Kernel Stein Discrepancy Descent

Anna Korba [1] Pierre-Cyril Aubin-Frankowski [2]
Szymon Majewski [3] Pierre Ablin[4]

[1]CREST, ENSAE, Institut Polytechnique de Paris

[2]CAS, MINES ParisTech, Paris, France

[3]CMAP, Ecole Polytechnique, Institut Polytechnique de Paris

[4]CNRS and DMA, Ecole Normale Supérieure, Paris, France

One world ML seminar

Outline

Introduction

Preliminaries on Kernel Stein Discrepancy

Sampling as optimization of the KSD

Experiments

Theoretical properties of the KSD flow

Conclusion

Problem : Sample from a target distribution π over $\mathbb{R}^d$, whose density w.r.t. Lebesgue is known up to a constant Z :

$$\pi(x) = \frac{\tilde{\pi}(x)}{Z}$$

where Z is the (untractable) normalization constant.

Problem : Sample from a target distribution π over $\mathbb{R}^d$, whose density w.r.t. Lebesgue is known up to a constant Z :

$$\pi(x) = \frac{\tilde{\pi}(x)}{Z}$$

where Z is the (untractable) normalization constant.

Motivation : Bayesian statistics.

- Let $\mathcal{D} = (w_i, y_i)_{i=1,\dots,N}$ observed data.
- Assume an underlying model parametrized by θ (e.g. $p(y|w,\theta)$ gaussian)

  $\Longrightarrow$ Likelihood: $p(\mathcal{D}|\theta) = \prod_{i=1}^N p(y_i|\theta, w_i)$.
- Assume also $\theta \sim p$ (prior distribution).

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

Sampling as optimization over distributions

Assume that $\pi \in \mathcal{P}_2(\mathbb{R}^d) = \{\mu \in \mathcal{P}(\mathbb{R}^d), \int \|x\|^2 d\mu(x) < \infty\}$. The sampling task can be recast as an optimization problem:

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

where D is a **dissimilarity functional.**

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

Choice of the loss function

Many possibilities for the choice of D among Wasserstein distances, *f*-divergences, Integral Probability Metrics...

For instance,

- D is the KL (Kullback-Leibler divergence):

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

- D is the MMD (Maximum Mean Discrepancy):

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

where $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a p.s.d. kernel.

Contributions of the paper

Here we choose *D* as the **Kernel Stein Discrepancy (KSD).**

We propose an algorithm that is:

- score-based (only requires $\nabla \log \pi$)
- using a set of particles whose empirical distribution minimizes the KSD
- easy to implement and to use (e.g. leverages L-BFGS) !

We study:

- its convergence properties (numerically and theoretically)
- its empirical performance compared to Stein Variational Gradient Descent

Outline

Introduction

Preliminaries on Kernel Stein Discrepancy

Sampling as optimization of the KSD

Experiments

Theoretical properties of the KSD flow

Conclusion

Kernel Stein Discrepancy [Chwialkowski et al., 2016, Liu et al., 2016]

For $\mu, \pi \in \mathcal{P}_2(\mathbb{R}^d)$, the KSD of μ relative to π is

$$\mathrm{KSD}(\mu|\pi) = \sqrt{\iint k_\pi(x,y) d\mu(x) d\mu(y)},$$

where $k_\pi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the **Stein kernel**, defined through

- the score function $s(x) = \nabla \log \pi(x)$,
- a p.s.d. kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, $k \in C^2(\mathbb{R}^d)$[1]

For $x, y \in \mathbb{R}^d$,

$$\begin{aligned} k_\pi(x,y) &= s(x)^T s(y) k(x,y) + s(x)^T \nabla_2 k(x,y) \\ &\quad + \nabla_1 k(x,y)^T s(y) + \boldsymbol{\nabla} \cdot_1 \nabla_2 k(x,y) \\ &= \sum_{i=1}^d \frac{\partial \log \pi(x)}{\partial x_i} . \frac{\partial \log \pi(y)}{\partial y_i} . k(x,y) + \frac{\partial \log \pi(x)}{\partial x_i} . \frac{\partial k(x,y)}{\partial y_i} \\ &\quad + \frac{\partial \log \pi(y)}{\partial y_i} . \frac{\partial k(x,y)}{\partial x_i} + \frac{\partial^2 k(x,y)}{\partial x_i \partial y_i} \in \mathbb{R}. \end{aligned}$$

