Inference with Scores: Slices, Diffusions and Flows

Yingzhen Li

yingzhen.li@imperial.ac.uk

Feat. Wenbo Gong (University of Cambridge)

Bayesian ML / Probability Theory

Decision making under uncertainty

The Central Computation for Inference

- Inference: infer the unknowns
  - Unobserved/latent variables in the model
  - Quantities depending on the latent variables in the model

$$\int F(\theta)\,\pi(\theta)\,d\theta$$

prob. density

Random variable (unobserved)

integrand function

probability measure

(For discrete probability measures, integration becomes discrete sum.)

(We will discuss continuous variables in the rest of the talk.)

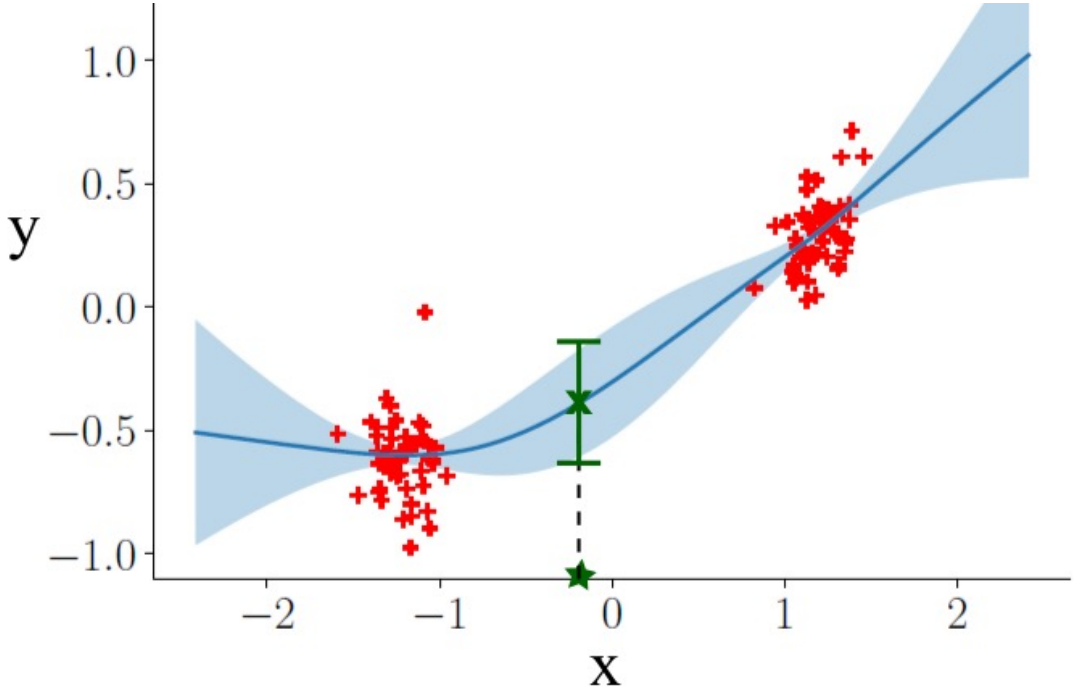
Computation Challenge

- The central equation for inference:

$$\int F(\theta)\,\pi(\theta)d\theta$$

"What is the prediction distribution of the test output given a test input?"

$F(\theta) = p(y|x,\theta), \pi(\theta) = p(\theta \mid D)$,
D = observed datapoints



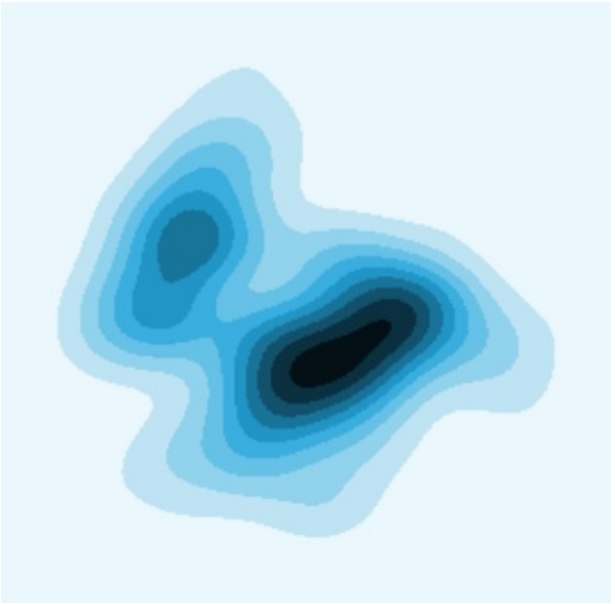
Computation Challenge

- The central equation for inference:

$$\int F(\theta)\,\pi(\theta)d\theta$$

"What is the mean of this distribution?"

$F(\theta) = \theta$, $\pi(\theta)$ can be complicated and high dimensional

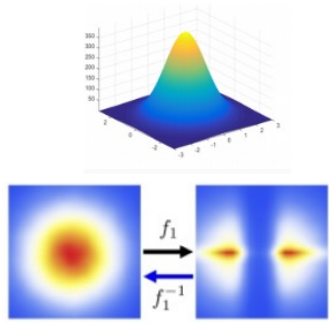


Approximate Inference

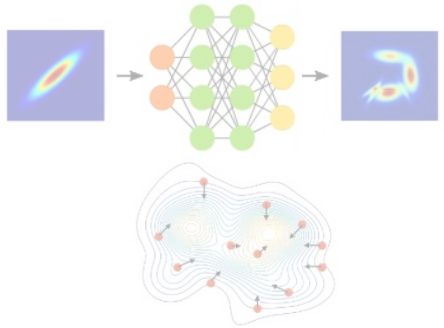
- Central task: approximate $\pi(\theta)$

$$q(\theta) \approx \pi(\theta)$$

Approximate distribution design



Explicit distributions



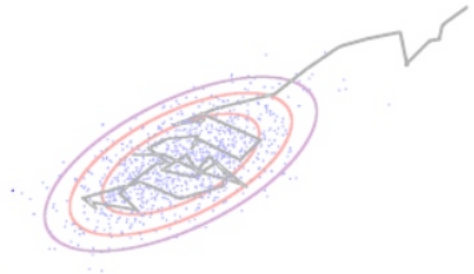
Implicit distributions

Today's focus

Algorithm for fitting $q(\theta)$ to $\pi(\theta)$

$$\min \ Loss(q(\theta), \pi(\theta))$$

Optimisation-based approaches



Sampling-based approaches

Approximating the Target Distribution

- Target distribution is usually intractable due to normalising constant:

$$p(\theta|D) = \frac{1}{p(D)} \exp[\log p(\theta) + \log p(D|\theta)]$$

Bayesian Posterior

$$p(x) = \frac{1}{Z} \exp[-E(x;\theta)]$$

Energy-based Model

- Approximation methods to side-step:

Variational Inference

$$\min_q KL[q\|p] = E_q[\log q(x) - \log p(x)]$$

"compare $\log q(x)$ with $\log p(x)$"

Using equiv. ELBO objective,
no need to evaluate $\log Z$

Score Matching

$$\min_q F(q,p) = \frac{1}{2} E_q[\|\nabla_x \log q(x) - \nabla_x \log p(x)\|_2^2]$$

"compare $\nabla_x \log q(x)$ with $\nabla_x \log p(x)$"

As $\nabla_x \log Z = 0$,
no need to evaluate $\log Z$

Comparing KL & Fisher Divergences

- Comparing 2 Gaussians with different mean, same variance ($x \in R^d$):

$$p(x) = N(x; \mu_p, \sigma^2 I_d)$$

$$q(x) = N(x; \mu_q, \sigma^2 I_d)$$

Variational Inference

$$KL[q\|p] = \frac{1}{2\sigma^2}\left\|\mu_q - \mu_p\right\|_2^2$$

Score Matching

$$F(q,p) = \frac{1}{2\sigma^4}\left\|\mu_q - \mu_p\right\|_2^2$$

Define $\mu_q := \mu_p + \sigma\gamma$, then $\left\|\mu_q - \mu_p\right\|_2^2 = \|\gamma\|_2^2$

Comparing KL & Fisher Divergences

- Comparing 2 Gaussians with different mean, same variance ($x \in R^d$):

$$p(x) = N(x; \mu_p, \sigma^2 I_d)$$

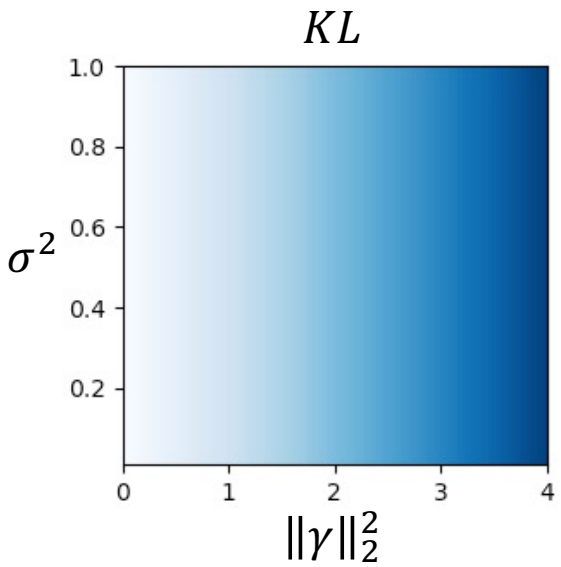
$$q(x) = N(x; \mu_p + \sigma\gamma, \sigma^2 I_d)$$

Variational Inference

$$KL[q\|p] = \frac{1}{2}\|\gamma\|_2^2$$

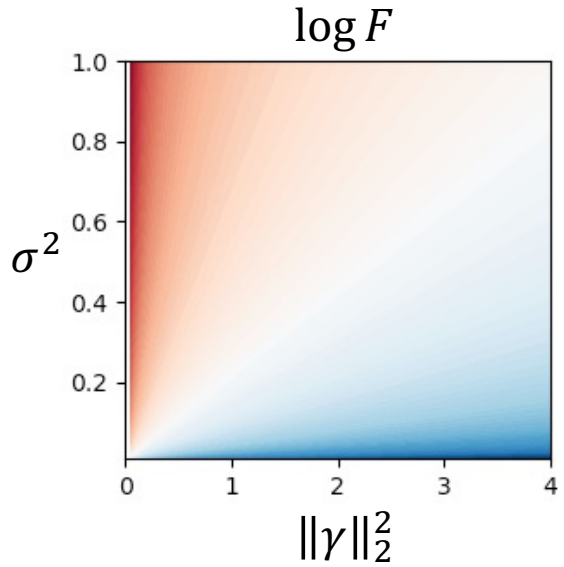
Score Matching

$$F(q,p) = \frac{1}{2\sigma^2}\|\gamma\|_2^2$$



high

low



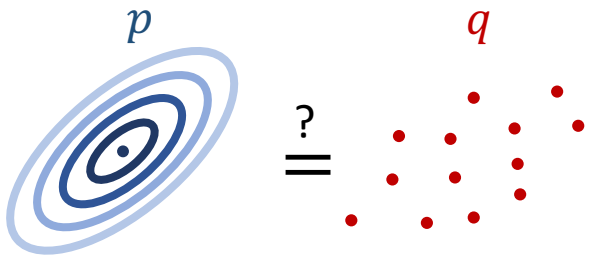
Very large if not capturing sharp peaks!

Stein's Method & Score Matching

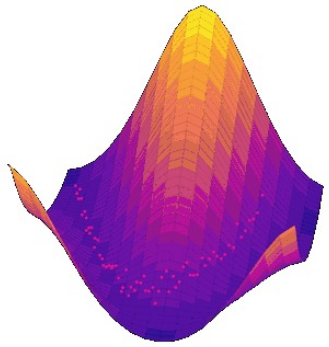
- Fisher divergence: a way to measure score difference:

$$F(p,q) = \frac{1}{2} E_q[\|\nabla_x \log q(x) - \nabla_x \log p(x)\|_2^2]$$

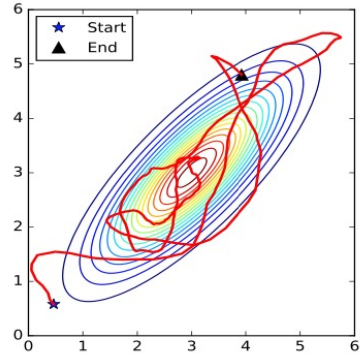
Unknown score (only have $x \sim q(x)$)



Goodness-of-fit Test
(q only has samples)



EBM learning
(q as data distribution)



Tuning MCMC
(q as MCMC samples)

Solution: Find a function to directly compute/approximate the score difference!

