Accurate Quantization of Measures via Interacting Particle-based Optimization

Lantian Xu[1], Anna Korba[2], Dejan Slepčev[1]

[1]Carnegie Mellon University
[2]ENSAE, CREST, IP Paris

ICML 2022

Quantization problem

Problem : Approximate a target distribution $\pi \in \mathcal{P}(\mathbb{R}^d)$ by a finite set of n points $x_1, \dots, x_n$,

Aim. Approximate integrals of functions f:

$$err(x_1, \dots, x_n) = \left| \frac{1}{n} \sum_{i=1}^{n} f(x_i) - \int_{\mathbb{R}^d} f(x) d\pi(x) \right| .$$

Several approaches, among which :

- MCMC methods : generate a Markov chain whose law converges to π. $err(x_1, \dots, x_n) = \mathcal{O}(n^{-1/2})$ [Łatuszyński et al., 2013]
- **interacting particle-based algorithms**. Goal: Smaller $err(x_1, \dots, x_n)$.

Sampling as optimization over distributions

4 algorithms/particle systems at study:

- Maximum Mean Discrepancy Descent [Arbel et al., 2019]
- Kernel Stein Discrepancy Descent [Korba et al., 2021]
- Stein Variational Gradient Descent [Liu and Wang, 2016]
- Normalized Stein Variational Gradient Descent

The sampling task can be recast as an optimization problem:

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

where D is a **dissimilarity functional** and $\mathcal{F}$ **"a loss"**.

Starting from an initial distribution $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$, one can consider the **gradient flow** of $\mathcal{F}$ to transport μ_0 to π.

MMD and KSD Descent

For MMD $$\mathcal{F}(\mu) = \sup_{\|f\|_{H_k} \leq 1} \int f \, d\mu - \int f \, d\pi$$

MMD/KSD are well defined for discrete measures $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X^i}$; let $F(X^1, \ldots, X^n) := \mathcal{F}(\mu)$. MMD descent is the gradient flow of F.

- If $\mathcal{F}$ is the MMD, the gradient of F is

$$\nabla_{X^i} F(X^1, \ldots, X^n) = \frac{1}{n} \sum_{j=1}^{n} \nabla_2 k(X^i, X^j) - \int \nabla_2 k(X^i, x) d\pi(x).$$

- If $\mathcal{F}$ is the KSD,

$$\nabla_{X^i} F(X^1, \ldots, X^n) = \frac{1}{n} \sum_{j=1}^{n} \nabla_2 k_\pi(X^i, X^j).$$

MMD/KSD Descent: at each time $l \geq 0$ and time step γ

$$X_{l+1}^i = X_l^i - \gamma \nabla_{X^i} F(X_l^1, \ldots, X_l^n) \qquad i = 1, \ldots, n.$$

Stein Variational Gradient Descent

Let $\pi \sim e^{-U}$. In continuum, SVGD flow is defined by the equation

$$\frac{\partial \mu_t}{\partial t} + \boldsymbol{\nabla} \cdot (\mu_t v_{\mu_t}) = 0, \;\; v_{\mu_t} = k \star (\mu_t \nabla U) - \nabla k \star \mu_t,$$

It is the gradient flow of the **KL divergence** with respect to **Stein metric**, studied by [Duncan et al., 2019]

SVGD: let $\gamma > 0$ be the step-size. Starting from $x_0^1, \dots, x_0^n \sim \mu_0$, SVGD algorithm updates the n particles as follows at each iteration :

$$x_{l+1}^i = x_l^i - \frac{\gamma}{n} \sum_{j=1}^{n} \left[-\nabla U(x_l^j) k(x_l^i, x_l^j) + \nabla_{x_l^j} k(x_l^i, x_l^j) \right].$$

Remark: SVGD flow is quadratic in density μ, which means the velocity would be small in low density regions.

Normalized Stein Variational Gradient Descent

Introduce another kernel of bandwidth $h > 0$: $\eta_h(x-y) = \frac{1}{h^d}\eta\left(\frac{x-y}{h}\right)$ and let $\mu_h = \mu * \eta_h$. We introduce the density-dependent kernel:

$$K_\mu(x,y) = K(x-y)\mu_h(x)^{-1/2}\mu_h(y)^{-1/2}$$

NSVGD: In the discrete setting where $\mu = {}^1/_n \sum_{i=1}^n \delta_{x_i}$, we can write the NSVGD vector field ruling the particle system as

$$\dot{x}_i = -\frac{1}{n}\sum_{j=1}^n \nabla K_\mu(x_i - x_j) - \frac{1}{n}\sum_{j=1}^n K_\mu(x_i - x_j)\nabla U(x_j),$$

where $$K_\mu(x_i - x_j) = K(x_i - x_j)\mu_h(x_i)^{-1/2}\mu_h(x_j)^{-1/2},$$

$$\mu_h(x_i) = \frac{1}{n}\sum_j \eta_h(x_i - x_j).$$

NSVGD behaves better than SVGD in low density regions.

Quantization problem review

We are interested in establishing bounds on the quantization error

$$Q_n = \inf_{X_n = x_1, \ldots, x_n} \mathrm{D}(\pi, \mu_n), \qquad \text{for } \mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i},$$

where D is the MMD or KSD.

