Sampling through Optimization of Divergences

Anna Korba

ENSAE, CREST, Institut Polytechnique de Paris

OT Berlin 2024

Joint work with many people cited on the flow.

Outline

Why sampling?

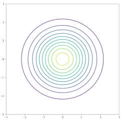
Suppose you are interested in some target probability distribution on $\mathbb{R}^d$, denoted μ^*, and you have access only to partial information, e.g.:

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...)

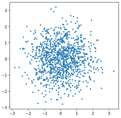
Problem: approximate $\mu^* \in \mathcal{P}(\mathbb{R}^d)$ by a finite set of n points $x_1, \dots, x_n$, e.g. to compute functionals $\int_{\mathbb{R}^d} f(x) d\mu^*(x)$.

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

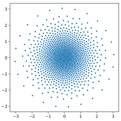
$$\left| \frac{1}{n}\sum_{i=1}^n f(x_i) - \int_{\mathbb{R}^d} f(x) d\mu^*(x) \right|.$$



a Gaussian density



I.i.d. samples.



Particle scheme (SVGD).

Example 1: Bayesian inference

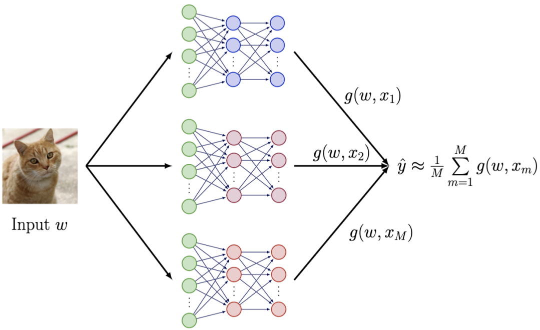
We want to sample from the posterior distribution

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

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 $\mu^*(x)$.

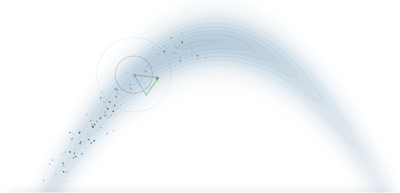


(Some, Non parametric, Unconstrained) Sampling methods

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

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

$$x_{t+1} = x_t - \gamma \nabla V(x_t) + \sqrt{2\gamma}\epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \mathrm{Id}_{\mathbb{R}^d}).$$



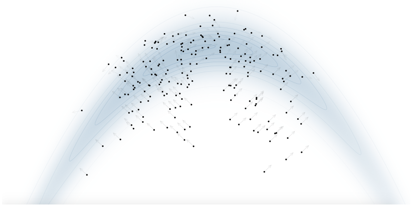
Picture from `https://chi-feng.github.io/mcmc-demo/app.html`.

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

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

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

where $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is a kernel (e.g. $k(x,y) = \exp\left(-\|x-y\|^2\right)$).



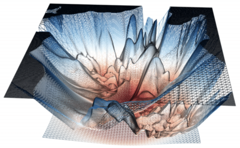
Picture from https://chi-feng.github.io/mcmc-demo/app.html.

Difficult cases (in practice and in theory)

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

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

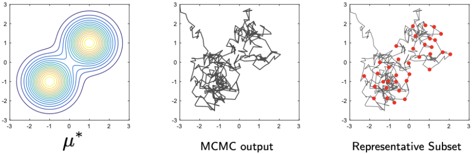
- if V is convex (e.g. $g(w,x) = \langle w, x\rangle$), these methods are known to work quite well [Durmus and Moulines, 2016, Vempala and Wibisono, 2019]
- but if its not (e.g. $g(w,x)$ is a neural network), the situation is much more delicate



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)

In an ideal world we would be able to post-process the MCMC output and keep only those states that are representative of the posterior μ^*.



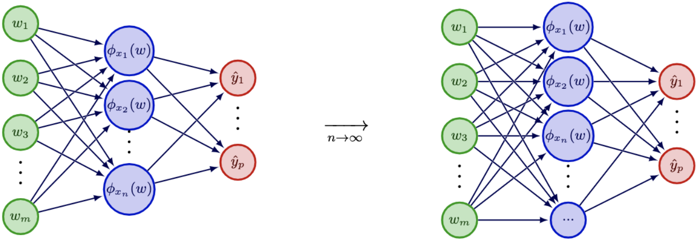
Picture from Chris Oates.

