Sampling through Optimization of Divergences

Anna Korba

ENSAE, CREST, Institut Polytechnique de Paris

Interacting Particle Systems: Analysis, Control, Learning and Computation, ICERM

Outline

1. Sampling as Optimization
2. Applications
3. Choice of the $\mathbb{D}$
4. De-regularized MMD
5. Mollified χ^2
6. Further connections with Optimization

Why sampling?

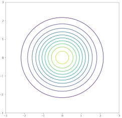
Suppose you are interested in some target probability distribution on $\mathbb{R}^d$, denoted μ^*, and you have access only to partial information, e.g.:

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

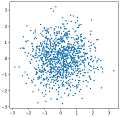
Problem: approximate $\mu^* \in \mathcal{P}(\mathbb{R}^d)$ by a finite set of n points $x_1, \ldots, x_n$, e.g. to compute functionals $\int_{\mathbb{R}^d} f(x) d\mu^*(x)$.

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

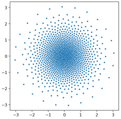
$$\left| \frac{1}{n}\sum_{i=1}^{n} f(x_i) - \int_{\mathbb{R}^d} f(x) d\mu^*(x) \right|.$$



a Gaussian density



i.i.d. samples.



Particle scheme (SVGD).

Sampling as optimization over probability distributions

Assume that $\mu^* \in \mathcal{P}_2(\mathbb{R}^d) = \left\{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\right\}$.

The sampling task can be recast as an optimization problem:

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

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

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

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

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

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

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

Particle system/Gradient descent approximating the WGF

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

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

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

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

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

$$x_{l+1}^i = x_l^i - \gamma \nabla_{W_2} \mathcal{F}(\hat{\mu}_l)(x_l^i) \quad \text{for } i = 1, \ldots, n, \text{ where } \hat{\mu}_l = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_l^i}. \qquad (1)$$

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

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

Some examples for $\mathcal{F} = \mathrm{D}(\cdot|\mu^*)$

- the Kullback-Leibler divergence

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

Pro: the normalization constant Z of $\mu^* = e^{-V}/Z$ is an additive constant; Con: $+\infty$ if $\mathrm{supp}(\mu) \not\subset \mathrm{supp}(\mu^*)$.

- the MMD (Maximum Mean Discrepancy)

$$\mathrm{MMD}^2(\mu,\mu^*) = \sup_{f\in\mathcal{H}_k, \|f\|_{\mathcal{H}_k}\le 1} \left| \int f d\mu - \int f d\mu^* \right| = \iint_{\mathbb{R}^d} k(x,y) d\mu(x) d\mu(y)$$
$$+ \iint_{\mathbb{R}^d} k(x,y) d\mu^*(x) d\mu^*(y) - 2 \iint_{\mathbb{R}^d} k(x,y) d\mu(x) d\mu^*(y).$$

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 $\mathcal{H}_k$ is the RKHS associated to k[‡] Pro: convenient for discrete measures. Con: requires access to samples of μ^*.

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

MMD Gradient flow in practice

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

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

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

- The WGF of the MMD can be implemented via :

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

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

$$x_{l+1}^i = x_l^i - \gamma\left(\sum_{j=1}^n \nabla_2 k(x_l^i, x_l^j) - \int \nabla_2 k(x_l^i, y)d\mu^*(y)\right)$$

which is gradient descent of $(x^1,\ldots,x^n) \mapsto \text{MMD}^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x^i}, \mu^*\right)$

KL Gradient flow in practice `https://chi-feng.github.io/mcmc-demo/app.html`

Take $\mathcal{F}(\mu) = \mathrm{KL}(\mu|\mu^*) = \int \log\left(\frac{\mu}{\mu^*}\right) d\mu$, we have $\nabla_{W_2}\mathcal{F}(\mu) = \nabla \log\left(\frac{\mu}{\mu^*}\right)$.

- The WGF of the KL can be written (rhs = Fokker-Planck equation)

$$\frac{\partial \mu_t}{\partial t} = \boldsymbol{\nabla} \cdot \left(\mu_t \nabla \log \frac{\mu_t}{\mu^*}\right) = \boldsymbol{\nabla} \cdot (\mu_t \nabla \log \mu^*) + \Delta \mu_t$$

- It can be implemented via ”Probability Flow” (2) or Langevin diffusion (3):

$$d\tilde{x}_t = -\nabla \log\left(\frac{\mu_t}{\mu^*}\right)(\tilde{x}_t)dt \tag{2}$$

$$dx_t = \nabla \log \mu^*(x_t)dt + \sqrt{2}dB_t \tag{3}$$

- (3) can be discretized in time as **Langevin Monte Carlo (LMC)**

$$x_{l+1} = x_l + \gamma \nabla \log \mu^*(x_l) + \sqrt{2\gamma}\epsilon_l, \quad \epsilon_l \sim \mathcal{N}(0, \mathrm{Id}_{\mathbb{R}^d}).$$

- (2) can be approximated by a particle system; e.g. **Stein Variational Gradient Descent**[§] [Liu, 2017, Duncan et al., 2019] for some kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$:

$$x_{l+1}^i = x_l^i + \frac{\gamma}{N}\sum_{j=1}^N \nabla \log \mu^*(x_l^j)k(x_l^i, x_l^j) + \nabla_2 k(x_l^i, x_l^j), \quad i = 1, \dots, N.$$

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

Other choices?