[1] e.g. : $k(x,y) = \exp\left(-\|x-y\|^2/h\right)$

Stein identity and link with MMD

Under mild assumptions on k and π, the Stein kernel k_π is p.s.d. and satisfies a **Stein identity** [Oates et al., 2017]

$$\boxed{\int_{\mathbb{R}^d} k_\pi(x,.)d\pi(x) = 0.}$$

Consequently, **KSD is a MMD** with kernel k_π, since:

$$\begin{aligned}\mathrm{MMD}^2(\mu|\pi) &= \int k_\pi(x,y)d\mu(x)d\mu(y) + \int k_\pi(x,y)d\pi(x)d\pi(y)\\ &\quad - 2\int k_\pi(x,y)d\mu(x)d\pi(y)\\ &= \int k_\pi(x,y)d\mu(x)d\mu(y)\\ &= \mathrm{KSD}^2(\mu|\pi)\end{aligned}$$

Rk : It is also as a kernelized Fisher divergence ($\left\|\nabla \log\left(\frac{\mu}{\pi}\right)\right\|^2_{L^2(\mu)}$):

$$\mathrm{KSD}^2(\mu|\pi) = \left\|S_{\mu,k}\nabla \log\left(\frac{\mu}{\pi}\right)\right\|^2_{\mathcal{H}_k} \text{ where } S_{\mu,k} : f \mapsto \int f(x)k(x,.)d\mu(x)$$

KSD benefits

KSD can be computed when

- one has access to the score of π
- μ is a discrete measure, e.g. $\mu = \frac{1}{N}\sum_{i=1}^{N} \delta_{x^i}$, then :

$$\mathrm{KSD}^2(\mu|\pi) = \frac{1}{N^2}\sum_{i,j=1}^{N} k_\pi(x^i, x^j).$$

KSD benefits

KSD can be computed when

- one has access to the score of π
- μ is a discrete measure, e.g. $\mu = \frac{1}{N}\sum_{i=1}^{N}\delta_{x^i}$, then :

$$\mathrm{KSD}^2(\mu|\pi) = \frac{1}{N^2}\sum_{i,j=1}^{N} k_\pi(x^i, x^j).$$

KSD is known to metrize weak convergence [Gorham and Mackey, 2017] when:

- π is strongly log-concave at infinity ("distantly dissipative", e.g. true for gaussian mixtures)
- k has a slow decay rate, e.g. true when k is the IMQ kernel defined by $k(x,y) = (c^2 + \|x-y\|_2^2)^\beta$ for $c > 0$ and $\beta \in (-1, 0)$.

KSD in the literature

The KSD has been used for

- nonparametric statistical tests for goodness-of-fit [Xu and Matsuda, 2020, Kanagawa et al., 2020]
- sampling tasks:
  - (greedy algorithms) to select a suitable set of static points to approximate π, adding a new one at each iteration [Chen et al., 2018, Chen et al., 2019]
  - to compress [Riabiz et al., 2020] or reweight [Hodgkinson et al., 2020] Markov Chain Monte Carlo (MCMC) outputs
  - to learn a static transport map from μ_0 to π [Fisher et al., 2020].
  - learn Energy-Based models $\pi \propto \exp(-V)$ from samples of π (use reverse KSD) [Domingo-Enrich et al., 2021]

Outline

Introduction

Preliminaries on Kernel Stein Discrepancy

Sampling as optimization of the KSD

Experiments

Theoretical properties of the KSD flow

Conclusion

Time/Space discretization of the KSD gradient flow

Let $\mathcal{F}(\mu) = \mathrm{KSD}^2(\mu|\pi)$.

- Its Wasserstein gradient flow on $\mathcal{P}_2(\mathbb{R}^d)$ finds a continuous path of distributions that decreases $\mathcal{F}$.
- Different algorithms to approximate π depend on the time and space discretization of this flow.

Forward discretization: Wasserstein gradient descent

Discrete measures: For discrete measures $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x^i}$, we have an explicit loss function

$$L([x^i]_{i=1}^N) := \mathcal{F}(\hat{\mu}) = \frac{1}{N^2} \sum_{i,j=1}^{N} k_\pi(x^i, x^j).$$

