Minimum Stein Discrepancy Estimators

Alessandro Barp[1,3], François-Xavier Briol[2,3], Andrew B. Duncan[1,3], Mark Girolami[2,3], and Lester Mackey[4]

[1]Department of Mathematics, Imperial College London.
[2]Department of Engineering, University of Cambridge.
[3]The Alan Turing Institute
[4]Microsoft Research

June 15, 2019

Abstract

When maximum likelihood estimation is infeasible, one often turns to score matching, contrastive divergence, or minimum probability flow learning to obtain tractable parameter estimates. We provide a unifying perspective of these techniques as *minimum Stein discrepancy estimators* and use this lens to design two new classes of estimators, called diffusion kernel Stein discrepancy (DKSD) and diffusion score matching (DSM), with complementary strengths. We establish the consistency, asymptotic normality, and robustness of DKSD and DSM estimators, derive stochastic Riemannian gradient descent algorithms for their efficient optimization, and demonstrate their advantages over score matching in models with non-smooth densities or heavy tailed distributions.

1 Introduction

Maximum likelihood estimation [9] is a de facto standard for estimating the unknown parameters in a statistical model $\{\mathbb{P}_\theta : \theta \in \Theta\}$. However, the computation and optimization of a likelihood typically requires access to the normalizing constants of the model distributions. This poses difficulties for complex statistical models for which direct computation of the normalisation constant would entail prohibitive multidimensional integration of an unnormalised density. Examples of such models arise naturally in modelling images [25, 37], natural language [46], Markov random fields [52] and nonparametric density estimation [54, 60]. To by-pass this issue, various approaches have been proposed to address parametric inference for unnormalised models, including Monte Carlo maximum likelihood [20], score matching (SM) estimators [32, 33], contrastive divergence [26], minimum probability flow learning [53] and noise-contrastive estimation [10, 24, 25].

The SM estimator is a minimum score estimator [14] based on the Hyvärinen scoring rule that avoids normalizing constants by depending on $\mathbb{P}_\theta$ only through the gradient of its log density $\nabla_x \log p_\theta$. SM estimators have proven to be a widely applicable method for estimation for models with unnormalised smooth positive densities, with generalisations to bounded domains [33] and compact Riemannian manifolds [45]. Despite the flexibility of this approach, SM has two important and distinct limitations. Firstly, as the Hyvärinen score depends on the Laplacian of the log-density, SM estimation will be expensive in high dimension and will break down for non-smooth models or for models in which the second derivative grows very rapidly. Secondly, as we shall demonstrate, SM estimators can behave poorly for models with heavy tailed distributions. Both of these situations arise naturally for energy models, particularly product-of-experts models and ICA models [31].

In separate strand of research, new approaches have been developed to measure discrepancy between an unnormalised distribution and a sample. In [21, 23, 44] it was shown that Stein's method can be used to construct discrepancy measures that control weak convergence of an empirical measure to a target. This was subsequently extended in [22] to encompass a family of discrepancy measures indexed by a reproducing kernel.

In this paper we consider minimum Stein discrepancy (SD) estimators and show that SM, minimum probability flow and contrastive divergence estimators are all special cases. Within this class we focus on SDs constructed from reproducing kernel Hilbert Spaces (RKHS), establishing the consistency, asymptotic normality and robustness of these estimators. We demonstrate that these SDs are appropriate for estimation of non-smooth distributions and heavy tailed distributions.

The remainder of the paper is organized as follows. In Section 2 we introduce the class of minimum SD estimators, including the subclass of diffusion kernel stein discrepancy estimators. In Section 3 we investigate asymptotic properties of these estimators, demonstrating consistency and asymptotic normality under general conditions, as well as conditions for robustness. Section 4 presents two toy problems where SM breaks down, but DKSD is able to recover the truth. All proofs are in the supplementary materials.

2 Minimum Stein Discrepancy Estimators

Let $\mathcal{P}_\mathcal{X}$ the set of Borel probability measures on $\mathcal{X}$. Given identical and independent (IID) realisations from a Borel measure $\mathbb{Q} \in \mathcal{P}_\mathcal{X}$ on an open subset $\mathcal{X} \subset \mathbb{R}^d$, the objective is to find a sequence of measures $\mathbb{P}_n$ that approximate $\mathbb{Q}$ in an appropriate sense. More precisely we will consider a family $\mathcal{P}_\Theta = \{\mathbb{P}_\theta : \theta \in \Theta\} \subset \mathcal{P}_\mathcal{X}$ together with a function $D : \mathcal{P}_\mathcal{X} \times \mathcal{P}_\mathcal{X} \to \mathbb{R}_+$ which quantifies the discrepancy between any two measures in $\mathcal{P}_\mathcal{X}$, and wish to estimate an optimal parameter θ^* satisfying $\theta^* \in \arg\min_{\theta\in\Theta} D(\mathbb{Q}\|\mathbb{P}_\theta)$. In practice, it is often difficult to compute the discrepancy D explicitly, and it is useful to consider a random approximation $\hat{D}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta)$ based on a IID sample $X_1, \dots, X_n \sim \mathbb{Q}$, such that $\hat{D}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta) \xrightarrow{a.s.} D(\mathbb{Q}\|\mathbb{P}_\theta)$ as $n \to \infty$. We then consider the sequence of estimators

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

The choice of discrepancy will impact the consistency, efficiency and robustness of the estimators. Examples of such estimators include minimum distance estimators [4, 50] where the discrepancy will be a metric on probability measures, including minimum maximum mean discrepancy (MMD) estimation [8, 16, 39] and minimum Wasserstein estimation [6, 17, 19].

More generally, minimum scoring rule estimators [14] arise from proper scoring rules, for example Hyvärinen, Bregman and Tsallis scoring rules. These discrepancies are often statistical divergences, i.e. $D(\mathbb{P}_\theta\|\mathbb{Q}) = 0 \Leftrightarrow \mathbb{P}_\theta = \mathbb{Q}$ for all $\mathbb{P}_\theta, \mathbb{Q}$ in a subset of $\mathcal{P}_\mathcal{X}$. Suppose that $\mathbb{P}_\theta$ and $\mathbb{Q}$ are absolutely continuous with respect to a common measure λ on $\mathcal{X}$, with respective densities p_θ and q. A well-known statistical divergence is the Kullback-Leibler (KL) divergence $\text{KL}(\mathbb{P}_\theta\|\mathbb{Q}) \equiv \int_\mathcal{X} \log(\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mathbb{Q})\mathrm{d}\mathbb{P}_\theta$ where $\mathrm{d}\mathbb{P}_\theta/\mathrm{d}\mathbb{Q}$ is the Radon-Nikodym derivative of $\mathbb{P}_\theta$ with respect to $\mathbb{Q}$. Since $\text{KL}(\mathbb{Q}\|\mathbb{P}_\theta) = \int_\mathcal{X} \log q \mathrm{d}\mathbb{Q} - \int_\mathcal{X} \log p_\theta \mathrm{d}\mathbb{Q}$, minimising $\text{KL}(\mathbb{Q}\|\mathbb{P}_\theta)$ is equivalent to maximising $\int_\mathcal{X} \log p_\theta \mathrm{d}\mathbb{Q}$, which can be estimated using the likelihood $\widehat{\text{KL}}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta) \equiv \frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i)$. Informally, we see that minimising the KL-divergence is equivalent to performing maximum likelihood estimation.

For our purposes we are interested in discrepancies that can be evaluated when $\mathbb{P}_\theta$ is only known up to normalisation, precluding the use of KL divergence. We instead consider a related class of discrepancies based on integral probability pseudometric (IPM) [47] and Stein's method [3, 11, 56]. Let $\Gamma(\mathcal{Y}) = \Gamma(\mathcal{X}, \mathcal{Y}) \equiv \{f : \mathcal{X} \to \mathcal{Y}\}$. A map $\mathcal{S}_{\mathbb{P}_\theta} : \mathcal{G} \subset \Gamma(\mathbb{R}^d) \to \Gamma(\mathbb{R}^d)$ is a Stein operator if:

$$\int_\mathcal{X} \mathcal{S}_{\mathbb{P}_\theta}[f]\mathrm{d}\mathbb{P}_\theta = 0 \quad \forall f \in \mathcal{G} \tag{1}$$

for any $\mathbb{P}_\theta$ and Stein class $\mathcal{G} \subset \Gamma(\mathbb{R}^d)$. We define a *Stein discrepancy* (SD) [21] to be the IPM with underlying function space $\mathcal{F} \equiv \mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]$. Using (1) this takes the form

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

We note that the Stein discrepancy depends on $\mathbb{Q}$ only through expectations, therefore permitting $\mathbb{Q}$ to be an empirical measure. If $\mathbb{P}_\theta$ has a C^1 density p_θ on $\mathcal{X}$, then one can consider the Langevin-Stein discrepancy arising from the Stein operator $\mathcal{T}_{p_\theta}[g] = \langle \nabla \log p_\theta, g\rangle + \nabla \cdot g$ defined on $g \in \Gamma(\mathbb{R}^d)$ [21, 23]. In this case, the Stein discrepancy will not depend on the normalising constant of p_θ. More general Langevin-Stein operators were considered in [23]:

$$\mathcal{S}_p^m[g] \equiv \frac{1}{p_\theta}\nabla \cdot (p_\theta m g), \quad \mathcal{S}_{p_\theta}^m[A] \equiv \frac{1}{p_\theta}\nabla \cdot (p_\theta m A), \tag{2}$$

where $g \in \Gamma(\mathbb{R}^d)$, $A \in \Gamma(\mathbb{R}^{d\times d})$, and $m \in \Gamma(\mathbb{R}^{d\times d})$ is a *diffusion matrix*. Several choices of Stein classes for this operator will be presented below. In this paper, we focus on obtaining the *minimum Stein discrepancy estimators* θ^{Stein} which minimises the criterion $\text{SD}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\mathbb{Q}\|\mathbb{P}_\theta)$. As we will only have access to the sample $\{X_i\}_{i=1}^n \sim \mathbb{Q}$, we will consider the estimators $\hat{\theta}_n^{\text{Stein}}$ minimising the approximation $\widehat{\text{SD}}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta)$ of $\text{SD}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\mathbb{Q}\|\mathbb{P}_\theta)$ based on a U-statistic of the $\mathbb{Q}$-integral, i.e. we seek

$$\hat{\theta}_n^{\text{Stein}} \equiv \text{argmin}_{\theta\in\Theta}\widehat{\text{SD}}_{\mathcal{S}_{\mathbb{P}_\theta}[\mathcal{G}]}(\{X_i\}_i^n\|\mathbb{P}_\theta).$$

2.1 Example 1: Diffusion Kernel Stein Discrepancy Estimators

A convenient choice of Stein class is the unit ball of reproducing kernel Hilbert spaces (RKHS) [5]. For the Stein operator $\mathcal{T}_p$ and kernel kI this was first introduced in [49], and considered extensively in the machine learning literature in the context of hypothesis testing, measuring sample quality and approximation of probability measures in [12, 13, 15, 22, 40, 42]. In this paper, we consider a more general class of discrepancies based on the diffusion Stein operator in (2).

Consider an RKHS $\mathcal{H}^d$ of functions $f \in \Gamma(\mathbb{R}^d)$ with (matrix-valued) kernel $K \in \Gamma(\mathcal{X} \times \mathcal{X}, \mathbb{R}^{d\times d})$, $K_x \equiv K(x,\cdot)$ (see Appendix A.3 and Appendix A.4 for further details). The Stein operator $\mathcal{S}_p^m[f]$ defined by (2) induces an operator $\mathcal{S}_p^{m,2}\mathcal{S}_p^{m,1} : \Gamma(\mathcal{X}\times\mathcal{X}, \mathbb{R}^{d\times d}) \to \Gamma(\mathbb{R})$ which acts first on the first variable and then on the second one. We shall consider two possible forms of the kernel K. Either the components of f are independent, in which case we have a diagonal kernel (i) $K = \text{diag}(\lambda_1 k^1, \ldots, \lambda_d k^d)$ where $\lambda_i > 0$ and k^i is a C^2 kernel on $\mathcal{X}$, for $i = 1, \ldots, d$; or (ii) $K = Bk$ where k is a (scalar) kernel on $\mathcal{X}$ and B is a constant symmetric positive definite matrix. In Appendix B we show that:

Theorem 1 (Diffusion Kernel Stein Discrepancy). *For either K, we find that $\mathcal{S}_p^m[f](x) = \langle \mathcal{S}_p^{m,1}K_x, f\rangle_{\mathcal{H}^d}$ for any $f \in \mathcal{H}^d$. Moreover if $x \mapsto \|\mathcal{S}_p^{m,1}K_x\|_{\mathcal{H}^d} \in L^1(\mathbb{Q})$, we have*

$$\text{DKSD}_{K,m}(\mathbb{Q}\|\mathbb{P})^2 \;\equiv\; \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\le 1}} \left|\int_{\mathcal{X}} \mathcal{S}_p^m[h]\mathrm{d}\mathbb{Q}\right|^2 = \int_{\mathcal{X}}\int_{\mathcal{X}} k^0(x,y)\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y)$$

$$k^0(x,y) \equiv \mathcal{S}_p^{m,2}\mathcal{S}_p^{m,1}K(x,y) = \tfrac{1}{p(y)p(x)}\nabla_y\cdot\nabla_x\cdot\big(p(x)m(x)K(x,y)m(y)^\top p(y)\big). \quad (3)$$

We call $\text{DKSD}_{K,m}$ the *diffusion kernel Stein discrepancy (DKSD)* and propose the following U-statistic approximation:

$$\widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta)^2 \;=\; \tfrac{2}{n(n-1)}\textstyle\sum_{1\le i<j\le n} k_\theta^0(X_i,X_j) \;=\; \tfrac{1}{n(n-1)}\textstyle\sum_{i\ne j} k_\theta^0(X_i,X_j) \quad (4)$$

with associated estimators: $\hat{\theta}_n^{\text{DKSD}} \in \text{argmin}_{\theta\in\Theta}\widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta)^2$. For $K = Ik$, $m = Ih$, DKSD is KSD with kernel $h(x)k(x,y)h(y)$, and if $h = 1$ we recover the usual definition of KSD considered in previous work (see Appendix B.4 for further details):

$$\text{DKSD}_{kI,I}(\mathbb{Q}\|\mathbb{P})^2 = \text{KSD}_k(\mathbb{Q}\|\mathbb{P})^2 = \int_{\mathcal{X}}\int_{\mathcal{X}} \tfrac{1}{p(y)p(x)}\nabla_y\cdot\nabla_x(p(x)k(x,y)p(y))\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y)$$

Now that our DKSD estimators are defined, an important question remaining is under which conditions can DKSD discriminate distinct probability measures. This will be dependent on the kernel and the model under consideration. We say a matrix kernel K is in the Stein class of $\mathbb{Q}$ if $\int_{\mathcal{X}} \mathcal{S}_q^{m,1}[K]\mathrm{d}\mathbb{Q} = 0$, and that it is strictly integrally positive definite (IPD) if $\int_{\mathcal{X}\times\mathcal{X}} \mathrm{d}\mu^\top(x)K(x,y)\mathrm{d}\mu(y) > 0$ for any finite non-zero signed vector Borel measure μ. From $\mathcal{S}_p^m[f](x) = \langle\mathcal{S}_p^{m,1}K_x, f\rangle_{\mathcal{H}^d}$ we have that $f \in \mathcal{H}^d$ is in the Stein class (i.e., $\int_{\mathcal{X}}\mathcal{S}_q^m[f]\mathrm{d}\mathbb{Q} = 0$) when K is also in the class. Setting $s_p \equiv m^\top\nabla\log p \in \Gamma(\mathbb{R}^d)$, we have:

Proposition 1 (DKSD as statistical divergence). *Suppose K is IPD and in the Stein class of $\mathbb{Q}$, and $m(x)$ is invertible. If $s_p - s_q \in L^1(\mathbb{Q})$, then $\text{DKSD}_{K,m}(\mathbb{Q}\|\mathbb{P})^2 = 0$ iff $\mathbb{Q} = \mathbb{P}$.*

See Appendix B.5 for the proof. Note that this proposition generalises Proposition 3.3 from [42] to a significantly larger class of SD. For the matrix kernels introduced above, the proposition below shows that K is IPD when its associated scalar kernels are; a well-studied problem [55].