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

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

  - It is not convenient neither when μ, μ^* are discrete
  - χ^2-gradient requires the normalizing constant of μ^*: $\nabla \frac{\mu}{\mu^*}$
  - However, the GF of χ^2 has interesting properties
    - we have $\chi^2(\mu|\mu^*) \geq \mathrm{KL}(\mu|\mu^*)$.
    - KL decreases exp. fast along CS flow/χ^2 decreases exp. fast along KL flow if μ^* satisfies Poincaré [Matthes et al., 2009]
    - χ^2 known to converge polynomially along its WGF [Dolbeault et al., 2007]

- If we pick $\mathcal{F} = W_2^2(\cdot, \mu^*)$, $\nabla_{W_2}\mathcal{F}(\mu) = \nabla f_{\mu,\mu^*}$ where f_{μ,μ^*} is the Kantorovitch potential between μ and μ^* (not closed-form, we need to solve an OT problem at each step: $\mathcal{O}(n^3)$ for n-sample distributions). Same story for Sinkhorn divergences ($\mathcal{O}(n^2/\epsilon^3)$).

Other choices?

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

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

  - It is not convenient neither when μ, μ^* are discrete
  - χ^2-gradient requires the normalizing constant of μ^*: $\nabla \frac{\mu}{\mu^*}$
  - However, the GF of χ^2 has interesting properties
    - we have $\chi^2(\mu|\mu^*) \geq \mathrm{KL}(\mu|\mu^*)$.
    - KL decreases exp. fast along CS flow/χ^2 decreases exp. fast along KL flow if μ^* satisfies Poincaré [Matthes et al., 2009]
    - χ^2 known to converge polynomially along its WGF [Dolbeault et al., 2007]

- If we pick $\mathcal{F} = W_2^2(\cdot, \mu^*)$, $\nabla_{W_2}\mathcal{F}(\mu) = \nabla f_{\mu,\mu^*}$ where f_{μ,μ^*} is the Kantorovitch potential between μ and μ^* (not closed-form, we need to solve an OT problem at each step: $\mathcal{O}(n^3)$ for n-sample distributions). Same story for Sinkhorn divergences ($\mathcal{O}(n^2/\epsilon^3)$).

Outline

1. Sampling as Optimization
2. Applications
3. Choice of the $\mathbb{D}$
4. De-regularized MMD
5. Mollified χ^2
6. Further connections with Optimization

Example 1: Bayesian (or Variational) inference

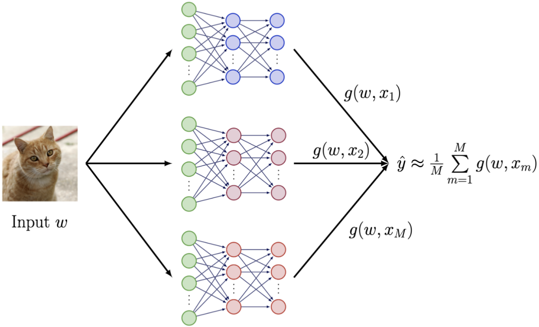
Given labelled data $(w_i, y_i)_{i=1}^m$, we want to sample from the posterior distribution over the parameters of a model $g(\cdot, x)$

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

Ensemble prediction for a new input w:

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

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



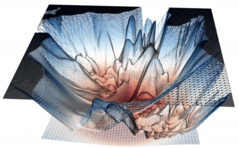
Sampling as minimization of the KL

$$\text{Recall } \mu^*(x) \propto \exp\left(-V(x)\right), \quad V(x) = \underbrace{\sum_{i=1}^{m} \|y_i - g(w_i, x)\|^2}_{\text{loss}} + \frac{\|x\|^2}{2}.$$

- LMC is known to be a GF of the KL w.r.t. the Wasserstein metric, while SVGD is w.r.t. to a "kernelized" Wasserstein metric, hence both solve

$$\min_{\mu} \mathrm{KL}(\mu|\mu^*)$$

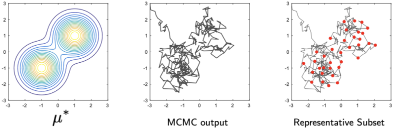
- if V is convex (e.g. $g(w,x) = \langle w, x\rangle$), these methods are known to work quite well [Durmus and Moulines, 2016, Vempala and Wibisono, 2019]
- but if its not (e.g. $g(w,x)$ is a neural network), the situation is much more delicate [Balasubramanian et al., 2022]



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

Example 2: Thinning (Postprocessing of MCMC output)

How can we post-process the MCMC output, and keep only the states that are representative of the posterior μ^* (e.g. to remove burn-in, correct time spent in each mode...)?



μ^* MCMC output Representative Subset

Picture from Chris Oates.

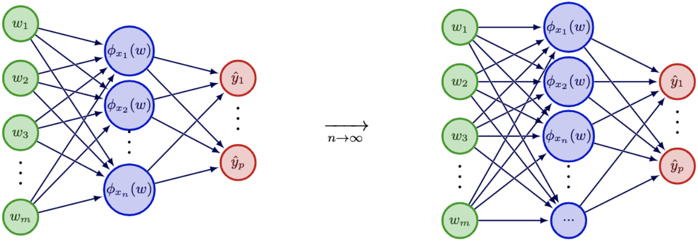
Idea: minimize a KSD divergence from the distribution of the states to μ^* [Riabiz et al., 2022], [KAMA21]:

$$\mu_n = \arg\min_{\mu} \mathrm{KSD}(\mu|\mu^*), \quad \mathrm{KSD}^2(\mu|\mu^*) = \iint k_{\mu^*}(x,y)d\mu(x)d\mu(y)$$

where $k_{\mu^*}(x,y) = k(x,y)\nabla \log \mu^*(x)^\top \nabla \log \mu^*(y) + \nabla_2 k(x,y)^\top \nabla \log \mu^*(x) + \nabla_1 k(x,y)^\top \nabla \log \mu^*(y) + \boldsymbol{\nabla} \cdot_1 \nabla_2 k(x,y)$, where k p.s.d. and smooth kernel e.g. $k(x,y) = e^{-\|x-y\|^2}$.

It is a specific case of MMD with kernel k_{μ^*}.

