Minimum Stein Discrepancy Estimators

Alessandro Barp, F.X. Briol, A. Duncan, M. Girolami, L. Mackey
a.barp16@imperial.ac.uk

Imperial College London

EPSRC Engineering and Physical Sciences Research Council

Minimum Stein Discrepancy Estimators

Aim: Given target distribution $\mathbb{Q}$ on $\mathbb{R}^d$, a statistical model $\mathcal{P}_\Theta \equiv \{\mathbb{P}_\theta\}_{\theta\in\Theta}$, and a measure of discrepancy $\theta \mapsto D(\mathbb{Q}\|\mathbb{P}_\theta)$, the aim is to estimate the best approximation to $\mathbb{Q}$ within the model

$$\theta^* \in \text{argmin}_{\theta\in\Theta} D(\mathbb{Q}\|\mathbb{P}_\theta) \tag{1}$$

Available Information: We are given i.i.d. sample $\{X_i\}_{i=1}^n$ from $\mathbb{Q}$, and differentiable **unnormalised** model $\mathbb{P}_\theta(\mathrm{d}x) \propto p_\theta(x)\mathrm{d}x$. We need a tractable approximation to $D(\mathbb{Q}\|\mathbb{P}_\theta)$, denoted $\hat{D}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta)$, and approximate (1) by

$$\hat{\theta}_n \in \text{argmin}_{\theta\in\Theta} \hat{D}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta)$$

Motivation: Such problems arise in modelling images, natural language, Markov random fields, nonparametric density estimation, statistical mechanics, sampling non-smooth distributions... but:

- The standard **MLE** requires normalised models.
- **Contrastive divergence** learning yields biased estimates.
- **Score Matching** (i) requires twice differentiable models, (ii) behaves poorly for models with heavy tailed distributions, (iii) is not robust to outliers.

Stein Discrepancy: We will need a Stein operator $\mathcal{S}_\mathbb{P}$ on Stein class $\mathcal{G}$, i.e., for any $\mathbb{P} \in \mathcal{P}_\Theta$

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

We then construct the Stein discrepancy using an integral probability metric

$$\text{SD}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\mathbb{Q}\|\mathbb{P}_\theta) \equiv \sup_{f\in\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]} \left|\int_\mathcal{X} f\mathrm{d}\mathbb{P}_\theta - \int_\mathcal{X} f\mathrm{d}\mathbb{Q}\right| = \sup_{g\in\mathcal{G}} \left|\int_\mathcal{X} \mathcal{S}_{\mathbb{P}_\theta}[g]\mathrm{d}\mathbb{Q}\right|, \tag{2}$$

and focus on **diffusion Stein operator** $\mathcal{S}_p^m[g] \equiv \frac{1}{p}\nabla\cdot(pmg)$

DKSD and DSM

Diffusions Kernel Stein Discrepancy:

- $\mathcal{G}$ is unit ball of RKHS $\mathcal{H}^d$ with matrix kernel K.
- Stein discrepancy becomes $\text{DKSD}_{K,m}(\mathbb{Q}\|\mathbb{P})^2 \equiv \sup_{\substack{h\in\mathcal{H}^d \\ \|h\|\leq 1}} \left|\int_\mathcal{X} \mathcal{S}_p^m[h]\mathrm{d}\mathbb{Q}\right|^2 = \int_\mathcal{X}\int_\mathcal{X} k^0(x,y)\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y)$ where k^0 Stein kernel with feature map $x \mapsto \mathcal{S}_p^{m,1}[K]|_x$.
- $K = Ik$, $m = I$ recovers KSD.

Diffusion Score Matching:

- $\mathcal{G}$ is unit ball in $L^2(\mathbb{Q})$ of $C^1(\mathbb{R}^d)\bigcap L^2(\mathbb{Q})$, using Stokes theorem

$$\text{DSM}_m(\mathbb{Q}\|\mathbb{P}) = \int_\mathcal{X} \left(\|m^\top\nabla_x \log p\|_2^2 + \|m^\top\nabla\log q\|_2^2 + 2\nabla\cdot\left(mm^\top\nabla\log p\right)\right)\mathrm{d}\mathbb{Q}$$

- recovers SM when $mm^\top = I$

Both can be approximated by U-statistics. Other choices lead to minimum probability flow and CD.