Hyvarinen. Estimation of Non-Normalized Statistical Models by Score Matching. JMLR 2005

Stein's Method & Score Matching

- Stein's identity:

$$E_{p(x)}[\nabla_x \log p(x)^\top f(x) + \nabla_x^\top f(x)] = 0$$

For all $f(x) \in F$ satisfying $\lim_{\|x\| \to \infty} p(x) f(x) = 0$

- Stein discrepancy:

$$S(q, p) = \sup_{f \in F} E_{q(x)}[\nabla_x \log p(x)^\top f(x) + \nabla_x^\top f(x)]$$

  - $q = p \Rightarrow S(q, p)$ (Stein's identity)
  - $S(q, p) = 0$, and if $f \in F$ is a rich function family $\Rightarrow q = p$

Gorham and Mackey. Measuring Sample Quality with Stein's Method. NIPS 2015
Liu et al. A Kernelized Stein Discrepancy for Goodness-of-fit Tests. ICML 2016
Chwialkowski et al. A kernel test of goodness of fit. ICML 2016

Stein's Method & Score Matching

- Stein discrepancy:

$$S(q,p) = \sup_{f \in F} E_{q(x)}[\nabla_x \log p(x)^\top f(x) + \nabla^\top f(x)]\,,$$

- Estimation requires:
  - Samples $x \sim q(x)$ ← enables estimation of $\nabla_x \log q(x)$ & posterior approximations
  - Score function $s_p(x)$ evaluated at x ← enables learning of $p(x)$ as an EBM

$$p(x) = \frac{1}{Z}\exp[-E_\theta(x)]$$

$$\nabla_x \log p(x) = -\nabla_x E_\theta(x) - \boxed{\nabla_x \log Z}$$

$$= 0$$

Gorham and Mackey. Measuring Sample Quality with Stein's Method. NIPS 2015
Li and Turner. Gradient Estimators for Implicit Models. ICLR 2018

Stein's Method & Score Matching

- Stein discrepancy:

$$S(q,p) = \sup_{f \in F} E_{q(x)}[\nabla_x \log p(x)^\top f(x) + \nabla^\top f(x)],$$

- Selecting test functions $f \in F$:
  - As L_2 integrable functions:
    - $f^*(x) \propto \nabla_x \log p(x) - \nabla_x \log q(x)$ (intractable)
    - Stein discrepancy ⇔ Fisher divergence (integration by parts)
    - Optimal test function approximated by a neural network $f_\phi(x)$: $R^D \to R^D$

Ranganath et al. Operator Variational Inference. NIPS 2016
Grathwohl et al. Learning the Stein Discrepancy for Training and Evaluating Energy-Based Models without Sampling. ICML 2020

Stein's Method & Score Matching

- Stein discrepancy:

$$S(q,p) = \sup_{f \in F} E_{q(x)}[\nabla_x \log p(x)^\top f(x) + \nabla^\top f(x)],$$

- Selecting test functions $f \in F$:
  - As functions in a unit ball of an RKHS:
    - Closed-form optimal test function as "smoothed" score difference
    - Optimal test function can be computed (integration by parts)
    - Test power depends on the choice of kernel

Liu et al. A Kernelized Stein Discrepancy for Goodness-of-fit Tests. ICML 2016
Chwialkowski et al. A kernel test of goodness of fit. ICML 2016

Stein's method & score matching

- Stein discrepancy:

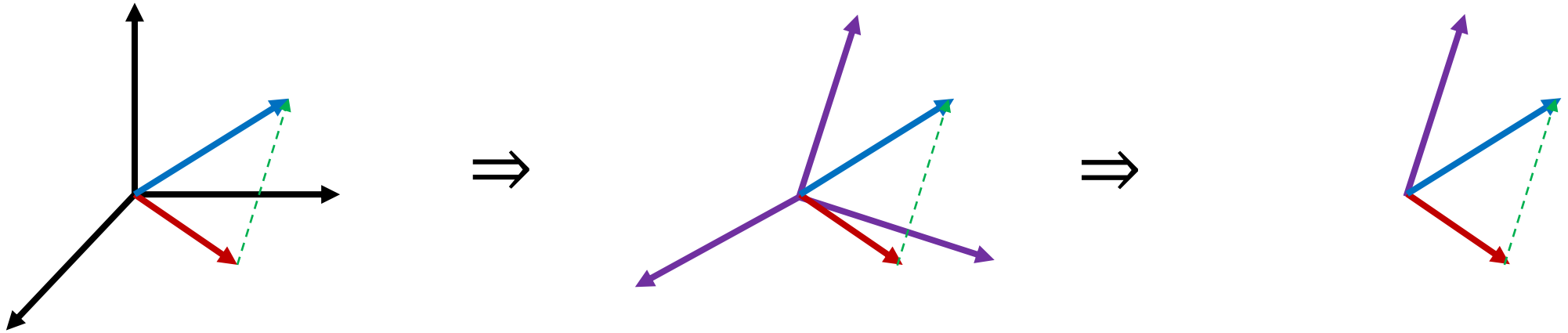
$$S(q,p) = \sup_{f \in F} E_{q(x)}[\nabla_x \log p(x)^\top f(x) + \nabla^\top f(x)],$$

- Curse-of-dimensionality problem:
  - $f \in F$ as L_2 integrable functions:
    - Optimizing $f_\phi(x)$: $R^D \to R^D$ is challenging in high dimensions
  - $f \in F$ as functions in a unit ball of an RKHS:
    - Open question of kernel choice in high dimensions

Gong et al. Sliced Kernelized Stein Discrepancy. ICLR 2021
Gong et al. Active Slices for Sliced Stein Discrepancy. ICML 2021

Sliced Stein Discrepancy

- 1st idea: find the projection direction where scores differ the most (on average)!



coordinate system with basis

$$I_d = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ & & \dots & \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

$$\nabla_x \log p(x) - \nabla_x \log q(x)$$

coordinate system with basis

$$O_r^\top = \begin{bmatrix} r_1^\top \\ \dots \\ r_d^\top \end{bmatrix}$$

$$O_r^\top (\nabla_x \log p(x) - \nabla_x \log q(x))$$

One best projection is enough!

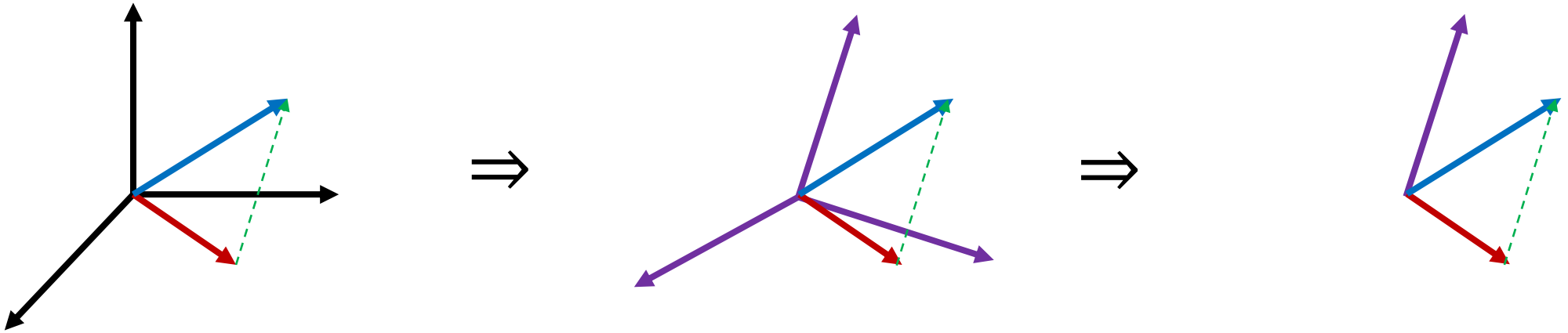
$$r_*$$

$$r_*^\top (\nabla_x \log p(x) - \nabla_x \log q(x))$$

Gong et al. Sliced Kernelized Stein Discrepancy. ICLR 2021
Gong et al. Active Slices for Sliced Stein Discrepancy. ICML 2021