Example 3 : Regression with infinite-width shallow NN



$$\min_{(x_i)_{i=1}^n \in \mathbb{R}^d} \mathbb{E}_{(w,y)\sim P_{data}} \left[\left\| y - \underbrace{\frac{1}{n}\sum_{i=1}^n \phi_{x_i}(w)}_{\hat{y}} \right\|^2 \right] \xrightarrow[n\to\infty]{} \min_{\mu \in \mathcal{P}(\mathbb{R}^d)} \underbrace{\mathbb{E}_{(w,y)\sim P_{data}} \left[\left\| y - \int_{\mathbb{R}^d} \phi_x(w) d\mu(x) \right\|^2 \right]}_{\mathcal{F}(\mu)}$$

Optimising the neural network $\iff$ approximating $\mu^* \in \arg\min \mathcal{F}(\mu)$ [Chizat and Bach, 2018, Mei et al., 2018, Rotskoff and Vanden-Eijnden, 2018]

If $y(w) = \frac{1}{m}\sum_{i=1}^m \phi_{x_i}(w)$ is generated by a neural network (as in the student-teacher network setting), then $\mu^* = \frac{1}{m}\sum_{i=1}^m \delta_{x_m}$ and $\mathcal{F}$ can be identified to an MMD [AKSG2019]:

$$\min_{\mu} \mathbb{E}_{w\sim P_{data}} \left[\|y_{\mu^*}(w) - y_\mu(w)\|^2 \right] = \mathrm{MMD}^2(\mu, \mu^*), \quad k(x,x') = \mathbb{E}_{w\sim P_{data}}[\phi_{x'}(w)^T \phi_x(w)].$$

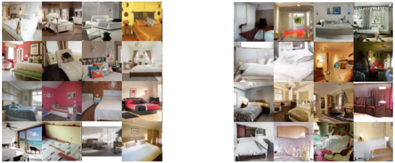
Example 4: Generative modelling

In generative modeling we want to generate novel samples from a distribution μ^* (given sample access).

Generative Adversarial Networks (GAN) or Normalizing Flows (NF) can be trained by minimizing specific distances or divergences:

$$\min_\theta \mathrm{D}(\mu_\theta | \mu^*)$$

where $\mu^* =$ distribution of the data samples, and $\mu_\theta =$ of the generative model.



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

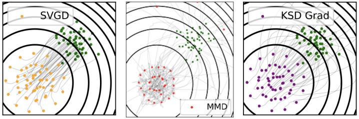
- for GANs: originally Jensen-Shannon [Goodfellow et al., 2014], but also MMD [Li et al., 2017], Sinkhorn divergence [Genevay et al., 2018],...
- for NF [Papamakarios et al., 2021], typically the likelihood (KL($\mu^*|\mu_\theta$)).

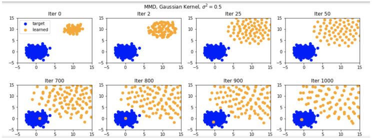
Outline

1. Sampling as Optimization
2. Applications
3. Choice of the D
4. De-regularized MMD
5. Mollified χ^2
6. Further connections with Optimization

Are all functionals good optimization objectives?

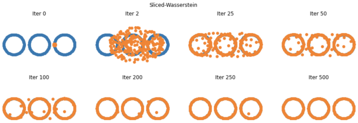
We already saw that **depending the application and the information on** μ^* (unnormalized density, samples...) we may pick the objective $\mathcal{F} = \mathrm{D}(\cdot|\mu^*)$ accordingly. But this is not all !

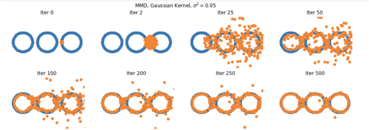




μ^*= 2d standard Gaussian μ^* (SVGD=KL objective, vs MMD/KSD)

Gradient flows of various $\mathrm{D}(\cdot|\mu^*)$ (to the same μ^*) behave very differently.





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

We want to study the convergence of Wasserstein gradient descent (Euler discretization of Wasserstein gradient flow)

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

In $\mathbb{R}^d$, fast rates are obtained if the objective function $f : \mathbb{R}^d \to \mathbb{R}$ is strongly convex and smooth, which is equivalent (if f twice differentiable) to lower and upper bounds on its Hessian:

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

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

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

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: $\mathrm{Hess}_\mu \mathcal{F}(\psi, \psi) := \left.\frac{\mathrm{d}^2}{\mathrm{d}t^2}\right|_{t=0} \mathcal{F}(\mu_t)$ and $(\mu_t, v_t)_{t\in[0,1]}$ is a Wasserstein geodesic with $\mu_0 = 0$, $v_0 = \nabla \psi$.

[¶] if $\lambda \geq 0$ we'll say $\mathcal{F}$ is geo. convex.

Convexity and Smoothness of KL and MMD

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

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

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

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

  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 [AKSG2019]

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

  It is M-smooth but not geodesically convex (Hessian lower bounded by a big negative constant). For KSD we obtain negative results even for strongly log concave μ^* [KAMA2021].

(Some) questions

1. what can we say on their geometrical properties?
2. are there IPMs (integral probability metrics) that enjoys a better behavior than the MMD?
3. are there good alternatives to the KL?

Outline

1 Sampling as Optimization

2 Applications

3 Choice of the $\mathbb{D}$

4 De-regularized MMD

5 Mollified χ^2

6 Further connections with Optimization

Discrete μ^*, and Variational formula of f-divergences

Assume we have sample access to μ^* (e.g. i.i.d. samples from μ^*).

Remember that MMD is convenient as an optimization objective but its WGF converges poorly, and KL is not well-suited for a discrete μ^*.

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

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

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

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

Examples:

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

Discrete μ^*, and Variational formula of f-divergences

Assume we have sample access to μ^* (e.g. i.i.d. samples from μ^*).

Remember that MMD is convenient as an optimization objective but its WGF converges poorly, and KL is not well-suited for a discrete μ^*.

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

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

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

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

Examples:

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

De-Regularized MMD:‖: Interpolate between MMD and χ^2

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

- **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**

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

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

where $\Sigma_{\mu^*} = \int k(\cdot,x)\otimes k(\cdot,x)d\mu^*(x)$, and h_{μ,μ^*} solves (4).

