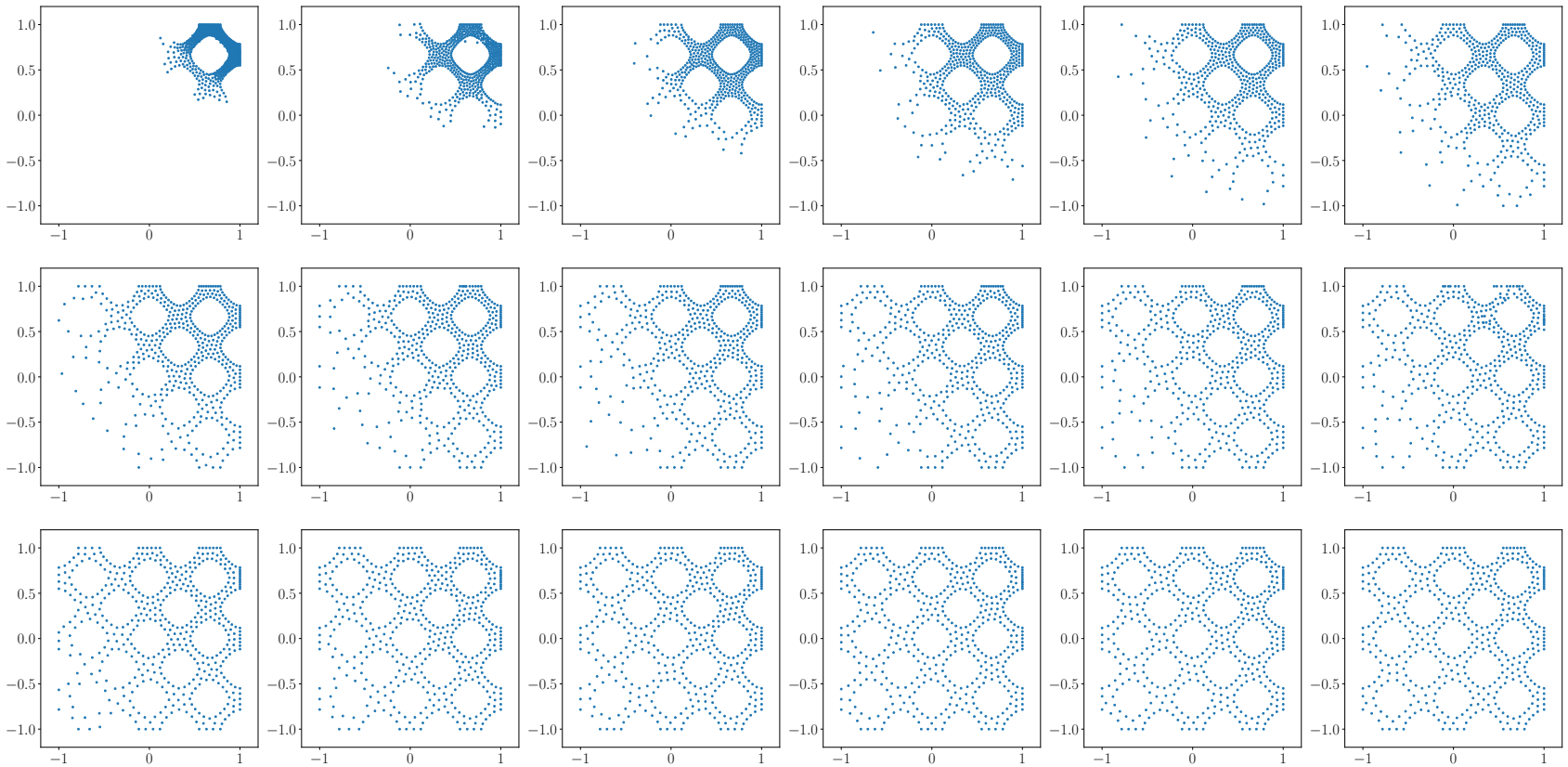
Samling uniformly on a box





We use the mirror map $\phi(\theta) = \sum_{i=1}^{n} \left((1+\theta_i)\log(1+\theta_i) + (1-\theta_i)\log(1-\theta_i)\right)$ or reparametrization using $f = tanh$.