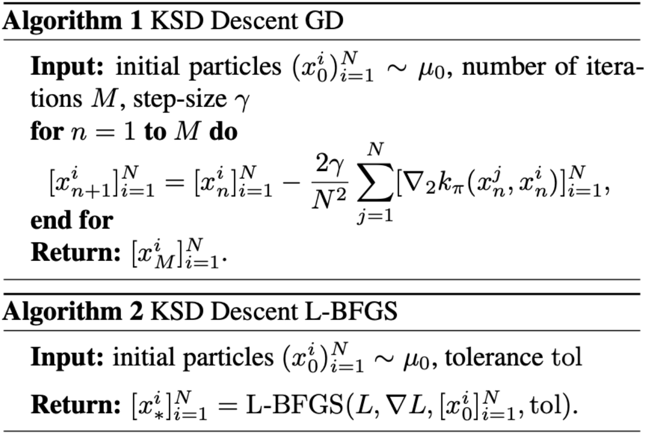
Then, Wasserstein gradient descent of $\mathcal{F}$ for discrete measures

$$\Updownarrow$$

(Euclidean) gradient descent of L on the particles.

KSD Descent - algorithms

We propose two ways to implement KSD Descent:



```
Algorithm 1 KSD Descent GD
Input: initial particles $(x_0^i)_{i=1}^N \sim \mu_0$, number of iterations $M$, step-size $\gamma$
for $n = 1$ to $M$ do
    $[x_{n+1}^i]_{i=1}^N = [x_n^i]_{i=1}^N - \frac{2\gamma}{N^2} \sum_{j=1}^N [\nabla_2 k_\pi(x_n^j, x_n^i)]_{i=1}^N,$
end for
Return: $[x_M^i]_{i=1}^N$.
```

```
Algorithm 2 KSD Descent L-BFGS
Input: initial particles $(x_0^i)_{i=1}^N \sim \mu_0$, tolerance tol
Return: $[x_*^i]_{i=1}^N = \text{L-BFGS}(L, \nabla L, [x_0^i]_{i=1}^N, \text{tol})$.
```

L-BFGS [Liu and Nocedal, 1989] is a quasi Newton algorithm that is faster and more robust than Gradient Descent, and **does not require the choice of step-size!**

L-BFGS

L-BFGS (Limited memory Broyden–Fletcher–Goldfarb–Shanno algorithm) is a quasi-Newton method:

$$x_{n+1} = x_n - \gamma_n B_n^{-1} \nabla L(x_n) := x_n + \gamma_n d_n \tag{1}$$

where B_n^{-1} is a p.s.d. matrix approximating the inverse Hessian at x_n.

Step1. (requires ∇L) It computes a cheap version of d_n based on BFGS recursion:

$$B_{n+1}^{-1} = \left(I - \frac{\Delta x_n y_n^T}{y_n^T \Delta x_n} \right) B_n^{-1} \left(I - \frac{y_n \Delta x_n^T}{y_n^T \Delta x_n} \right) + \frac{\Delta x_n \Delta x_n^T}{y_n^T \Delta x_n}$$

where

$$\begin{aligned} \Delta x_n &= x_{n+1} - x_n \\ y_n &= \nabla L(x_{n+1}) - \nabla L(x_n) \end{aligned}$$

Step2. (requires L and ∇L) A line-search is performed to find the best step-size in (1) :

$$\begin{aligned} L(x_n + \gamma_n d_n) &\leq L(x_n) + c_1 \gamma_n \nabla L(x_n)^T d_n \\ \nabla L(x_n + \gamma_n d_n)^T d_n &\geq c_2 \nabla L(x_n)^T d_n \end{aligned}$$

See [Nocedal and Wright, 2006].

Related work

1. minimize the **KL divergence** (requires $\nabla \log \pi$), e.g. with **Stein Variational Gradient descent** (SVGD, [Liu and Wang, 2016]).

Uses a set of N interacting particles and a p.s.d. kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ to approximate π:

$$x_{n+1}^i = x_n^i - \gamma \left[\frac{1}{N} \sum_{j=1}^{N} k(x_n^i, x_n^j) \nabla \log \pi(x_n^j) + \nabla_1 k(x_n^j, x_n^i) \right],$$

Does not minimize a closed-form functional for discrete measures! $\implies$ cannot use L-BFGS.