Sliced Stein Discrepancy

- 1st idea: find the projection direction where scores differ the most (on average)!



coordinate system with basis

$$I_d = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ & & \dots & \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

$$\sup_{f \in F} E_{q(x)}[\nabla_x \log p(x)^\top f(x) + \nabla^\top f(x)]$$

$$f: R^d \to R^d$$

coordinate system with basis

$$O_r^\top = \begin{bmatrix} r_1^\top \\ \dots \\ r_d^\top \end{bmatrix}$$

$$\sum_{r \in O_r} \sup_{f_r \in F_r} E_{q(x)}[r^\top \nabla_x \log p(x)\, f_r(x) + r^\top \nabla f_r(x)]$$

$$f_r: R^d \to R, \forall r \in O_r$$

One best projection is enough!

$$r_* = \sup_r S_r$$

$$:= S_r$$

$$\sup_r \sup_{f_r \in F_r} E_{q(x)}[r^\top \nabla_x \log p(x)\, f_r(x) + r^\top \nabla f_r(x)]$$

$$f_r: R^d \to R$$

Gong et al. Sliced Kernelized Stein Discrepancy. ICLR 2021
Gong et al. Active Slices for Sliced Stein Discrepancy. ICML 2021

Sliced Stein Discrepancy

- 2nd idea: apply Radon transform

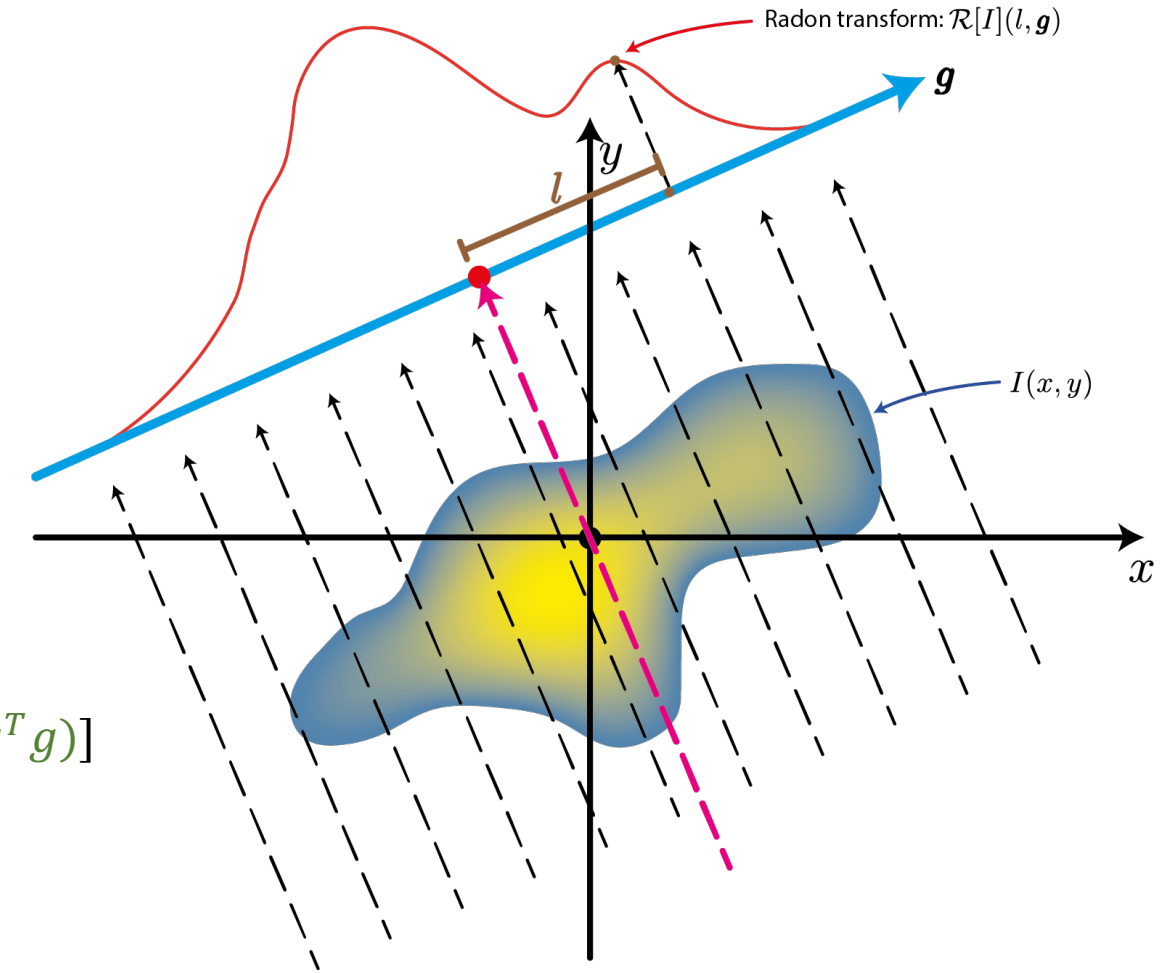
$$f_r: R^d \to R \quad \overset{\text{Radon transform}}{\Leftrightarrow} \quad f_{rg}: R \to R, \forall g \in S^{d-1}$$