- In particular for μ,μ^* discrete (supported on N, M samples respectively), it writes with kernel Gram matrices over samples of μ,μ^* in complexity $\mathcal{O}(M^3+NM)$.

‖ with H. Chen, A. Gretton, P. Glaser (UCL), A. Mustafi, B. Sriperumbudur (CMU). Soon on arxiv.

Properties of DMMD

1. It is a reweighted χ^2-divergence: for $\mu \ll \mu^*$

$$\mathrm{DMMD}(\mu \| \mu^*) = (1+\lambda) \sum_{i \geq 1} \frac{\varrho_i}{\varrho_i + \lambda} \left\langle \frac{d\mu}{d\mu^*} - 1, e_i \right\rangle^2_{L^2(\mu^*)},$$

where (ρ_i, e_i) is the eigendecomposition of $\mathcal{T}_{\mu^*} : f \in L^2(\mu^*) \mapsto \int k(x,\cdot) f(x) d\mu^*(x) \in L^2(\mu^*)$.

2. It is an MMD with a regularized kernel:

$$\tilde{k}(x,x') = \sum_{i \geq 1} \frac{\varrho_i}{\varrho_i + \lambda} e_i(x) e_i(x')$$

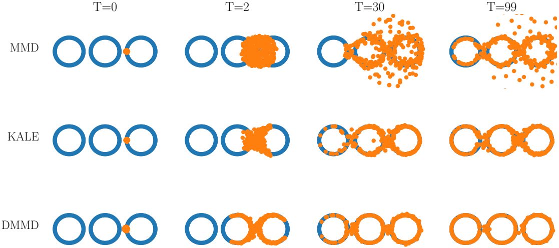
which is a regularized version of the original kernel $k\left(x,x'\right) = \sum_{i\geq 1} \varrho_i e_i(x) e_i(x')$.

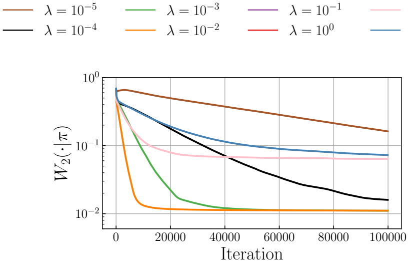
3. We can prove that for μ^* m-strongly log-concave, DMMD is strongly convex (it is always $\frac{1}{\lambda}$ smooth!)

Related work:

- Regularized MMD's ($\mathrm{DMMD}(\mu || \mu + \mu^*)$) appeared in: Eric, M., Bach, F., Harchaoui, Z. (2007). Testing for homogeneity with kernel Fisher discriminant analysis. Neurips
- Kernelization of KL divergence variational formulation (but is not closed-form !): Glaser, P., Arbel, M., Gretton, A. (2021). Kale flow: A relaxed kl gradient flow for probabilities with disjoint support. Neurips.
- Kernelization of f-divergences variational formulation in : Neumayer, S., Stein, V., Steidl, G. (2024). Wasserstein Gradient Flows for Moreau Envelopes of f-Divergences in Reproducing Kernel Hilbert Spaces. arXiv preprint arXiv:2402.04613.

Ring Experiment





Outline

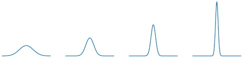
1. Sampling as Optimization
2. Applications
3. Choice of the $\mathbb{D}$
4. De-regularized MMD
5. Mollified χ^2
6. Further connections with Optimization

Another idea - "Mollified" discrepancies [LLKYS2022]

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

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

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



We propose the **Mollified chi-square**:

$$\begin{aligned}\mathcal{E}_\epsilon(\mu) &= \iint k_\epsilon(x-y)(\mu^*(x)\mu^*(y))^{-1/2}\mu(x)\mu(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int \left(k_\epsilon * \frac{\mu}{\sqrt{\mu^*}}\right)(x)\frac{\mu}{\sqrt{\mu^*}}(x)\,\mathrm{d}x \xrightarrow[\varepsilon\to 0]{} \chi^2(\mu|\mu^*)+1\end{aligned}$$

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

Sampling/Optimization with constraints

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

$$\min_{\mu \in \mathcal{P}_2(X)} \mathrm{D}(\mu \| \mu^*)$$

  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 $X = \{x \in R^d, \ g(x) \leq 0\}$ (e.g. fairness constraints)

  For (1) **”projected/mirror” methods:** Projected LMC [Bubeck et al., 2018], Mirror LMC [Ahn and Chewi, 2021], Mirror SVGD [Shi et al., 2022], for (2) we can use dynamic barrier [LLKYS2022]

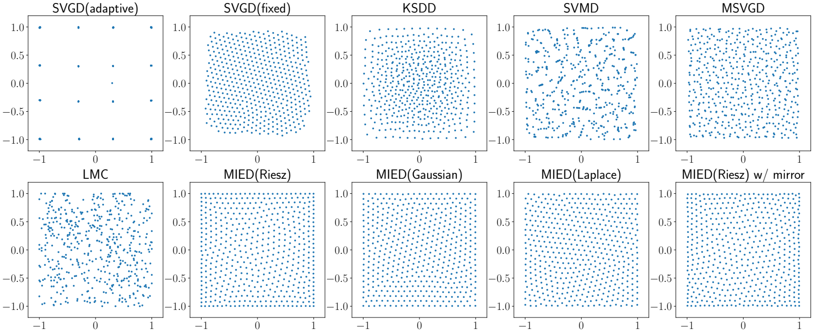
- Sampling with (population) inequality constraints [Liu et al., 2021]

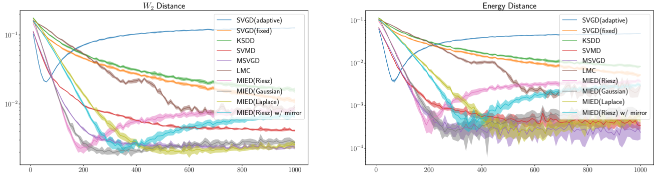
$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathrm{KL}(\mu \| \mu^*)$$

$$\text{subject to}\, \mathbb{E}_{x \sim \mu}\left[g(x)\right] \leq 0$$

using primal-dual optimization.

