Minimum Stein Discrepancy Estimators

François-Xavier Briol
University of Cambridge & The Alan Turing Institute

ICML Workshop on
"Stein's Method for Machine Learning and Statistics"

Collaborators

Alessandro Barp
ICL

Andrew Duncan
ICL

Mark Girolami
U. Cambridge

Lester Mackey
Microsoft

Barp, A., Briol, F-X., Duncan, A., Girolami, M., Mackey, L. (2019)
Minimum Stein Discrepancy Estimators.

(preprint available here: `https://fxbriol.github.io`)

Statistical Inference for Unnormalised Models

- Motivation: Suppose we observe some data $\{x_1, \ldots, x_n\}$.

  Given a parametric family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$ with densities denoted p_θ, we seek $\theta^* \in \Theta$ which best approximates the empirical distribution:

$$\mathbb{Q}^n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

- Challenge: For complex models, we often only have access to the likelihood in unnormalised form:

$$p_\theta(x) = \frac{\tilde{p}_\theta(x)}{C}$$

  where $C > 0$ is unknown and $\tilde{p}$ can be evaluated pointwise.

- Examples include models of natural images, large graphical models, deep energy models, etc...

Statistical Inference for Unnormalised Models

- Motivation: Suppose we observe some data $\{x_1, \ldots, x_n\}$.

  Given a parametric family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$ with densities denoted p_θ, we seek $\theta^* \in \Theta$ which best approximates the empirical distribution:

$$\mathbb{Q}^n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

- Challenge: For complex models, we often only have access to the likelihood in unnormalised form:

$$p_\theta(x) = \frac{\tilde{p}_\theta(x)}{C}$$

  where $C > 0$ is unknown and $\tilde{p}$ can be evaluated pointwise.

- Examples include models of natural images, large graphical models, deep energy models, etc...

Statistical Inference for Unnormalised Models

- Motivation: Suppose we observe some data $\{x_1, \ldots, x_n\}$.

  Given a parametric family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$ with densities denoted p_θ, we seek $\theta^* \in \Theta$ which best approximates the empirical distribution:

$$\mathbb{Q}^n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

- Challenge: For complex models, we often only have access to the likelihood in unnormalised form:

$$p_\theta(x) = \frac{\tilde{p}_\theta(x)}{C}$$

  where $C > 0$ is unknown and $\tilde{p}$ can be evaluated pointwise.

- Examples include models of natural images, large graphical models, deep energy models, etc...

Minimum Discrepancy Estimators

- Let D be a function such that $D(\mathbb{Q}||\mathbb{P}_\theta) \geq 0$ measures the discrepancy between the empirical distribution $\mathbb{Q}$ and $\mathbb{P}_\theta$.
- We say that $\hat{\theta} \in \Theta$ is a minimum discrepancy estimator if:

$$\hat{\theta}_n \in \text{argmin}_{\theta \in \Theta} D(\mathbb{Q}^n||\mathbb{P}_\theta)$$

- This includes, but is not limited to:
  1. KL-divergence or other Bregman Divergence
  2. Wasserstein distance or Sinkhorn Divergence
  3. Maximum Mean Discrepancy
  4. ...
- Question: Which discrepancy should we use for unnormalised models?

Minimum Discrepancy Estimators

- Let D be a function such that $D(\mathbb{Q}||\mathbb{P}_\theta) \geq 0$ measures the discrepancy between the empirical distribution $\mathbb{Q}$ and $\mathbb{P}_\theta$.
- We say that $\hat{\theta} \in \Theta$ is a minimum discrepancy estimator if:

$$\hat{\theta}_n \in \text{argmin}_{\theta \in \Theta} D(\mathbb{Q}^n||\mathbb{P}_\theta)$$

- This includes, but is not limited to:
  1. KL-divergence or other Bregman Divergence
  2. Wasserstein distance or Sinkhorn Divergence
  3. Maximum Mean Discrepancy
  4. ...
- Question: Which discrepancy should we use for unnormalised models?