Proposition 2 (IPD matrix kernels). *(i) Let $K = diag(k^1, \dots, k^d)$. 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.*

The remainder of the paper will focus on properties of DKSD estimators, but before proceeding further we introduce alternative minimum SD estimators.

2.2 Example 2: Diffusion Score Matching Estimators

A well-known family of statistical estimators due to [32, 33] are the score matching (SM) estimators (based on the Fisher or Hyvarinen divergence). As will be shown in this section, these can be seen as special cases of minimum SD estimators. The SM divergence is computable for unnormalised models with sufficiently smooth densities:

$$\mathrm{SM}(\mathbb{Q}\|\mathbb{P}) \equiv \int_{\mathcal{X}} \|\nabla \log p - \nabla \log q\|_2^2 \, \mathrm{d}\mathbb{Q} = \int_{\mathcal{X}} \big(\|\nabla \log q\|_2^2 + \|\nabla \log p\|_2^2 + 2\Delta \log p\big)\mathrm{d}\mathbb{Q}$$

where Δ denotes the Laplacian and we have used the divergence theorem. If $\mathbb{P} = \mathbb{P}_\theta$, the first integral above does not depend on θ, and the second one does not depend on the density of $\mathbb{Q}$, so we consider the approximation $\widehat{\mathrm{SM}}(\{X_i\}_{i=1}^n \|\mathbb{P}_\theta) \equiv \frac{1}{n}\sum_{i=1}^n \Delta \log p_\theta(X_i) + \frac{1}{2}\|\nabla \log p_\theta(X_i)\|_2^2$ based on an unbiased estimation for the minimiser of the SM divergence, and its estimators $\hat{\theta}_n^{\mathrm{SM}} \equiv \mathrm{argmin}_{\theta\in\Theta}\widehat{\mathrm{SM}}(\{X_i\}_{i=1}^n\|\mathbb{P}_\theta)$, for independent random vectors $X_i \sim \mathbb{Q}$. The SM divergence can also be generalised to include higher-order derivatives of the log-likelihood [43] and does not depend on the normalised likelihood. We will now introduce a further generalisation that we call *diffusion score matching (DSM)* and is a SD constructed from the Stein operator (2) (see Appendix B.6 for proof):

Theorem 2 (Diffusion Score Matching). *Let $\mathcal{X} = \mathbb{R}^d$ and consider the Stein operator $\mathcal{S}_p$ in (2) for some function $m \in \Gamma(\mathbb{R}^{d\times d})$ and the Stein class $\mathcal{G} \equiv \{g = (g_1, \dots, g_d) \in C^1(\mathcal{X}, \mathbb{R}^d) \cap L^2(\mathcal{X};\mathbb{Q}) : \|g\|_{L^2(\mathcal{X};\mathbb{Q})} \le 1\}$. If $p, q > 0$ are differentiable and $s_p - s_q \in L^2(\mathbb{Q})$, then we define the* diffusion score matching *divergence as the Stein discrepancy,*

$$\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}) \equiv \sup_{f\in\mathcal{S}_p[\mathcal{G}]} \left|\int_{\mathcal{X}} f \mathrm{d}\mathbb{Q} - \int_{\mathcal{X}} f \mathrm{d}\mathbb{P}\right|^2 = \int_{\mathcal{X}} \left\|m^\top(\nabla \log q - \nabla \log p)\right\|_2^2 \mathrm{d}\mathbb{Q}.$$

This satisfies $\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}) = 0$ iff $\mathbb{Q} = \mathbb{P}$ when $m(x)$ is invertible. Moreover, if p is twice-differentiable, and $qmm^\top \nabla \log p, \nabla \cdot (qmm^\top \nabla \log p) \in L^1(\mathbb{R}^d)$, then Stoke's theorem gives

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

Notably, DSM_m recovers SM when $m(x)m(x)^\top = I$, and the (generalised) non-negative score matching estimator of [43] with the choice $m(x) \equiv \mathrm{diag}(h_1(x_1)^{1/2}, \dots, h_d(x_d)^{1/2})$. Like standard SM, DSM is only defined for distributions with sufficiently smooth densities. However the θ-dependent part of $\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}_\theta)$ does not depend on the density of $\mathbb{Q}$, and can be estimated using an empirical mean, which leads to a sequence of estimators

$$\hat{\theta}_n^{\mathrm{DSM}} \equiv \mathrm{argmin}_{\theta\in\Theta} \tfrac{1}{n}\textstyle\sum_{i=1}^n \Big(\|m^\top(X_i)\nabla_x \log p_\theta(X_i)\|_2^2 + 2\nabla \cdot \big(m(X_i)m^\top(X_i)\nabla \log p_\theta(X_i)\big)\Big)$$

where $\{X_i\}_{i=1}^n$ is a sample from $\mathbb{Q}$. Note that this is only possible if m is independent of θ, in contrast to DKSD where m can depend on $\mathcal{X} \times \Theta$, thus leading to a more flexible class of estimators. An interesting remark is that the DSM_m discrepancy may in fact be obtained as a limit of DKSD:

Theorem 3 (DSM as a limit of DKSD). *Let $\mathbb{Q}(\mathrm{d}x) \equiv q(x)\mathrm{d}x$ be a probability measure on $\mathbb{R}^d$ with $q > 0$. Suppose that $f, g \in C(\mathbb{R}^d) \cap L^2(\mathbb{Q})$, $\Phi \in L^1(\mathbb{R}^d)$, $\Phi > 0$ and $\int_{\mathbb{R}^d} \Phi(s)\,\mathrm{d}s = 1$, define $K_\gamma \equiv B\Phi_\gamma$ where $\Phi_\gamma(s) \equiv \gamma^{-d}\Phi(s/\gamma)$ and $\gamma > 0$. Let $k_\gamma^q(x,y) = k_\gamma(x,y)/\sqrt{q(x)q(y)} = \Phi_\gamma(x-y)/\sqrt{q(x)q(y)}$, and set $K_\gamma^p \equiv Bk_\gamma^q$. Then,*

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(x)^\top K_\gamma^q(x,y) g(y)\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) \;\to\; \int_{\mathbb{R}^d} f(x)^\top Bg(x)\mathrm{d}\mathbb{Q}(x), \qquad \textit{as} \qquad \gamma \to 0.$$

In particular choosing $f, g = s_p - s_q$ shows $\mathrm{DKSD}_{K_\gamma^q, m}(\mathbb{Q}\|\mathbb{P})^2$ converge to $\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P})$.

See Appendix B.6 for a proof. Note that this theorem corrects, and significantly generalises, previously established connections between the SM divergence and KSD (such as in Sec. 5 of [42]).

We conclude by commenting on the computational complexity of evaluating the DKSD loss function. The most general formulation requires $\mathcal{O}(n^2d^2)$ computational cost due to computation of a U-statistic and a matrix-matrix product. However, if $K = \mathrm{diag}(\lambda_1 k^1, \ldots, \lambda_d k^d)$ or $K = Bk$, and if m is a diagonal matrix, then we can by-pass expensive matrix products and reduce the computational cost to $\mathcal{O}(n^2d)$, making it comparable to that of standard KSD. Although we do not consider these in this paper, recent approximations to KSD could also be adapted to DKSD to reduce the computational cost to $\mathcal{O}(nd)$ [30, 34]. For the DSM loss function, the computational cost is of order $\mathcal{O}(nd^2)$, making the cost comparable to that of the SM loss. From a purely computational viewpoint, DSM will hence be preferable to DKSD for large n, whilst DKSD will be preferable to DSM for large d.

2.3 Further Examples: Contrastive Divergence and Minimum Probability Flow

The class of minimum SD estimators also includes other well-known estimators for unnormalised models. Let X_θ^n, $n \in \mathbb{N}$ be a Markov process with unique invariant probality measure $\mathbb{P}_\theta$, for example a Metropolis-Hastings chain. Let P_θ^n be the associated transition semigroup, i.e. $(P_\theta^n f)(x) = \mathbb{E}[f(X_\theta^n)|X_\theta^0 = x]$. Choosing the Stein operator $\mathcal{S}_p = I - P_\theta^n$ and Stein class $\mathcal{G} = \{\log p_\theta + c : c \in \mathbb{R}\}$, leads to the following SD:

$$\mathrm{CD}(\mathbb{Q}\|\mathbb{P}_\theta) = \int_{\mathcal{X}} (\log p_\theta(x) - P_\theta^n \log p_\theta(x)) \mathrm{d}\mathbb{Q}(x) = \mathrm{KL}(\mathbb{Q}\|\mathbb{P}_\theta) - \mathrm{KL}(\mathbb{Q}_\theta^n\|\mathbb{P}_\theta),$$

where $\mathbb{Q}_\theta^n$ is the law of $X_\theta^n|X_\theta^0 \sim \mathbb{Q}$ and assuming that $\mathbb{Q} \ll \mathbb{P}_\theta$ and $\mathbb{Q}_\theta^n \ll \mathbb{P}_\theta$, which is the loss function associated with contrastive divergence (CD) [26, 41]. Suppose now that $\mathcal{X}$ is a finite set. Given $\theta \in \Theta$ let P_θ be the transition matrix for a Markov process with unique invariant distribution $\mathbb{P}_\theta$. Suppose we observe data $\{x_i\}_{i=1}^n$ and let q be the corresponding empirical distribution. Choosing the Stein operator $\mathcal{S}_p = I - P_\theta$ and the Stein set $\mathcal{G} = \{f \in \Gamma(\mathbb{R}) : \|f\|_\infty \le 1\}$. Note that, $g \in \arg\sup_{g \in G} |\mathbb{Q}(\mathcal{S}_p[g])|$ will satisfy $g(i) = \mathrm{sgn}(q^\top (I - P_\theta)_i)$, and the resulting Stein discrepancy is the minimum probability flow loss objective function [53]:

$$\mathrm{MPFL}(\mathbb{Q}\|\mathbb{P}) = \sum_y \left|((I - P_\theta)^\top q)_y\right| = \sum_{y \notin \{x_i\}_{i=1}^n} \left| \frac{1}{n} \sum_{x \in \{x_i\}_{i=1}^n} (I - P_\theta)_{xy} \right|.$$

2.4 Implementing Minimum SD Estimators: Stochastic Riemannian Gradient Descent

In order to implement the minimum SD estimators, we propose to use a stochastic gradient descent (SGD) algorithm associated to the information geometry induced by the SD on the parameter space. More precisely, consider a parametric family $\mathcal{P}_\Theta$ of probability measures on $\mathcal{X}$ with $\Theta \subset \mathbb{R}^m$. Given a discrepancy $D : \mathcal{P}_\Theta \times \mathcal{P}_\Theta \to \mathbb{R}$ satisfying $D(\mathbb{P}_\alpha\|\mathbb{P}_\theta) = 0$ iff $\mathbb{P}_\alpha = \mathbb{P}_\theta$ (called a statistical divergence), its associated information tensor on Θ is defined as the map $\theta \mapsto g(\theta)$, where $g(\theta)$ is the symmetric bilinear form $g(\theta)_{ij} = -\frac{1}{2}(\partial^2/\partial\alpha^i\partial\theta^j) D(\mathbb{P}_\alpha\|\mathbb{P}_\theta)|_{\alpha=\theta}$ [2]. When g is positive definite, we can use it to perform (Riemannian) gradient descent on the parameter space Θ. Under conditions stated in Proposition 4, DKSD is a statistical divergence. We provide below its metric tensor:

Proposition 3 (Information Tensor DKSD). *Assume the conditions of Proposition 4 hold. The information associated to DKSD is positive semi-definite and has components*

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

See Appendix C for a proof of this result, and the corresponding expression and proof for DSM (which extends the result for SM of [35]). Given an (information) Riemannian metric, recall the gradient flow of a curve θ on the Riemannian manifold Θ is the solution to $\dot{\theta}(t) = -\nabla_{\theta(t)} \mathrm{SD}(\mathbb{Q}\|\mathbb{P}_\theta)$, where ∇_θ denotes the Riemannian gradient at θ. It is the curve that follows the direction of steepest decrease (measured with respect to the Riemannian metric) of the function $\mathrm{SD}(\mathbb{Q}\|\mathbb{P}_\theta)$ (see Appendix A.5).

The well-studied natural gradient descent [1, 2] corresponds to the case in which the Riemannian manifold is $\Theta = \mathbb{R}^m$ equipped with the Fisher metric and SD is replaced by KL. When Θ is a linear manifold with coordinates (θ^i) we have $\nabla_\theta \operatorname{SD}(\mathbb{Q}\|\mathbb{P}_\theta) = g(\theta)^{-1}\mathrm{d}_\theta \operatorname{SD}(\mathbb{Q}\|\mathbb{P}_\theta)$, where $\mathrm{d}_\theta f$ denotes the tuple $(\partial_{\theta^i} f)$, which we will approximate at step t of the descent using the biased estimator $\hat{g}_{\theta_t}(\{X_i^t\}_i)^{-1}\mathrm{d}_{\theta_t}\widehat{\operatorname{SD}}(\{X_i^t\}_{i=1}^n\|\mathbb{P}_\theta)$, where $\hat{g}_{\theta_t}(\{X_i^t\}_{i=1}^n)$ is an unbiased estimator for the information matrix $g(\theta_t)$ and $\{X_i^t \sim \mathbb{Q}\}_i$ is a sample at step t. Given a sequence (γ_t) of step sizes we will approximate the gradient flow with

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \gamma_t \hat{g}_{\theta_t}(\{X_i^t\}_{i=1}^n)^{-1}\mathrm{d}_{\theta_t}\widehat{\operatorname{SD}}(\{X_i^t\}_{i=1}^n\|\mathbb{P}_\theta).$$

3 Theoretical Properties for Minimum Stein Discrepancy Estimators

In this section we show that the minimum $\operatorname{DKSD}_{K,m}$ estimators have many desirable properties. We begin by establishing strong consistency under simple assumptions, $\hat{\theta}_n^{\mathrm{DKSD}} \xrightarrow{a.s.} \theta^* \equiv \operatorname{argmin}_{\theta\in\Theta}\operatorname{DKSD}_{K,m}(\mathbb{Q},\mathbb{P}_\theta)^2$. We will assume we are in the specified setting, $\mathbb{Q} = \mathbb{P}_{\theta^*} \in \mathcal{P}_\Theta$. In the misspecified setting will need to also assume the existence of a unique minimiser. The derivations of the results are given in Appendix D.