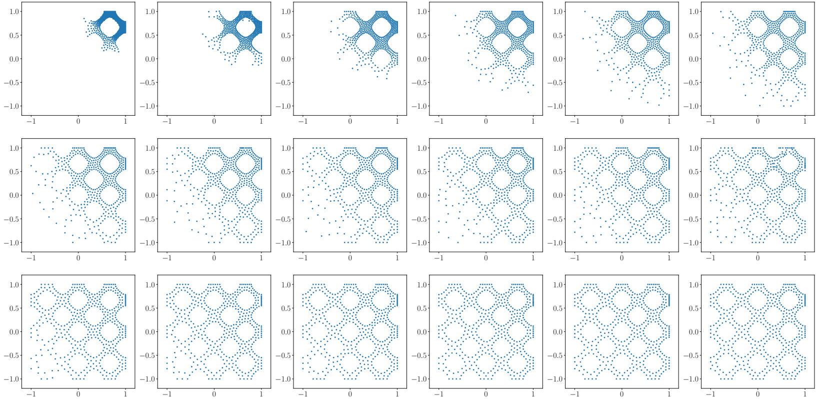
A numerical example from [LLKYS2022]





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 $f = tanh$.

A numerical example from [LLKYS2022]



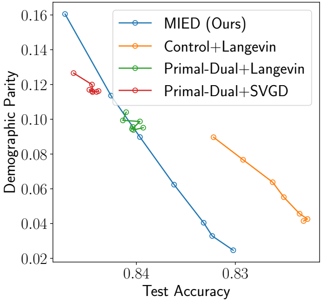
Uniform distribution on $X = \{(x, y) \in [-1, 1]^2 : (\cos(3\pi x) + \cos(3\pi y))^2 < 0.3\}$.
Mirror LMC/SVGD cannot be applied due to non convexity of the constraints.

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 $[0.5, 1.0]^2$.

Still [LLKYS2022] (Fair Bayesian Neural Network)

Given a dataset $\mathcal{D} = \{w^{(i)}, y^{(i)}, z^{(i)}\}_{i=1}^{|\mathcal{D}|}$ 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

1. Sampling as Optimization
2. Applications
3. Choice of the D
4. De-regularized MMD
5. Mollified χ^2
6. Further connections with Optimization

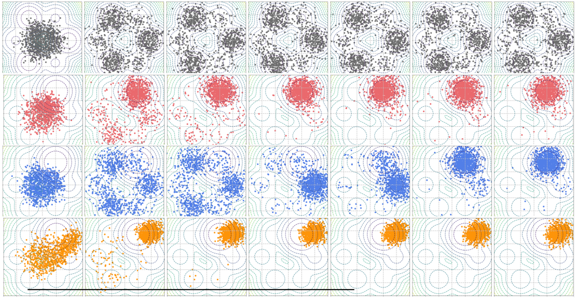
(1) Bilevel Sampling - Optimize while Sampling **

$$\min_{\theta \in \mathbb{R}^p} \ell(\theta) := \min_{\theta \in \mathbb{R}^p} \mathcal{F}(\mu^*(\theta))$$

where for instance $\mu^*(\theta)$ is a Gibbs distribution, minimizing the KL

$$\mu^*(\theta)[x] = \exp(-V(x,\theta))/Z_\theta \, .$$

Example: Reward training ($R(x) = \mathbf{1}_{x_1>0} \exp(-\|x-\mu\|^2)$) of Langevin diffusions, $V(\cdot,\theta)$ potential of a mixture of Gaussians parametrized by θ.



Sampling from $V(\cdot,\theta_0)$.

Sampling from $V(\cdot,\theta_{opt})$.

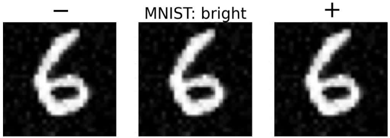
Bilevel approach.

$V = V(\cdot,\theta_{opt}) - \lambda R_{smooth}$

**Implicit Diffusion: Efficient Optimization through Stochastic Sampling. Pierre Marion, Anna Korba, Peter Bartlett, Mathieu Blondel, Valentin De Bortoli, Arnaud Doucet, Felipe Llinares-Lopez, Courtney Paquette, Quentin Berthet. `https://arxiv.org/abs/2402.05468`.

Extension to diffusion models [††]

$\mu^*(\theta)$: output of a diffusion model whose neural network is parametrized by θ. Different rewards are optimized while training.

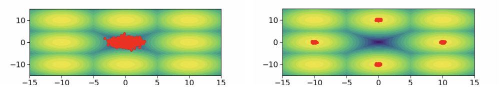


[††] Implicit Diffusion: Efficient Optimization through Stochastic Sampling. Pierre Marion, Anna Korba, Peter Bartlett, Mathieu Blondel, Valentin De Bortoli, Arnaud Doucet, Felipe Llinares-Lopez, Courtney Paquette, Quentin Berthet. `https://arxiv.org/abs/2402.05468`.

(2) The issue of multimodality and tempering

Langevin Monte Carlo, which is a discrete-time implementation of the Wasserstein gradient flow of the $\mathrm{KL}(\cdot|\mu^*)$.

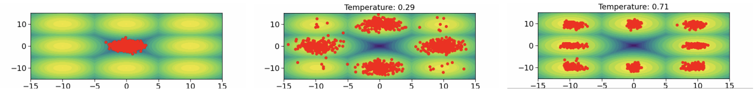
On a μ^* a mixture of Gaussians, it does not manage to target all modes in reasonable time, even in low dimensions.