- Fix problems with MCMC (automatic identification of burn-in; number of points proportional to the probability mass in a region; etc.)
- Compressed representation of the posterior, to reduce any downstream computational load.

Idea: minimize a divergence from the distribution of the states to μ^* [Korba et al., 2021]:

$$\mu_n = \arg\min_{\mu} \mathrm{KSD}(\mu|\mu^*)$$

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]

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 [Arbel et al., 2019]:

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

Outline

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:

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

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...

Starting from an initial distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, one can then consider the **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.

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

The first variation of $\mu \mapsto \mathcal{F}(\mu)$ evaluated at $\mu \in \mathcal{P}(\mathbb{R}^d)$ is the unique function $\frac{\partial \mathcal{F}(\mu)}{\partial \mu} : \mathbb{R}^d \to \mathbb{R}$ s. t. for any $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, $\nu - \mu \in \mathcal{P}(\mathbb{R}^d)$:

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

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:

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

where $\nabla_{W_2} \mathcal{F}(\mu) := \nabla \frac{\partial \mathcal{F}(\mu)}{\partial \mu}$ denotes the Wasserstein gradient of $\mathcal{F}$.

It can be implemented by the deterministic process:

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

Particle system/Gradient descent approximating the WGF

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

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

In particular, if $\mathcal{F}$ is well-defined for discrete measures, the algorithm above 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 $\mu^N = \frac{1}{N} \sum_{i=1}^N \delta_{x^i}$.

We consider several questions:

- what can we say as time goes to infinity ? (optimization error) $\Longrightarrow$ heavily linked with the geometry (convexity, smoothness in the Wasserstein sense) of the loss
- (for minimizers) what can we say as the number of particles grow ? (quantization error)

Outline

Loss function for the unnormalized densities - the KL

Recall that we want to minimize $\mathcal{F} = \mathrm{D}(\cdot|\mu^*)$. Which D can we choose? For instance, D could be the Kullback-Leibler divergence:

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

The KL as an objective is convenient when the unnormalized density of μ^* is known since it **does not depend on the normalization constant!**

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

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

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

KL Gradient flow in practice

- The gradient flow of the KL can be implemented via the Probability Flow (ODE):

$$d\tilde{x}_t = -\nabla \log\left(\frac{\mu_t}{\mu^*}\right)(\tilde{x}_t)dt \tag{1}$$

or the Langevin diffusion (SDE):

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

(they share the same marginals $(\mu_t)_{t\geq 0}$)

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

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

- (1) can be approximated by a particle system (e.g. Stein Variational Gradient Descent [Liu, 2017, He et al., 2022])
- however MCMC methods suffer an integral approximation error of order $\mathcal{O}(n^{-1/2})$ if we use $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ (x_i iterates of MCMC) [Łatuszyński et al., 2013], and for SVGD we don't know [Xu et al., 2022].

Another f-divergence?

Consider the chi-square (CS) divergence:

$$\chi^2(\mu|\mu^*) := \begin{cases} \int \left(\frac{\mathrm{d}\mu}{\mathrm{d}\mu^*} - 1\right)^2 \mathrm{d}\mu^* & \mu \ll \mu^* \\ +\infty & \text{otherwise.} \end{cases}$$

- 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
  - KL decreases exp. fast along CS flow/χ^2 decreases exp. fast along KL flow if μ^* satisfies Poincaré
  - we have $\chi^2(\mu|\mu^*) \geq \mathrm{KL}(\mu|\mu^*)$.

$\Longrightarrow$ distinguishing whether KL or χ^2 GF is more favorable is an active area of research†

†see [Chewi et al., 2020, Craig et al., 2022] for a discussion, results from [Matthes et al., 2009, Dolbeault et al., 2007]

Losses for the discrete case

When we have a discrete approximation of μ^*, it is convenient to choose D as an integral probability metric (to approximate integrals).

For instance, D could be the MMD (Maximum Mean Discrepancy):

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

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 $\mathcal{H}_k$ is the RKHS associated to k‡.

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

Why we care about the loss

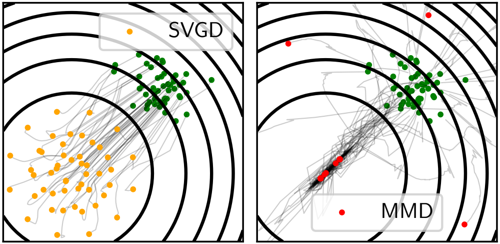


Figure: Toy example with 2D standard Gaussian. The green points represent the initial positions of the particles. The light grey curves correspond to their trajectories.