Sampling with inequality constraints



Uniform distribution on $X = \{(x, y) \in [-1, 1]^2 : (\cos(3\pi x) + \cos(3\pi y))^2 < 0.3\}$. Mirror LMC/SVGD cannot be applied due to non convexity of the constraints.

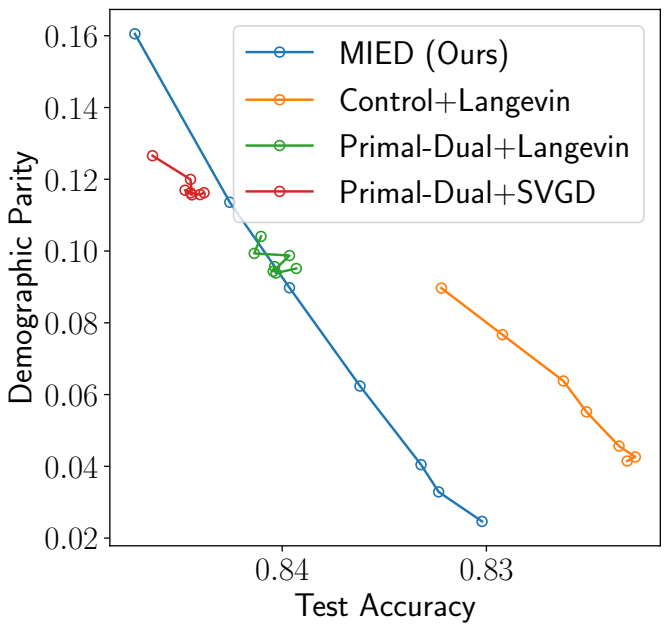
MIED with a Riesz mollifier ($s = 3$) where the constraint is enforced using the dynamic barrier method. The plot in row i column j shows the samples at iteration $100 + 200(6i + j)$. The initial samples are drawn uniformly from the top-right square $[0.5, 1.0]^2$.

Fair Bayesian Neural Network

Let $\mathcal{D} = \{w^{(i)}, y^{(i)}, z^{(i)}\}_{i=1}^{|\mathcal{D}|}$ a dataset consisting of features $w^{(i)}$, labels $y^{(i)}$ (whether the income is $\geq \$50,000$), and genders $z^{(i)}$ (protected attribute).

We set the target density to be the posterior of a logistic regression using a 2-layer Bayesian neural network $\hat{y}(\cdot; x)$. Given $t > 0$, the fairness constraint is

$$g(x) = \left(\mathrm{cov}_{(w,y,z)\sim\mathcal{D}}[z, \hat{y}(w; x)]\right)^2 - t \leq 0.$$



Other methods come from [Liu et al. [2021]].

Outline

Introduction
- Motivation for sampling
- Sampling as optimization: context
- Choice of the divergence

Some recent work
- De-regularized MMD
- Mollified χ^2
- Quantization

Main ingredients
- Background on Wasserstein gradient flow
- Geometry of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Conclusion
- Conclusion

Quantization - classical results

What can we say on $\inf_{x_1,\dots,x_n} \mathrm{D}(\mu_n|\pi)$ where $\mu_n = \sum_{i=1}^n \delta_{x_i}$?

