Geometry of Stein variational gradient descent (Dunkan,2019)

Nikita Mashalov

Plan

- Motivation
- Recap on Stein Gradient
- Geometry: inequalities on GF with Sobolev and Poincare inequalities
- Questions

Motivation

- Suppose we have oracle x –> ∇log p(x), suggest method to **sample**
- Schemes like HMC and Langevin suffers from **correlation issues** between samples
- Sample **in group** and give **penalty for correlation** -> Stein Gradient Descent (Liu,2015)
- Method suffers from **curse of dimension.** Need **convergence guarantee** -> Geometry of SGVD (Duncan, 2019)

Setting

Stein Framework

- Stein (1991) advised **Stein identity** operator for metric between prob measures (legandre formalism)

$$A_p\phi(x) = \phi(x)\nabla_x \log p(x) + \nabla_x \phi(x)$$

- Metric is called **Stein Disrepancy**

$$S(p,q) = \left[\max_{\phi} \int (\phi(x)\nabla_x \log p(x) + \nabla_x \phi(x))q(x)dx\right]^2$$

- On RKHS we have extremum in closed form (kernel K is given by domain)

$$\phi(x) = \int \left(k(x,*)\nabla_{x_i'} \log p(x_i') + \nabla_{x_i'} k(x, x_i')\right) q(x)dx$$

Sampling strategy

- So phi feels like gradient, why not to follow it like we do in GD

$$x \to x + \epsilon\phi(x)$$

- Expectation can be replaced with many other particles

$$\phi(x) = \frac{1}{n}\sum_i k(x, x_i')\nabla_{x_i'} \log p(x_i') + \nabla_{x_i'} k(x, x_i')$$

- Mean field approximation (Vlasov Equation)

$$\frac{\partial \rho}{\partial t} = \nabla_x \rho(x) \int k(x, y)\left(\nabla_y \rho + \rho_y \nabla_y \log \rho_y\right) dy$$

Convergence of Mean Field

Interrelation between sampling strategies

Recall natural gradient (parametrization free)

- Langevin

$$\rho_{n+1} = argmin_{\rho}\left(KL(\rho|\pi) + \frac{1}{\varepsilon} d_{OT}^2(\rho_n, \rho)\right)$$

- Stein (low index of d denotes kernel K)

$$\rho_{n+1} = argmin_{\rho}\left(KL(\rho|\pi) + \frac{1}{\varepsilon} d_{K}^2(\rho_n, \rho\ )\right)$$

Otto formalism for Langevin dynamics

- Otto 1991 developed approach. **Langevin dynamics induce gradient flow on KL**:

$$\frac{dKL(p(t)|q)}{dt} = I(p|q) = \int \left| \nabla_{\mathrm{x}} \log \frac{dp}{dq} \right|^2 p(dx)$$

- **Log-Sobolev inequalities** give lover bound on fisher information matrix:

$$\exists \lambda \colon\ I(p|q) > \lambda KL(p|q)$$

- We have **exponential decay** -> method converge

$$KL(p(t)|q)\ = e^{-\lambda t} KL(p(0)|q)$$

Log-Sobolev for Stein Discrepancy?