2. minimize the **MMD** [Arbel et al., 2019]

$$x_{n+1}^i = x_n^i - \gamma \left[\frac{1}{N} \sum_{j=1}^{N} \nabla_2 k(x_n^j, x_n^i) - \nabla_2 k(y^j, x_n^i) \right].$$

(requires samples $(y_j)_{j=1}^N \sim \pi$)

Outline

Introduction

Preliminaries on Kernel Stein Discrepancy

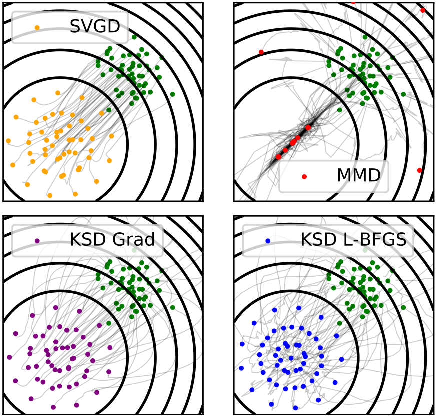
Sampling as optimization of the KSD

Experiments

Theoretical properties of the KSD flow

Conclusion

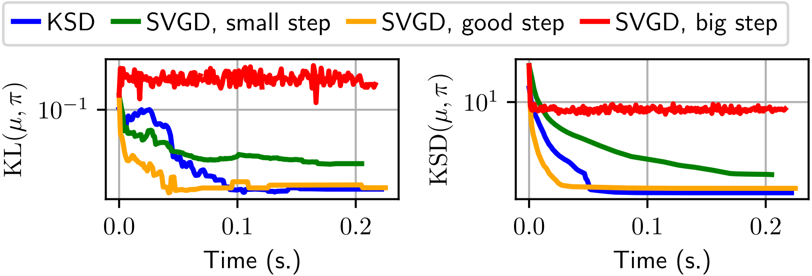
Toy experiments - 2D standard gaussian



The green points represent the initial positions of the particles.

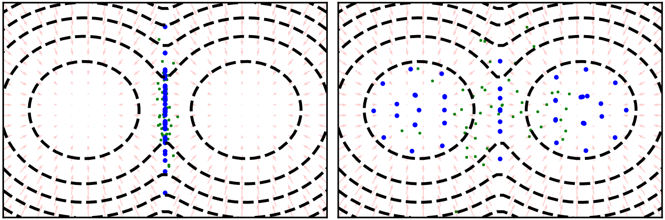
The light grey curves correspond to their trajectories.

SVGD vs KSD Descent - importance of the step-size



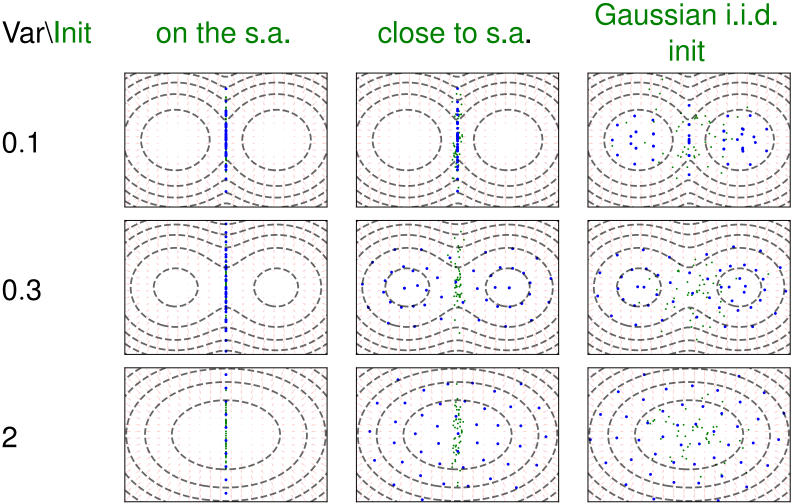
Convergence speed of KSD and SVGD on a Gaussian problem in 1D, with 30 particles.

2D mixture of (isolated) Gaussians - failure cases



The green crosses indicate the initial particle positions
the blue ones are the final positions
The light red arrows correspond to the score directions.

More initializations



Var\Init

on the s.a.

close to s.a.

Gaussian i.i.d. init

0.1

0.3

2

Green crosses : initial particle positions

Blue crosses : final positions

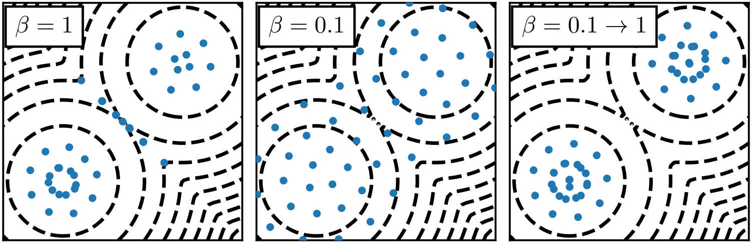
Stationary measures - some explanations

In the paper, we explain how particles can get stuck in planes of symmetry of the target π.