- Quantization rates for the Wasserstein distance [Kloeckner [2012], Mérigot et al. [2021]]

$$W_2(\mu_n, \pi) \sim O(n^{-\frac{1}{d}})$$

- Forward KL [Li and Barron [1999]]: for every $g_P = \int k_\epsilon(\cdot - w)dP(w)$,

$$\arg\min_{\mu_n} \mathrm{KL}(\pi|k_\epsilon \star \mu_n) \leq \mathrm{KL}(\pi|g_P) + \frac{C_{\pi,P}^2 \gamma}{n}$$

where $C_{\pi,P}^2 = \int \frac{\int k_\epsilon(x-m)^2 dP(m)}{(\int k_\epsilon(x-w)dP(w))^2} d\pi(x)$, and $\gamma = 4\log\left(3\sqrt{e} + a\right)$ is a constant depending on ϵ with $a = \sup_{z,z' \in \mathbb{R}^d} \log\left(k_\epsilon(x-z)/k_\epsilon(x-z')\right)$.

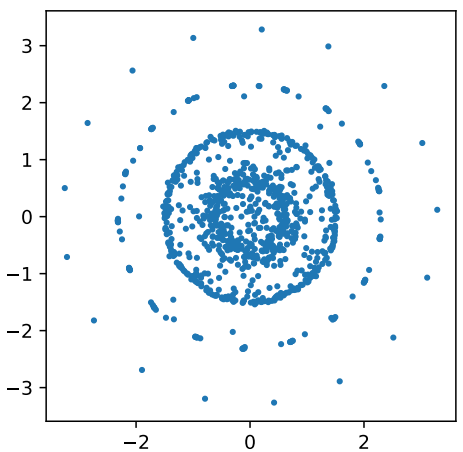
Quantization - Recent results

For smooth and bounded kernels in [Xu et al. [2022]] and π with exponential tails, we get using Koksma-Hlawka inequality

$$\min_{\mu_n} \mathrm{MMD}(\mu_n, \pi) \leq C_d \frac{(\log n)^{\frac{5d+1}{2}}}{n},$$

and similar rates for KSD. This bounds the integral error for $f \in \mathcal{H}_k$ (by Cauchy-Schwartz):

$$\left| \int_{\mathbb{R}^d} f(x) d\pi(x) - \int_{\mathbb{R}^d} f(x) d\mu(x) \right| \leq \|f\|_{\mathcal{H}_k} \mathrm{MMD}(\mu, \pi).$$



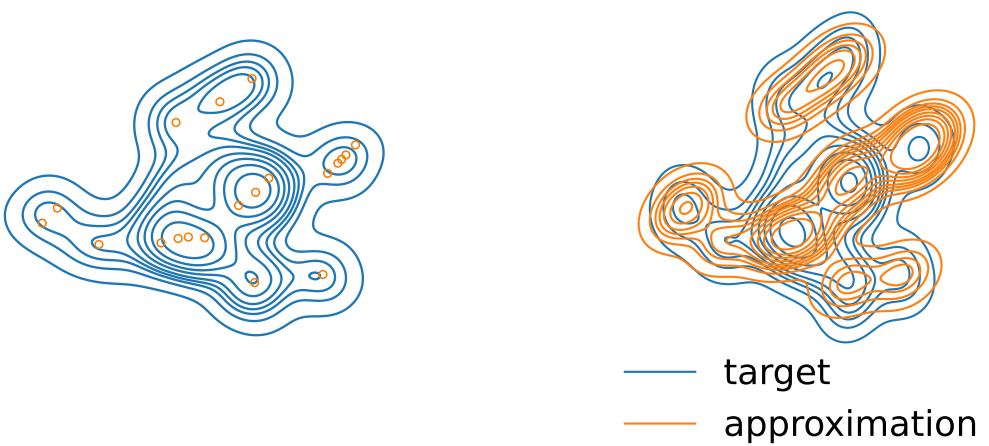
Quantization - Recent results

For the reverse KL, in [Huix et al. [2024]] we get (in the well-specified case where $\mu^* = g_P = \int k_\epsilon(\cdot - w)dP(w)$ for some P):

$$\min_{\mu_n} \mathrm{KL}(k_\epsilon \star \mu | \pi) \leq C_\pi^2 \frac{\log(n) + 1}{n}.$$