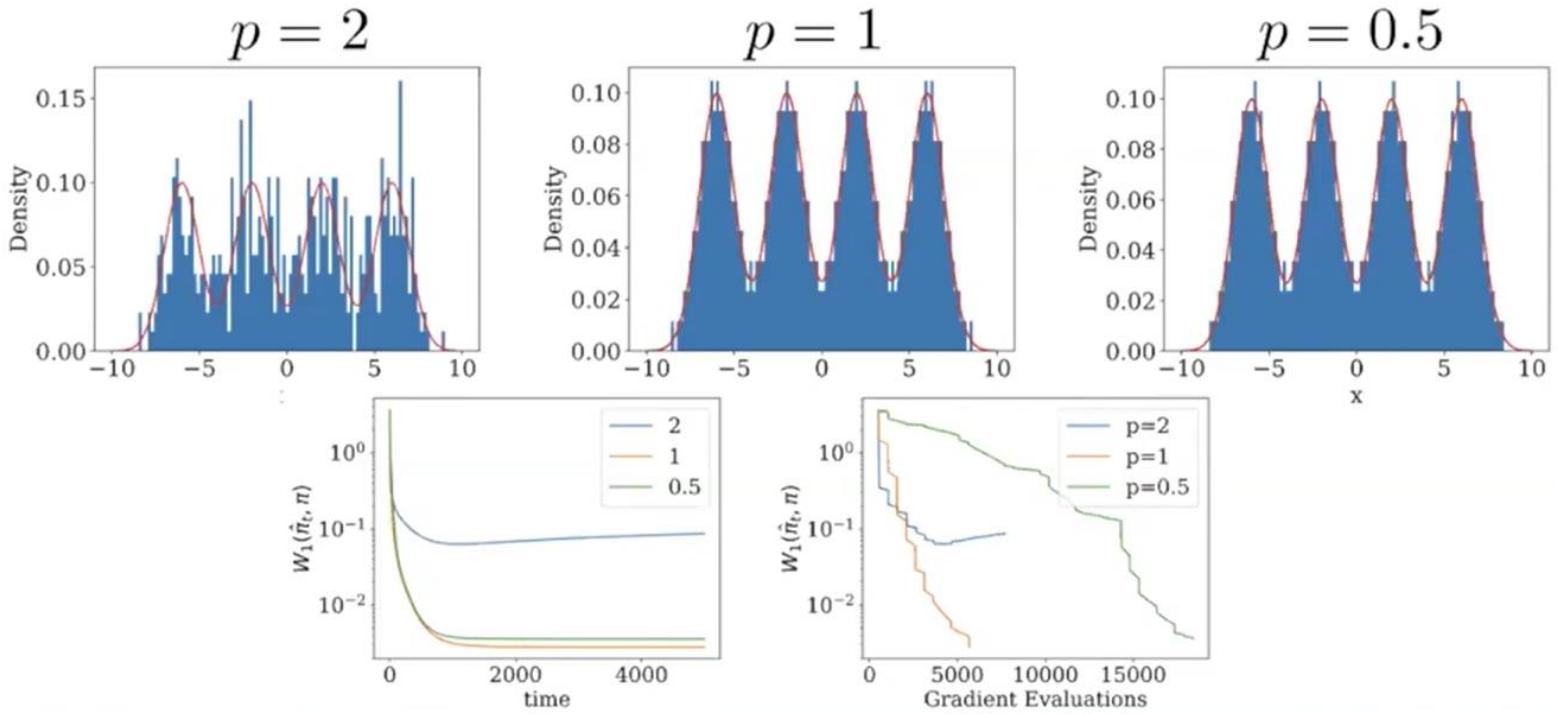
- For Stein flow (Liu,2019) and Dunkan(2019):
$$\frac{d}{dt}KL(p(t)|q) = -S^2(p(t)|q), \qquad \frac{d^2}{dt^2}KL(p(t)|q) = Hess_\rho(v,v)$$
- Existence of inequality comes from study of vector fields (velocities). Dunkan comes to fact, that **even for log-concave measures we don't have convergence**.
- Heuristic explanation(recall kinematic $v = v_{||} + v_{\perp}$:
$$Hess_\rho(v,u) = Hess_{reg}(v,u) + Hess_{cost}(v,u)$$

Dunkan(2019) shown that $Hess_{reg}(v,v) < 0$ for all linear vector fields. That means, SGVD **dissipates** energy of particle, contrary to Langevin

Mattern kernel $k(x,y) = \exp(-\frac{|x-y|^{\sigma}}{\sigma^{p}})$

Example: In one dimension

$$p = \frac{1}{4}\mathcal{N}(2,1) + \frac{1}{4}\mathcal{N}(-2,1) + \frac{1}{4}\mathcal{N}(6,1) + \frac{1}{4}\mathcal{N}(-6,1)$$



Adjourn

- For **arbitrary kernel doen't have convergence guarantees** even in Mean field and infinitesimal settings
- So we need to pay attention to **selection of kernel** and selection of **prior, which can reach**
- Advised approach seems *reductionist* in sense of optimal transport. Juveniles researches works on more grounded approaches as projection, natural gradient and functional

Rewind author video:
https://youtu.be/2tiu3HDJjE4?t=2037

Question

Recall, you mentioned problem with curse of dimension. Elaborate on it.

Kernel methods is known to work bad in high dimensional settings, due to **tractability** and **distance concentration.** If your weights are standard normally distributed, with increase of dimension distance between points will become N(1,O(1/d)). So kernel lose their sensitivity.