Gradient flow of the KL to a Gaussian $\mu^*(x) \propto e^{-\frac{\|x\|^2}{2}}$ is well-behaved, but not the MMD.

Question: is there an IPM (integral probability metric) that enjoys a better behavior?

Variational formula of f-divergences

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]:

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

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

Examples:

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

A proposal[§]: Interpolate between MMD and χ^2

"De-Regularized MMD" leverages the variational formulation of χ^2:

$$\mathrm{DMMD}(\mu||\mu^*) = (1+\lambda)\Big\{ \max_{h\in\mathcal{H}_k} \int h d\mu - \int (h + \frac{1}{4}h^2) d\mu^* - \frac{1}{4}\lambda \|h\|^2_{\mathcal{H}_k} \Big\} \quad (3)$$

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, in particular if μ, μ^* are discrete** (depends on λ and kernel matrices over samples of μ, μ^*):

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

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

where $\Sigma_{\mu^*} = \int k(\cdot, x) \otimes k(\cdot, x) d\mu^*(x)$, where $(a \otimes b)c = \langle b, c\rangle_{\mathcal{H}_k} a$; and h^*_{μ,μ^*} solves (3).

Complexity: $\mathcal{O}(M^3 + NM)$ for μ^*, μ supported on M, N atoms, can be decreased to $\mathcal{O}(M + N)$ with random features.

A similar idea was proposed for the KL, yielding Kale divergence [Glaser et al., 2021] but was not closed-form.

[§] with H. Chen, A. Gretton, P. Glaser (UCL), A. Mustafi, B. Sriperumbudur (CMU)

Several interpretations of DMMD

DMMD can be seen as:

- A reweighted χ^2-divergence: for $\mu \ll \pi$

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

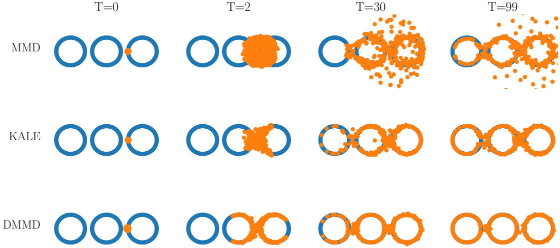
  where (ρ_i, e_i) is the eigendecomposition of $\mathcal{T}_\pi : f \in L^2(\pi) \mapsto \int k(x,\cdot) f(x) d\pi(x) \in L^2(\pi)$.

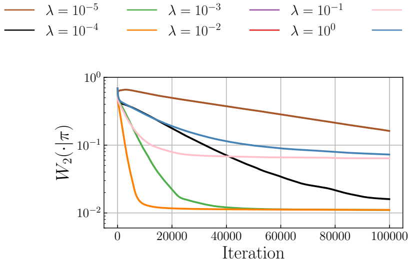
- An MMD with the kernel:

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

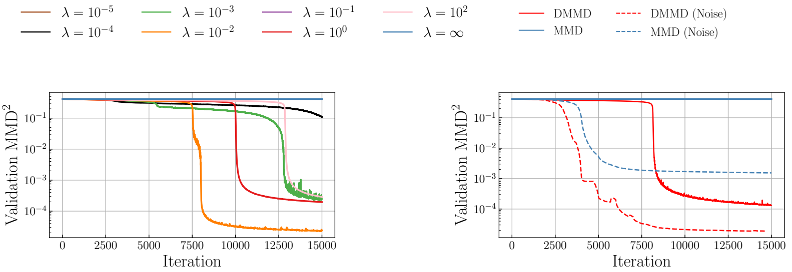
  which is a regularized version of the original kernel $k(x, x') = \sum_{i \geq 1} \varrho_i e_i(x) e_i(x')$.

Ring Experiment





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 $\Longrightarrow \mu^* = \frac{1}{M}\sum_{j=1}^{M} \delta_{\xi_j}$
- the student network $w \mapsto y_{\mu}(w)$ has n particles $(x_1, \ldots, x_n)$ that are initialized randomly $\Longrightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_j}$

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

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

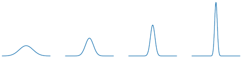
¶Same setting as [Arbel et al., 2019].