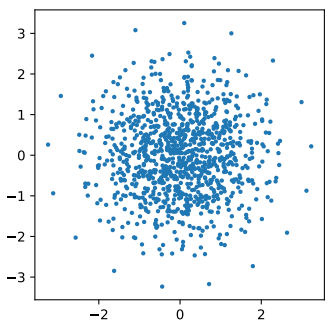
This bounds the integral error for measurable $f : \mathbb{R}^d \to [-1, 1]$ (by Pinsker):

$$\left| \int f d(k_\epsilon \star \mu_n) - \int f d\pi \right| \leq \sqrt{\frac{C_\pi^2 (\log(n) + 1)}{2n}}.$$

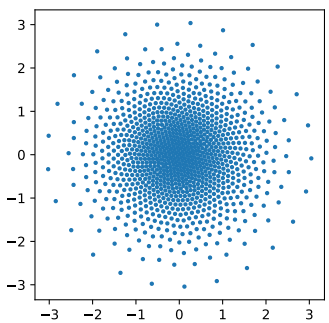


What is missing

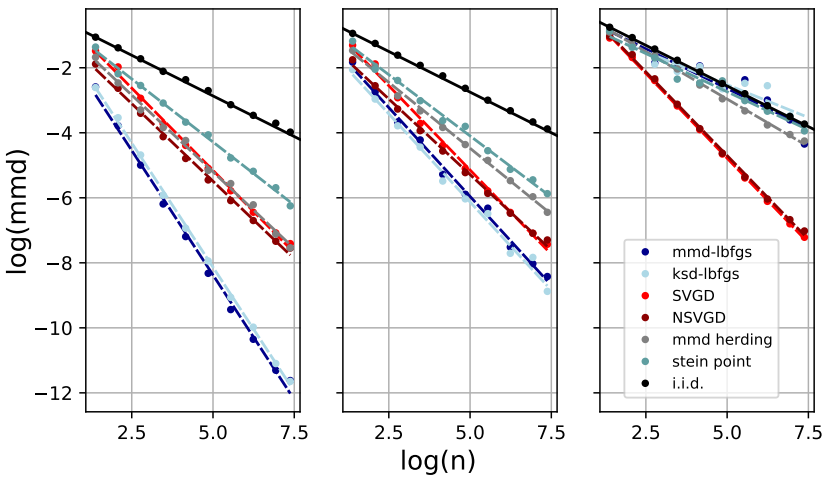
Both rates for SVGD, mollified chi-squares, MMD with non smooth kernels (e.g., $k(x, y) = -\|x - y\|$) are missing.



(a) SVGD Gaussian kernel



(b) SVGD Laplace kernel



(c) Importance of the choice of the bandwidth in the MMD evaluation metric when evaluating the final states, in 2D. From Left to Right: (evaluation) MMD bandwidth $= 1, 0.7, 0.3$.

Outline

Introduction
- Motivation for sampling
- Sampling as optimization: context
- Choice of the divergence

Some recent work
- De-regularized MMD
- Mollified χ^2
- Quantization

Main ingredients
- Background on Wasserstein gradient flow
- Geometry of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Conclusion
- Conclusion

Gradient flows on probability distributions?

Recall that we want to approximate a distribution π by solving

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

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

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

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

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

RHS: Wasserstein gradient = Gradient of First Variation

Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. Under regularity assumptions[1],

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

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

- Consider a linear perturbation $\mu + \varepsilon\xi \in \mathcal{P}_2(\mathbb{R}^d)$ for a perturbation ξ. The First Variation of $\mathcal{F}$ at μ, denoted $\mathcal{F}'(\mu) : \mathbb{R}^d \to \mathbb{R}$ (if it $\exists$) verifies:

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

- Consider a perturbation on the Wasserstein space $(\mathrm{Id} + \varepsilon h)_{\#}\mu$[2] for $h \in L^2(\mu)$. The Wasserstein gradient of $\mathcal{F}$ at μ (if it $\exists$), denoted $\nabla_{W_2}\mathcal{F}(\mu) \in L^2(\mu)$ verifies:

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

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

[2] If $x \sim \mu$, $x + \epsilon h(x) \sim (\mathrm{Id} + \epsilon h)_{\#}\mu$.