Minimum Discrepancy Estimators

- Let D be a function such that $D(\mathbb{Q}||\mathbb{P}_\theta) \geq 0$ measures the discrepancy between the empirical distribution $\mathbb{Q}$ and $\mathbb{P}_\theta$.
- We say that $\hat{\theta} \in \Theta$ is a minimum discrepancy estimator if:

$$\hat{\theta}_n \in \text{argmin}_{\theta \in \Theta} D(\mathbb{Q}^n||\mathbb{P}_\theta)$$

- This includes, but is not limited to:
  1. KL-divergence or other Bregman Divergence
  2. Wasserstein distance or Sinkhorn Divergence
  3. Maximum Mean Discrepancy
  4. ...
- Question: Which discrepancy should we use for unnormalised models?

Score Matching Estimators

- The score matching estimator [Hyvarinen, 2006] is based on the Fisher Divergence:

$$\begin{aligned} \mathrm{SM}(\mathbb{Q}||\mathbb{P}_\theta) &:= \int_{\mathcal{X}} \|\nabla \log q(x) - \nabla \log p_\theta(x)\|_2^2 \mathbb{Q}(dx) \\ &= \int_{\mathcal{X}} (\|\nabla \log p_\theta(x)\|_2^2 + 2\Delta \log p_\theta(x)) \mathbb{Q}(dx) + Z \end{aligned}$$

  where $Z \in \mathbb{R}$ is independent of θ

- This is one of the most competitive methods to date with applications for inference in natural images, deep energy models and directional statistics.

- Several Failure Modes: This approach requires second-order derivatives and struggles with heavy-tailed data [Swersky, 2011].

Score Matching Estimators

- The score matching estimator [Hyvarinen, 2006] is based on the Fisher Divergence:

$$\begin{aligned} \mathrm{SM}(\mathbb{Q}||\mathbb{P}_\theta) &:= \int_{\mathcal{X}} \|\nabla \log q(x) - \nabla \log p_\theta(x)\|_2^2 \mathbb{Q}(dx) \\ &= \int_{\mathcal{X}} (\|\nabla \log p_\theta(x)\|_2^2 + 2\Delta \log p_\theta(x)) \mathbb{Q}(dx) + Z \end{aligned}$$

  where $Z \in \mathbb{R}$ is independent of θ

- This is one of the most competitive methods to date with applications for inference in natural images, deep energy models and directional statistics.

- Several Failure Modes: This approach requires second-order derivatives and struggles with heavy-tailed data [Swersky, 2011].

Score Matching Estimators

- The score matching estimator [Hyvarinen, 2006] is based on the Fisher Divergence:

$$\begin{aligned} \mathrm{SM}(\mathbb{Q}||\mathbb{P}_\theta) &:= \int_{\mathcal{X}} \|\nabla \log q(x) - \nabla \log p_\theta(x)\|_2^2 \mathbb{Q}(dx) \\ &= \int_{\mathcal{X}} (\|\nabla \log p_\theta(x)\|_2^2 + 2\Delta \log p_\theta(x)) \mathbb{Q}(dx) + Z \end{aligned}$$

  where $Z \in \mathbb{R}$ is independent of θ

- This is one of the most competitive methods to date with applications for inference in natural images, deep energy models and directional statistics.

- Several Failure Modes: This approach requires second-order derivatives and struggles with heavy-tailed data [Swersky, 2011].

Minimum Stein Discrepancy Estimators

- Let $\Gamma(\mathcal{Y}) := \{f : \mathcal{X} \to \mathcal{Y}\}$. A function class $\mathcal{G} \subset \Gamma(\mathbb{R}^d)$ is a Stein class, with corresponding Stein operator $\mathcal{S}_{\mathbb{P}_\theta} : \mathcal{G} \subset \Gamma(\mathbb{R}^d) \to \Gamma(\mathbb{R}^d)$ if:

$$\int_{\mathcal{X}} \mathcal{S}_{\mathbb{P}_\theta}[f] d\mathbb{P}_\theta = 0 \quad \forall f \in \mathcal{G}$$

- This leads to the notion of Stein discrepancy (SD) [Gorham, 2015]:

$$\begin{aligned} \mathrm{SD}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\mathbb{Q}||\mathbb{P}_\theta) &:= \sup_{f \in \mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]} \left| \int_{\mathcal{X}} f d\mathbb{P}_\theta - \int_{\mathcal{X}} f d\mathbb{Q} \right| \\ &= \sup_{g \in \mathcal{G}} \left| \int_{\mathcal{X}} \mathcal{S}_{\mathbb{P}_\theta}[g] d\mathbb{Q} \right|, \end{aligned}$$

on which we base our minimum Stein discrepancy estimators:

$$\hat{\theta}_n \in \mathrm{argmin}_{\theta \in \Theta} \mathrm{SD}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\mathbb{Q}^n||\mathbb{P}_\theta).$$