Theorem 4 (Strong Consistency of DKSD). *Let $\mathcal{X} = \mathbb{R}^d$, $\Theta \subset \mathbb{R}^m$. Suppose that K is bounded with bounded derivatives up to order 2, that $k^0(x,y)$ is continuously-differentiable on an $\mathbb{R}^m$-open neighbourhood of Θ, and that for any compact subset $C \subset \Theta$ there exist functions f_1, f_2, g_1, g_2 s.t.*

1. $\left\|m^\top(x)\nabla\log p_\theta(x)\right\| \leq f_1(x)$, *where $f_1 \in L^1(\mathbb{Q})$ and continuous,*
2. $\left\|\nabla_\theta\big(m(x)^\top\nabla\log p_\theta(x)\big)\right\| \leq g_1(x)$, *where $g_1 \in L^1(\mathbb{Q})$ is continuous,*
3. $\|m(x)\| + \|\nabla_x m(x)\| \leq f_2(x)$ *where $f_2 \in L^1(\mathbb{Q})$ and continuous,*
4. $\|\nabla_\theta m(x)\| + \|\nabla_\theta\nabla_x m(x)\| \leq g_2(x)$ *where $g_2 \in L^1(\mathbb{Q})$ is continuous.*

Assume further that $\theta \mapsto \mathbb{P}_\theta$ is injective. Then we have a unique minimiser θ^*, and if either Θ is compact, or $\theta^* \in int(\Theta)$ and Θ and $\theta \mapsto \widehat{\operatorname{DKSD}}_{K,m}(\{X_i\}_{i=1}^n,\mathbb{P}_\theta)^2$ are convex, then $\hat{\theta}_n^{\mathrm{DKSD}}$ is strongly consistent.

Given consistency of the estimators, we now characterise their oscillations around θ^*.

Theorem 5 (Central Limit Theorem for DKSD). *Let $\mathcal{X}$ and Θ be open subsets of $\mathbb{R}^d$ and $\mathbb{R}^m$ respectively. Let K be a bounded kernel with bounded derivatives up to order 2 and suppose that $\hat{\theta}_n^{\mathrm{DKSD}} \xrightarrow{p} \theta^*$ and that there exists a compact neighbourhood $\mathcal{N} \subset \Theta$ of θ^* such that $\theta \to \widehat{\operatorname{DKSD}}_{K,m}(\{X_i\}_{i=1}^n,\mathbb{P}_\theta)^2$ is twice continuously for $\theta \in \mathcal{N}$,*

1. $\|m^\top(x)\nabla\log p_\theta(x)\| + \|\nabla_\theta\big(m(x)^\top\nabla\log p_\theta(x)\big)\| \leq f_1(x)$,
2. $\|m(x)\| + \|\nabla_x m(x)\| + \|\nabla_\theta m(x)\| + \|\nabla_\theta\nabla_x m(x)\| \leq f_2(x)$,
3. $\|\nabla_\theta\nabla_\theta\big(m(x)^\top\nabla\log p_\theta(x)\big)\| + \|\nabla_\theta\nabla_\theta\nabla_\theta\big(m(x)^\top\nabla\log p_\theta(x)\big)\| \leq g_1(x)$,
4. $\|\nabla_\theta\nabla_\theta m(x)\| + \|\nabla_\theta\nabla_\theta\nabla_x m(x)\| + \|\nabla_\theta\nabla_\theta\nabla_\theta m(x)\| + \|\nabla_\theta\nabla_\theta\nabla_\theta\nabla_x m(x)\| \leq g_2(x)$,

where $f_1, f_2 \in L^2(\mathbb{Q})$,$g_1, g_2 \in L^1(\mathbb{Q})$ are continuous. Suppose also that the information tensor g is invertible at θ^*. Then

$$\sqrt{n}\Big(\hat{\theta}_n^{\mathrm{DKSD}} - \theta^*\Big) \xrightarrow{d} \mathcal{N}\big(0, g_{\mathrm{DKSD}}^{-1}(\theta^*)\Sigma g_{\mathrm{DKSD}}^{-1}(\theta^*)\big)$$

where $\Sigma = \int_{\mathcal{X}}\big(\int_{\mathcal{X}}\nabla_{\theta^*}k_\theta^0(x,y)\mathrm{d}\mathbb{Q}(y)\big)\otimes\big(\int_{\mathcal{X}}\nabla_{\theta^*}k_\theta^0(x,z)\mathrm{d}\mathbb{Q}(z)\big)\mathrm{d}\mathbb{Q}(x)$.

For both results, the assumptions on the kernel are satisfied by most kernels common in the literature, such as Gaussian, inverse-multiquadric (IMQ) and any Matérn kernels with smoothness greater than 2. Similarly, the assumptions on the model are very weak given that the diffusion tensor can be adapted to guarantee consistency and asymptotic normality.

In Appendix D.2 we also prove the consistency and asymptotic normality of DSM_m. In the important case in which the density p_θ lies in an exponential family, i.e. $p_\theta(x) \propto \exp(\langle \theta, T(x)\rangle_{\mathbb{R}^m} - c(\theta))\exp(b(x))$, (here $\theta \in \mathbb{R}^m$ and $T \in \Gamma(\mathbb{R}^m)$ is the sufficient statistic) the Stein kernel is a quadratic $k^0 = \theta^\top A\theta + v^\top \theta + c$. Note $\nabla_x \log p_\theta = \nabla_x b + \theta \cdot \nabla_x T$ and $\nabla_\theta \nabla_x \log p_\theta = \nabla_x T^\top$. If K is IPD with bounded derivative up to order 2, ∇T has linearly independent rows, m is invertible, and $\|\nabla T m\|, \|\nabla_x b\|\|m\|, \|\nabla_x m\| + \|m\| \in L^2(\mathbb{Q})$, then the sequence of Stein estimators is strongly consistent and asymptotically normal (see Appendix D.3).

3.1 Robustness of Diffusion Stein Discrepancies

A concept of importance to practical inference is robustness when subjected to corrupted data [29]. In this section we quantify the robustness of DKSD estimators in terms of their influence function, which can be interpreted as measuring the impact of an infinitesimal perturbation of a distribution $\mathbb{P}$ by a Dirac located at a point $z \in \mathcal{X}$ on the estimator. If $\theta_{\mathbb{Q}}$ denotes the unique minimum DKSD estimator for $\mathbb{Q}$, then the influence functions is given by $\mathrm{IF}(z,\mathbb{Q}) \equiv \partial_t \theta_{\mathbb{Q}_t}|_{t=0}$ 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 $\mathrm{IF}(z,\mathbb{Q})$ is bounded in z.

Proposition 6 (Robustness of DKSD estimators)**.** *Suppose that the map $\theta \to \mathbb{P}_\theta$ is injective, for $\theta \in \Theta$, then $\mathrm{IF}(z,\mathbb{P}_\theta) = g_{\mathrm{DKSD}}(\theta)^{-1}\int_X \nabla_\theta k^0(z,y)\mathrm{d}\mathbb{P}_\theta(y)$. Moreover, suppose that the kernel is bounded with bounded derivatives and that for $s_p = m^\top \nabla \log p_\theta$, we have $\|s_p\| + \|\nabla_\theta s_p\| \in L^1(\mathbb{Q})$, and $\|m\| + \|\nabla_x m\| + \|\nabla_\theta \nabla_x m\| \in L^1(\mathbb{Q})$. If $F(x,y) \to 0$ as $|y| \to \infty$ for all $x \in \mathcal{X}$, for each $F(x,y) = \|K(x,y)s_p(y)\|$, $\|K(x,y)\nabla_\theta s_p(y)\|$, $\|\nabla_x K(x,y)s_p(y)\|$, $\|\nabla_x K(x,y)\nabla_\theta s_p(y)\|$, $\|\nabla_y \nabla_x (K(x,y)m(y))\|$,$\|\nabla_y \nabla_x(K(x,y)\nabla_\theta m(y))\|$, then $\sup_{z\in\mathcal{X}}\|\mathrm{IF}(z,\mathbb{P}_\theta)\| < \infty$.*

The analogous results for DSM estimators can be found in Appendix E. Consider a Gaussian location model, i.e. $p_\theta \propto \exp(-\|x-\theta\|_2^2)$, for $\theta \in \mathbb{R}^d$. The Gaussian scalar kernel $k(x,y)$ satisfies the assumptions of Proposition 6 so that $\sup_z \|\mathrm{IF}(z,\mathbb{P}_\theta)\| < \infty$, even when $m = I$. The classical score matching estimator θ_{SM} for θ is the arithmetic mean $\int_{\mathcal{X}} x\mathrm{d}\mathbb{Q}(x)$, for which the corresponding influence function is $\mathrm{IF}(z,\mathbb{Q}) = z - \int_{\mathcal{X}} x\mathrm{d}\mathbb{Q}(x)$ which is unbounded with respect to z, and thus not robust. This clearly demonstrates the importance of carefully selecting a Stein class for use in minimum SD estimators.

On the other hand, introducing a spatially decaying diffusion matrix in DSM can induce robustness. To this end, consider the minimum DSM estimator with scalar diffusion coefficient m. Then $\theta_{\mathrm{DSM}} = \left(\int_{\mathcal{X}} m^2(x)\mathrm{d}\mathbb{Q}(x)\right)^{-1}\left(\int_{\mathcal{X}} m^2(x)x\mathrm{d}\mathbb{Q}(x) + \int_{\mathcal{X}} \nabla m^2(x)\mathrm{d}\mathbb{Q}(x)\right)$. A straightforward calculation yields that the associated influence function will be bounded if both $m(x)$ and $\|\nabla m(x)\|$ decay as $\|x\| \to \infty$. This clearly demonstrates another significant advantage provided by the flexibility of our family of diffusion SD, where the Stein operator also plays an important role.

4 Numerical Experiments

In this section, we explore advantages of DKSD and KSD over SM for two toy estimation problems. These examples demonstrate worrying breakpoints for SM, and highlight how these can be straightforwardly handled using DKSD. In all experiments, the kernel is fixed to an IMQ kernel $k(x,y;c,\beta) = (c^2 + \|x-y\|_2^2)^\beta$ with $c = 1.$ and $\beta = -0.5$.

4.1 Rough densities: the symmetric Bessel distributions

A major drawback of SM is the smoothness requirement on the target density. However, this can be remedied by choosing alternative Stein classes, as will be demonstrated below in the case of the symmetric multivariate Bessel distributions. Let $K_{s-d/2}$ denote the modified Bessel function of the second kind with parameter $s - d/2$, which is real-valued whenever the input is real and positive. This distribution is a generalization of the Laplace distribution [38]

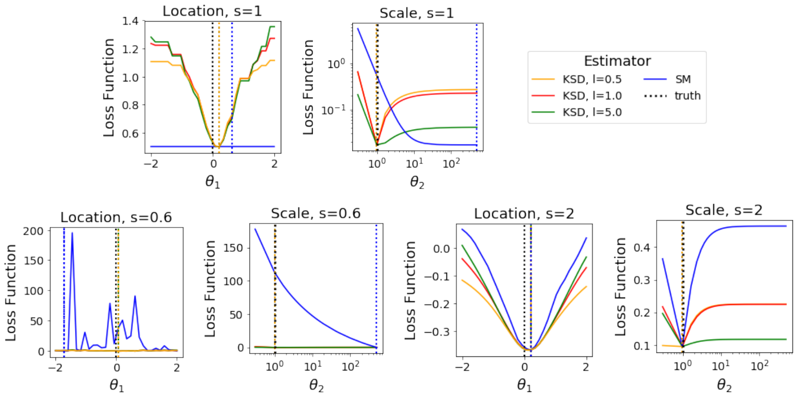

Figure 1: *Minimum SD Estimators for the Symmetric Bessel Distribution.* We consider the case where $\theta_1^* = 0$ and $\theta_2^* = 1$ and $n = 100$ realisations are available from the distribution for a range of smoothness parameter values s in $d = 1$.

and has log-density: $p_\theta(x) \propto (\|x - \theta_1\|_2/\theta_2)^{(s-d/2)} K_{s-d/2}(\|x - \theta_1\|_2/\theta_2)$ where $\theta = (\theta_1, \theta_2)$ consists of location parameter $\theta_1 \in \mathbb{R}^d$ and a scale parameter $\theta_2 > 0$. The parameter $s \geq \frac{d}{2}$ encodes smoothness.

We compared SM with KSD based on a Gaussian kernel and a range of lengthscale values in Fig. 1. These results are based on $n = 100$ IID realisations in $d = 1$. The case $s = 1$ corresponds to a Laplace distribution, and we notice that although both SM and KSD are able to obtain a reasonable estimate of the location parameter θ_1, SM is not able to recover the scale parameter θ_2. For rougher values, for example $s = 0.6$, we notice that the same behaviour of SM also occurs for the location parameter, even though KSD is still able to recover it. Finally, when $s = 2$, SM and KSD are both able to recover θ_1^* and θ_2^* up to some error due to the finite number of data points available.

4.2 Heavy-tailed distributions: the non-standardised student-t distribution

A second drawback of standard SM is that it is inefficient for heavy-tailed distributions. To demonstrate this, we focus on the following family of non-standardised student-t distributions: $p_\theta(x) \propto (1/\theta_2)(1 + (1/\nu)(\|x - \theta_1\|_2/\theta_2)^2)^{-(\nu+1)/2}$. Once again, $\theta = (\theta_1, \theta_2)$, where θ_1 is a location parameter and θ_2 a scale parameter. Furthemore, ν is an additional parameter determining the degree's of freedom. When $\nu = 1$, this correspond to the Cauchy distribution, whereas $\nu = \infty$ gives the Gaussian distribution. For small values of ν, the student-t distribution is heavy-tailed.

We illustrate SM and KSD for $\nu = 5$ with $(\theta_1^*, \theta_2^*) = (25, 10)$ in Figure 2. This choice of ν is large enough so that the first two moments exist, but also guarantees that the distribution is heavy-tailed. As observed in the far left plot, both SM and KSD struggle to recover θ_1^* when $n = 100$, and the loss functions are far from convex in this case. However, DKSD with matrix $m_\theta(x) = 1 = \|x - \theta_1\|^2/\theta_2^2$ is able to obtain a very accurate estimate of θ_1. In the middle left plot, we reproduce the same experiment but for θ_2 with SM, KSD and their correponding non-negative version (NNSM & NNKSD), which are particularly well suited for scale parameters. However, DKSD with $m_\theta(x) = ((x - \theta_1)/\theta_2)(1 + (1/\nu)\|x - \theta_1\|_2^2/\theta_2^2)$ provides significant further gains. On the right-hand side, we also consider the advantage of the Riemannian SGD algorithm over SGD for this same experiment by illustrating both methods on the KSD loss function, but with $n = 1000$. Both algorithms use constant stepsizes optimised for the experiment and minibatches of size 50. As demonstrated, for both θ_1 and θ_2, Riemmannian SGD converges within a few dozen iterations, whereas SGD hasn't converged after 1000 iterations.

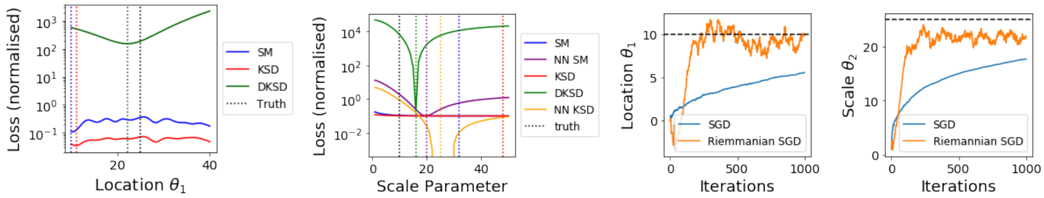

Figure 2: *Minimum SD Estimators for Non-standardised Student-t Distributions.* We consider several minimum Stein discrepancy estimators for a student-t problem with $\nu = 5, \theta_1^* = 25, \theta_2^* = 10$.