- … again, pick best g

Resulting Stein discrepancy: maxSSD-rg

$$\sup_{r,g,f_{rg} \in F_{rg}} E_{q(x)}[r^\top \nabla_x \log p(x)\, f_{rg}(x^T g) + r^T g \nabla_{x^T g} f_{rg}(x^T g)]$$

$$f_{rg}: R \to R$$



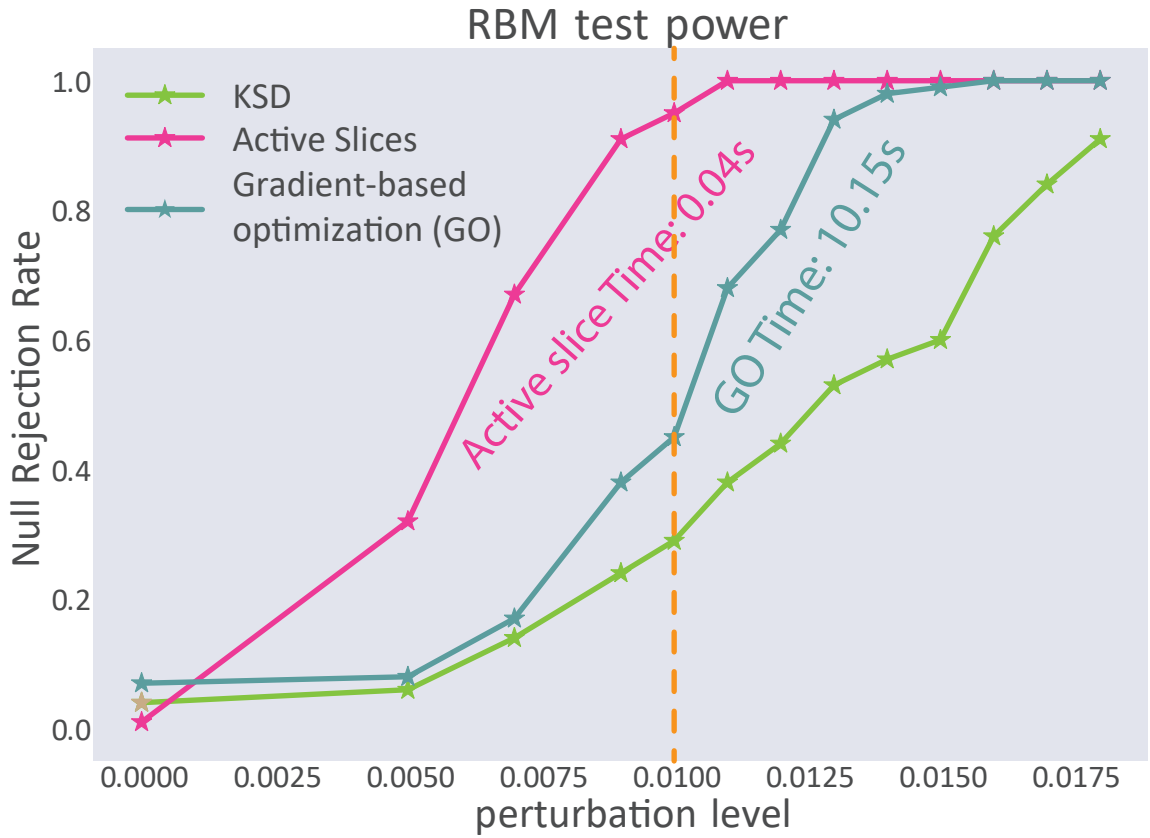
Gong et al. Sliced Kernelized Stein Discrepancy. ICLR 2021
Gong et al. Active Slices for Sliced Stein Discrepancy. ICML 2021

Goodness-of-fit Test with SSD

Goodness-of-fit test:

- Both p and q are Gaussian-RBMs
- q weights as perturbed from p weights

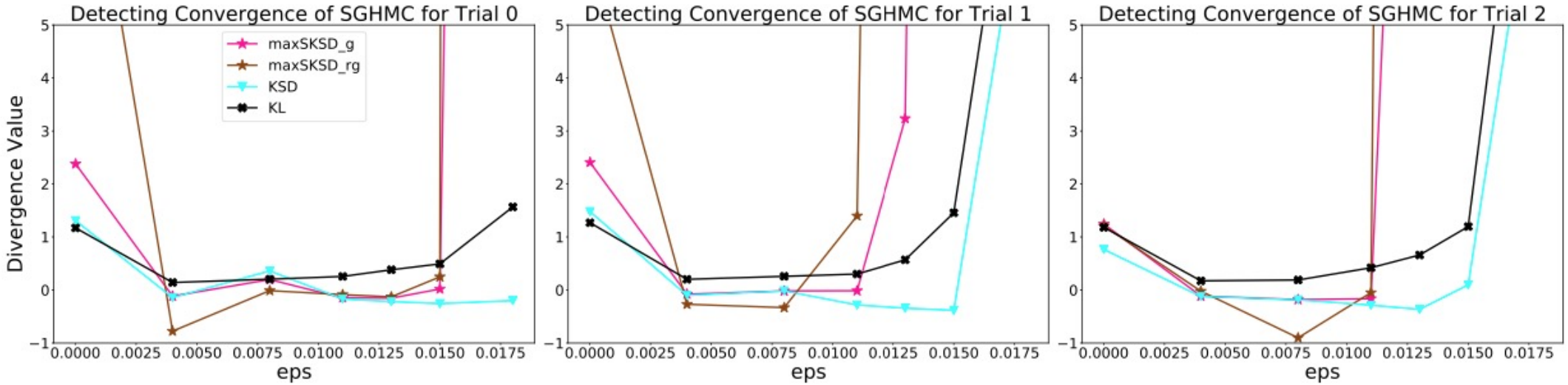
- Better test power:
  - at perturbation level 0.01, rejection rate 95% (Active slice) vs 45%(GO)
- Significant speed-up:
  - 0.04s (Active slice) vs 10.15s (GO)
  - 254x times faster



SSD variants: active slice vs GO

Gong et al. Sliced Kernelized Stein Discrepancy. ICLR 2021
Gong et al. Active Slices for Sliced Stein Discrepancy. ICML 2021

Selecting SGHMC Step-sizes with SD



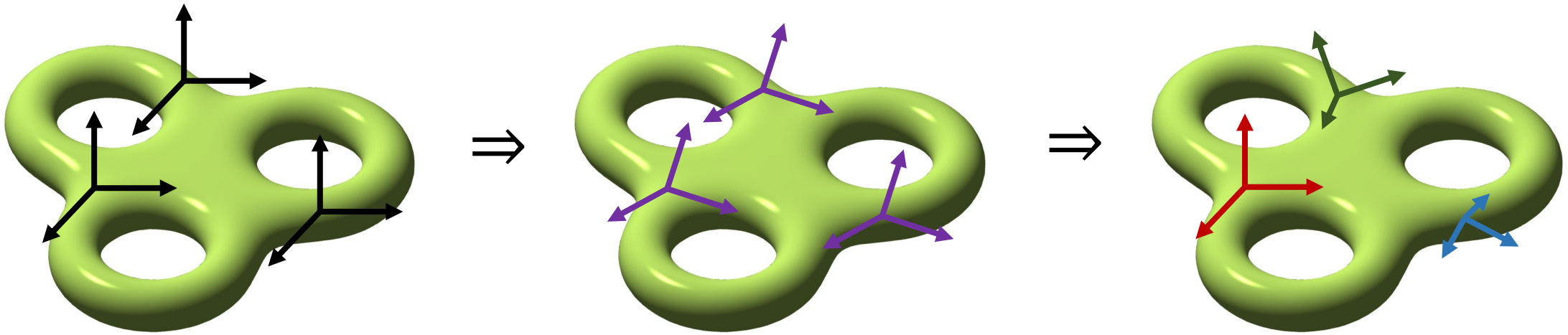
(a) Random seed 0 (b) Random seed 1 (c) Random seed 2