- we show that if a stationary measure μ_∞ is full support, then $\mathcal{F}(\mu_\infty) = 0$.
- but we also show that if *supp*$(\mu_0) \subset \mathcal{M}$, where $\mathcal{M}$ is a plane of symmetry of π, then for any time *t* it remains true for μ_t: *supp*$(\mu_t) \subset \mathcal{M}$.

Isolated Gaussian mixture - annealing

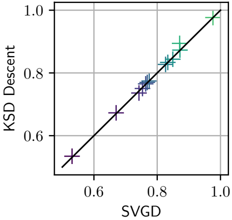
Add an inverse temperature variable β : $\pi^{\beta}(x) \propto \exp(-\beta V(x))$, with $0 < \beta \leq 1$ (i.e. multiply the score by β.)



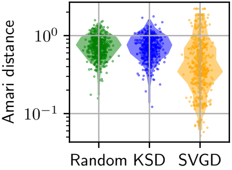
This is a hard problem, even for Langevin diffusions, where tempering strategies also have been proposed.

Beyond Log-concavity: Provable Guarantees for Sampling Multi-modal Distributions using Simulated Tempering Langevin Monte Carlo. Rong Ge, Holden Lee, Andrej Risteski. 2017.

Real world experiments (10 particles)

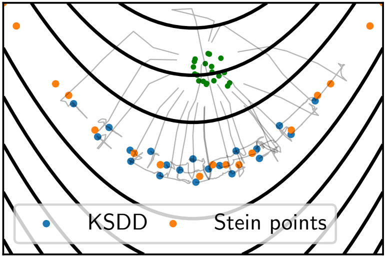


Bayesian logistic regression.
Accuracy of the KSD descent and SVGD for 13 datasets ($d \sim 50$).
Both methods yield similar results. KSD is better by 2% on one dataset.
Hint: convex likelihood.



Bayesian ICA.
Each dot is the Amari distance between an estimated matrix and the true unmixing matrix ($d \leq 8$).
KSD is not better than random.
Hint: highly non-convex likelihood.

So.. when does it work?



Comparison of KSD Descent and Stein points on a “banana” distribution. Green points are the initial points for KSD Descent. Both methods work successfully here, **even though it is not a log-concave distribution.**

We posit that KSD Descent succeeds because **there is no saddle point in the potential.**

Outline

Introduction

Preliminaries on Kernel Stein Discrepancy

Sampling as optimization of the KSD

Experiments

Theoretical properties of the KSD flow

Conclusion

First strategy : functional inequality?

$\mathcal{F}(\mu|\pi) = \iint k_\pi(x,y)d\mu(x)d\mu(y)$.

We have

$$\frac{\partial \mathcal{F}(\mu)}{\partial \mu} = \int k_\pi(x,.)d\mu(x) = \mathbb{E}_{x\sim\mu}[k_\pi(x,.)]$$

and under appropriate growth assumptions on k_π:

$$\nabla_{W_2}\mathcal{F}(\mu) = \mathbb{E}_{x\sim\mu}[\nabla_2 k_\pi(x,\cdot)],$$

Hence

$$\begin{aligned}\frac{d\mathcal{F}(\mu_t)}{dt} &= \langle \nabla_{W_2}\mathcal{F}(\mu_t), -\nabla_{W_2}\mathcal{F}(\mu_t)\rangle_{L^2(\mu_t)} \\ &= -\,\mathbb{E}_{y\sim\mu_t}\left[\|\mathbb{E}_{x\sim\mu_t}[\nabla_2 k_\pi(x,y)]\|^2\right] \leq 0.\end{aligned}$$

$\Longrightarrow$**Difficult to identify a functional inequality to relate** $d\mathcal{F}(\mu_t)/dt$ **to** $\mathcal{F}(\mu_t)$, and establish convergence in continuous time (similar to [Arbel et al., 2019]).

Second strategy : geodesic convexity of the KSD?

Let $\psi \in C_c^\infty(\mathbb{R}^d)$ and the path $\rho_t = (I + t\nabla\psi)_{\#}\mu$ for $t \in [0,1]$.

Define the quadratic form $\operatorname{Hess}_\mu \mathcal{F}(\psi,\psi) := \frac{d^2}{dt^2}\Big|_{t=0}\mathcal{F}(\rho_t)$, which is related to the W_2 **Hessian of $\mathcal{F}$ at μ.**

For $\psi \in C_c^\infty(\mathbb{R}^d)$, we have

$$\operatorname{Hess}_\mu \mathcal{F}(\psi,\psi) = \mathbb{E}_{x,y\sim\mu}\left[\nabla\psi(x)^T\nabla_1\nabla_2 k_\pi(x,y)\nabla\psi(y)\right] + \mathbb{E}_{x,y\sim\mu}\left[\nabla\psi(x)^T H_1 k_\pi(x,y)\nabla\psi(x)\right].$$

The first term is always positive but not the second one.