Examples below: $\mathcal{F}(\mu) \longrightarrow \mathcal{F}'(\mu) \longrightarrow \nabla \mathcal{F}'(\mu) = \nabla_{W_2} \mathcal{F}(\mu)$.

1. Potential energy (linear function of μ)

$$\mathcal{F}_1(\mu) = \int V(x) d\mu(x) \longrightarrow V \longrightarrow \nabla V$$

2. Negative entropy (assuming μ has a density, using $(y \log y)' = \log y + 1$):

$$\mathcal{F}_2(\mu) = \int \log(\mu(x)) d\mu(x) \longrightarrow \log(\mu) + 1 \longrightarrow \nabla \log \mu.$$

3. for $\pi \propto \exp(-V)$, denoting $\mathcal{F} = \mathcal{F}_1 + \mathcal{F}_2$, $\mathrm{KL}(\mu|\pi) = \mathcal{F}(\mu) - \underbrace{\mathcal{F}(\pi)}_{\text{Constant}}$.

   By additivity, $\nabla_{W_2} \mathrm{KL}(\mu|\pi) = \nabla V + \nabla \log(\mu) = \nabla \log\left(\frac{\mu}{\pi}\right)$.

4. Interaction energy (quadratic function of μ):

$$\mathcal{F}_3(\mu) = \iint W(x,y) d\mu(x) d\mu(y) \longrightarrow \int W(x,\cdot) d\mu(x) \longrightarrow \int \nabla W(x,\dot{)} d\mu(x).$$

   By additivity,

$$\mathrm{MMD}^2(\mu,\pi) = \iint_{\mathbb{R}^d} k(x,y) d\mu(x) d\mu(y) - 2 \iint_{\mathbb{R}^d} k(x,y) d\mu(x) d\pi(y) + \mathcal{F}_3(\pi)$$

$$\longrightarrow \nabla_{W_2} \mathrm{MMD}(\mu|\pi) = \int \nabla k(x,\cdot) d\mu(x) - \int \nabla k(x,\cdot) d\pi(x).$$

Wasserstein gradient flow dynamic

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

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

where $\nabla_{W_2}\mathcal{F}(\mu) := \nabla \mathcal{F}'(\mu) : \mathbb{R}^d \to \mathbb{R}^d$ [1], and where $\boldsymbol{\nabla} \cdot A(x) = \sum_{i=1}^{d} \frac{\partial A_i(x)}{\partial x_i}$ for $A(x) = (A_1(x), \ldots, A_d(x))$, $A : \mathbb{R}^d \to \mathbb{R}^d$.

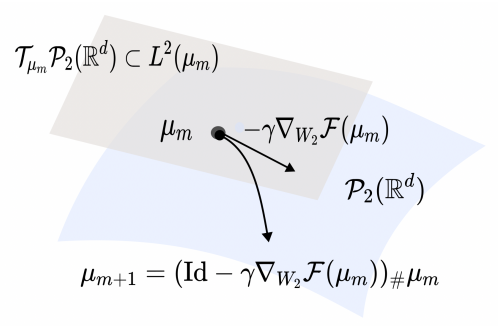
It can be implemented by the deterministic process in $\mathbb{R}^d$:

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

[1] recall $\lim_{\epsilon \to 0} \frac{1}{\epsilon}(\mathcal{F}(\mu + \epsilon(\nu - \mu)) - \mathcal{F}(\mu)) = \int_{\mathbb{R}^d} \mathcal{F}'(\mu)(x)(d\nu - d\mu)(x)$, $\mathcal{F}'(\mu) : \mathbb{R}^d \to \mathbb{R}$.

Time and space discretization of the WGF = particle scheme

(Explicit) Time discretization: Let $\gamma > 0$ a step-size. **Wasserstein gradient descent** is:



Problem: Recall that if $x \sim \mu_m$, $x - \gamma\nabla_{W_2}\mathcal{F}(\mu_m)(x) \sim \mu_{m+1}$. Generally requires the knowledge of the density μ_m.