Minimum Stein Discrepancy Estimators

- Let $\Gamma(\mathcal{Y}) := \{f : \mathcal{X} \to \mathcal{Y}\}$. A function class $\mathcal{G} \subset \Gamma(\mathbb{R}^d)$ is a Stein class, with corresponding Stein operator $\mathcal{S}_{\mathbb{P}_\theta} : \mathcal{G} \subset \Gamma(\mathbb{R}^d) \to \Gamma(\mathbb{R}^d)$ if:

$$\int_{\mathcal{X}} \mathcal{S}_{\mathbb{P}_\theta}[f] d\mathbb{P}_\theta = 0 \quad \forall f \in \mathcal{G}$$

- This leads to the notion of Stein discrepancy (SD) [Gorham, 2015]:

$$\begin{aligned} \mathrm{SD}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}\left(\mathbb{Q}||\mathbb{P}_\theta\right) &:= \sup_{f \in \mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]} \left| \int_{\mathcal{X}} f d\mathbb{P}_\theta - \int_{\mathcal{X}} f d\mathbb{Q} \right| \\ &= \sup_{g \in \mathcal{G}} \left| \int_{\mathcal{X}} \mathcal{S}_{\mathbb{P}_\theta}[g] d\mathbb{Q} \right|, \end{aligned}$$

  on which we base our minimum Stein discrepancy estimators:

$$\hat{\theta}_n \in \mathrm{argmin}_{\theta \in \Theta} \mathrm{SD}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}\left(\mathbb{Q}^n||\mathbb{P}_\theta\right).$$

Score Matching Estimators are Minimum Stein Discrepancy Estimators

- Consider the Stein operator $\mathcal{S}_p^m[g] := \frac{1}{p_\theta} \nabla \cdot (p_\theta g)$ and the Stein class:

$$\mathcal{G} = \left\{ g = (g_1, \ldots, g_d) \in C^1(\mathcal{X}, \mathbb{R}^d) \cap L^2(\mathcal{X}; \mathbb{Q}) : \|g\|_{L^2(\mathcal{X};\mathbb{Q})} \leq 1 \right\}.$$

In this case, the Stein discrepancy is the Score Matching divergence:

$$\mathrm{SD}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\mathbb{Q}||\mathbb{P}_\theta) = \mathrm{SM}(\mathbb{Q}||\mathbb{P}_\theta).$$

- Our paper also shows that several other popular estimators for unnormalised, including contrastive divergence and minimum probability flow are minimum SD estimators.

Score Matching Estimators are Minimum Stein Discrepancy Estimators

- Consider the Stein operator $\mathcal{S}_p^m[g] := \frac{1}{p_\theta}\nabla \cdot (p_\theta g)$ and the Stein class:

$$\mathcal{G} = \left\{ g = (g_1, \ldots, g_d) \in C^1(\mathcal{X}, \mathbb{R}^d) \cap L^2(\mathcal{X}; \mathbb{Q}) : \|g\|_{L^2(\mathcal{X};\mathbb{Q})} \leq 1 \right\}.$$

  In this case, the Stein discrepancy is the Score Matching divergence:

$$\mathrm{SD}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}\left(\mathbb{Q}||\mathbb{P}_\theta\right) = \mathrm{SM}(\mathbb{Q}||\mathbb{P}_\theta).$$

- Our paper also shows that several other popular estimators for unnormalised, including contrastive divergence and minimum probability flow are minimum SD estimators.

Score Matching Estimators are Minimum Stein Discrepancy Estimators

- Consider the Stein operator $\mathcal{S}_p^m[g] := \frac{1}{p_\theta}\nabla \cdot (p_\theta g)$ and the Stein class:

$$\mathcal{G} = \left\{ g = (g_1, \ldots, g_d) \in C^1(\mathcal{X}, \mathbb{R}^d) \cap L^2(\mathcal{X}; \mathbb{Q}) : \|g\|_{L^2(\mathcal{X};\mathbb{Q})} \leq 1 \right\}.$$

  In this case, the Stein discrepancy is the Score Matching divergence:

$$\mathrm{SD}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\mathbb{Q}||\mathbb{P}_\theta) = \mathrm{SM}(\mathbb{Q}||\mathbb{P}_\theta).$$

- Our paper also shows that several other popular estimators for unnormalised, including contrastive divergence and minimum probability flow are minimum SD estimators.