$\Longrightarrow$ **the KSD is not convex w.r.t. W_2 geodesics.**

Third strategy : curvature near equilibrium?

What happens near equilibrium π? the second term vanishes due to the Stein property of k_π and :

$$\operatorname{Hess}_\pi \mathcal{F}(\psi,\psi) = \|S_{\pi,k_\pi}\mathcal{L}_\pi\psi\|^2_{\mathcal{H}_{k_\pi}} \geq 0$$

where

$$\begin{aligned} \mathcal{L}_\pi &: f \mapsto -\Delta f - \langle \nabla \log \pi, \nabla f\rangle_{\mathbb{R}^d} \\ S_{\mu,k_\pi} &: f \mapsto \int k_\pi(x,.)f(x)d\mu(x) \in \mathcal{H}_{k_\pi} = \overline{\left\{k_\pi(x,.), x \in \mathbb{R}^d\right\}} \end{aligned}$$

Question: can we bound from below the Hessian at π by a quadratic form on the tangent space of $\mathcal{P}_2(\mathbb{R}^d)$ at π ($\subset L^2(\pi)$)?

$$\boxed{\|S_{\pi,k_\pi}\mathcal{L}_\pi\psi\|^2_{\mathcal{H}_{k_\pi}} = \operatorname{Hess}_\pi \mathcal{F}(\psi,\psi) \geq \lambda\|\nabla\psi\|^2_{L^2(\pi)} \text{ ?}}$$

That would imply exponential decay of $\mathcal{F}$ near π.

Curvature near equilibrium - negative result

The previous inequality

$$\boxed{\|S_{\pi,k_\pi}\mathcal{L}_\pi\psi\|^2_{\mathcal{H}_{k_\pi}} \geq \lambda\|\nabla\psi\|^2_{L^2(\pi)}}$$

- can be seen as a kernelized version of the Poincaré inequality for π :

$$\|\mathcal{L}_\pi\psi\|^2_{L_2(\pi)} \geq \lambda_\pi\|\nabla\psi\|^2_{L_2(\pi)}.$$

- can be written:

$$\langle\psi, T_{\pi,k_\pi}\psi\rangle_{L_2(\pi)} \geq \lambda\langle\psi, \mathcal{L}_\pi^{-1}\psi\rangle_{L_2(\pi)},$$

where $T_{\pi,k_\pi} : L^2(\pi) \to L^2(\pi), f \mapsto \int k_\pi(x,.)f(x)d\pi(x).$

Theorem : Let $\pi \propto e^{-V}$. Assume that $V \in C^2(\mathbb{R}^d)$, ∇V is Lipschitz and $\mathcal{L}_\pi$ has discrete spectrum. Then exponential decay near equilibium does not hold.

Outline

Introduction

Preliminaries on Kernel Stein Discrepancy

Sampling as optimization of the KSD

Experiments

Theoretical properties of the KSD flow

Conclusion

Conclusion

Pros:

- KSD Descent is a very simple algorithm, and can be used with L-BFGS [Liu and Nocedal, 1989] (fast, and does not require the choice of a step-size as in SVGD)
- works well on log-concave targets (unimodal gaussian, Bayesian logistic regression with gaussian priors) or "nice" distributions (banana)

Cons:

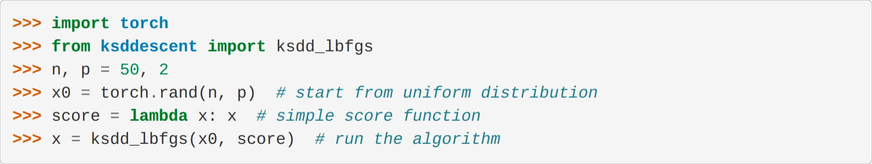
- KSD is not convex w.r.t. W_2, and no exponential decay near equilibrium holds
- does not work well on non log-concave targets (mixture of isolated gaussians, Bayesian ICA)

Open questions

- explain the convergence of KSD Descent when π is log-concave?
- quantify propagation of chaos ? (KSD for a finite number of particles vs infinite)
- how good is KSD quantisation?