5 Conclusion

This paper introduced a general approach for constructing minimum distance estimators based on Stein's method, and demonstrated that many popular inference schemes can be recovered as special cases, including SM [32, 33], contrastive divergence [26] and minimum probability flow [53]. This class of algorithms gives us additional flexibility through the choice of an operator and function space (the Stein operator and Stein class), which can be used to tailor the inference scheme to the model at hand, and we illustrated this through simple examples including distributions with heavy tails or rough densities for which SM breaks down.

However, this paper only scratches the surface of what is possible with minimum SD estimators. Looking ahead, it will be interesting to identify diffusion tensors which increase efficiency for important classes of problems in machine learning. One example on which we foresee progress are the product of student-t experts models [36, 57, 59], whose heavy tails render estimation challenging for SM. Advantages could also be found for other energy models, such as large graphical models where the kernel could be adapted to the graph [58].

Acknowledgments

AB was supported by a Roth scholarship from the Department of Mathematics at Imperial College London. FXB was supported by the EPSRC grants [EP/L016710/1, EP/R018413/1]. AD and MG were supported by the Lloyds Register Foundation Programme on Data-Centric Engineering, the UKRI Strategic Priorities Fund under the EPSRC Grant [EP/T001569/1] and the Alan Turing Institute under the EPSRC grant [EP/N510129/1]. MG was supported by the EPSRC grants [EP/J016934/3, EP/K034154/1, EP/P020720/1, EP/R018413/1].

References

[1] S.-I. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2): 251–276, 1998.

[2] S.-I. Amari. *Information Geometry and Its Applications*, volume 194. Springer, 2016.

[3] A. Barbour and L. H. Y. Chen. *An introduction to Stein's method.* Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore, 2005.

[4] A. Basu, H. Shioya, and C. Park. *Statistical Inference: The Minimum Distance Approach.* CRC Press, 2011.

[5] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Springer Science+Business Media, New York, 2004.

[6] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert. Approximate Bayesian computation with the Wasserstein distance. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 81(2):235–269, 2019.

[7] S. Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

[8] F.-X. Briol, A. Barp, A. B. Duncan, and M. Girolami. Statistical inference for generative models with maximum mean discrepancy. 2019.

[9] G. Casella and R. Berger. *Statistical Inference*. 2001.

[10] C. Ceylan and M. U. Gutmann. Conditional noise-contrastive estimation of unnormalised models. *arXiv preprint arXiv:1806.03664*, 2018.

[11] L. H. Y. Chen, L. Goldstein, and Q.-M. Shao. *Normal Approximation by Stein's Method*. Springer, 2011.

[12] W. Y. Chen, L. Mackey, J. Gorham, F.-X. Briol, and C. J. Oates. Stein points. In *Proceedings of the International Conference on Machine Learning, PMLR 80:843-852*, 2018.

[13] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, pages 2606–2615, 2016.

[14] A. P. Dawid, M. Musio, and L. Ventura. Minimum scoring rule inference. *Scandinavian Journal of Statistics*, 43(1):123–138, 2016.

[15] G. Detommaso, T. Cui, Y. Marzouk, A. Spantini, and R. Scheichl. A stein variational newton method. In *Advances in Neural Information Processing Systems 31*, pages 9169–9179. 2018.

[16] G. K. Dziugaite, D. M. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence*, 2015.

[17] C. Frogner, C. Zhang, H. Mobahi, M. Araya-Polo, and T. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

[18] D. Gabay. Minimizing a differentiable function over a differential manifold. *Journal of Optimization Theory and Applications*, 37(2):177–219, 1982.

[19] A. Genevay, G. Peyré, and M. Cuturi. Learning generative models with Sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics, PMLR 84*, pages 1608–1617, 2018.

[20] C. J. Geyer. On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):261–274, 1994.

[21] J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234, 2015.

[22] J. Gorham and L. Mackey. Measuring sample quality with kernels. In *Proceedings of the International Conference on Machine Learning*, pages 1292–1301, 2017.

[23] J. Gorham, A. Duncan, L. Mackey, and S. Vollmer. Measuring sample quality with diffusions. *arXiv:1506.03039. To appear in Annals of Applied Probability.*, 2016.

[24] M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[25] M. U. Gutmann and A. Hyvarinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13:307–361, 2012.

[26] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[27] W. Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325, 1948.

[28] W. Hoeffding. The strong law of large numbers for U-statistics. Technical report, North Carolina State University Department of Statistics, 1961.

[29] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley, 2009.

[30] J. Huggins and L. Mackey. Random feature stein discrepancies. In *Advances in Neural Information Processing Systems*, pages 1899–1909, 2018.

[31] A. Hyvärinen. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural computation*, 11(7):1739–1768, 1999.

[32] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–708, 2006.

[33] A. Hyvärinen. Some extensions of score matching. *Computational Statistics and Data Analysis*, 51(5):2499–2512, 2007.

[34] W. Jitkrittum, W. Xu, Z. Szabo, K. Fukumizu, and A. Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pages 261–270, 2017.

[35] R. Karakida, M. Okada, and S.-I. Amari. Adaptive natural gradient learning algorithms for unnormalized statistical models. *Artificial Neural Networks and Machine Learning - ICANN*, 2016.

[36] D. P. Kingma and Y. LeCun. Regularized estimation of image statistics by score matching. In *Advances in Neural Information Processing Systems*, pages 1126–1134, 2010.

[37] U. Köster and A. Hyvärinen. A two-layer model of natural stimuli estimated with score matching. *Neural Computation*, 22(9):2308–2333, 2010.

[38] S. Kotz, T. J. Kozubowski, and K. Podgorski. *The Laplace Distribution and Generalizations*. Springer, 2001.

[39] Y. Li, K. Swersky, and R. Zemel. Generative moment matching networks. In *Proceedings of the International Conference on Machine Learning*, volume 37, pages 1718–1727, 2015.

[40] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, 2016.

[41] Q. Liu and D. Wang. Learning deep energy models: Contrastive divergence vs. amortized mle. *arXiv preprint arXiv:1707.00797*, 2017.

[42] Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the International Conference on Machine Learning*, pages 276–284, 2016.

[43] S. Lyu. Interpretation and generalization of score matching. In *Conference on Uncertainty in Artificial Intelligence*, pages 359–366, 2009.

[44] L. Mackey and J. Gorham. Multivariate Stein factors for a class of strongly log-concave distributions. *Electronic Communications in Probability*, 21, 2016.

[45] K. V. Mardia, J. T. Kent, and A. K. Laha. Score matching estimators for directional distributions. *arXiv preprint arXiv:1604.08470*, 2016.

[46] A. Mnih and Y. W. Teh. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the International Conference on Machine Learning*, pages 419–426, 2012.

[47] A. Muller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[48] W. K. Newey and D. McFadden. Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4:2111–2245, 1994.

[49] C. J. Oates, M. Girolami, and N. Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society B: Statistical Methodology*, 79(3):695–718, 2017.

[50] L. Pardo. *Statistical Inference Based on Divergence Measures*, volume 170. Chapman and Hall/CRC, 2005.

[51] S. Pigola and A. G. Setti. Global divergence theorems in nonlinear PDEs and geometry. *Ensaios Matemáticos*, 26:1–77, 2014.

[52] S. Roth and M. J. Black. Fields of experts. *International Journal of Computer Vision*, 82(2):205, 2009.

[53] J. Sohl-dickstein, P. Battaglino, and M. R. DeWeese. Minimum probability flow learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 905–912, 2011.

[54] B. Sriperumbudur, K. Fukumizu, A. Gretton, A. Hyvärinen, and R. Kumar. Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(1):1830–1888, 2017.

[55] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

[56] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of 6th Berkeley Symposium on Mathematical Statistics and Probability*, pages 583–602. University of California Press, 1972.

[57] K. Swersky, M. A. Ranzato, D. Buchman, B. M. Marlin, and N. de Freitas. On autoencoders and score matching for energy based models. In *International Conference on Machine Learning*, pages 1201–1208, 2011.

[58] S. V. N. Vishwanathan, N. Schraudolph, R. Kondor, and K. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, pages 1201–1242, 2010.

[59] M. Welling, G. Hinton, and S. Osindero. Learning sparse topographic representations with products of student-t distributions. In *Advances in Neural Information Processing Systems*, pages 1383–1390, 2003.

[60] L. Wenliang, D. Sutherland, H. Strathmann, and A. Gretton. Learning deep kernels for exponential family densities. *arXiv:1811.08357*, 2018.

[61] I.-K. Yeo and R. A. Johnson. A uniform strong law of large numbers for U-statistics with application to transforming to near symmetry. *Statistics & Probability Letters*, 51 (1):63–69, 2001.

Supplementary Material

This document provides additional details for the paper "Minimum Stein Discrepancy Estimators". Appendix A contains background technical material required to understand the paper, Appendix B derives the minimum SD estimators from first principles and Appendix C derives the information metrics for DKSD and DSM. Appendix D contains proof of all asymptotic results including consistency and central limit theorems for DKSD and DSM, whilst Appendix E discusses their robustness.

Our derivations will use standard operators from vector calculus which we summarise in Appendix A.1. We will additionally introduce the following notation. We write $f \lesssim g$ if there is a constant $C > 0$ for which $f(x) \leq Cg(x)$ for all x. We set $\mathbb{Q}f \equiv \int f \mathrm{d}\mathbb{Q}$ and use $\Gamma(\mathcal{W}, \mathcal{Y})$ for the set of maps $\mathcal{W} \to \mathcal{Y}$ when $\mathcal{W} \neq \mathcal{X}$.

A Background Material

In this section, we provide background material which is necessary to follow the proofs in the following sections. This includes background in vector calculus, stochastic optimisation over manifolds and vector-valued reproducing kernel Hilbert spaces.

A.1 Background on Vector Calculus

The following section contains background and important identities from vector calculus. For a function $g \in \Gamma(\mathcal{X}, \mathbb{R})$, $v \in \Gamma(\mathcal{X}, \mathbb{R}^d)$ and $A \in \Gamma(\mathcal{X}, \mathbb{R}^{d\times d})$ with components A_{ij}, v_i, g, we have $(\nabla g)_i = \partial_i g$, $(v \cdot A)_i = v_j A_{ji} = (v^\top A)_i$, $(\nabla \cdot A)_i = \partial_j A_{ji}$ which must be interpreted as the components of row-vectors; $(Av)_i = A_{ij} v_j$ which are the components of a column vector. Moreover $(\nabla v)_{ij} = \partial_j v_i$, $\nabla^2 f \equiv \nabla(\nabla f)$, $A : B \equiv \langle A, B\rangle = \mathrm{Tr}(A^\top B) = A_{ij}B_{ij}$. We have the following identities (where in the last equality we treat $\nabla \cdot A$ and ∇g as column vectors)

$$\begin{aligned}
\nabla \cdot (gv) &= \partial_i(gv_i) = v_i \partial_i g + g\partial_i v_i = (\nabla g)v + g\nabla \cdot v = \nabla g \cdot v + g\nabla \cdot v, \\
\nabla \cdot (gA) &= \partial_i(gA_{ij})e_j = (A_{ij}\partial_i g + g\partial_i A_{ij})e_j = \nabla g \cdot A + g\nabla \cdot A = \nabla g^\top A + g\nabla \cdot A, \\
\nabla \cdot (Av) &= \partial_i(A_{ij}v_j) = (\nabla \cdot A)v + \mathrm{Tr}[A\nabla v] = (\nabla \cdot A) \cdot v + \mathrm{Tr}[A\nabla v].
\end{aligned}$$

A.2 Background on Norms

For $F \in \Gamma(\mathcal{X}, \mathbb{R}^{n_1\times n_2})$ we set $\|F\|_p^p \equiv \int \|F(x)\|_p^p \mathrm{d}\mathbb{Q}(x)$, where $\|F(x)\|_p$ is the vector p-norm on $\mathbb{R}^{n_1\times n_2}$ when $n_2 = 1$, else it is the induced operator norm. If $v \in \Gamma(\mathcal{X}, \mathbb{R}^{n_1})$, then $\|v\|_p^p = \int \|v(x)\|_p^p \mathrm{d}x = \int \sum_i |v_i(x)|^p \mathrm{d}x = \sum_i \|v_i\|_p^p$, hence $v \in L_p(\mathbb{Q})$ iff $v_i \in L_p(\mathbb{Q})$ for all i, and similarly $F \in L_p(\mathbb{Q})$ iff $F_{ij} \in L_p(\mathbb{Q})$ for all i, j since the induced norm $\|F(x)\|_p$ and the vector norm $\|F\|_{vec}^p \equiv \sum_{ij} |F_{ij}(x)|^p$ are equivalent.

A.3 Background on Vector-valued RKHS

A Hilbert space $\mathcal{H}$ of functions $\mathcal{X} \to \mathbb{R}^d$ is a RKHS if $\|f(x)\|_{\mathbb{R}^d} \leq C_x \|f\|_{\mathcal{H}}$. It follows that the evaluation "functional" $\delta_x : \mathcal{H} \to \mathbb{R}^d$ is continuous, for any x. Moreover for any $x \in \mathcal{X}, v \in \mathbb{R}^d$, the linear map $f \mapsto v \cdot f(x)$ is cts. By the Riesz representation theorem, there exists $K_x v \in \mathcal{H}$ s.t. $v \cdot f(x) = \langle K_x v, f\rangle$. From this we see that $K_x v$ is linear in v (turns out linear combinations of $K_{x_i} v_i$ are dense in $\mathcal{H}$), and $K_x^* = \delta_x$. We define $K : \mathcal{X} \times \mathcal{X} \to \mathrm{End}(\mathbb{R}^d)$ by

$$K(x, y)v \equiv (K_y v)(x) = \delta_x \delta_y^* v.$$

It follows that $K(x, y) = K(y, x)^*$ and $u \cdot K(x, y)v = \langle K_y v, K_x u\rangle$. Denote by e_i the i^{th} vector in the standard basis of $\mathbb{R}^d$. From this we can get the components of the matrix:

$$(K(x, y))_{ij} = \langle K_x e_i, K_y e_j\rangle.$$

We have for any v_i, x_j, $\sum_{j,k} v_j \cdot K(x_j, x_k)v_k \geq 0$.

A.4 Background on Separable Kernels

Consider the d dimensional product space $\mathcal{H}^d$ of function $f : \mathcal{X} \to \mathbb{R}^d$ with components $f_i \in \mathcal{H}_i$ and $\mathcal{H}_i$ is a RKHS with kernel C^2 kernel $k^i : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Let $K : \mathcal{X} \times \mathcal{X} \to \mathrm{End}(\mathbb{R}^d) \cong \mathbb{R}^{d\times d}$ be the kernel of $\mathcal{H}^d$ (see Appendix A.3). Note if $K_x \equiv K(x,\cdot) : \mathcal{X} \to \mathrm{End}(\mathbb{R}^d)$, and if $v \in \mathbb{R}^d$, then $K_x v \in \mathcal{H}^d$. The reproducing property then states that $\forall f \in \mathcal{H}^d$: $\langle f(x), v\rangle_{\mathbb{R}^d} = \langle f, K(\cdot, x)v\rangle_{\mathcal{H}^d}$. Moreover for the kernel $K = \mathrm{diag}(\lambda_1 k^1, \ldots, \lambda_d k^d)$ we will prove below that $\langle f, g\rangle_{\mathcal{H}^d} = \frac{1}{\lambda_i}\sum_i \langle f_i, g_i\rangle_{\mathcal{H}_i}$, whereas for $K = Bk$ where B is symmetric and invertible we should have $\langle f, g\rangle_{\mathcal{H}^d} = \sum_{ij} B^{-1}_{ij}\langle f_i, g_j\rangle_{\mathcal{H}}$.