Consider the sequence of tempered targets as:

$$\mu^*_\beta \propto \mu_0^\beta (\mu^*)^{1-\beta}, \quad \beta \in [0,1]$$

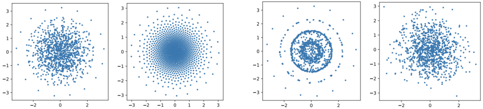
It corresponds to a **discretized Fisher-Rao gradient flow** of the KL [CCK2023].



Future directions

- Other divergences from the field of information/quantum theory? MMD with non-smooth/psd kernels [‡‡] e.g. $k(x,y) = -\|x-y\|^r$, $0 < r < 2$?
- How to improve the performance of the algorithms for highly non-log concave targets? e.g. through interpolations between μ_0 and μ^*?
- Shape of the trajectories? change the underlying metric and consider W_c gradient flows (e.g. like in SVGD)
- Derive theoretical guarantees
  - on the optimization error (how many iterations needed?)
  - on the quantization error (how many particles?)
  - on critical points, e.g. their stability

Some results exist for specific D but a lot remains to be done.



(a) SVGD Gaussian (b) NSVGD Laplace (c) MMD-lbfgs (d) i.i.d.

[‡‡] Hertrich, J., Wald, C., Altekrüger, F., Hagemann, P. (2023). Generative sliced MMD flows with Riesz kernels. arXiv preprint arXiv:2305.11463.

Main references

(code available):

- Maximum Mean Discrepancy Gradient Flow. Arbel, M., Korba, A., Salim, A., and Gretton, A. (Neurips 2019).
- Korba, A., Aubin-Frankowski, P. C., Majewski, S., Ablin, P. (2021, July). Kernel stein discrepancy descent. In International Conference on Machine Learning (ICML 2021).
- Accurate quantization of measures via interacting particle-based optimization. Xu, L., Korba, A., and Slepcev, D. (ICML 2022).
- Sampling with mollified interaction energy descent. Li, L., Liu, Q., Korba, A., Yurochkin, M., and Solomon, J. (ICLR 2023).
- Chopin, N., Crucinio, F. R., Korba, A. A connection between Tempering and Entropic Mirror Descent. arXiv preprint arXiv:2310.11914 (accepted to ICML 2024).
- Marion, P., Korba, A., Bartlett, P., Blondel, M., De Bortoli, V., Doucet, A., ... Berthet, Q. (2024). Implicit Diffusion: Efficient Optimization through Stochastic Sampling. arXiv preprint arXiv:2402.05468
- (De)-regularized Maximum Mean Discrepancy Gradient Flow. Chen, H., Mustafi, A., Glaser, P., Korba, A., Gretton, A., Sriperumbudur, B. (Submitted 2024)

References I

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

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

Balasubramanian, K., Chewi, S., Erdogdu, M. A., Salim, A., and Zhang, S. (2022).
Towards a theory of non-log-concave sampling: first-order stationarity guarantees for langevin monte carlo.
In *Conference on Learning Theory*, pages 2896–2923. PMLR.

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

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

Craig, K., Elamvazhuthi, K., Haberland, M., and Turanova, O. (2022).
A blob method method for inhomogeneous diffusion with applications to multi-agent control and sampling.
arXiv preprint arXiv:2202.12927.

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

References II

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

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

Genevay, A., Peyré, G., and Cuturi, M. (2018).
Learning generative models with sinkhorn divergences.
In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617. PMLR.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014).
Generative adversarial nets.
Advances in neural information processing systems, 27.

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

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

Li, J. and Barron, A. (1999).
Mixture density estimation.
Advances in neural information processing systems, 12.

References III

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

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

Liu, X., Tong, X., and Liu, Q. (2021).
Sampling with trusthworthy constraints: A variational gradient framework.
Advances in Neural Information Processing Systems, 34:23557–23568.

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

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

Mérigot, Q., Santambrogio, F., and Sarrazin, C. (2021).
Non-asymptotic convergence bounds for wasserstein approximation using point clouds.
Advances in Neural Information Processing Systems, 34:12810–12821.

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

References IV

Ohta, S.-i. and Takatsu, A. (2011).
Displacement convexity of generalized relative entropies.
Advances in Mathematics, 228(3):1742–1787.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021).
Normalizing flows for probabilistic modeling and inference.
Journal of Machine Learning Research, 22(57):1–64.

Riabiz, M., Chen, W. Y., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., and Oates, C. J. (2022).
Optimal thinning of mcmc output.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(4):1059–1081.

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

Shi, J., Liu, C., and Mackey, L. (2022).
Sampling with mirrored stein operators.
International Conference of Learning Representations.

Vempala, S. and Wibisono, A. (2019).
Rapid convergence of the unadjusted langevin algorithm: Isoperimetry suffices.
Advances in neural information processing systems, 32.

Villani, C. (2009).
Optimal transport: old and new, volume 338.
Springer.

References V

Xu, L., Korba, A., and Slepčev, D. (2022).

Accurate quantization of measures via interacting particle-based optimization.
International Conference on Machine Learning.

Other tempered paths

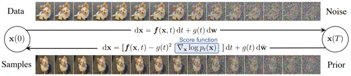
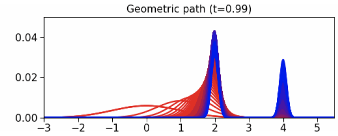

Data
Noise
$\mathbf{x}(0)$
$\mathbf{x}(T)$
$\mathrm{d}\mathbf{x} = \boldsymbol{f}(\mathbf{x},t)\,\mathrm{d}t + g(t)\,\mathrm{d}\mathbf{w}$
Score function
$\mathrm{d}\mathbf{x} = [\boldsymbol{f}(\mathbf{x},t) - g(t)^2 \boxed{\nabla_{\mathbf{x}} \log p_t(\mathbf{x})}\,]\,\mathrm{d}t + g(t)\,\mathrm{d}\bar{\mathbf{w}}$
Samples
Prior

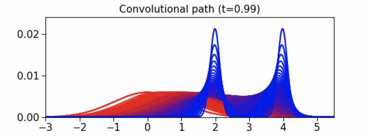
Figure by S. Coste, available at `https://scoste.fr/posts/diffusion/`.