Minimum Diffusion Kernel Stein Discrepancy Estimators

- More general Stein operators were considered in [Gorham, 2016]:

$$\mathcal{S}_p^m[g] := \frac{1}{p_\theta}\nabla \cdot (p_\theta m g)\,, \quad \mathcal{S}_{p_\theta}^m[A] := \frac{1}{p_\theta}\nabla \cdot (p_\theta m A)\,,$$

  where $g \in \Gamma(\mathbb{R}^d)$, $A \in \Gamma(\mathbb{R}^{d\times d})$, and $m \in \Gamma(\mathbb{R}^{d\times d})$.

- Taking $\mathcal{G}$ to be the unit ball of a vector-valued RKHS $\mathcal{H}_K$, we get a diffusion kernel Stein discrepancy, which generalises the KSD:

$$\mathrm{DKSD}_{K,m}(\mathbb{Q}\|\mathbb{P})^2 = \int_{\mathcal{X}}\int_{\mathcal{X}} k_0(x,y)d\mathbb{Q}(x)d\mathbb{Q}(y)$$

  where

$$\begin{aligned} k_0(x,y) &:= \mathcal{S}_p^{m,2}\mathcal{S}_p^{m,1}K(x,y) \\ &= \frac{1}{p(y)p(x)}\nabla_y \cdot \nabla_x \cdot \Big(p(x)m(x)K(x,y)m(y)^\top p(y)\Big)\,. \end{aligned}$$

Diffusion Kernel Stein Discrepancy Estimators

We therefore end up with the following estimators:

$$\hat{\theta}_n^{\text{DKSD}} \in \text{argmin}_{\theta \in \Theta} \widehat{\text{DKSD}}_{K,m}(\mathbb{Q}^n \| \mathbb{P}_\theta)^2$$

where $\widehat{\text{DKSD}}_{K,m}(\mathbb{Q}^n \| \mathbb{P}_\theta)^2 = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} k_0(x_i, x_j)$.

Proposition (DKSD as statistical divergence)

Suppose K is IPD and in the Stein class of $\mathbb{Q}$, and $m(x)$ is invertible. If $\nabla \log p - \nabla \log q \in L^1(\mathbb{Q})$, then $DKSD_{K,m}(\mathbb{Q} \| \mathbb{P})^2 = 0$ iff $\mathbb{Q} = \mathbb{P}$.

Proposition (IPD matrix kernels)

(i) Let $K = diag(k^1, \ldots, k^d)$. Then K is IPD iff each kernel k^i is IPD. (ii) Let $K = Bk$ for B be symmetric positive definite. Then K is IPD iff k is IPD.

Diffusion Kernel Stein Discrepancy Estimators

We therefore end up with the following estimators:

$$\hat{\theta}_n^{\text{DKSD}} \in \text{argmin}_{\theta \in \Theta} \widehat{\text{DKSD}}_{K,m}(\mathbb{Q}^n \| \mathbb{P}_\theta)^2$$

where $\widehat{\text{DKSD}}_{K,m}(\mathbb{Q}^n \| \mathbb{P}_\theta)^2 = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} k_0(x_i, x_j)$.

Proposition (DKSD as statistical divergence)

Suppose K is IPD and in the Stein class of $\mathbb{Q}$, and $m(x)$ is invertible. If $\nabla \log p - \nabla \log q \in L^1(\mathbb{Q})$, then $DKSD_{K,m}(\mathbb{Q} \| \mathbb{P})^2 = 0$ iff $\mathbb{Q} = \mathbb{P}$.

Proposition (IPD matrix kernels)

(i) Let $K = diag(k^1, \ldots, k^d)$. Then K is IPD iff each kernel k^i is IPD. (ii) Let $K = Bk$ for B be symmetric positive definite. Then K is IPD iff k is IPD.

Diffusion Kernel Stein Discrepancy Estimators

We therefore end up with the following estimators:

$$\hat{\theta}_n^{\text{DKSD}} \in \text{argmin}_{\theta \in \Theta} \widehat{\text{DKSD}}_{K,m}(\mathbb{Q}^n \| \mathbb{P}_\theta)^2$$

where $\widehat{\text{DKSD}}_{K,m}(\mathbb{Q}^n \| \mathbb{P}_\theta)^2 = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} k_0(x_i, x_j)$.