But if output of net is low dimensional (scoring), we can imagine particle to be a function instead of weights of net.See Function space particle optimization for bayesian neural networks (Wang, 2019) for elaboration on idea

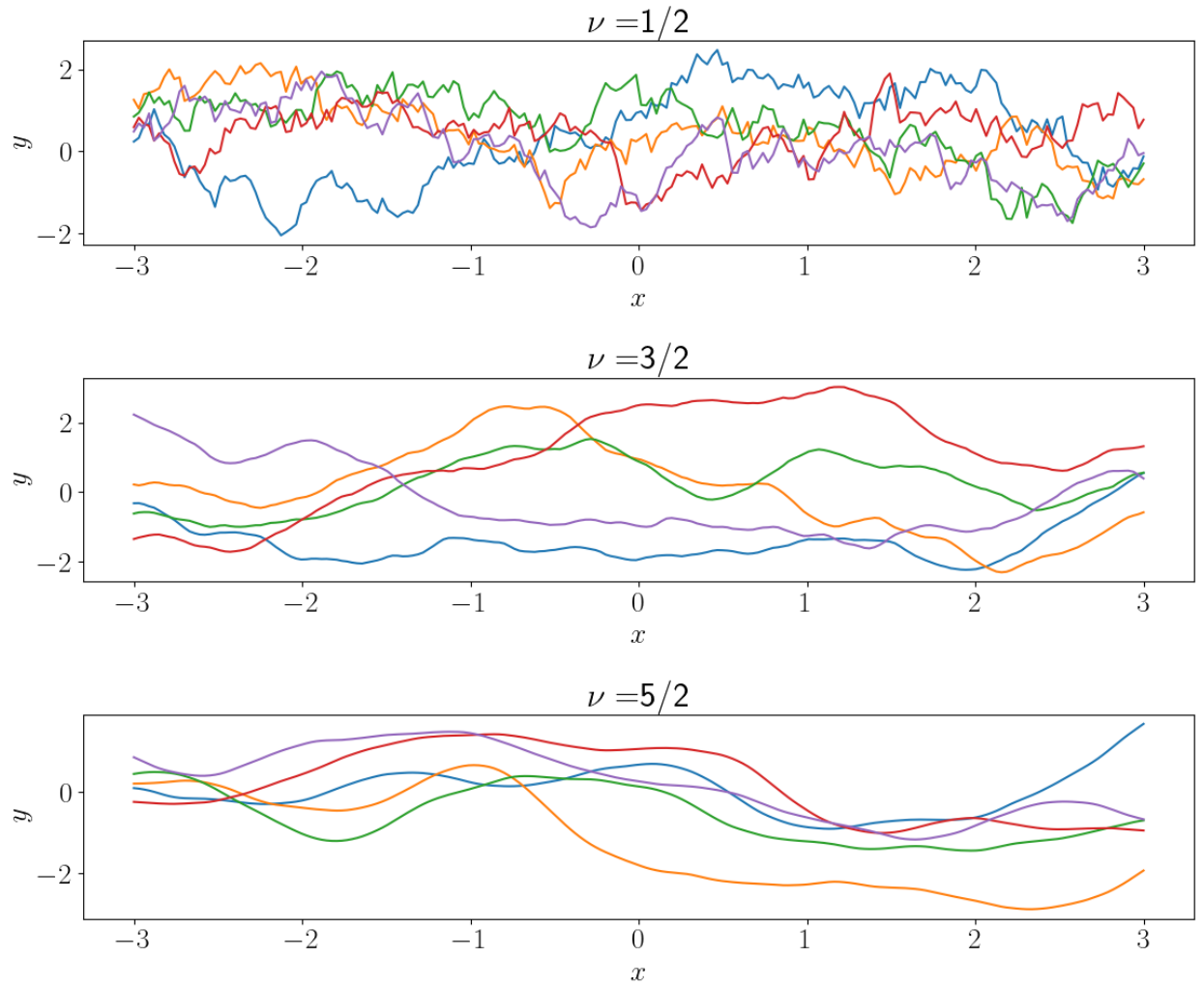
Article link https://arxiv.org/pdf/1902.09754

Can I use this sampling approach to llm and stable diffusion models?

Definitely for SD. Research community knows it as ParVIs methods. Check Prolific and Stein Dreamer articles.

To my best knowledge kernelized sampling methods are unknown for transformer llm. Yet you can use stein distance for alignment (but why?)

Prolific https://arxiv.org/abs/2305.16213
Stein https://arxiv.org/abs/2401.00604



https://andrewcharlesjones.github.io/journal/matern-kernels.html