Selection result in Trial 1 (random seed 1)

Method	Metric				
	step size	KSD	maxSKSD_rg	maxSKSD_g	KL
KSD	0.015	-0.384	902	66.0	1.46
maxSKSD_rg	0.008	-0.0166	-0.332	-0.018	0.257
maxSKSD_g	0.004	-0.0100	-0.269	-0.079	0.201

Gong et al. Sliced Kernelized Stein Discrepancy. ICLR 2021

From Global to Local Coordinates



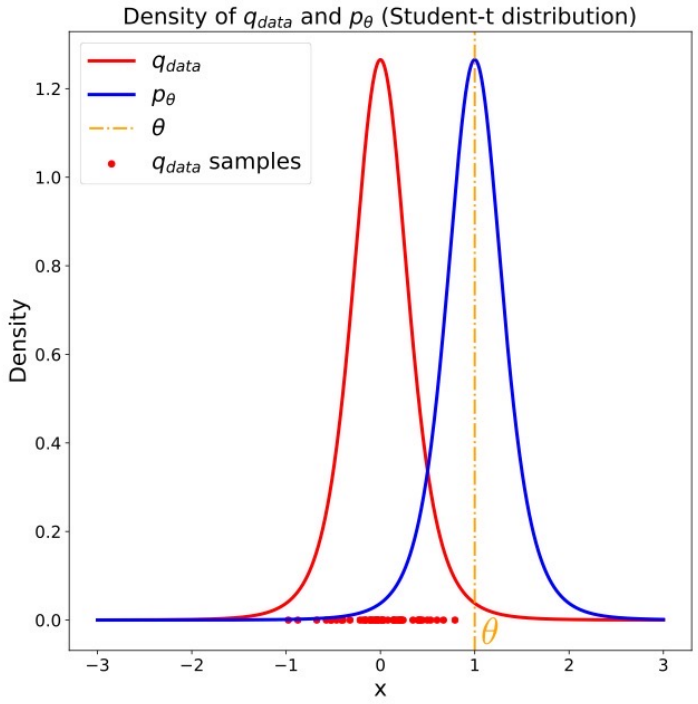
Stein discrepancy (SD)
Global Coordinate system I_d

Sliced Stein discrepancy (SSD)
Global Coordinate system O_r

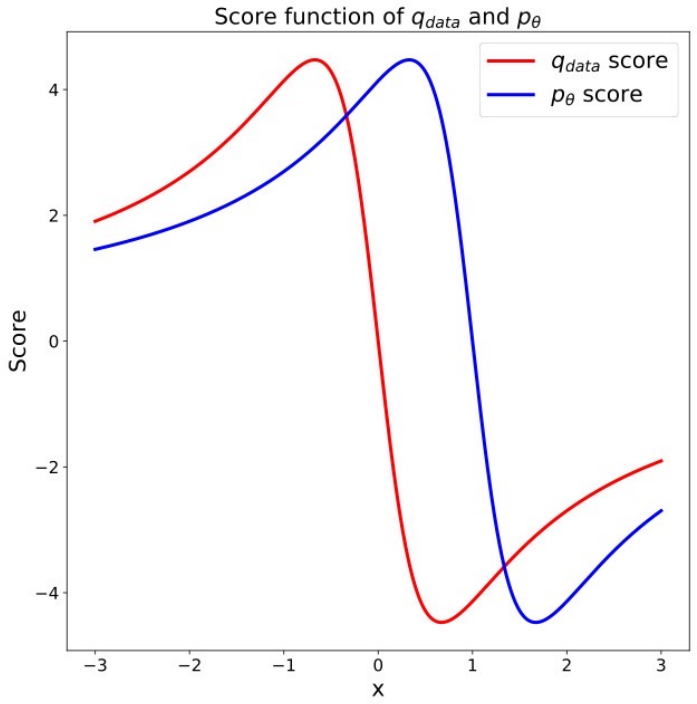
Diffusion score matching & SD
Local Coordinate system

Barp et al. Minimum Stein Discrepancy Estimators. NeurIPS 2019
Gong and Li. Interpreting Diffusion Score Matching using Normalizing Flows. ICML 2021 INNF+ workshop

Fitting Student-t with Score Matching



Train p_θ to fit q_{data}



Score of p_θ and q_{data}

Fisher Divergence

$$\mathcal{F}(q_{\text{data}}, p_\theta) = \mathbb{E}_{q_{\text{data}}}[|| \underbrace{\nabla_{\boldsymbol{x}} \log p_\theta(\boldsymbol{x})}_{\boldsymbol{s}_{p_\theta}(\boldsymbol{x})} - \underbrace{\nabla_{\boldsymbol{x}} \log q_{\text{data}}(\boldsymbol{x})}_{\boldsymbol{s}_{q_{\text{data}}}(\boldsymbol{x})} ||^2]$$