Another idea - "Mollified" discrepancies [Li et al., 2022a]‖

What if we don't have access to samples of μ^*? (recall that DMMD involves an integral over μ^*)

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

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



- Mollified chi-square (differs from $\chi^2(k_\epsilon \star \mu|\mu^*)$ as in [Craig et al., 2022]):

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

It writes as an interaction energy, allowing to consider μ discrete and μ^* with a density.

‖Sampling with mollified interaction energy descent. Li, L., Liu, Q., Korba, A., Yurochkin, M., and Solomon, J. (ICLR 2023).

Outline

Background on convexity and smoothness in $\mathbb{R}^d$

Recall that if $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable,

- f is λ-convex

$$\forall x, y \in \mathbb{R}^d, t \in [0,1] :$$

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{\lambda}{2} t(1-t)\|x-y\|^2$$
$$\iff v^T \nabla f(x) v \geq \lambda \|v\|_2^2 \quad \forall x, v \in \mathbb{R}^d.$$

- f is M-smooth

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

(Geodesically)-convex and smooth losses

$\mathcal{F}$ is said to be λ-displacement convex if along W_2 geodesics $(\rho_t)_{t\in[0,1]}$:

$$\mathcal{F}(\rho_t) \leq (1-t)\mathcal{F}(\rho_0) + t\mathcal{F}(\rho_1) - \frac{\lambda}{2}t(1-t)W_2^2(\rho_0,\rho_1) \qquad \forall\, t \in [0,1].$$

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:

$$\mathrm{Hess}_\mu \mathcal{F}(\psi,\psi) := \left.\frac{\mathrm{d}^2}{\mathrm{d}t^2}\right|_{t=0} \mathcal{F}(\mu_t)$$

where $(\mu_t, v_t)_{t\in[0,1]}$ is a Wasserstein geodesic with $\mu_0 = 0$, $v_0 = \nabla\psi$.

$$\boxed{\mathcal{F} \text{ is } \lambda\text{-displacement convex} \iff \mathrm{Hess}_\mu \mathcal{F}(\psi,\psi) \geq \lambda \|\nabla\psi\|^2_{L^2(\mu)}}$$

(See [Villani, 2009, Proposition 16.2]). In an analog manner we can define smooth functionals as functionals with upper bounded Hessians.

Guarantees for Wasserstein gradient descent

Consider Wasserstein gradient descent (Euler discretization of Wasserstein gradient flow)

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

Assume $\mathcal{F}$ is M-smooth. Then, we have a descent lemma (if $\gamma < \frac{2}{M}$):

$$\mathcal{F}(\mu_{l+1}) - \mathcal{F}(\mu_l) \leq -\gamma \left(1 - \frac{\gamma}{2} M\right) \|\nabla \mathcal{F}'(\mu_l)\|^2_{L^2(\mu_l)}.$$

Moreover, if $\mathcal{F}$ is λ-convex, we have the global rate

$$\mathcal{F}(\mu_L) \leq \frac{W_2^2(\mu_0, \mu^*)}{2\gamma L} - \frac{\lambda}{L} \sum_{l=0}^{L} W_2^2(\mu_l, \mu^*).$$

(so the barrier term degrades with λ).

Some examples

- Let $\mu^* \propto e^{-V}$, we have [Wibisono, 2018]

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

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

$$\text{Hess}_\mu \, \text{MMD}^2(\psi, \psi) = \int \nabla \psi(x)^\top \nabla_1 \nabla_2 k(x, y) \nabla \psi(y) d\mu(x) d\mu(y) +$$

$$2 \int \nabla \psi(x)^\top \left(\int \mathrm{H}_1 k(x, z) \, d\mu(z) - \int \mathrm{H}_1 k(x, z) \, d\mu^*(z) \right) \nabla \psi(x) d\mu(x)$$

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

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

- Th1, informal: for step size γ small enough (depending on λ, k) we get a $\mathcal{O}(1/L)$ rate
- Th2, informal: we can obtain a linear $\mathcal{O}(e^{-L})$ rate if we have a lower bound on the density ratios and a source condition ($\frac{\mu}{\pi} \in Ran(\mathcal{T}_\pi^r)$, $0 < r \leq \frac{1}{2}$)

Idea:

1. We can write Hessian of χ^2

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

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

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

Outline

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):

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

- we can apply these results to DMMD which is a regularized MMD with kernel $\tilde{k}$, replacing C_d by $\frac{C_d}{\lambda}$.