Given a real-valued kernel k_i on $\mathcal{X}$, consider $K = \mathrm{diag}(\lambda_1 k_1, \ldots, \lambda_n k_n)$. Let $f = \sum_j \delta^*_{x_j} v_j$. Recall this is a dense subset of $\mathcal{H}^d$: we will derive the RKHS norm for this dense subset and by continuity this will hold for any function. Given the norm, the formula for the inner product will follow by the polarization identity. We have

$$\begin{aligned} f_i(x) &= \delta_x(f)\cdot e_i = \delta_x\delta^*_{x_j} v_j \cdot e_i = K(x, x_j)v_j \cdot e_i \\ &= \mathrm{diag}(\lambda_1 k_1, \ldots, \lambda_n k_n)(x, x_j)v_j \cdot e_i = \lambda_i k_i(x, x_j)v^i_j \end{aligned}$$

$$\|f\|^2_{\mathcal{H}_K} = \langle \delta^*_{x_j} v_j, \delta^*_{x_l} v_l\rangle_{\mathcal{H}_K} = v_j \cdot K(x_j, x_l)v_l = v^i_j \lambda_i k_i(x_j, x_l)v^i_l$$

On the other hand, $\sum_i \frac{1}{\lambda_i}\langle f_i, f_i\rangle_{k_i} = \sum_i \frac{1}{\lambda_i}\lambda_i^2 v^i_j v^i_l k_i(x_j, x_l)$. Thus $\|f\|^2_{\mathcal{H}_K} = \frac{1}{\lambda_i}\sum_i \langle f_i, f_i\rangle_{k_i}$.

For a symmetric positive definite matrix B, consider the kernel on $\mathcal{H}$ $K(x,y) \equiv k(x,y)B$. Let $f = \sum_j \delta^*_{x_j} v_j$. We have:

$$f_i(x) = \delta_x(f)\cdot e_i = \delta_x\delta^*_{x_j} v_j \cdot e_i = K(x, x_j)v_j \cdot e_i = Bv_j \cdot e_i k_{x_j}(x)$$

This implies $f_i \in \mathcal{H}_k$. Then

$$\|f\|^2_{\mathcal{H}_K} = \langle \delta^*_{x_j} v_j, \delta^*_{x_l} v_l\rangle_{\mathcal{H}_K} = v_j \cdot K(x_j, x_l)v_l = k(x_j, x_l)v_j \cdot Bv_l.$$

On the other hand $\langle f_i, f_j\rangle_k = e_i^\top B v_r e_j^\top B v_s k(x_s, x_r)$. Notice

$$B^{-1}_{ij} e_i^\top B v_r = B^{-1}_{ij} B_{il} v^l_r = \delta_{lj} v^l_r = v^j_r.$$

So we have:

$$B^{-1}_{ij}\langle f_i, f_j\rangle_k = v^j_r e_j^\top B v_s k(x_s, x_r) = v^j_r B_{ja} v^a_s k(x_s, x_r) = v_r \cdot B v_s k(x_s, x_r)$$

A.5 Background on Stochastic Optimisation on Riemannian Manifolds

The gradient flow of a curve θ on a complete connected Riemannian manifold Θ (for example a Hilbert space) is the solution to $\dot\theta(t) = -\nabla_{\theta(t)}\mathrm{SD}(\mathbb{Q}\|\mathbb{P}_\theta)$, where ∇_θ is the Riemannian gradient at θ. Typically [1] the gradient flow is approximated by the update equation $\theta(t+1) = \exp_{\theta(t)}(-\gamma_t H(Z_t, \theta))$ where exp is the Riemannian exponential map, (γ_t) is a sequence of step sizes with $\sum \gamma_t^2 < \infty$, $\sum \gamma_t = +\infty$, and H is an unbiased estimator of the loss gradient, $\mathbb{E}[H(Z_t, \theta)] = \nabla_\theta \mathrm{SD}(\mathbb{Q}\|\mathbb{P}_\theta)$. When the Riemannian exponential is computationally expensive, it is convenient to replace it by a retration $\mathcal{R}$, that is a first-order approximation which stays on the manifold. This leads to the update $\theta(t+1) = \mathcal{R}_{\theta(t)}(-\gamma_t H(Z_t, \theta))$ [7]. When Θ is a linear manifold it is common to take $\mathcal{R}_{\theta(t)}(-\gamma_t H(Z_t, \theta)) \equiv \theta(t) - \gamma_t H(Z_t, \theta(t))$. In local coordinates (θ^i) we have $\nabla_\theta \mathrm{SD}(\mathbb{Q}\|\mathbb{P}_\theta) = g(\theta)^{-1}\mathrm{d}_\theta \mathrm{SD}(\mathbb{Q}\|\mathbb{P}_\theta)$, where $\mathrm{d}_\theta f$ denotes the tuple $(\partial_{\theta^i} f)$, which we will approximate using the biased estimator $H(\{X^t_i\}_i, \theta) \equiv \hat g_{\theta(t)}(\{X^t_i\}^n_{i=1})^{-1}\mathrm{d}_\theta\widehat{\mathrm{SD}}(\{X^t_i\}^n_{i=1}\|\mathbb{P}_\theta)$, where $\hat g_{\theta(t)}(\{X^t_i\}^n_{i=1})$ is an unbiased estimator for the information matrix $g(\theta(t))$ using a sample $\{X^t_i\}^n_{i=1} \sim \mathbb{Q}$. We thus obtain the following Riemannian gradient descent algorithm

$$\theta(t+1) = \theta(t) - \gamma_t \hat g_{\theta(t)}(\{X^t_i\}^n_{i=1})^{-1}\mathrm{d}_{\theta(t)}\widehat{\mathrm{SD}}(\{X^t_i\}^n_{i=1}\|\mathbb{P}_\theta).$$

When $\Theta = \mathbb{R}^m$, $\gamma_t = \frac{1}{t}$, g is the Fisher metric and $\widehat{\mathrm{SD}}(\{X^t_i\}^n_{i=1}\|\mathbb{P}_\theta)$ is replaced by $\widehat{\mathrm{KL}}(\{X^t_i\}^n_{i=1}\|\mathbb{P}_\theta)$ this recovers the natural gradient descent algorithm [1].

[1] See sec 4.4 [18] for Riemannian Newton method

B Derivation of Diffusion Stein Discrepancies

In this appendix, we carefully derive the diffusion SD studied in this paper. We begin by providing details on the diffusion Stein operator, then move on to the DKSD and DSM divergences and corresponding estimators.

For either matrix kernels introduced in Appendix A.4, we will show in Appendix B.1 that $\forall f \in \mathcal{H}^d$: $\mathcal{S}_p^m[f](x) = \langle \mathcal{S}_p^{m,1} K_x, f\rangle_{\mathcal{H}^d}$. In Appendix B.2 we prove that if $x \mapsto \|\mathcal{S}_p^{m,1} K_x\|_{\mathcal{H}^d} \in L^1(\mathbb{Q})$, then

$$\mathrm{DKSD}_{K,m}(\mathbb{Q}\|\mathbb{P})^2 \;\equiv\; \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\leq 1}} \left|\int_{\mathcal{X}} \mathcal{S}_p^m[h]\mathrm{d}\mathbb{Q}\right|^2 = \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{S}_p^{m,2}\mathcal{S}_p^{m,1} K(x,y)\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y).$$

In Appendix B.3 we further show the Stein kernel satisfies

$$k^0(x,y) \equiv \mathcal{S}_p^{m,2}\mathcal{S}_p^{m,1} K(x,y) = \tfrac{1}{p(y)p(x)} \nabla_y \cdot \nabla_x \cdot \big(p(x)m(x)K(x,y)m(y)^\top p(y)\big).$$

B.1 Stein Operator

By definition for $f \in \Gamma\big(\mathcal{X},\mathbb{R}^d\big)$ and $A \in \Gamma\big(\mathcal{X},\mathbb{R}^{d\times d}\big)$

$$\begin{aligned}\mathcal{S}_p[f] &= \tfrac{1}{p}\nabla\cdot(pmf) = m^\top \nabla \log p \cdot f + \nabla\cdot(mf),\\ \mathcal{S}_p[A] &= \tfrac{1}{p}\nabla\cdot(pmA) = m^\top \nabla \log p \cdot A + \nabla\cdot(mA)\end{aligned}$$

which are operators $\Gamma\big(\mathcal{X},\mathbb{R}^d\big) \to \Gamma\big(\mathcal{X},\mathbb{R}\big)$ and $\Gamma\big(\mathcal{X},\mathbb{R}^{d\times d}\big) \to \Gamma\big(\mathcal{X},\mathbb{R}^d\big)$ respectively.

Proposition 7. *Let $\mathcal{X}$ be an open subset of $\mathbb{R}^d$, $K = \mathrm{diag}(\lambda_1 k^1, \ldots, \lambda_d k^d)$ or $K = Bk$. Suppose m and k^i are continuously differentiable. Then for any $f \in \mathcal{H}^d$*

$$\mathcal{S}_p[f](x) = \langle \mathcal{S}_p^1[K]|_x, f\rangle_{\mathcal{H}^d}$$

Proof

For any $f \in \mathcal{H}^d$

$$\begin{aligned}\langle f(x), m(x)^\top \nabla\log p(x)\rangle_{\mathbb{R}^d} &= \langle f, K(\cdot,x)m(x)^\top\nabla\log p(x)\rangle_{\mathcal{H}^d}\\ &= \langle f, K_x^\top m(x)^\top \nabla \log p(x)\rangle_{\mathcal{H}^d}\\ &= \langle f, m(x)^\top\nabla\log p(x)\cdot K_x\rangle_{\mathcal{H}^d}.\end{aligned}$$

Moreover if $K = Bk$, then $\nabla\cdot(mBk) = B_{ji}\big(k\partial_r m_{rj} + m_{jr}^\top \partial_r k\big)e_i$ so that

$$\begin{aligned}\langle f, \nabla\cdot(mK)|_x\rangle_{\mathcal{H}^d} &= \langle f, B_{ji}\big(k\partial_r m_{rj} + m^\top_{jr}\partial_r k\big)|_x e_i\rangle_{\mathcal{H}^d}\\ &= B_{sl}^{-1}\langle f_s, B_{jl}\big(k_x\partial_r m_{rj}|_x + m^\top_{jr}(x)\partial_r k|_x\big)\rangle_{\mathcal{H}}\\ &= \delta_{js}\langle f_s, k_x\partial_r m_{rj}|_x + m^\top_{jr}(x)\partial_r k|_x\rangle_{\mathcal{H}}\\ &= \langle f_s, k_x\partial_r m_{rs}|_x + m^\top_{sr}(x)\partial_r k|_x\rangle_{\mathcal{H}}\\ &= \partial_r m_{rs}|_x\langle f_s, k_x\rangle_{\mathcal{H}} + m^\top_{sr}(x)\langle f_s, \partial_r k|_x\rangle_{\mathcal{H}}\\ &= \partial_r m_{rs}|_x f_s(x) + m^\top_{sr}(x)\partial_r f_s|_x\\ &= \nabla\cdot m|_x\cdot f + \mathrm{Tr}[m\nabla f]\\ &= \nabla\cdot(mf)|_x.\end{aligned}$$

Similarly if $K = \mathrm{diag}(\lambda_1 k^1,\ldots,\lambda_d k^d)$ and $B = \mathrm{diag}(\lambda_1,\ldots,\lambda_d)$, then $\nabla\cdot(mK) = \lambda_i\partial_s(k^i m_{si})e_i$ and hence

$$\begin{aligned}\langle f, \nabla\cdot(mK)|_x\rangle_{\mathcal{H}^d} &= \langle f, \lambda_i\partial_s(m_{si}k^i)|_x e_i\rangle_{\mathcal{H}^d}\\ &= \tfrac{1}{\lambda_i}\langle f_i, \lambda_i\partial_s(m_{si}k^i)|_x\rangle_{\mathcal{H}}\\ &= m_{si}(x)\langle f_i, \partial_s k^i|_x\rangle_{\mathcal{H}} + \partial_s m_{si}|_x\langle f_i, k^i_x\rangle_{\mathcal{H}}\\ &= m_{si}(x)\langle f_i, \partial_s k^i|_x\rangle_{\mathcal{H}} + \partial_s m_{si}|_x\langle f_i, k^i_x\rangle_{\mathcal{H}}\\ &= m_{si}(x)\partial_s f_i(x) + \partial_s m_{si} f_i(x)\\ &= \mathrm{Tr}[m(x)\nabla f|_x] + \nabla\cdot m|_x\cdot f(x)\\ &= \nabla\cdot(mf)|_x.\end{aligned}$$

Therefore, we conclude that $\mathcal{S}_p[f](x) = \langle \mathcal{S}_p^1 K_x, f\rangle_{\mathcal{H}^d}$ where $\mathcal{S}_p^1 K_x \equiv \mathcal{S}_p^1[K]|_x$ means applying $\mathcal{S}_p$ to the first entry of K and evaluate it x, so informally $\mathcal{S}_p^1[K]|_x : y \mapsto \frac{1}{p}\nabla_x \cdot (p(x)m(x)K(x,y))$. ■

B.2 Diffusion Kernel Stein Discrepancies

Proposition 8. *Suppose $\mathcal{S}_p[f](x) = \langle \mathcal{S}_p^1[K]|_x, f\rangle_{\mathcal{H}^d}$ for any $f \in \mathcal{H}^d$. Let m and K be C^2, and $x \mapsto \mathcal{S}_p K_x$ be $\mathbb{Q}$-Bochner integrable. Then*

$$\mathrm{DKSD}_{K,m}(\mathbb{Q},\mathbb{P})^2 = \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{S}_p^2\mathcal{S}_p^1 K(x,y)\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y).$$

Proof

Let us identify $\mathcal{H}_1 \otimes \mathcal{H}_2 \cong L(\mathcal{H}_1 \times \mathcal{H}_2, \mathbb{R}) \cong L(\mathcal{H}_2, \mathcal{H}_1)$ with $(v_1 \otimes v_2) \sim v_1\langle v_2, \cdot\rangle_{\mathcal{H}_2}$ (since $\mathcal{H}_2 \cong \mathcal{H}_2^*$), so that $(v_1 \otimes v_2)u_2 \equiv v_1\langle v_2, u_2\rangle_{\mathcal{H}_2}$ (here $L(V,W)$ is the space of linear maps from V to W). Then

$$\langle u_1 \otimes u_2, v_1 \otimes v_2\rangle_{HS} \equiv \langle u_1, v_1\rangle_{\mathcal{H}_1}\langle u_2, v_2\rangle_{\mathcal{H}_2} = \langle u_1, (v_1 \otimes u_2)v_2\rangle_{\mathcal{H}_1}.$$