Space discretization: Introduce a particle system $x_0^1, \dots, x_0^n \sim \mu_0$:

$$x_{m+1}^i = x_m^i - \gamma\nabla_{W_2}\mathcal{F}(\hat{\mu}_m)(x_m^i) \quad \text{for } i = 1, \dots, n, \text{ where } \hat{\mu}_m = \frac{1}{n}\sum_{i=1}^n \delta_{x_m^i}. \qquad (2)$$

In particular, if $\mathcal{F}(\mu) = \mathrm{D}(\mu|\pi)$ is well-defined for discrete measures μ, Algorithm (2) **simply corresponds to gradient descent of** $F : \mathbb{R}^{N\times d} \to \mathbb{R}$, $F(x^1, \dots, x^n) := \mathcal{F}(\mu^n)$ where $\mu^n = \frac{1}{n}\sum_{i=1}^n \delta_{x^i}$.

Outline

Introduction
- Motivation for sampling
- Sampling as optimization: context
- Choice of the divergence

Some recent work
- De-regularized MMD
- Mollified χ^2
- Quantization

Main ingredients
- Background on Wasserstein gradient flow
- Geometry of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Conclusion
- Conclusion

Wasserstein geodesics between μ, ν

Recall

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

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

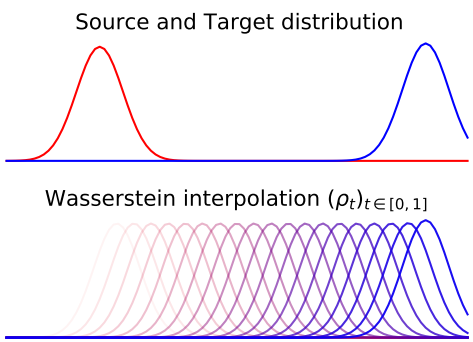
1. $T_{\mu\#}^\nu \mu = \nu$
2. $W_2^2(\mu,\nu) = \| \mathrm{Id} - T_\mu^\nu \|_\mu^2 \overset{\text{def.}}{=} \int \|x - T_\mu^\nu(x)\|^2 d\mu(x)$.

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

The path

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

is the Wasserstein geodesic between $\rho_0 = \mu$ and $\rho_1 = \nu$. Can also be written as a continuity equation with vector field $\nabla\psi \in L^2(\mu)$.



Convexity and Smoothness (in $\mathbb{R}^d$ and $\mathcal{P}_2(\mathbb{R}^d)$)

We want to study the convergence of Wasserstein gradient descent

$$\mu_{l+1} = (\mathrm{Id} - \gamma \nabla \mathcal{F}'(\mu_l))_{\#}\mu_l$$

In $\mathbb{R}^d$, fast rates were obtained when the objective function $V : \mathbb{R}^d \to \mathbb{R}$ was strongly convex and smooth:

$$\lambda \|v\|_2^2 \leq v^T \nabla^2 f(x) v \leq M \|v\|_2^2 \quad \forall x, v \in \mathbb{R}^d.$$

In $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, the same story holds [Villani [2009]] (Prop 16.2):

$\mathcal{F}$ is λ-convex and M-smooth $\iff \lambda \|\nabla \psi\|^2_{L^2(\mu)} \leq \mathrm{Hess}_\mu \mathcal{F}(\psi, \psi) \leq M \|\nabla \psi\|^2_{L^2(\mu)}$,

where the **Wasserstein Hessian** of a functional $\mathcal{F} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ at μ is defined for any $\psi \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ as: $\mathrm{Hess}_\mu \mathcal{F}(\psi, \psi) := \frac{\mathrm{d}^2}{\mathrm{d}t^2}\Big|_{t=0} \mathcal{F}(\mu_t)$ and $(\mu_t, v_t)_{t\in[0,1]}$ is a Wasserstein geodesic with $\mu_0 = \mu$, $v_0 = \nabla \psi$.