"Convolutional path" ($\beta \in [0, +\infty[$) frequently used in Diffusion Models

$$\mu^*_\beta = \frac{1}{\sqrt{1-\beta}}\mu_0\left(\frac{\cdot}{\sqrt{1-\beta}}\right) * \frac{1}{\sqrt{\beta}}\mu^*\left(\frac{\cdot}{\sqrt{\beta}}\right)$$

(vs "geometric path" $\mu^*_\beta \propto \mu_0^\beta (\mu^*)^{1-\beta}$)





Strong convexity of DMMD

Let $\mu^* \propto e^{-V}$.

If V is m-strongly convex, for λ small enough, we can lower bound $\text{Hess}_\mu \, \mathrm{DMMD}(\cdot||\mu^*)(\psi, \psi)$ by a positive constant times $\|\nabla \psi\|^2_{L^2(\mu)}$, and obtain:

- a general existence result for $\mu \ll \mu^*$

$$\left|\text{Hess}_\mu \, \mathrm{DMMD}(\cdot||\mu^*)(\psi,\psi) - \text{Hess}_\mu \, \chi^2(\cdot||\mu^*)(\psi,\psi)\right| \\ \leq \sum_{i \geq 1} \frac{\lambda}{\varrho_i + \lambda} \left(K_{1d} + \sqrt{K_{2d}} \left\| \frac{\mu}{\mu^*} - 1 \right\|_{L^2(\mu^*)} \right) \|\nabla \psi\|^2_{L^2(\mu)}$$

- a "non-asymptotic" result wrt λ if we have a lower bound on the density ratios and a source condition ($\frac{\mu}{\mu^*} \in Ran(\mathcal{T}_\pi^r)$, $0 < r \leq \frac{1}{2}$)

$$\left|\text{Hess}_\mu \, \mathrm{DMMD}(\cdot||\mu^*)(\psi,\psi) - \text{Hess}_\mu \, \chi^2(\cdot||\mu^*)(\psi,\psi)\right| \\ \leq \left(K_{1d} + \lambda^r \sqrt{K_{2d}} \, \|q\|_{L^2(\mu^*)} \right) \|\nabla \psi\|^2_{L^2(\mu)}$$

where K_{1d} and K_{2d} are constants bounding the first and second derivatives of the kernel, and q is the preimage of $\frac{\mu}{\mu^*}$.

Idea of the proof

1. We can write Hessian of χ^2

$$\mathrm{Hess}_\mu\, \chi^2(\mu\|\mu^*) = \int \frac{\mu(x)^2}{\mu^*(x)} (L_{\mu^*}\psi(x))^2 dx$$
$$+ \int \frac{\mu(x)^2}{\mu^*(x)} \left\langle \mathrm{H}_V(x)\nabla\psi(x), \nabla\psi(x)\right\rangle dx + \int \frac{\mu(x)^2}{\mu^*(x)} \|\mathrm{H}\psi(x)\|_{HS}^2\, dx$$

   where L_{μ^*} is the Langevin diffusion operator
   $L_{\mu^*}\psi = \langle \nabla V(x), \nabla\psi(x)\rangle - \Delta\psi(x)$.

2. $\mathrm{DMMD}(\mu\|\mu^*) = (1+\lambda)\sum_{i\geq 1} \frac{\varrho_i}{\varrho_i+\lambda} \left\langle \frac{d\mu}{d\mu^*} - 1, e_i \right\rangle_{L^2(\mu^*)}^2$, where (ρ_i, e_i) eigendecomposition of
   $\mathcal{T}_{\mu^*} : f \in L^2(\mu^*) \mapsto \int k(x,\cdot)f(x)d\mu^*(x) \in L^2(\mu^*)$

Quantization - classical results

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

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

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

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

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

where $C^2_{\mu^*,P} = \int \frac{\int k_\epsilon(x-m)^2 dP(m)}{(\int k_\epsilon(x-w)dP(w))^2} d\mu^*(x)$, 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 μ^* with exponential tails, we get using Koksma-Hlawka inequality

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

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

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

- For the reverse KL (joint work with Tom Huix) we get (in the well-specified case) adapting the proof of [Li and Barron, 1999]:

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

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

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

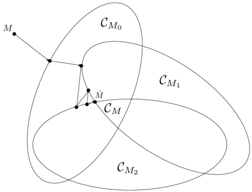
Generalized dynamic barrier: Dykstra's algorithm

Observe that

$$\min_{v \in \mathbb{R}^d} \left\| v - \nabla_{x_i} E_\epsilon(\omega_N^t) \right\|^2 \text{ s.t. } \forall j = 1, \dots, m, \nabla g_j(x_i^t)^\top v \geq \alpha_i g_j(x_i^t),$$

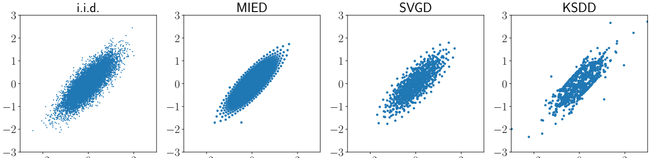
is the same as projecting $\nabla_{x_i} E_\epsilon(\omega_N^t)$ on $\cap_{i=1}^m \{x \in \mathbb{R}^d, \nabla g_i(x^t)^\top v \geq \alpha_i g_i(x^t)\}$.

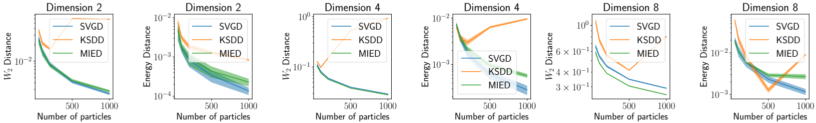
we use Dykstra's projection algorithm which in this case is the same as running coordinate descent on the dual problem, and hence with fast linear convergence rate.