For simplicity we will write $\mathcal{S}_p K_x \equiv \mathcal{S}_p^1[K]|_x$. Using the fact $x \mapsto \mathcal{S}_p K_x$ is $\mathbb{Q}$-Bochner integrable, then by Cauchy-Schwartz $x \mapsto \langle h, \mathcal{S}_p K_x\rangle_{\mathcal{H}^d}$ is $\mathbb{Q}$-integrable. Then

$$\begin{aligned}
\mathrm{DKSD}_{K,m}(\mathbb{Q},\mathbb{P})^2 &= \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\le 1}} \left\langle \int_{\mathcal{X}} \mathcal{S}_p[h](x)\mathrm{d}\mathbb{Q}(x), \int_{\mathcal{X}} \mathcal{S}_p[h](y)\mathrm{d}\mathbb{Q}(y)\right\rangle_{\mathbb{R}} \\
&= \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\le 1}} \int_{\mathcal{X}}\langle h, \mathcal{S}_p K_x\rangle_{\mathcal{H}^d}\mathrm{d}\mathbb{Q}(x) \int_{\mathcal{X}}\langle h, \mathcal{S}_p K_y\rangle_{\mathcal{H}^d}\mathrm{d}\mathbb{Q}(y) \\
&= \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\le 1}} \int_{\mathcal{X}}\int_{\mathcal{X}}\langle h, \mathcal{S}_p K_x\rangle_{\mathcal{H}^d}\langle h, \mathcal{S}_p K_y\rangle_{\mathcal{H}^d}\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) \\
&= \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\le 1}} \int_{\mathcal{X}}\int_{\mathcal{X}}\langle h, \mathcal{S}_p K_x \otimes \mathcal{S}_p K_y h\rangle_{\mathcal{H}^d}\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) \\
&= \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\le 1}} \int_{\mathcal{X}}\int_{\mathcal{X}}\langle h \otimes h, \mathcal{S}_p K_x \otimes \mathcal{S}_p K_y\rangle_{HS}\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y)
\end{aligned}$$

Moreover $\int_{\mathcal{X}} \|\mathcal{S}_p K_x \otimes \mathcal{S}_p K_y\|_{HS}\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) < \infty$, since

$$\begin{aligned}
&\int_{\mathcal{X}} \|\mathcal{S}_p K_x \otimes \mathcal{S}_p K_y\|_{HS}\mathrm{d}\mathbb{Q}(x) \otimes \mathrm{d}\mathbb{Q}(y) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \sqrt{\langle \mathcal{S}_p K_x, \mathcal{S}_p K_x\rangle_{\mathcal{H}^d}\langle \mathcal{S}_p K_y, \mathcal{S}_p K_y\rangle_{\mathcal{H}^d}}\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) \\
&= \left(\int_{\mathcal{X}} \sqrt{\langle \mathcal{S}_p K_x, \mathcal{S}_p K_x\rangle_{\mathcal{H}^d}}\mathrm{d}\mathbb{Q}(x)\right)^2 \\
&= \left(\int_{\mathcal{X}} \|\mathcal{S}_p K_x\|_{\mathcal{H}^d}\mathrm{d}\mathbb{Q}(x)\right)^2 < \infty
\end{aligned}$$

since by assumption $x \mapsto \mathcal{S}_p K_x$ is $\mathbb{Q}$-Bochner integrable. Thus

$$\begin{aligned}
\mathrm{DKSD}_{K,m}(\mathbb{Q},\mathbb{P})^2 &= \sup_{\substack{h\in\mathcal{H}^d\\ \|h\|\le 1}} \left\langle h \otimes h, \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{S}_p K_x \otimes \mathcal{S}_p K_y\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y)\right\rangle_{HS} \\
&= \left\|\int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{S}_p K_x \otimes \mathcal{S}_p K_y\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y)\right\|_{HS} \\
&= \left\|\int_{\mathcal{X}} \mathcal{S}_p K_x\mathrm{d}\mathbb{Q}(x) \otimes \int_{\mathcal{X}} \mathcal{S}_p K_y\mathrm{d}\mathbb{Q}(y)\right\|_{HS} \\
&= \left\|\int_{\mathcal{X}} \mathcal{S}_p K_x\mathrm{d}\mathbb{Q}(x)\right\|_{\mathcal{H}^d}^2 \\
&= \left\langle \int_{\mathcal{X}} \mathcal{S}_p K_x\mathrm{d}\mathbb{Q}(x), \int_{\mathcal{X}} \mathcal{S}_p K_y\mathbb{Q}(\mathrm{d}y)\right\rangle_{\mathcal{H}^d} \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}}\langle \mathcal{S}_p K_x, \mathcal{S}_p K_y\rangle_{\mathcal{H}^d}\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{S}_p^2\mathcal{S}_p^1 K(x,y)\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y).
\end{aligned}$$

To show the penultimate equality (exchange integral and inner product), we use the fact $\mathcal{S}_p K_x$ is $\mathbb{Q}$-Bochner integrable, and that the operator $W : f \mapsto \langle f, \int_{\mathcal{X}} \mathcal{S}_p K_y \mathbb{Q}(\mathrm{d}y)\rangle_{\mathcal{H}^d}$ is bounded, from which it follows that

$$\begin{aligned}
\left\langle \int_{\mathcal{X}} \mathcal{S}_p K_x\mathrm{d}\mathbb{Q}(x), \int_{\mathcal{X}} \mathcal{S}_p K_y\mathbb{Q}(\mathrm{d}y)\right\rangle_{\mathcal{H}^d} &= W\left[\int_{\mathcal{X}} \mathcal{S}_p K_x\mathrm{d}\mathbb{Q}(x)\right] = \int_{\mathcal{X}} W[\mathcal{S}_p K_x\mathrm{d}\mathbb{Q}(x)] \\
&= \int_{\mathcal{X}}\left\langle \mathcal{S}_p K_x, \int_{\mathcal{X}} \mathcal{S}_p K_y\mathrm{d}\mathbb{Q}(y)\right\rangle_{\mathcal{H}^d}\mathrm{d}\mathbb{Q}(x) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}}\langle \mathcal{S}_p K_x, \mathcal{S}_p K_y\rangle_{\mathcal{H}^d}\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y)
\end{aligned}$$

Hence $\mathrm{DKSD}_{K,m}(\mathbb{Q},\mathbb{P})^2 = \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{S}_p^2\mathcal{S}_p^1 K(x,y)\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y)$.

■

B.3 The Stein Kernel Corresponding to the Diffusion Kernel Stein Discrepancy

Note the Stein kernel satisfies

$$k^0 = \tfrac{1}{p(y)p(x)}\nabla_y \cdot \nabla_x \cdot \big(p(x)m(x)Km(y)^\top p(y)\big)$$

since

$$\begin{aligned}
k^0 &= \mathcal{S}_p^2\mathcal{S}_p^1 K(x,y) = \tfrac{1}{p(y)p(x)}\nabla_y \cdot (p(y)m(y)\nabla_x \cdot (p(x)m(x)K)) \\
&= \tfrac{1}{p(y)p(x)}\nabla_y \cdot (p(y)m(y)\partial_{x^i}(p(x)m(x)_{ir}K_{rs})e_s) \\
&= \tfrac{1}{p(y)p(x)}\nabla_y \cdot (p(y)m(y)_{ls}\partial_{x^i}(p(x)m(x)_{ir}K_{rs})e_l) \\
&= \tfrac{1}{p(y)p(x)}\partial_{y^l}(p(y)m(y)_{ls}\partial_{x^i}(p(x)m(x)_{ir}K_{rs})) \\
&= \tfrac{1}{p(y)p(x)}\partial_{y^l}\partial_{x^i}\big(p(x)m(x)_{ir}K_{rs}m(y)^\top_{sl}p(y)\big) \\
&= \tfrac{1}{p(y)p(x)}\nabla_y \cdot \nabla_x \cdot \big(p(x)m(x)Km(y)^\top p(y)\big).
\end{aligned}$$

Note it is also possible to view $m(x)Km(y)^\top$ as a new matrix kernel. That is the matrix field m defines a new kernel $K_m : (x,y) \mapsto m(x)K(x,y)m^\top(y)$, since $K_m(y,x)^\top = m(x)K(y,x)m(y)^\top = K_m(x,y)$ and for any $v_j \in \mathbb{R}^d, x_i \in \mathcal{X}$,

$$v_j \cdot K_m(x_j,x_l)v_l = v_j \cdot m(x_j)K(x_j,x_l)m(x_l)^\top v_l = \big(m(x_j)^\top v_j\big) \cdot K(x_j,x_l)\big(m(x_l)^\top v_l\big) \geq 0$$

We can expand the Stein kernel using the following expressions:

$$\begin{aligned}
&\nabla_y \cdot (p(y)m(y)\nabla_x \cdot (p(x)m(x)K)) \\
&= \nabla_y \cdot \big(p(y)m(y)\big(Km(x)^\top\nabla_x p + p(x)\nabla_x \cdot (m(x)K)\big)\big).
\end{aligned}$$

$$\begin{aligned}
&\nabla_y \cdot \big(p(y)m(y)Km(x)^\top\nabla_x p\big) \\
&= m^\top(x)\nabla_x p \cdot Km(y)^\top\nabla_y p + p(y)\nabla_y \cdot \big(m(y)Km(x)^\top\nabla_x p\big) \\
&= m^\top(x)\nabla_x p \cdot Km(y)^\top\nabla_y p + p(y)\nabla_y \cdot (m(y)K) \cdot m(x)^\top\nabla_x p,
\end{aligned}$$

$$\begin{aligned}
&\nabla_y \cdot (p(y)m(y)p(x)\nabla_x \cdot (m(x)K)) \\
&= p(x)(\nabla_y \cdot (p(y)m(y)) \cdot \nabla_x \cdot (m(x)K) + p(y)\mathrm{Tr}[m(y)\nabla_y\nabla_x \cdot (m(x)K)]) \\
&= p(x)p(y)\mathrm{Tr}[m(y)\nabla_y\nabla_x \cdot (m(x)K)] \\
&\quad + p(x)\nabla_x \cdot (m(x)K) \cdot \big(m(y)^\top\nabla_y p + p(y)\nabla_y \cdot m\big).
\end{aligned}$$

Hence

$$\begin{aligned}
k^0 &= m^\top(x)\nabla_x \log p \cdot Km(y)^\top\nabla_y \log p \\
&\quad + \nabla_y \cdot (m(y)K) \cdot m(x)^\top\nabla_x \log p + \nabla_x \cdot (m(x)K) \cdot m(y)^\top\nabla_y \log p \\
&\quad + \nabla_x \cdot (m(x)K) \cdot \nabla_y \cdot m + \mathrm{Tr}[m(y)\nabla_y\nabla_x \cdot (m(x)K)] \\
&= \langle s_p(x), Ks_p(y)\rangle + \langle\nabla_y \cdot (m(y)K), s_p(x)\rangle + \langle\nabla_x \cdot (m(x)K), s_p(y)\rangle \\
&\quad + \langle\nabla_x \cdot (m(x)K), \nabla_y \cdot m\rangle + \mathrm{Tr}[m(y)\nabla_y\nabla_x \cdot (m(x)K)]
\end{aligned}$$

B.4 Special Cases of Diffusion Kernel Stein Discrepancy

Consider

$$k^0 = \tfrac{1}{p(y)p(x)}\nabla_y \cdot \nabla_x \cdot \big(p(x)m(x)K(x,y)m(y)^\top p(y)\big)$$

and decompose $m(x)K(x,y)m(y)^\top \equiv gA$ where g is scalar and A is matrix-valued. Then we

$$\begin{aligned} k^0 &= g\langle \nabla_y \log p, A\nabla_x \log p\rangle + \langle \nabla_y \log p, A\nabla_x g\rangle + \langle \nabla_y g, A\nabla_x \log p\rangle \\ &+ \mathrm{Tr}[A\nabla_x\nabla_y g] + g\nabla_y\cdot\nabla_x\cdot A + \langle \nabla_x\cdot A, \nabla_y g\rangle + \left\langle \nabla_y\cdot A^\top, \nabla_x g\right\rangle \\ &+ g\left\langle \nabla_y\cdot A^\top, \nabla_x \log p\right\rangle + g\langle \nabla_x\cdot A, \nabla_y \log p\rangle. \end{aligned}$$

For the case, $K = \mathrm{diag}(k^1,\dots,k^d)$, setting $\mathcal{T}_i^x \equiv \frac{1}{p(x)}\partial_{x^i}(p(x)\cdot)$ then

$$\mathcal{S}_p^2\mathcal{S}_p^1[\mathrm{diag}(k^1,\dots,k^d)] = \mathcal{T}_l^y\big(m_{li}(y)\mathcal{T}_c^x\big(k^i(x,y)m_{ic}^\top(x)\big)\big) = \mathcal{T}_l^y\mathcal{T}_c^x\big(m_{li}(y)k^i(x,y)m_{ci}(x)\big).$$

If $K = Ik$ in components

$$\begin{aligned} \mathcal{S}_p^2\mathcal{S}_p^1[Ik] &= (s_p(x))_i k(x,y)(s_p(y))_i + \partial_{y^i}(m_{ir}k)(s_p(x))_r + \partial_{x^i}(m(x)_{ir}k)(s_p(y))_r \\ &+ \partial_{x^i}(m(x)_{ir}k)\partial_{y^l}(m_{lr}) + m(y)_{ir}\partial_{y^i}\partial_{x^s}(m(x)_{sr}k) \end{aligned}$$

When $p = p_\theta$ we are often interested in the gradient $\nabla_\theta k_\theta^0$. Note $\nabla_y\cdot(m(y)K) = k\nabla_y\cdot m + \nabla_y k\cdot m(y)$, so [2]

$$\begin{aligned} \partial_{\theta^i}[k\langle\nabla_y\cdot m, s_p(x)\rangle] &= k\partial_{\theta^i}\langle\nabla_y\cdot m, s_p(x)\rangle \\ \partial_{\theta^i}[\langle\nabla_y k\cdot m(y), s_p(x)\rangle] &= \langle\nabla_y k, \partial_{\theta^i}[m(y)s_p(x)]\rangle \\ \mathrm{Tr}[m(y)\nabla_y\nabla_x\cdot(m(x)K)] &= \nabla_y k^\top m(y)\nabla_x\cdot m + \mathrm{Tr}[m(y)m(x)^\top\nabla_y\nabla_x k] \end{aligned}$$

and the terms in $\partial_{\theta^i}k^0$ reduce to

$$\begin{aligned} \partial_{\theta^i}\langle s_p(x), Ks_p(y)\rangle &= k\partial_{\theta^i}\langle s_p(x), s_p(y)\rangle \\ \partial_{\theta^i}\langle\nabla_y\cdot(m(y)K), s_p(x)\rangle &= k\partial_{\theta^i}\langle\nabla_y\cdot m, s_p(x)\rangle + \langle\nabla_y k, \partial_{\theta^i}[m(y)s_p(x)]\rangle \\ \partial_{\theta^i}\langle\nabla_x\cdot(m(x)K), s_p(y)\rangle &= k\partial_{\theta^i}\langle\nabla_x\cdot m, s_p(y)\rangle + \langle\nabla_x k, \partial_{\theta^i}[m(x)s_p(y)]\rangle \\ \partial_{\theta^i}\langle\nabla_x\cdot(m(x)K), \nabla_y\cdot m\rangle &= k\partial_{\theta^i}\langle\nabla_x\cdot m, \nabla_y\cdot m\rangle + \partial_{\theta^i}\langle\nabla_x k\cdot m(x), \nabla_y\cdot m\rangle. \end{aligned}$$