Outline

More ideas can be borrowed to optimization (but there are limitations)

- Sampling with inequality constraints
  [Liu et al., 2021, Li et al., 2022b]

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathrm{KL}(\mu \| \mu^*)$$
$$\text{subject to } \mathbb{E}_{x\sim\mu}\big[g(x)\big] \leq 0$$

- Bilevel 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
  $$\mu^*(\theta)[x] = \exp(-V(x,\theta))/Z_\theta \,.$$
  - $\mu^*(\theta)$ is the output of a Diffusion model parametrized by θ, this does not minimize a divergence on $\mathcal{P}(\mathbb{R}^d)$

**with P. Marion, Q. Berthet, P. Bartlett, M. Blondel, V. Bortoli, A. Doucet, F. Llinares-Lopez, C. Paquette

A numerical example from [Li et al., 2022a]

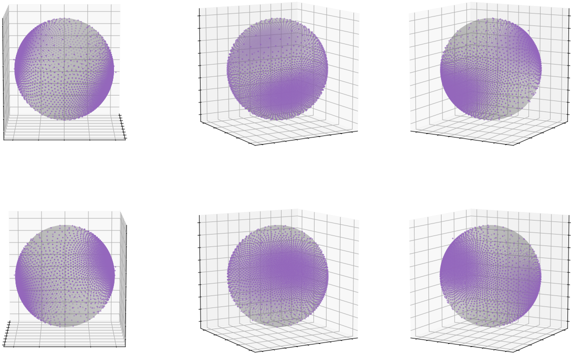
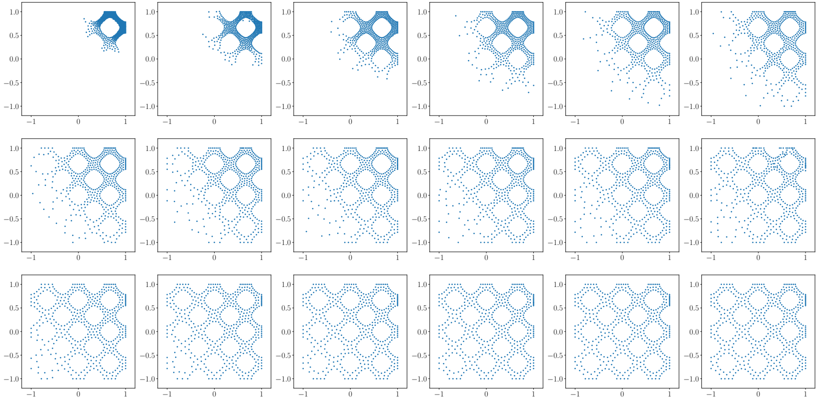


Figure: Sampling from the von Mises-Fisher distribution obtained by constraining a 3-dimensional Gaussian to the unit sphere. The unit-sphere constraint is enforced using the dynamic barrier method and the shown results are obtained using MIED with Riesz kernel and $s = 3$. The six plots are views from six evenly spaced angles.

A numerical example from [Li et al., 2022a]

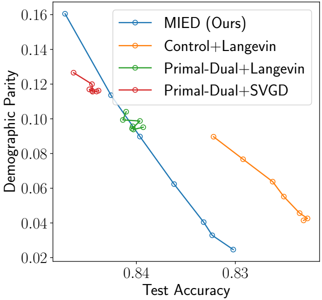


Uniform sampling of the region $\{(x,y) \in [-1,1]^2 : (\cos(3\pi x) + \cos(3\pi y))^2 < 0.3\}$ using MIED with a Riesz mollifier ($s = 3$) where the constraint is enforced using the dynamic barrier method.

IV - Fairness Bayesian neural network

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

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



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

Open questions, directions

- Finite-particle/quantization guarantees are still missing for many losses (e.g. mollified chi-square)

$$\mathrm{D}(\mu_n||\mu^*) \leq \mathsf{error}(n, \mu^*)?$$

- How to improve the performance of the algorithms for highly non-log concave targets? e.g. through sequence of targets $(\mu^*)_{t\in[0,1]}$ interpolating between μ_0 and μ^*?
- Shape of the trajectories? change the underlying metric and consider W_c gradient flows

Main references

(code available):

- Maximum Mean Discrepancy Gradient Flow. Arbel, M., Korba, A., Salim, A., and Gretton, A. (Neurips 2019).
- 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).
- (De)-regularized Maximum Mean Discrepancy Gradient Flow. Chen, H., Mustafi, A., Glaser, P., Korba, A., Gretton, A., Sriperumbudur, B. (Submitted 2024)

References I

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

Arbel, M., Korba, A., Salim, A., and Gretton, A. (2019).
Maximum mean discrepancy gradient flow.
In *Advances in Neural Information Processing Systems*, pages 6481–6491.