Proposition (DKSD as statistical divergence)

Suppose K is IPD and in the Stein class of $\mathbb{Q}$, and $m(x)$ is invertible. If $\nabla \log p - \nabla \log q \in L^1(\mathbb{Q})$, then $DKSD_{K,m}(\mathbb{Q} \| \mathbb{P})^2 = 0$ iff $\mathbb{Q} = \mathbb{P}$.

Proposition (IPD matrix kernels)

(i) Let $K = diag(k^1, \ldots, k^d)$. Then K is IPD iff each kernel k^i is IPD. (ii) Let $K = Bk$ for B be symmetric positive definite. Then K is IPD iff k is IPD.

Consistency & Asymptotic Normality

Theorem (Consistency and Asymptotic Normality of DKSD)

Under smoothness and integrability conditions on K, m and $\theta \to \mathbb{P}_\theta$ and their derivatives, we have that θ_n^{DKSD} converges to θ^* a.s. Furthermore,

$$\sqrt{n}\left(\hat{\theta}_n^{DKSD} - \theta^*\right) \to \mathcal{N}\left(0, g_{DKSD}^{-1}(\theta^*)\Sigma g_{DKSD}^{-1}(\theta^*)\right)$$

where $\Sigma = \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \nabla_{\theta^*} k_0(x,y) d\mathbb{Q}(y)\right) \otimes \left(\int_{\mathcal{X}} \nabla_{\theta^*} k_0(x,z) d\mathbb{Q}(z)\right) d\mathbb{Q}(x)$ *and:*

$$g_{DKSD}(\theta)_{ij} = \int_{\mathcal{X}} \int_{\mathcal{X}} \left(\nabla_x \partial_{\theta^j} \log p_\theta\right)^\top m_\theta(x) K(x,y) m_\theta^\top(y) \nabla_y \partial_{\theta^i} \log p_\theta \, d\mathbb{P}_\theta(x) d\mathbb{P}_\theta(y).$$

- Important Remark: The choice of kernel K and diffusion matrix m will have a significant impact on the performance of these estimators!

Consistency & Asymptotic Normality

Theorem (Consistency and Asymptotic Normality of DKSD)

Under smoothness and integrability conditions on K, m and $\theta \to \mathbb{P}_\theta$ and their derivatives, we have that θ_n^{DKSD} converges to θ^* a.s. Furthermore,

$$\sqrt{n}\left(\hat{\theta}_n^{DKSD} - \theta^*\right) \to \mathcal{N}\left(0, g_{DKSD}^{-1}(\theta^*)\Sigma g_{DKSD}^{-1}(\theta^*)\right)$$

where $\Sigma = \int_{\mathcal{X}} \left(\int_{\mathcal{X}} \nabla_{\theta^*} k_0(x,y) d\mathbb{Q}(y)\right) \otimes \left(\int_{\mathcal{X}} \nabla_{\theta^*} k_0(x,z) d\mathbb{Q}(z)\right) d\mathbb{Q}(x)$ *and:*

$$g_{DKSD}(\theta)_{ij} = \int_{\mathcal{X}} \int_{\mathcal{X}} \left(\nabla_x \partial_{\theta^j} \log p_\theta\right)^\top m_\theta(x) K(x,y) m_\theta^\top(y) \nabla_y \partial_{\theta^i} \log p_\theta \, d\mathbb{P}_\theta(x) d\mathbb{P}_\theta(y).$$

- Important Remark: The choice of kernel K and diffusion matrix m will have a significant impact on the performance of these estimators!

Robustness of DKSD

The influence function describes infinitesimal corruption of the data and is given by $\text{IF}(z, \mathbb{Q}) := \partial_t \theta_{\mathbb{Q}_t}|_{t=0}$ if it exists, where $\mathbb{Q}_t = (1-t)\mathbb{Q} + t\delta_z$, for $t \in [0,1]$. An estimator is said to be bias robust if $\text{IF}(z, \mathbb{Q})$ is bounded in z.