When $K = kI$ and we further have a diagonal matrix $m = \mathrm{diag}(f_i)$, $m(y)m(x)^\top = \mathrm{diag}(f_i(y)f_i(x))$. If $u\odot v$ denotes the vector given by the pointwise product of vectors, i.e., $(u\odot v)_i = u_iv_i$, and f is the vector, then $m(x)\nabla_x\log p = f(x)\odot\nabla_x\log p$ and $(\nabla_y\cdot m)_i = \partial_{y^i}f_i$, $(\nabla_x\cdot(mk))_i = \partial_{x^i}(f_ik)$,

$$\begin{aligned} s_p(x)\cdot Ks_p(y) &= k(x,y)f_i(x)\partial_{x^i}\log p f_i(y)\partial_{y^i}\log p \\ \nabla_y\cdot(m(y)K)\cdot s_p(x) &= \partial_{y^i}(f_i(y)k)f_i(x)\partial_{x^i}\log p \\ \nabla_x\cdot(m(x)K)\cdot\nabla_y\cdot m &= \partial_{x^i}(f_i(x)k)\partial_{y^i}(f_i(y)) \\ \mathrm{Tr}[m(y)\nabla_y\nabla_x\cdot(mk)] &= f_i(y)\partial_{x^i}\big(f_i(x)\partial_{y^i}k\big) \end{aligned}$$

and if $m\mapsto mI$ (is scalar), (this is just KSD with $k(x,y)\mapsto m(x)k(x,y)m(y)$):

$$\begin{aligned} k^0 &= m(x)m(y)k(x,y)\nabla_x\log p\cdot\nabla_y\log p \\ &+ m(x)\nabla_y(m(y)k)\cdot\nabla_x\log p + m(y)\nabla_x(m(x)k)\cdot\nabla_y\log p \\ &+ \nabla_x(m(x)k)\cdot\nabla_y m + m(y)\nabla_x\cdot(m(x)\nabla_y k), \end{aligned}$$

When $m = I$, we recover the usual definition of kernel-Stein discrepancy (KSD):

$$\mathrm{KSD}(\mathbb{Q}\|\mathbb{P})^2 = \int_{\mathcal{X}}\int_{\mathcal{X}}\frac{1}{p(y)p(x)}\nabla_y\cdot\nabla_x(p(x)k(x,y)p(y))\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y).$$

[2] More generally $\nabla_y\cdot(m(y)K) = (\nabla_y\cdot m)\cdot K + \mathrm{Tr}[\nabla_y K\otimes m(y)]$ where $\mathrm{Tr}[\nabla_y K\otimes m]_r = \partial_{y^i}K_{jr}m_{ij}$ and if $K = Bk$

$$\begin{aligned} \partial_{\theta^i}[(\nabla_y\cdot m)\cdot Ks_p(x)] &= kB_{sr}\partial_{\theta^i}((\nabla_y\cdot m)_s(s_p(x))_r) = k\,\mathrm{Tr}[B\partial_{\theta^i}(s_p(x)\otimes\nabla_y\cdot m)] \\ \partial_{\theta^i}\left[\nabla_y k^\top m(y)Bs_p(x)\right] &= \partial_{y^s}kB_{jr}\partial_{\theta^i}[m_{sj}(y)(s_p(x))_r] \end{aligned}$$

B.5 Diffusion Kernel Stein Discrepancies as Statistical Divergences

In the following section, we prove that DKSD is a statistical divergence, and provide sufficent conditions on the (matrix-valued) kernel.

Proposition 4 (**DKSD as statistical divergence**). *Suppose K is IPD and in the Stein class of $\mathbb{Q}$, and $m(x)$ is invertible. If $s_p - s_q \in L^1(\mathbb{Q})$, then $\mathrm{DKSD}_{K,m}(\mathbb{Q}\|\mathbb{P})^2 = 0$ iff $\mathbb{Q} = \mathbb{P}$.*

Proof By Stoke's theorem $\int_{\mathcal{X}} \mathcal{S}_q[v]\mathrm{d}\mathbb{Q} = \int_{\mathcal{X}} \nabla\cdot(qmv)\mathrm{d}x = 0$, thus $\int_{\mathcal{X}} \mathcal{S}_p[v]\mathrm{d}\mathbb{Q} = \int_{\mathcal{X}}(\mathcal{S}_p[v] - \mathcal{S}_q[v])\mathrm{d}\mathbb{Q} = \int_{\mathcal{X}}(s_p - s_q)\cdot v\mathrm{d}\mathbb{Q}$, and by assumption $\int_{\mathcal{X}} \mathcal{S}_q[K]\mathrm{d}\mathbb{Q} = \int_{\mathcal{X}} \nabla\cdot(qmK)\mathrm{d}x = 0$. Moreover, with $s_p = m^\top \nabla \log p$, and $\delta_{p,q} \equiv s_p - s_q$. Hence

$$\begin{aligned}
\mathrm{DKSD}_{K,m}(\mathbb{Q},\mathbb{P})^2 &= \int_{\mathcal{X}}\int_{\mathcal{X}} \mathcal{S}_p^2\big[\mathcal{S}_p^1 K(x,y)\big]\mathrm{d}\mathbb{Q}(y)\mathrm{d}\mathbb{Q}(x) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}}(s_p(y) - s_p(y))\cdot\big[\mathcal{S}_p^1 K(x,y)\big]\mathrm{d}\mathbb{Q}(y)\mathrm{d}\mathbb{Q}(x) \\
&= \int_{\mathcal{X}}(s_p(y) - s_p(y))\mathrm{d}\mathbb{Q}(y)\cdot\int_{\mathcal{X}}\big[\mathcal{S}_p^1 K(x,y)\big]\mathrm{d}\mathbb{Q}(x) \\
&= \int_{\mathcal{X}}(s_p(y) - s_p(y))\mathrm{d}\mathbb{Q}(y)\cdot\int_{\mathcal{X}}\big[\mathcal{S}_p^1 K(x,y) - \mathcal{S}_q^1 K(x,y)\big]\mathrm{d}\mathbb{Q}(x) \\
&= \int_{\mathcal{X}}(s_p(y) - s_p(y))\mathrm{d}\mathbb{Q}(y)\cdot\int_{\mathcal{X}}[(s_p(x) - s_p(x))\cdot K(x,y)]\mathrm{d}\mathbb{Q}(x) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} q(x)\delta_{p,q}(x)^\top K(x,y)\delta_{p,q}(y)q(y)\mathrm{d}x\mathrm{d}y \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \mathrm{d}\mu^\top(x)K(x,y)\mathrm{d}\mu(y).
\end{aligned}$$

where $\mu(\mathrm{d}x) \equiv q(x)\delta_{p,q}(x)\mathrm{d}x$, which is a finite measure by assumption. If $\mathcal{S}(q,p) = 0$, then since K is IPD we have $q\delta_{p,q} \equiv 0$, and since $q > 0$ and m is invertible we must have $\nabla \log p = \nabla \log q$ and thus $q = p$. ■

Proposition 5 (**IPD matrix kernels**). *(i) When $K = diag(k^1,\ldots,k^d)$, 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.*

Proof Let μ be a finite signed vector measure. (i) If each k^i is IPD, then $\int \mathrm{d}\mu^\top K\mathrm{d}\mu = \int k^i(x,y)\mathrm{d}\mu_i(x)\mathrm{d}\mu_i(y) \geq 0$ with equality iff $\mu_i \equiv 0$ for all i. Conversely suppose $\int k^i(x,y)\mathrm{d}\mu_i(x)\mathrm{d}\mu_i(y) \geq 0$ with equality iff $\mu_i \equiv 0$ for all i . Suppose k^j is not IPD for some j, then there exists a finite non-zero signed measure ν s.t., $\int k^j \mathrm{d}\nu\otimes\mathrm{d}\nu \leq 0$, so if we define the vector measure $\mu_i \equiv \delta_{ij}\nu$, which is non-zero and finite, then $\int k^i(x,y)\mathrm{d}\mu_i(x)d\mu_i(y) \leq 0$ which contradicts the assumption. For (ii), we first diagonalise $B = R^\top DR$ where R is orthogonal and D diagonal with positive entries $\lambda_i > 0$. Then

$$\int \mathrm{d}\mu^\top K\mathrm{d}\mu = \int k\mathrm{d}\mu^\top R^\top DR\mathrm{d}\mu = \int k(R\mathrm{d}\mu)^\top D(R\mathrm{d}\mu) = \int k(x,y)\lambda_i\mathrm{d}\nu_i(x)\mathrm{d}\nu_i(y),$$

where $\nu \equiv R\mu$ is finite and non-zero, since μ is non-zero and R is invertible, thus maps non-zero vectors to non-zero vectors. Clearly if k is IPD then $\int \mathrm{d}\mu^\top K\mathrm{d}\mu \geq 0$ with equality iff $\nu_i \equiv 0$ for all i. Suppose K is IPD but k is not, then there exists finite non-zero signed measure ν for which $\int k\mathrm{d}\nu\otimes\mathrm{d}\nu \leq 0$, but then setting $\mu \equiv R^\top\xi$, with $\xi_i \equiv \delta_{ij}\nu$ which is finite and non-zero, implies $\int \mathrm{d}\mu^\top K\mathrm{d}\mu = \int k\mathrm{d}\xi^\top D\mathrm{d}\xi = \lambda_j\int k\mathrm{d}\nu\otimes\mathrm{d}\nu \leq 0$. ■

B.6 Diffusion Score Matching

Another example of SD is the diffusion score matching (DSM), as introduced below:

Theorem 9 (**Diffusion Score Matching**). *Let $\mathcal{X} = \mathbb{R}^d$ and consider the Stein operator $\mathcal{S}_p$ in (2) for some function $m \in \Gamma(\mathbb{R}^{d\times d})$ and the Stein class $\mathcal{G} \equiv \{g = (g_1,\ldots,g_d) \in C^1(\mathcal{X},\mathbb{R}^d)\cap L^2(\mathcal{X};\mathbb{Q}) : \|g\|_{L^2(\mathcal{X};\mathbb{Q})} \leq 1\}$. If $p, q > 0$ are differentiable and $s_p - s_q \in L^2(\mathbb{Q})$, then we define the* diffusion score matching *divergence as the Stein discrepancy,*

$$\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}) \equiv \sup_{f\in\mathcal{S}_p[\mathcal{G}]}\left|\int_{\mathcal{X}} f\mathrm{d}\mathbb{Q} - \int_{\mathcal{X}} f\mathrm{d}\mathbb{P}\right|^2 = \int_{\mathcal{X}}\left\|m^\top(\nabla\log q - \nabla\log p)\right\|_2^2\mathrm{d}\mathbb{Q}.$$

This satisfies $\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}) = 0$ iff $\mathbb{Q} = \mathbb{P}$ when $m(x)$ is invertible. Moreover, if p is twice-differentiable, and $qmm^\top\nabla\log p, \nabla\cdot(qmm^\top\nabla\log p) \in L^1(\mathbb{R}^d)$, then Stoke's theorem gives

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

Proof Note that the Stein operator satisfies

$$\mathcal{S}_p[g] = \frac{\nabla\cdot(pmg)}{p} = \frac{\langle\nabla p, mg\rangle + p\nabla\cdot(mg)}{p} = \langle\nabla\log p, mg\rangle + \nabla\cdot(mg) = \langle m^\top\nabla\log p, g\rangle + \nabla\cdot(mg).$$

Since $\int_\mathcal{X}\mathcal{S}_q[g]\mathrm{d}\mathbb{Q} = 0$, we have

$$\begin{aligned}
D(\mathbb{Q}\|\mathbb{P}) &= \sup_{g\in\mathcal{G}}\left|\int_\mathcal{X}\mathcal{S}_p[g](x)\mathbb{Q}(\mathrm{d}x)\right|^2 = \sup_{g\in\mathcal{G}}\left|\int_\mathcal{X}(\mathcal{S}_p[g](x) - \mathcal{S}_q[g](x))\mathbb{Q}(\mathrm{d}x)\right|^2 \\
&= \sup_{g\in\mathcal{G}}\left|\int_\mathcal{X}((\nabla\log p - \nabla\log q)\cdot(mg))\mathrm{d}\mathbb{Q}\right|^2, \\
&= \sup_{g\in\mathcal{G}}\left|\left\langle m^\top(\nabla\log p - \nabla\log q), g\right\rangle_{L^2(\mathbb{Q})}\right|^2 \\
&= \left\|m^\top(\nabla\log p - \nabla\log q)\right\|^2_{L^2(\mathbb{Q})} \\
&= \int_\mathcal{X}\left\|m^\top(\nabla\log p - \nabla\log q)\right\|_2^2\mathrm{d}\mathbb{Q},
\end{aligned}$$

where we have used the fact that $\mathcal{G}$ is dense in the unit ball of $L^2(\mathbb{Q})$ (since smooth functions with compact support are dense in $L^2(\mathbb{Q})$), and that the supremum over a dense subset of the continuous functional $F(\cdot) \equiv \left\langle m^\top(\nabla\log p - \nabla\log q), \cdot\right\rangle_{L^2(\mathbb{Q})}$ is equal to the supremum over the closure, $\sup_\mathcal{G}F = \sup_{\overline{\mathcal{G}}}F$. Suppose $D(\mathbb{Q}\|\mathbb{P}) = 0$. Then since $q > 0$ we must have $\left\|m^\top(\nabla\log p - \nabla\log q)\right\|_2^2 = 0$, i.e., $m^\top(\nabla\log p - \nabla\log q) = 0$, i.e., $\nabla(\log p - \log q) = 0$. Thus $\log(p/q) = c$, so $p = qe^c$ and integrating implies $c = 0$, so $D(\mathbb{Q}\|\mathbb{P}) = 0$ iff $\mathbb{Q} = \mathbb{P}$ a.e..