Chewi, S., Le Gouic, T., Lu, C., Maunu, T., and Rigollet, P. (2020).
Svgd as a kernelized wasserstein gradient flow of the chi-squared divergence.
Advances in Neural Information Processing Systems, 33:2098–2109.

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.

Chopin, N., Crucinio, F. R., and Korba, A. (2023).
A connection between tempering and entropic mirror descent.
arXiv preprint arXiv:2310.11914.

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.

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

References II

Glaser, P., Arbel, M., and Gretton, A. (2021).
Kale flow: A relaxed kl gradient flow for probabilities with disjoint support.
Advances in Neural Information Processing Systems, 34:8018–8031.

He, Y., Balasubramanian, K., Sriperumbudur, B. K., and Lu, J. (2022).
Regularized stein variational gradient flow.
arXiv preprint arXiv:2211.07861.

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

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. (2021).
Kernel stein discrepancy descent.
arXiv preprint arXiv:2105.09994.

Łatuszyński, K., Miasojedow, B., and Niemiro, W. (2013).
Nonasymptotic bounds on the estimation error of mcmc algorithms.
Bernoulli, 19(5A):2033–2066.

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

Li, L., Liu, Q., Korba, A., Yurochkin, M., and Solomon, J. (2022a).
Sampling with mollified interaction energy descent.
arXiv preprint arXiv:2210.13400.

References III

Li, R., Tao, M., Vempala, S. S., and Wibisono, A. (2022b).
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, Q. and Wang, D. (2016).
Stein variational gradient descent: A general purpose bayesian inference algorithm.
In *Advances in neural information processing systems*, pages 2378–2386.

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.

References IV

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.

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

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

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.

Wibisono, A. (2018).
Sampling as optimization in the space of measures: The langevin dynamics as a composite optimization problem.
In *Conference on Learning Theory*, pages 2093–3027. PMLR.

Xu, L., Korba, A., and Slepčev, D. (2022).
Accurate quantization of measures via interacting particle-based optimization.
International Conference on Machine Learning.

Outline

What is known

What can we say on $\inf_{x_1,\dots,x_n} \mathrm{D}(\mu_n|\mu^*)$ 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 $\gamma = 4\log(3\sqrt{e} + a)$ is a constant depending on ϵ with $a = \sup_{z,z' \in \mathbb{R}^d} \log\left(k_\epsilon(x-z)/k_\epsilon(x-z')\right)$.

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]:

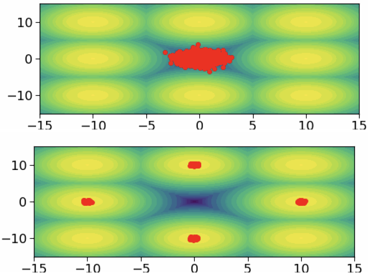
$$\min_{\mu_n} \mathrm{KL}(k_\epsilon \star \mu | \mu^*) \leq C_{\mu^*}^2 \frac{\log(n) + 1}{n}.$$

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

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

Mixture of Gaussians

Langevin Monte Carlo on a mixture of Gaussians does not manage to target all modes in reasonable time, even in low dimensions.



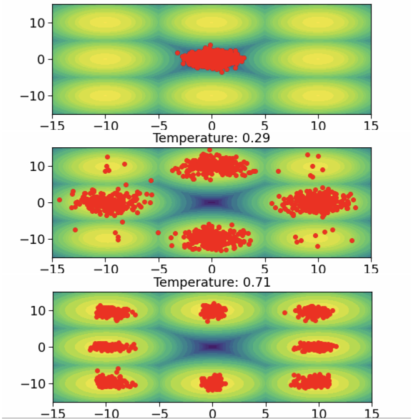
Picture from O. Chehab.

Annealing

One possible fix : sequence of tempered targets as:

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

It is **discretized Fisher-Rao gradient flow** [Chopin et al., 2023].



Other tempered path

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