Code

- Python package to try KSD descent yourself:
  pip install ksddescent
- website: `pierreablin.github.io/ksddescent/`
- It also features pytorch/numpy code for SVGD.



```
>>> import torch
>>> from ksddescent import ksdd_lbfgs
>>> n, p = 50, 2
>>> x0 = torch.rand(n, p)  # start from uniform distribution
>>> score = lambda x: x  # simple score function
>>> x = ksdd_lbfgs(x0, score)  # run the algorithm
```

Thank you for listening!

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.

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., Oates, C., et al. (2019).
Stein point Markov Chain Monte Carlo.
Proceedings of the 36th International Conference on Machine Learning,.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. J. (2018).
Stein points.
Proceedings of the 35th International Conference on Machine Learning,.

References II

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

Domingo-Enrich, C., Bietti, A., Vanden-Eijnden, E., and Bruna, J. (2021).
On energy-based models with overparametrized shallow neural networks.
arXiv preprint arXiv:2104.07531.

Fisher, M. A., Nolan, T., Graham, M. M., Prangle, D., and Oates, C. J. (2020).
Measure transport with kernel Stein discrepancy.
arXiv preprint arXiv:2010.11779.

References III

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

Hodgkinson, L., Salomone, R., and Roosta, F. (2020).
The reproducing Stein kernel approach for post-hoc corrected sampling.
arXiv preprint arXiv:2001.09266.

Kanagawa, H., Jitkrittum, W., Mackey, L., Fukumizu, K., and Gretton, A. (2020).
A kernel Stein test for comparing latent variable models.
arXiv preprint arXiv:1907.00586.

References IV

Liu, D. C. and Nocedal, J. (1989).
On the limited memory BFGS method for large scale optimization.
Mathematical programming, 45(1-3):503–528.

Liu, Q., Lee, J., and Jordan, M. (2016).
A kernelized stein discrepancy for goodness-of-fit tests.
In *International conference on machine learning*, pages 276–284.

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

References V

Nocedal, J. and Wright, S. (2006).
Numerical optimization.
Springer Science & Business Media.

Oates, C. J., Girolami, M., and Chopin, N. (2017).
Control functionals for monte carlo integration.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3):695–718.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., Oates, C., et al. (2020).
Optimal thinning of MCMC output.
arXiv preprint arXiv:2005.03952.

Steinwart, I. and Christmann, A. (2008).
Support vector machines.
Springer Science & Business Media.

References VI

Xu, W. and Matsuda, T. (2020).
A Stein goodness-of-fit test for directional distributions.
In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pages 320–330. PMLR.

W2 GF of KSD

Let $\mathcal{F}(\mu) = \frac{1}{2}\,\mathrm{KSD}^2(\mu|\pi)$. The KSD gradient flow is defined as the flow induced by the continuity equation:

$$\frac{\partial \mu_t}{\partial t} + div(\mu_t v_{\mu_t}) = 0,\ v_{\mu_t} := -\nabla_{W_2}\mathcal{F}(\mu_t).$$

For μ_t regular enough,

$$\boxed{\nabla_{W_2}\mathcal{F}(\mu_t) = \nabla \frac{\partial \mathcal{F}(\mu_t)}{\partial \mu}}$$

$\frac{\partial \mathcal{F}(\mu)}{\partial \mu} : \mathbb{R}^d \to \mathbb{R}$ is the differential of $\mu \mapsto \mathcal{F}(\mu)$, evaluated at μ.

It is the unique function such that for any $\mu, \mu' \in \mathcal{P}$, $\mu' - \mu \in \mathcal{P}$:

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

Stationary measures of the KSD flow

Consider a stationary measure μ_∞ of the KSD flow, i.e **the dissipation** is null:

$$\frac{d\mathcal{F}(\mu_\infty)}{dt} = 0$$

$\implies \int k_\pi(x,.)d\mu_\infty(x)$ is μ_∞-a.e equal to a constant function c.

If μ_∞ has full support, since we can prove $\mathcal{H}_{k_\pi}$ does not contain non-zero constant functions, **then** $\mathcal{F}(\mu_\infty) = 0$.

If μ_∞ is a discrete measure (as in practice) the dissipation can vanish even for $\mu \neq \pi$ because μ is not full-support.

Some results on stationary measures of the KSD flow

Lemma

Let x_0 such that $s(x_0) = 0$ and $J(s)(x_0)$ is invertible, and consider a translation-invariant kernel $k(x, y) = \phi(x - y)$, for $\psi \in C^1(\mathbb{R}^d)$. Then δ_{x_0} is a stable fixed measure of the KSD flow.