Proposition (Robustness of DKSD estimators)

The influence function of DKSD is given by:

$$\text{IF}(z, \mathbb{P}_\theta) = g_{DKSD}(\theta)^{-1} \int_{\mathcal{X}} \nabla_\theta k_0(z, y) d\mathbb{P}_\theta(y).$$

In particular, there are various conditions on m and K which can guarantee that $\sup_{z \in \mathcal{X}} \|IF(z, \mathbb{P}_\theta)\| < \infty$.

- Important Remark: Once again, carefully choosing K and m can lead to good robustness properties.

Robustness of DKSD

The influence function describes infinitesimal corruption of the data and is given by $\text{IF}(z, \mathbb{Q}) := \partial_t \theta_{\mathbb{Q}_t}|_{t=0}$ if it exists, where $\mathbb{Q}_t = (1-t)\mathbb{Q} + t\delta_z$, for $t \in [0,1]$. An estimator is said to be bias robust if $\text{IF}(z, \mathbb{Q})$ is bounded in z.

Proposition (Robustness of DKSD estimators)

The influence function of DKSD is given by:

$$\text{IF}(z, \mathbb{P}_\theta) = g_{DKSD}(\theta)^{-1} \int_{\mathcal{X}} \nabla_\theta k_0(z, y) d\mathbb{P}_\theta(y).$$

In particular, there are various conditions on m and K which can guarantee that $\sup_{z \in \mathcal{X}} \|IF(z, \mathbb{P}_\theta)\| < \infty$.

- Important Remark: Once again, carefully choosing K and m can lead to good robustness properties.

Robustness of DKSD

The influence function describes infinitesimal corruption of the data and is given by $\text{IF}(z, \mathbb{Q}) := \partial_t \theta_{\mathbb{Q}_t}|_{t=0}$ if it exists, where $\mathbb{Q}_t = (1-t)\mathbb{Q} + t\delta_z$, for $t \in [0,1]$. An estimator is said to be bias robust if $\text{IF}(z, \mathbb{Q})$ is bounded in z.

Proposition (Robustness of DKSD estimators)

The influence function of DKSD is given by:

$$\text{IF}(z, \mathbb{P}_\theta) = g_{DKSD}(\theta)^{-1} \int_{\mathcal{X}} \nabla_\theta k_0(z, y) d\mathbb{P}_\theta(y).$$

In particular, there are various conditions on m and K which can guarantee that $\sup_{z \in \mathcal{X}} \|IF(z, \mathbb{P}_\theta)\| < \infty$.

- Important Remark: Once again, carefully choosing K and m can lead to good robustness properties.

Implementation of Minimum DKSD Estimators

In order to implement our DKSD estimators

$$\hat{\theta}_n^{\mathrm{DKSD}} \in \mathrm{argmin}_{\theta \in \Theta} \widehat{\mathrm{DKSD}}_{K,m}(\mathbb{Q}^n \| \mathbb{P}_\theta)^2,$$

we propose to make use of stochastic optimisation. In particular, we can make use of the geometry induced by DKSD to obtain an efficient algorithm akin to stochastic natural gradient descent [Amari, 1998]:

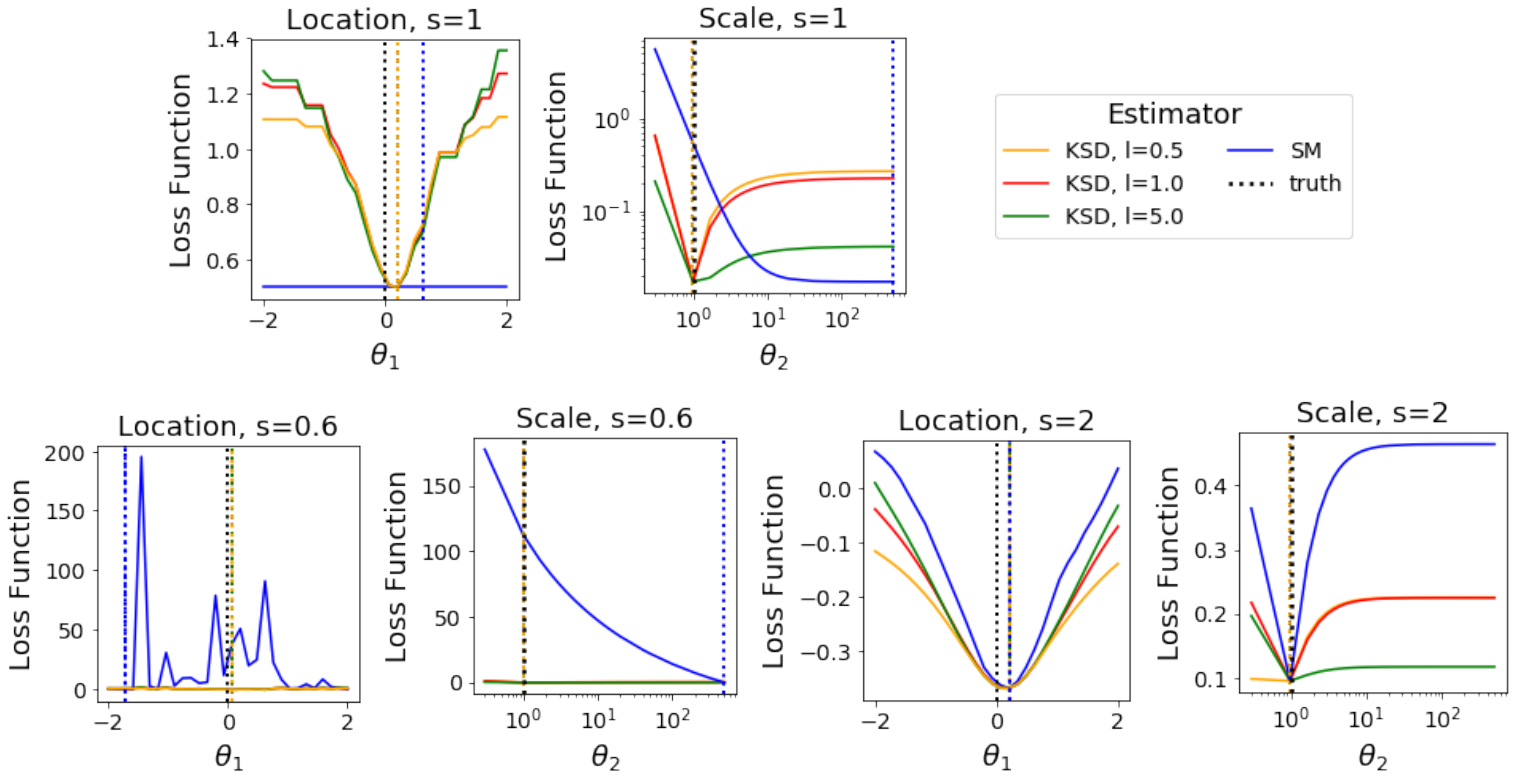
$$\hat{\theta}_{t+1} = \hat{\theta}_t - \gamma_t \widehat{g}^{-1}_{\mathrm{DKSD}}(\theta_t) \nabla_{\theta_t} \widehat{\mathrm{DKSD}}(\mathbb{Q}^n \| \mathbb{P}_\theta).$$

which approximates the gradient flow

$$\dot{\theta}(t) = -g^{-1}_{\mathrm{DKSD}}(\theta(t)) \nabla_\theta \mathrm{DKSD}(\mathbb{Q} \| \mathbb{P}_{\theta(t)}),$$

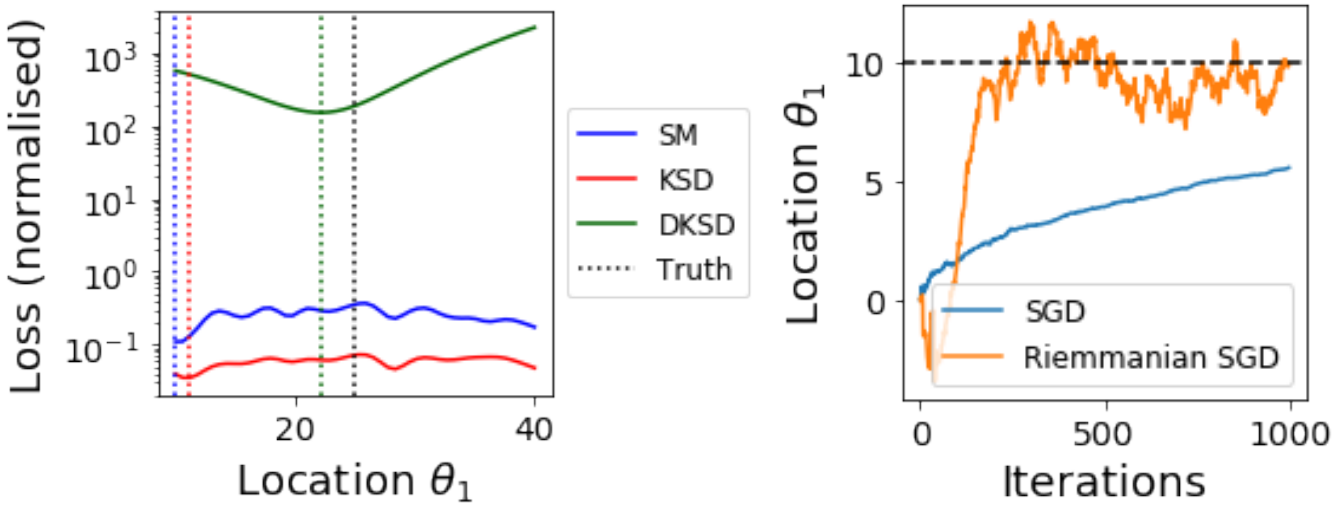
using U-statistics estimates of the metric tensor and gradient.