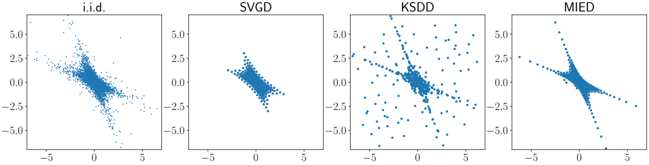
Since the constraints are the same for all particles, we can parallelize Dykstra's algorithm by using a fixed maximum number of iterations for all particles to find the update direction v_i^* for each i.

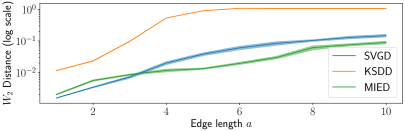
Unconstrained examples I - Gaussian

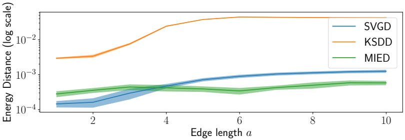




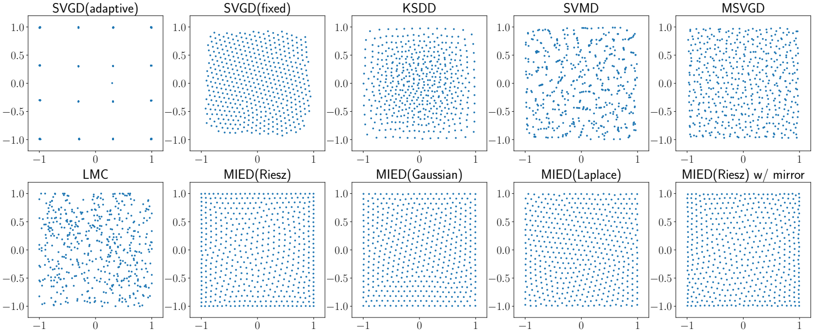
II - Product of two Student's t-distributions (heavy tail)

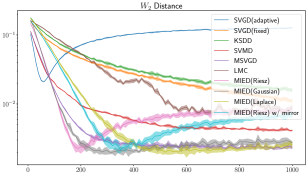


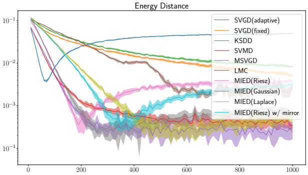




Constrained example I - Uniform sampling in a box







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 $f = tanh$.

Sensitivity to the mirror map

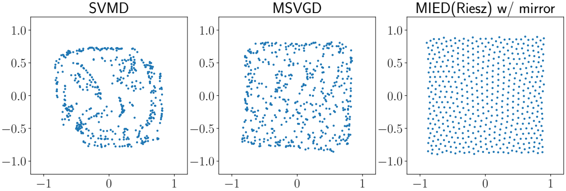


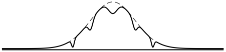
Figure: Visualization of samples for uniform sampling from a 2D box when using a suboptimal mirror map. All three methods fail to draw samples near the boundary of the box $[-1, 1]^2$.

Here we use the mirror map $\phi(\theta) = \sum_{i=1}^{n} \left(\log \frac{1}{1-\theta_i} + \log \frac{1}{1+\theta_i} \right)$ as in [Ahn and Chewi, 2021].

Relaxation of convexity: functional inequalities

It is also possible to show fast rates of convergence for Wasserstein gradient descent (or related schemes) if we have inequalities of the form $\mathcal{F}(\mu) \leq \frac{1}{\lambda}\|\nabla_{W_2}\mathcal{F}(\mu)\|^2_{L^2(\mu)}$ where the r.h.s. corresponds to the dissipation of $\mathcal{F}$ along the flow.

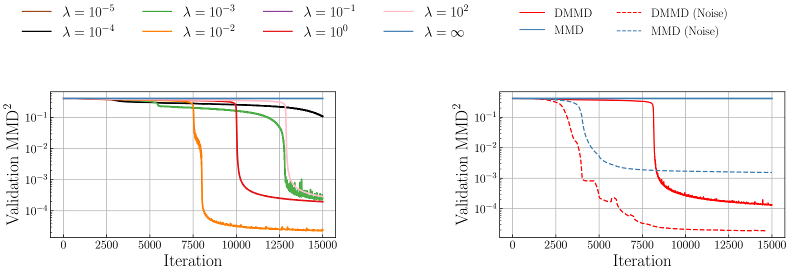
- For the KL along its WGF it corresponds to the log-Sobolev inequality



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

- for SVGD on the r.h.s. we have $\mathrm{KSD}^2(\mu|\mu^*)$, which is hard to achieve for smooth kernels [Duncan et al., 2019]
- for MMD we can obtain a functional inequality, but where λ depends on the whole trajectory, and may be vacuous for discrete measures [AKSG2019]

Student-teacher networks experiment



- the teacher network $w \mapsto y_{\mu^*}(w)$ is given by M particles $(\xi_1, \ldots, \xi_M)$ which are fixed during training $\implies \mu^* = \frac{1}{M}\sum_{j=1}^{M} \delta_{\xi_j}$
- the student network $w \mapsto y_{\mu}(w)$ has n particles $(x_1, \ldots, x_n)$ that are initialized randomly $\implies \mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_j}$

$$\min_{\mu} \mathbb{E}_{w \sim P_{data}} \left[(y_{\mu^*}(w) - y_{\mu}(w)^2 \right]$$

$$\iff \min_{\mu} \mathrm{MMD}(\mu, \mu^*) \text{ with } k(x, x') = \mathbb{E}_{w \sim P_{data}}[\phi_{x'}(w)\phi_x(w)].$$

Same setting as [AKSG2019].