Remark: For $x_1, \ldots, x_n \sim \pi$ i.i.d., the rate is known to be $\mathcal{O}(n^{-1/2})$

Quantization result for the MMD

Theorem 1: Suppose K is sufficiently smooth. Then, there exists a constant C_d depending on d, such that for all $n \geq 2$,

- If π is Lebesgue on $[0,1]^d$, there exist points $x_1, \ldots, x_n$ such that

$$\mathrm{MMD}(\pi, \mu_n) \leq C_d \frac{(\log n)^{d-1}}{n}.$$

- If $\pi \in mathcalP([0,1]^d)$ there exist points $x_1, \ldots, x_n$ such that

$$\mathrm{MMD}(\pi, \mu_n) \leq C_d \frac{(\log n)^{\frac{3d+1}{2}}}{n}.$$

Proposition 1: Suppose K is sufficiently smooth. Assume π is a light-tailed distribution on $\mathbb{R}^d$. Then, for $n \geq 2$ there exist points $x_1, ..., x_n$ such that

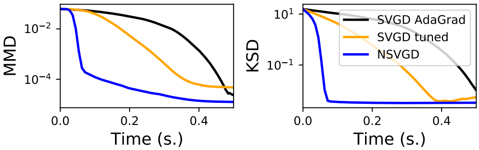
$$\mathrm{MMD}(\pi, \mu_n) \leq C_d \frac{(\log n)^{\frac{5d+1}{2}}}{n}.$$

Experiments

1. We compare the practical behavior of SVGD & NSVGD

2. We investigate numerically the quantization properties of :
   - SVGD & NSVGD
   - MMD & KSD Descent
   - Kernel Herding (KH) & Stein points (SP) :
     greedy minimization of the MMD & KSD

Practical behavior of SVGD & NSVGD



(a) Gaussian mixture sampling task

Figure: Convergence speed of SVGD (tuned time-step or Ada- Grad) and Normalized SVGD (fixed time-step) on a 2D mixture of Gaussians, with 128 particles.

Quantization rates of the algorithms, $\pi = \mathcal{N}(0, 1/d I_d)$

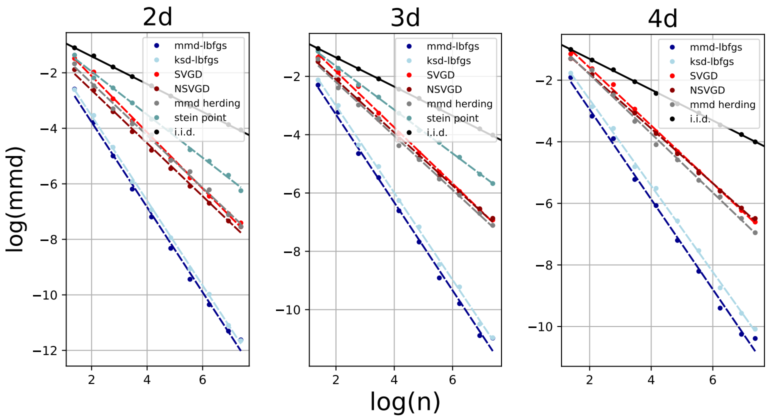


Figure: Averaged over 10 runs of each algorithm. Initial particles are i.i.d. samples of π. We use MMD with Gaussian kernel to evaluate; MMD/KSD Descent use bandwidth 1; SVGD and NSVGD use Laplace kernel.

d	Eval.	SVGD	MMD-lbfgs	KSD-lbfgs	KH	SP
2	**KSD**	-0.98	-1.48	-1.46	-0.84	-0.77
	MMD	-1.04	-1.60	-1.54	-0.93	-0.77
3	**KSD**	-0.91	-1.38	-1.44	-0.84	-0.78
	MMD	-0.96	-1.51	-1.49	-0.92	-0.75
4	**KSD**	-0.91	-1.35	-1.39	-0.89	–
	MMD	-0.94	-1.46	-1.40	-0.95	–
8	**KSD**	-0.84	-1.14	-1.16	–	–
	MMD	-0.77	-1.25	-1.13	–	–

Some remarks:

- The slopes remain much steeper than the Monte Carlo rate, even when the dimension increases
- MMD/KSD slopes are better than our theoretical upper bounds

Robustness to evaluation discrepancy

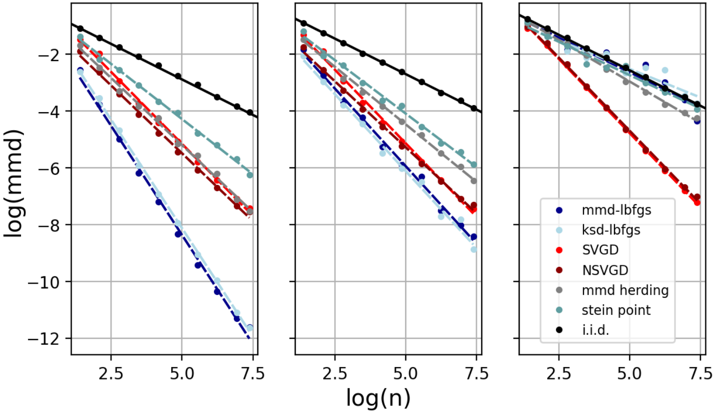


Figure: Fragility of MMD and KSD based quantization with respect to bandwidth of the MMD evaluation metric, in 2D. From Left to Right: evaluation MMD bandwidth = 1, 0.7, 0.3.

Conclusion

Contributions:

- Optimization: NSVGD accelerates the dynamics
- Quantization: Interacting-particle based sampling algorithms can create "super samples"

Future work/open questions:

- Improve our quantization bounds for MMD/KSD (dependence in dimension, Laplace kernel?)
- Obtain quantization bounds for SVGD
- What is a robust way to measure quantization error?
- What are good ensemble based algorithms to quantize a measure?

Thank you !

References I

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.

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

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. (2021).
Kernel Stein discrepancy descent.
International Conference of Machine Learning.

References II

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

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.