Integration by parts

Treat q_{data} as constant

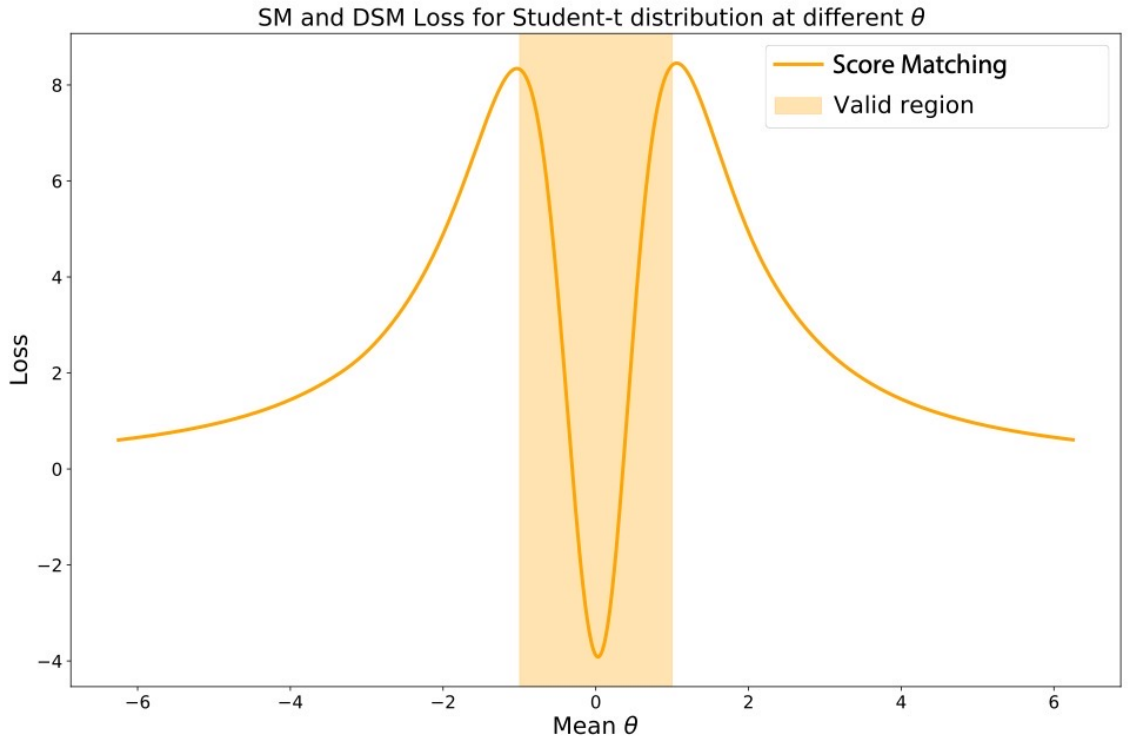
Score Matching

$$\mathcal{L}(\theta) = \mathbb{E}_{q_{\text{data}}}[||\boldsymbol{s}_{p_\theta}(\boldsymbol{x})||^2 + 2\nabla_{\boldsymbol{x}}^T \boldsymbol{s}_{p_\theta}(\boldsymbol{x})]$$

$$\text{argmin}_\theta \mathcal{F}(q_{\text{data}}, p_\theta) = \text{argmin}_\theta \mathcal{L}(\theta)$$

Fitting Student-t with Score Matching

$$p_\theta(\boldsymbol{x}) = \text{Student-t}(\theta, 0.3) \qquad q_{\text{data}}(\boldsymbol{x}) = \text{Student-t}(0, 0.3)$$



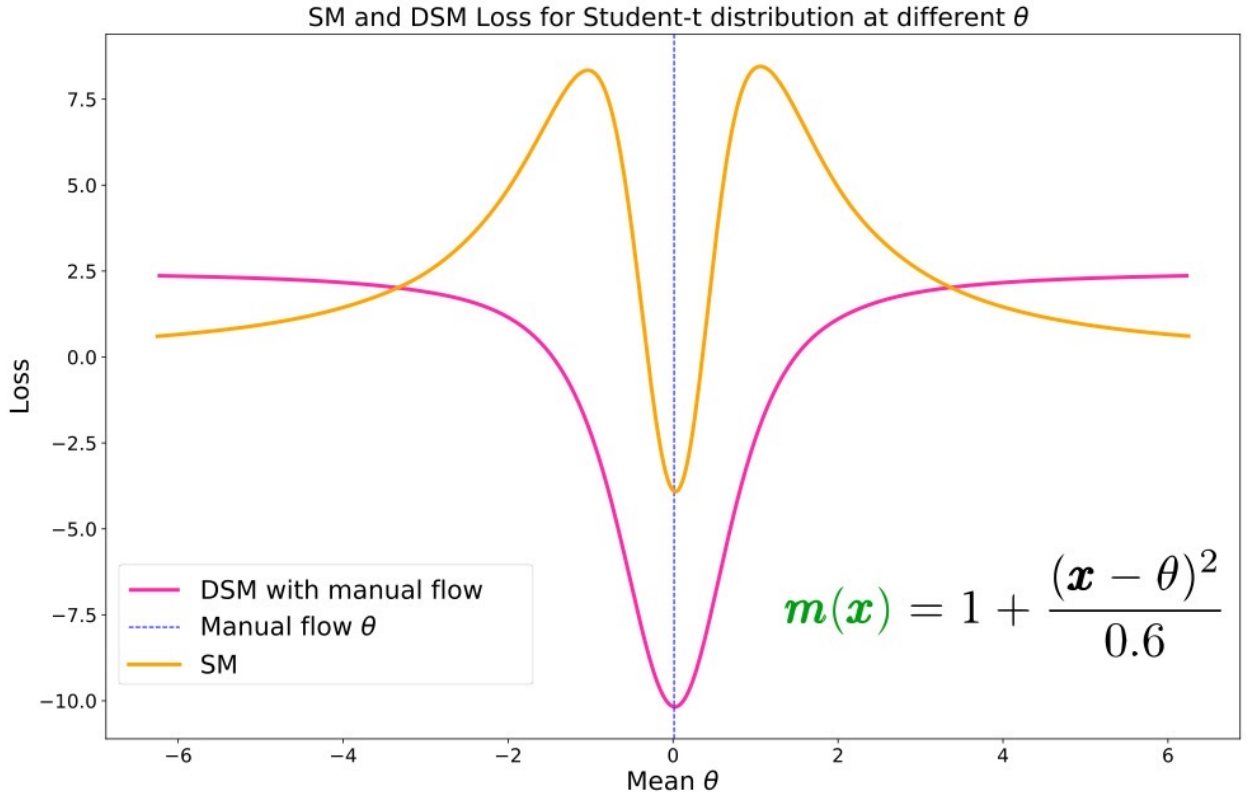
- θ initialized in valid region ⇒ optimisation converges to $\theta^* = 0$ (global optimum)
- Otherwise, optimisation converges to $\pm\infty$ (local optimum)

Fitting Student-t with Score Matching

Fisher Divergence → (With diffusion matrix $\boldsymbol{m}(\boldsymbol{x})$; $\boldsymbol{m}(\boldsymbol{x})$ is invertible) → **Diffusion Score Matching**