Application 1: Models with Rough Densities



- Model: $p_\theta(x) \propto (\|x - \theta_1\|_2/\theta_2)^{(s-d/2)} K_{s-d/2}(\|x - \theta_1\|_2/\theta_2)$
- Parameters: $(\theta_1^*, \theta_2^*) = (0, 1)$ and s varies.
- Number of samples: $n = 100$.

Application 2: Models with Heavy-Tails



- Model: $p_\theta(x) \propto (1/\theta_2)(1 + (1/\nu)(\|x - \theta_1\|_2/\theta_2)^2)^{-(\nu+1)/2}$.
- Diffusion Matrix: $m_\theta(x) = 1 = \|x - \theta_1\|^2/\theta_2^2$.
- Parameters: $\nu = 5$, $(\theta_1^*, \theta_2^*) = (25, 10)$.
- Number of Samples: Left: $n = 100$, Right: $n = 1000$.

Summary & Conclusions

In this talk, we have:

- Introduced a class of minimum Stein discrepancy estimators, and focused on a particular subclass called minimum DKSD.
- Shown that this class includes many popular estimators for unnormalised models including score-matching, contrastive divergence and minimum probability flow.
- Discussed consistency, a CLT, and robustness of minimum DKSD estimators, and discussed the importance of the kernel and operator.
- Demonstrated the advantage of the estimators for rough densities or heavy-tailed distributions.

Take home message: The flexibility offered by the choice of Stein class and operator allows us to tailor the estimators to the model of interest.

Barp, A., Briol, F-X., Duncan, A., Girolami, M., Mackey, L. (2019) Minimum Stein Discrepancy Estimators. (preprint: `https://fxbriol.github.io`)

Summary & Conclusions

In this talk, we have:

- Introduced a class of minimum Stein discrepancy estimators, and focused on a particular subclass called minimum DKSD.
- Shown that this class includes many popular estimators for unnormalised models including score-matching, contrastive divergence and minimum probability flow.
- Discussed consistency, a CLT, and robustness of minimum DKSD estimators, and discussed the importance of the kernel and operator.
- Demonstrated the advantage of the estimators for rough densities or heavy-tailed distributions.

Take home message: The flexibility offered by the choice of Stein class and operator allows us to tailor the estimators to the model of interest.

Barp, A., Briol, F-X., Duncan, A., Girolami, M., Mackey, L. (2019) Minimum Stein Discrepancy Estimators. (preprint: https://fxbriol.github.io)

Summary & Conclusions

In this talk, we have:

- Introduced a class of minimum Stein discrepancy estimators, and focused on a particular subclass called minimum DKSD.
- Shown that this class includes many popular estimators for unnormalised models including score-matching, contrastive divergence and minimum probability flow.
- Discussed consistency, a CLT, and robustness of minimum DKSD estimators, and discussed the importance of the kernel and operator.
- Demonstrated the advantage of the estimators for rough densities or heavy-tailed distributions.

Take home message: The flexibility offered by the choice of Stein class and operator allows us to tailor the estimators to the model of interest.

Barp, A., Briol, F-X., Duncan, A., Girolami, M., Mackey, L. (2019) Minimum Stein Discrepancy Estimators. (preprint: `https://fxbriol.github.io`)