Theoretical Properties and Natural Gradient Descent

- We prove **consistency**

$$\hat{\theta}_n^{\text{DKSD}} \xrightarrow{a.s.} \text{argmin}_{\theta\in\Theta} \text{DKSD}_{K,m}(\mathbb{Q}\|\mathbb{P}_\theta)^2, \qquad \hat{\theta}_n^{\text{DSM}} \xrightarrow{p} \text{argmin}_{\theta\in\Theta} \text{DSM}_m(\mathbb{Q}\|\mathbb{P}_\theta).$$

- We derive **asymptotic normality** results of the form (with g the information metric)

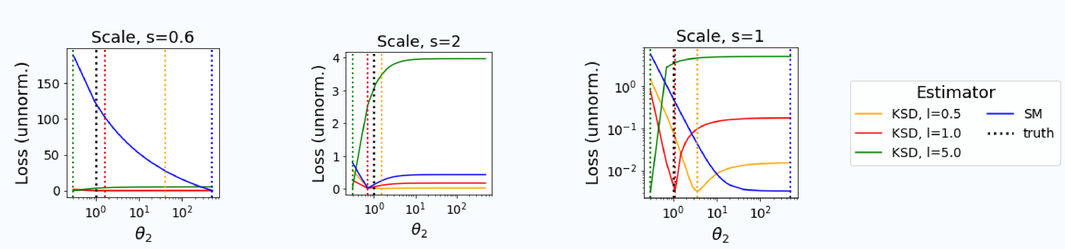
$$\sqrt{n}\left(\hat{\theta}_n - \theta^*\right) \xrightarrow{d} \mathcal{N}\left(0, g^{-1}(\theta^*)\Sigma g^{-1}(\theta^*)\right).$$

- We obtain conditions for **robustness** to corrupted data that improve on SM learning.
- For exponential models, our objectives are convex quadratics with closed-form solutions.

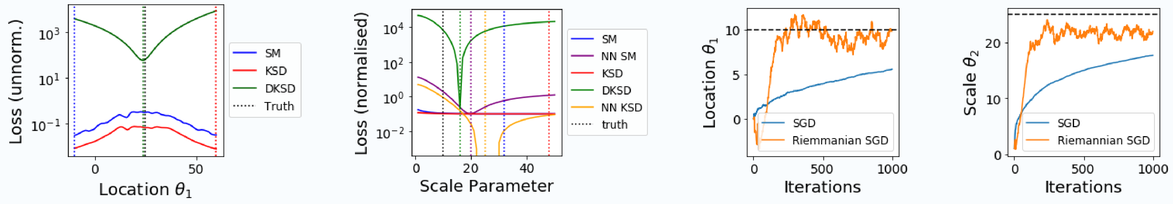
To take into account induced geometry of statistical model we minimise our objective functions using stochastic natural gradient descent with the information metric g.

Numerical Simulations

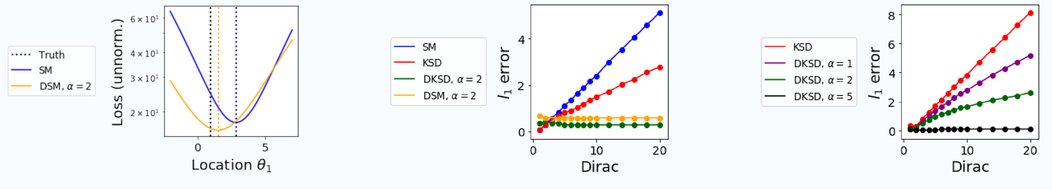
Rough Densities: learn scale symmetric Bessel distribution. Smoothness s, $\theta_2^* = 1$ and $n = 500$



Heavy-tailed Distributions: Non-standardised student-t distributions. $\theta_1^* = 25$, $\theta_2^* = 10$, $n = 300$



Robust Estimators: Generalised gamma location models, $m(x) = 1/(1+\|x\|^\alpha)$, corrupt 80 samples by setting their value to $x = 8$. $\theta_1^* = 0$ and $\theta_2^* = 2$ (left/middle) or $\theta_2^* = 5$ (right)



Acknowledgements

AB was supported by ICL. FXB was supported by [EP/L016710/1]. AD and MG were supported by [EP/T001569/1] and [EP/N510129/1]. MG was supported by [EP/J016934/3, EP/K034154/1, EP/P020720/1, EP/R018413/1].