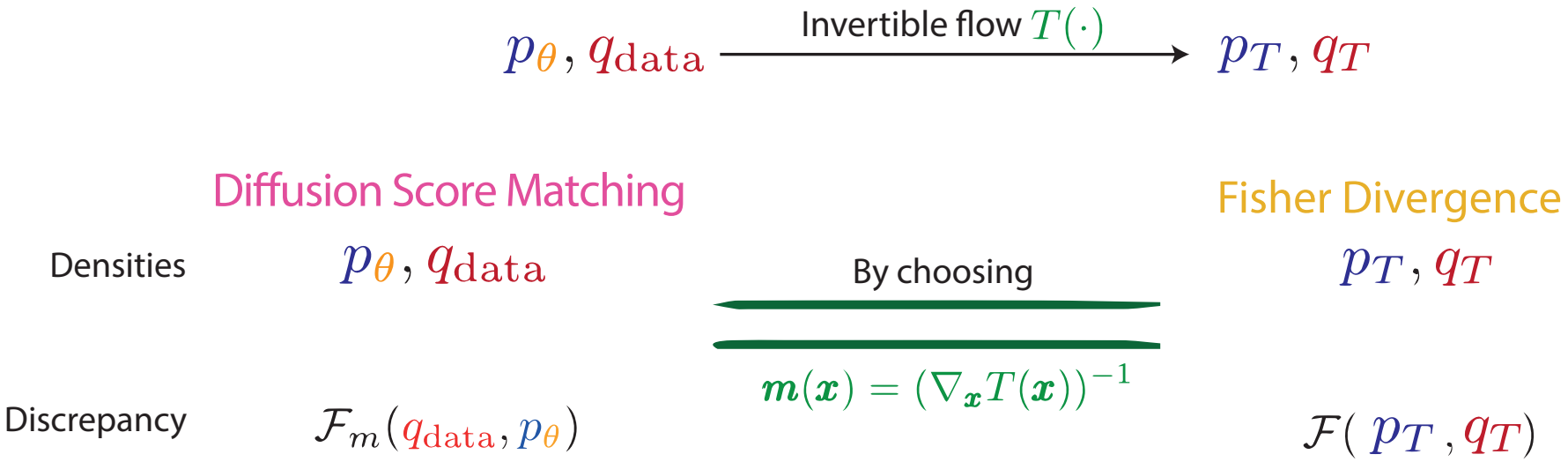
$$\mathcal{F}(q_{\text{data}}, p_\theta) = \mathbb{E}_{q_{\text{data}}}[|| \underbrace{\nabla_{\boldsymbol{x}} \log p_\theta(\boldsymbol{x})}_{\boldsymbol{s}_{p_\theta}(\boldsymbol{x})} - \underbrace{\nabla_{\boldsymbol{x}} \log q_{\text{data}}(\boldsymbol{x})}_{\boldsymbol{s}_{q_{\text{data}}}(\boldsymbol{x})} ||^2]$$

$$\mathcal{F}_m(q_{\text{data}}, p_\theta) = \mathbb{E}_{q_{\text{data}}}[||\boldsymbol{m}(\boldsymbol{x})^T(\boldsymbol{s}_{p_\theta}(\boldsymbol{x}) - \boldsymbol{s}_{q_{\text{data}}}(\boldsymbol{x}))||^2]$$



How to interpret & choose $m(x)$?

Interpreting DSM using Flows



Similar idea applies to Diffusion Stein Discrepancy

Better Flow Design

Flow (Barp et al.)

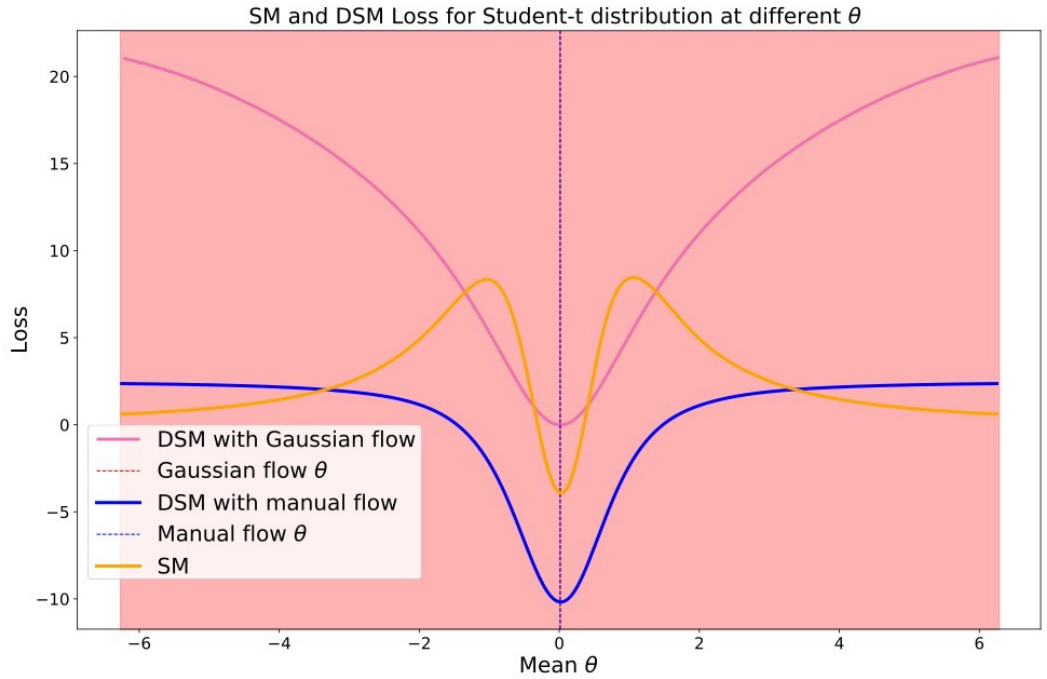
$\boldsymbol{m(x)}$ $\quad m(x) = 1 + \dfrac{(x-\theta)^2}{0.6}$

Gaussian Flow

$$m(x) = (\nabla_x T(x))^{-1}$$

where $T(\cdot)$ transforms p_θ to standard Gaussian

$T(\cdot)$ is composition of CDF_{p_θ} and $\mathrm{CDF}_{\mathcal{N}}^{-1}$



Faster convergence with Gaussian Flow compared to Flow by Barp et al. 2019

The Name "Diffusion"

- Stein operator ⇒ Stein discrepancy:
  - Find Stein operator $A_p[f]$ such that $E_{p(x)}\left[A_p[f](x)\right] = 0$ for $\forall f \in F$
  - Construct the corresponding Stein discrepancy

$$S(q,p) = \sup_{f \in F} E_{q(x)}[A_p[f](x)]$$

$$dx = b(x)dt + \sqrt{2D(x)}dW(t)$$
$$b(x) = [D(x) + Q(x)]\nabla_x \log p(x) + \Gamma(x)$$

Complete SG-MCMC recipe
Ito diffusion SDE

generator method ⇒

$$A_p[f](x) = \frac{1}{p(x)}\langle \nabla, p(x)(D(x) + Q(x))f(x)\rangle$$

Diffusion Stein discrepancy
with $m(x) = D(x) + Q(x)$

Gorham et al. Measuring Sample Quality with Diffusions. Annuals of Applied Probability 2019
Ma et al. A Complete Recipe for Stochastic Gradient MCMC. NIPS 2015

Conclusion

- Score-matching & Stein discrepancy as promising alternatives
  - Still facing challenges in high dimensions
- Ideas from flows can help improve Stein's method
- How about the other direction?

Thank you!