Convexity and Smoothness of KL and MMD

- Let $\pi \propto e^{-V}$, we have [Villani [2009]]

$$\text{Hess}_\mu \, \mathrm{KL}(\cdot||\pi)(\psi,\psi) = \int \left[\langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\rangle + \|\mathrm{H}\psi(x)\|^2_{HS}\right] \mu(x)\mathrm{d}x.$$

If V is λ-strongly convex, then **the KL is λ-geo. convex**:

$$\langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\rangle \geq m\|\nabla\psi(x)\|^2 \Longrightarrow \text{Hess}_\mu \, \mathrm{KL}(\cdot||\pi)(\psi,\psi) \geq \lambda\|\nabla\psi\|^2_{L^2(\mu)}.$$

However it is **not smooth** (Hessian is unbounded wrt $\|\nabla\psi\|^2_{L^2(\mu)}$). Similar story for χ^2-square [Ohta and Takatsu [2011]].

- For a M-smooth kernel k [Arbel et al. [2019]]

$$\text{Hess}_\mu \, \mathrm{MMD}^2(\cdot||\pi)(\psi,\psi) = \int \nabla\psi(x)^\top \nabla_1\nabla_2 k(x,y)\nabla\psi(y) d\mu(x) d\mu(y) +$$
$$2\int \nabla\psi(x)^\top \left(\int \mathrm{H}_1 k\,(x,z)\, d\mu(z) - \int \mathrm{H}_1 k\,(x,z)\, d\pi(z)\right) \nabla\psi(x) d\mu(x)$$

The MMD is M-smooth but not geodesically convex (Hessian lower bounded by a big negative constant).

Functional inequalities

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

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

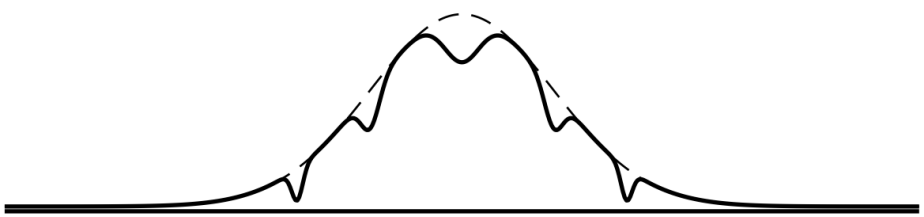
- V is λ-strongly convex $\Rightarrow$ $\pi \propto \exp(-V)$ satisfies Log Sobolev with λ (Bakry–Emery theorem)
- Log Sobolev $\not\Rightarrow$ V convex.

Example: Consider a standard Gaussian distribution

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

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

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



About the losses I presented

Such functional inequalities enable to obtain fast rates for the gradient flows:

- for instance, Log-sobolev enables to obtain

$$\mathrm{KL}(\mu_t|\pi) \leq \exp(-(2/C_{LS})t)\,\mathrm{KL}(\mu_0|\pi)$$

  where C_{LS} denotes the Log-Sobolev constant.
- Poincaré inequality plays a similar role for χ^2 flow measured in KL
- For DMMD, we can leverage its link with χ^2 to obtain rates of convergence

$$\mathrm{KL}(\mu_t|\pi) \leq \exp(-(2(1+\lambda)/C_p)t)\,\mathrm{KL}(\mu_0|\pi) + \text{bias}$$

  denoting C_p the Poincaré constant of π, and bias increases with λ [Chen et al. [2024]].
- For the mollified chi-square, we did not obtain such results.