Lemma

Let $\mathcal{M}$ be a plane of symmetry of π and consider a radial kernel $k(x, y) = \phi(\|x - y\|^2/2)$ with $\phi \in C^2$, then, for all $(x, y) \in \mathcal{M}^2$, $\nabla_2 k_\pi(x, y) \in T_{\mathcal{M}}(x)$ and $\mathcal{M}$ is flow-invariant for the KSD flow, i.e. : for any μ_0 s.t. $\mathrm{supp}(\mu_0) \subset \mathcal{M}$, then $\mathrm{supp}(\mu_t) \subset \mathcal{M}$ for all $t \geq 0$.

Background on kernels and RKHS [Steinwart and Christmann, 2008]

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

$$k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}, \quad \phi : \mathbb{R}^d \to \mathcal{H}$$

- $\mathcal{H}_k$ its corresponding RKHS (Reproducing Kernel Hilbert Space):

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

- $\mathcal{H}_k$ is a Hilbert space with inner product $\langle ., . \rangle_{\mathcal{H}_k}$ and norm $\|.\|_{\mathcal{H}_k}$. It satisfies the reproducing property:

$$\forall \quad f \in \mathcal{H}_k,\ x \in \mathbb{R}^d, \quad f(x) = \langle f, k(x, .) \rangle_{\mathcal{H}_k}$$

We assume $\int_{\mathbb{R}^d \times \mathbb{R}^d} k(x, x) d\mu(x) < \infty$ for any $\mu \in \mathcal{P}$.
$\Longrightarrow \mathcal{H}_k \subset L^2(\mu)$.

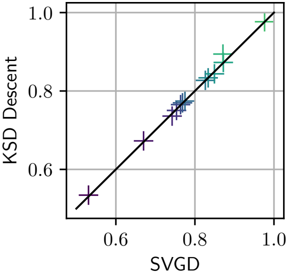
1 - Bayesian Logistic regression

Datapoints $d_1, \ldots, d_q \in \mathbb{R}^p$, and labels $y_1, \ldots, y_q \in \{\pm 1\}$.

Labels y_i are modelled as $p(y_i = 1|d_i, w) = (1 + \exp(-w^\top d_i))^{-1}$ for some $w \in \mathbb{R}^p$.

The parameters w follow the law $p(w|\alpha) = \mathcal{N}(0, \alpha^{-1} I_p)$, and $\alpha > 0$ is drawn from an exponential law $p(\alpha) = \mathrm{Exp}(0.01)$.

The parameter vector is then $x = [w, \log(\alpha)] \in \mathbb{R}^{p+1}$, and we use KSD-LBFGS to obtain samples from $p(x|\left(d_i, y_i\right)_{i=1}^q)$ for 13 datasets, with $N = 10$ particles for each.



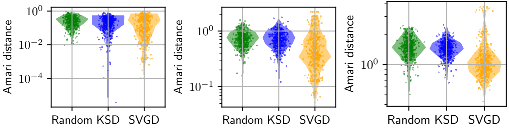
Accuracy of the KSD descent and SVGD on bayesian logistic regression for 13 datasets.
Both methods yield similar results. KSD is better by 2% **on one dataset.**

2 - Bayesian Independent Component Analysis

ICA: $x = W^{-1}s$, where x is an observed sample in $\mathbb{R}^p$, $W \in \mathbb{R}^{p \times p}$ is the unknown square unmixing matrix, and $s \in \mathbb{R}^p$ are the independent sources.

1)Assume that each component has the same density $s_i \sim p_s$.
2) The likelihood of the model is $p(x|W) = \log|W| + \sum_{i=1}^{p} p_s([Wx]_i)$.
3)Prior: W has i.i.d. entries, of law $\mathcal{N}(0, 1)$.

The posterior is $p(W|x) \propto p(x|W)p(W)$, and the score is given by $s(W) = W^{-\top} - \psi(Wx)x^{\top} - W$, where $\psi = -\frac{p_s'}{p_s}$. In practice, we choose p_s such that $\psi(\cdot) = \mathtt{tanh}(\cdot)$. We then use the presented algorithms to draw 10 particles $W \sim p(W|x)$ on 50 experiments.



Left: $p = 2$. Middle: $p = 4$. Right: $p = 8$.
Each dot = Amari distance between an estimated matrix and the true unmixing matrix.
KSD Descent is not better than random. Explanation: ICA likelihood is highly non-convex.