Outline

Introduction
- Motivation for sampling
- Sampling as optimization: context
- Choice of the divergence

Some recent work
- De-regularized MMD
- Mollified χ^2
- Quantization

Main ingredients
- Background on Wasserstein gradient flow
- Geometry of $(\mathcal{P}_2(\mathbb{R}^d), W_2)$

Conclusion
- Conclusion

Conclusion:

- Sampling can be seen as an optimization problem on a "Wasserstein manifold", and we can consider Wasserstein gradient flows, that decrease a loss (e.g. KL, MMD, etc)
- Their discretizations (space/time) lead to different algorithms

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...
- We did not talk about practical considerations, e.g. improving convergence (e.g., how to handle multimodality of π)

References I

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

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

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

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

S. Bubeck, R. Eldan, and J. Lehec. Sampling from a log-concave distribution with projected Langevin Monte Carlo. *Discrete & Computational Geometry*, 59(4):757–783, 2018.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 2024.

Katy Craig, Karthik Elamvazhuthi, Matt Haberland, and Olga Turanova. A blob method method for inhomogeneous diffusion with applications to multi-agent control and sampling. *arXiv preprint arXiv:2202.12927*, 2022.

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

Pierre Glaser, Michael Arbel, and Arthur Gretton. Kale flow: A relaxed kl gradient flow for probabilities with disjoint support. *Advances in Neural Information Processing Systems*, 34:8018–8031, 2021.

Chengyue Gong, Xingchao Liu, and Qiang Liu. Automatic and harmless regularization with constrained and lexicographic optimization: A dynamic barrier approach. *Advances in Neural Information Processing Systems*, 34:29630–29642, 2021.

Johannes Hertrich, Christian Wald, Fabian Altekrüger, and Paul Hagemann. Generative sliced mmd flows with riesz kernels. *arXiv preprint arXiv:2305.11463*, 2023.

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

Benoit Kloeckner. Approximation by finitely supported measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 18(2): 343–359, 2012.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017.

Jonathan Li and Andrew Barron. Mixture density estimation. *Advances in neural information processing systems*, 12, 1999.

References II

Qiang Liu. Stein variational gradient descent as gradient flow. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/17ed8abedc255908be746d245e50263a-Paper.pdf.

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

Xingchao Liu, Xin Tong, and Qiang Liu. Sampling with trusthworthy constraints: A variational gradient framework. *Advances in Neural Information Processing Systems*, 34:23557–23568, 2021.

Haihao Lu, Robert M Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018.

Quentin Mérigot, Filippo Santambrogio, and Clément Sarrazin. Non-asymptotic convergence bounds for wasserstein approximation using point clouds. *Advances in Neural Information Processing Systems*, 34:12810–12821, 2021.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Shin-ichi Ohta and Asuka Takatsu. Displacement convexity of generalized relative entropies. *Advances in Mathematics*, 228(3): 1742–1787, 2011.

Felix Otto. The Geometry of Dissipative Evolution Equations: The Porous Medium Equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, January 2001. ISSN 0360-5302. doi: 10.1081/PDE-100002243. URL https://doi.org/10.1081/PDE-100002243.

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

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

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

Jiaxin Shi, Chang Liu, and Lester Mackey. Sampling with mirrored stein operators. *arXiv preprint arXiv:2106.12506*, 2021.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

A. Wibisono. 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, 2018.

References I

Michael Arbel, Anna Korba, Adil Salim, and Arthur Gretton. Maximum mean discrepancy gradient flow. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6484–6494, 2019.

Zonghao Chen, Aratrika Mustafi, Pierre Glaser, Anna Korba, Arthur Gretton, and Bharath K Sriperumbudur. (de)-regularized maximum mean discrepancy gradient flow. *arXiv preprint arXiv:2409.14980*, 2024.

Tom Huix, Anna Korba, Alain Durmus, and Eric Moulines. Theoretical guarantees for variational inference with fixed-variance mixture of gaussians. *International Conference on Machine Learning*, 2024.

Lingxiao Li, Qiang Liu, Anna Korba, Mikhail Yurochkin, and Justin Solomon. Sampling with mollified interaction energy descent. *arXiv preprint arXiv:2210.13400*, 2022.

Lantian Xu, Anna Korba, and Dejan Slepčev. Accurate quantization of measures via interacting particle-based optimization. *International Conference on Machine Learning*, 2022.