To obtain the estimator we will use the divergence theorem, which holds for example if $X, \nabla\cdot X \in L^1(\mathbb{R}^d)$ for $X = qmm^\top\nabla\log p$ (see theorem 2.36, 2.28 [51] or theorem 2.38 for weaker conditions). Note

$$\left\|m^\top(\nabla\log p - \nabla\log q)\right\|_2^2 = \|m^\top\nabla\log p\|_2^2 + \|m^\top\nabla\log q\|_2^2 - 2m^\top\nabla\log p\cdot m^\top\nabla\log q$$

thus we have

$$\begin{aligned}
\int_\mathcal{X}\langle m^\top\nabla\log p, m^\top\nabla\log q\rangle\mathrm{d}\mathbb{Q} &= \int_\mathcal{X}\langle\nabla\log q, mm^\top\nabla\log p\rangle\mathrm{d}Q \\
&= \int_\mathcal{X}\langle\nabla q, mm^\top\nabla\log p\rangle\mathrm{d}x \\
&= \int_\mathcal{X}\left(\nabla\cdot\left(qmm^\top\nabla\log p\right) - q\nabla\cdot\left(mm^\top\nabla\log p\right)\right)\mathrm{d}x \\
&= -\int_\mathcal{X} q\nabla\cdot\left(mm^\top\nabla\log p\right)\mathrm{d}x \\
&= -\int_\mathcal{X}\nabla\cdot\left(mm^\top\nabla\log p\right)\mathrm{d}\mathbb{Q}.
\end{aligned}$$

■

As for the standard SM estimator, the DSM is only defined for distributions with sufficiently smooth densities. However the θ-dependent part of $\mathrm{DSM}_m(\mathbb{Q}, \mathbb{P}_\theta)$ [3]

$$\begin{aligned}
&\int_\mathcal{X}\left(\left\|m^\top\nabla_x\log p_\theta\right\|_2^2 + 2\nabla\cdot\left(mm^\top\nabla\log p_\theta\right)\right)\mathrm{d}\mathbb{Q} \\
&= \int_\mathcal{X}\left(\left\|m^\top\nabla_x\log p_\theta\right\|_2^2 + 2\left(\left\langle\nabla\cdot(mm^\top), \nabla\log p\right\rangle + \mathrm{Tr}\left[mm^\top\nabla^2\log p\right]\right)\right)\mathrm{d}\mathbb{Q},
\end{aligned}$$

does not depend on the density of $\mathbb{Q}$. An unbiased estimator for this quantity follows by replacing $\mathbb{Q}$ with the empirical random measure $\mathbb{Q}_n \equiv \frac{1}{n}\sum_i\delta_{X_i}$ where $X_i \sim \mathbb{Q}$ are independent. Hence we consider the estimator

$$\hat{\theta}_n^{\mathrm{DSM}} \equiv \mathrm{argmin}_{\theta\in\Theta}\mathbb{Q}_n\left(\left\|m^\top\nabla_x\log p_\theta\right\|_2^2 + 2\left(\left\langle\nabla\cdot(mm^\top), \nabla\log p_\theta\right\rangle + \mathrm{Tr}\left[mm^\top\nabla^2\log p_\theta\right]\right)\right).$$

In components, this corresponds to:

$$\begin{aligned}
\hat{\theta}_n^{\mathrm{DSM}} = \mathrm{argmin}_{\theta\in\Theta}\int_\mathcal{X}\mathrm{d}\mathbb{Q}(x)\|m(x)^\top\nabla_x\log p(x|\theta)\|_2^2 + 2\textstyle\sum_{j,k,l=1}^d\partial_{x^j}\partial_{x^k}\log p(x|\theta)m_{kl}(x)m_{jl}(x) \\
+ 2\textstyle\sum_{j,k,l=1}^d\partial_{x^k}\log p(x|\theta)(\partial_{x^j}m_{kl}(x)m_{jl}(x) + m_{kl}(x)\partial_{x^j}m_{jl}(x))
\end{aligned}$$

We now consider the the limit in which DKSD converges to DSM. We use the following lemma as a stepping stone.

[3] Here we use $\nabla\cdot\left(mm^\top\nabla\log p\right) = \left\langle\nabla\cdot(mm^\top), \nabla\log p\right\rangle + \mathrm{Tr}\left[mm^\top\nabla^2\log p\right]$

Lemma 1. *Suppose* $\Phi \in L^1(\mathbb{R}^d)$, $\Phi > 0$ *and* $\int \Phi(s)\,\mathrm{d}s = 1$. *Let* $f, g \in C(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, *then defining* $K_\gamma \equiv B\Phi_\gamma$ *where* $\Phi_\gamma(s) \equiv \gamma^{-d}\Phi(s/\gamma)$ *and* $\gamma > 0$, *we have*

$$\int\int f(x)^\top K_\gamma(x,y)g(y)\mathrm{d}x\,\mathrm{d}y \;\to\; \int f(x)^\top Bg(x)\,\mathrm{d}x, \quad \text{as} \quad \gamma \to 0.$$

Proof We rewrite

$$\int_{\mathcal{X}}\int_{\mathcal{X}} f(x)^\top B\Phi_\gamma(x-y)g(y)\,\mathrm{d}x\,\mathrm{d}y = \int_{\mathcal{X}}\int_{\mathcal{X}} f(x)^\top Bg(x-s)\mathrm{d}x\,\Phi_\gamma(s)\,\mathrm{d}s = \int_{\mathcal{X}} H(s)\Phi_\gamma(s)\,\mathrm{d}s,$$

where $H : \mathcal{X} \to \mathbb{R}$ is defined by

$$H(s) \equiv \int_{\mathcal{X}} f(x)^\top Bg(x-s)\,\mathrm{d}x = \int_{\mathcal{X}} \langle f(x), Bg(x-s)\rangle_{\mathbb{R}^d}\mathrm{d}x \equiv \int_{\mathcal{X}} \langle f(x), g(x-s)\rangle_B \mathrm{d}x.$$

Since $f, g \in C(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$, the function $H(s)$ is continuous, bounded, $|H(s)| \le A\|f\|_{L^2(\mathbb{R}^d)}\|g\|_{L^2(\mathbb{R}^d)}$ for a constant $A > 0$ depending only on B, and $H(0) = \int f(x)^\top Bg(x)\,\mathrm{d}x$. Given $\delta > 0$, we can split the integral as follows:

$$\int_{|s|<\delta} H(s)\Phi_\gamma(s)\,\mathrm{d}s \;+\; \int_{|s|>\delta} H(s)\Phi_\gamma(s)\,\mathrm{d}s \equiv I_1 + I_2.$$

By continuity, given $\epsilon \in (0,1)$ there exists $\delta > 0$ such that $|H(s) - H(0)| < \epsilon$ for all $|s| < \delta$. Let $I_{<\delta} \equiv \int_{|y|<\delta} \Phi_\gamma(y)\,\mathrm{d}y > 0$ since $\Phi > 0$. Consider

$$\begin{aligned} I_1 - H(0) &= \int_{|s|<\delta} \Phi_\gamma(s)H(s)\mathrm{d}s - H(0) \;=\; \int_{|s|<\delta} \Phi_\gamma(s)\Big(H(s) - \tfrac{H(0)}{I_{<\delta}}\Big)\mathrm{d}s \\ &= \int_{|s|<\delta} \tfrac{\Phi_\gamma(s)}{I_{<\delta}}(H(s)I_{<\delta} \;-\; H(0))\,\mathrm{d}s. \end{aligned}$$

Clearly $\int \Phi_\gamma(s)\mathrm{d}s = \int \gamma^{-d}\Phi(s/\gamma)\mathrm{d}s = \int \Phi(z)\mathrm{d}z = 1$, since $z \equiv s/\gamma$ implies $\mathrm{d}z = \gamma^{-d}\mathrm{d}s$, so

$$I_{<\delta} \;=\; 1 - I_{>\delta} \;=\; 1 - \int_{|y|>\delta/\gamma} \Phi(y)\,\mathrm{d}y.$$

Then since Φ is integrable, there exists $\gamma_0(\delta) > 0$ s.t. for $\gamma < \gamma_0(\delta)$ we have $\int_{|y|>\delta/\gamma} \Phi(y)\,\mathrm{d}y < \epsilon$ and thus $0 < 1 - \epsilon < I_{<\delta} < 1$. Therefore, for $\gamma < \gamma_0(\delta)$:

$$\begin{aligned} |I_1 - H(0)| &= \Big|\int_{|s|<\delta} \tfrac{\Phi_\gamma(s)}{I_{<\delta}}(H(s)I_{<\delta} - H(0))\mathrm{d}s\Big| \\ &\le \int_{|s|<\delta} \tfrac{\Phi_\gamma(s)}{I_{<\delta}}|((H(s) - H(0))I_{<\delta} + H(0)(I_{<\delta} - 1))|\mathrm{d}s \\ &\le \int_{|s|<\delta} \tfrac{\Phi_\gamma(s)}{I_{<\delta}}(|H(s) - H(0)|I_{<\delta} + |1 - I_{<\delta}|H(0))\mathrm{d}s \\ &\le \int_{|s|<\delta} \tfrac{\Phi_\gamma(s)}{I_{<\delta}}(\epsilon I_{<\delta} + \epsilon H(0))\mathrm{d}s \\ &\le \epsilon \int_{|z|<\delta/\gamma} \Phi(z)\mathrm{d}z + H(0)\epsilon \le (1 + H(0))\epsilon. \end{aligned}$$

For the second term, since H is bounded we have

$$I_2 = \int_{|s|>\delta} H(s)\Phi_\gamma(s)\mathrm{d}s = \int_{|s|>\delta/\gamma} H(\gamma s)\Phi(s)\mathrm{d}s \le \|H\|_\infty \int_{|s|>\delta/\gamma} \Phi(s)\mathrm{d}s,$$

so that, $|I_2| \le \|H\|_\infty\epsilon$, for $\gamma < \gamma_0(\delta)$. It follows that

$$\begin{aligned} \Big|\int\int f(x)^\top K_\gamma(x,y)g(y)\mathrm{d}x\,\mathrm{d}y - \int f(x)^\top Bg(x)\,\mathrm{d}x\Big| &= \Big|\int H(s)\Phi_\gamma(s)\mathrm{d}s - H(0)\Big| \\ &= |I_1 + I_2 - H(0)| \\ &\le |I_1 - H(0)| + |I_2| \;\to\; 0, \end{aligned}$$

as $\gamma \to 0$ as required. ■

Theorem 1 (DSM as a limit of DKSD). *Let* $\mathbb{Q}(\mathrm{d}x) \equiv q(x)\mathrm{d}x$ *be a probability measure on* $\mathbb{R}^d$ *with* $q > 0$. *Suppose that* $f, g \in C(\mathbb{R}^d) \cap L^2(\mathbb{Q})$, $\Phi \in L^1(\mathbb{R}^d)$, $\Phi > 0$ *and* $\int_{\mathbb{R}^d} \Phi(s)\,\mathrm{d}s = 1$, *define* $K_\gamma \equiv B\Phi_\gamma$ *where* $\Phi_\gamma(s) \equiv \gamma^{-d}\Phi(s/\gamma)$ *and* $\gamma > 0$. *Let* $k^q_\gamma(x,y) = k_\gamma(x,y)/\sqrt{q(x)q(y)} = \Phi_\gamma(x-y)/\sqrt{q(x)q(y)}$, *and set* $K^p_\gamma \equiv Bk^q_\gamma$. *Then,*

$$\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(x)^\top K^q_\gamma(x,y)g(y)\mathrm{d}\mathbb{Q}(x)\mathrm{d}\mathbb{Q}(y) \;\to\; \int_{\mathbb{R}^d} f(x)^\top Bg(x)\mathrm{d}\mathbb{Q}(x), \qquad \text{as} \qquad \gamma \to 0.$$

In particular choosing $f, g = \delta_{p,q}$ *shows* $\mathrm{DKSD}_{K^q_\gamma,m}(\mathbb{Q}\|\mathbb{P})^2$ *converge to* $\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P})$ *with inner product* $\langle\cdot,\cdot\rangle_B \equiv \langle\cdot, B\cdot\rangle_2$

Proof We note that $f \in L^2(\mathbb{Q})$ if and only if $f\sqrt{q} \in L^2(\mathbb{R}^d)$. Therefore applying the previous result, we have that

$$\int_{\mathcal{X}}\int_{\mathcal{X}} f(x)^\top K^q_\gamma(x,y)g(y)\,\mathrm{d}\mathbb{Q}(x)\,\mathrm{d}\mathbb{Q}(y) = \int_{\mathcal{X}}\int_{\mathcal{X}}\Big(\sqrt{q(x)}f(x)\Big)^\top K_\gamma(x,y)\Big(g(y)\sqrt{q(y)}\Big)\mathrm{d}x\mathrm{d}y$$
$$\to \int_{\mathcal{X}} f(x)^\top B g(x)\mathrm{d}\mathbb{Q}(x), \quad \text{as} \quad \gamma \to 0.$$

Note that if k is a (scalar) kernel function, then $(x,y) \mapsto r(x)k(x,y)r(y)$ is a kernel for any function $r : \mathcal{X} \to \mathbb{R}$, and thus k^q_γ defines a sequence of kernels parametrised by a scale parameter $\gamma > 0$. It follows that the sequence of DKSD paramaterised by K^q_γ

$$\mathrm{DKSD}_{K^q_\gamma,m}(\mathbb{Q}\|\mathbb{P})^2 = \int_{\mathcal{X}}\int_{\mathcal{X}} q(x)\delta_{p,q}(x)^\top K^q_\gamma(x,y)\delta_{p,q}(y)q(y)\mathrm{d}x\mathrm{d}y$$

converges to DSM with inner product $\langle\cdot,\cdot\rangle_B \equiv \langle\cdot, B\cdot\rangle_2$ on $\mathbb{R}^d$.

$$\mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}) = \int_{\mathcal{X}} \delta_{q,p}(x)^\top B\delta_{q,p}(x)\mathrm{d}\mathbb{Q} = \int_{\mathcal{X}} \|m^\top(\nabla\log p - \nabla \log q)\|^2_B \mathrm{d}\mathbb{Q}$$

■

C Information Semi-Metrics of Minimum Stein Discrepancy Estimators

In this section, we derive expressions for the metric tensor of DKSD and DSM.

C.1 Information Semi-Metric of Diffusion Kernel Stein Discrepancy

Let $\mathcal{P}_\Theta$ be a parametric family of probability measures on $\mathcal{X}$. Given a map $D : \mathcal{P}_\Theta \times \mathcal{P}_\Theta \to \mathbb{R}$, for which $D(\mathbb{P}_1\|\mathbb{P}_2) = 0$ iff $\mathbb{P}_1 = \mathbb{P}_2$, its associated information semi-metric is defined as the map $\theta \mapsto g(\theta)$, where $g(\theta)$ is the symmetric bilinear form $g(\theta)_{ij} = -\frac{1}{2}\frac{\partial^2}{\partial\alpha^i\partial\theta^j}D(\mathbb{P}_\alpha\|\mathbb{P}_\theta)|_{\alpha=\theta}$. When g is positive definite, we can use it to perform (Riemannian) gradient descent on $\mathcal{P}_\Theta \cong \Theta$.

Proposition 6 (**Information Tensor DKSD**)**.** *Assume the conditions of Proposition 4 hold. The information semi-metric associated to DKSD is*

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

Proof From Proposition 4 we have

$$\mathrm{DKSD}_{K,m}(\mathbb{P}_\alpha,\mathbb{P}_\theta)^2 = \int_{\mathcal{X}}\int_{\mathcal{X}} p_\alpha(x)\delta_{p_\theta,p_\alpha}(x)^\top K(x,y)\delta_{p_\theta,p_\alpha}(y)p_\alpha(y)\mathrm{d}x\mathrm{d}y$$

where $\delta_{p_\theta,p_\alpha} = m_\theta^\top(\nabla\log p_\theta - \nabla\log p_\alpha)$. Thus

$$\begin{aligned}\partial_{\alpha^i}\partial_{\theta^j}\,\mathrm{DKSD}_{K,m}(\mathbb{P}_\alpha,\mathbb{P}_\theta)^2 &= \partial_{\alpha^i}\partial_{\theta^j}\int_{\mathcal{X}}\int_{\mathcal{X}} p_\alpha(x)\delta_{p_\theta,p_\alpha}(x)^\top K(x,y)\delta_{p_\theta,p_\alpha}(y)p_\alpha(y)\mathrm{d}x\mathrm{d}y\\ &= \partial_{\alpha^i}\int_{\mathcal{X}}\int_{\mathcal{X}} p_\alpha(x)\partial_{\theta^j}\delta_{p_\theta,p_\alpha}(x)^\top K(x,y)\delta_{p_\theta,p_\alpha}(y)p_\alpha(y)\mathrm{d}x\mathrm{d}y\\ &+ \partial_{\alpha^i}\int_{\mathcal{X}}\int_{\mathcal{X}} p_\alpha(x)\delta_{p_\theta,p_\alpha}(x)^\top K(x,y)\partial_{\theta^j}\delta_{p_\theta,p_\alpha}(y)p_\alpha(y)\mathrm{d}x\mathrm{d}y,\end{aligned}$$

and using $\delta_{p_\theta,p_\theta} = 0$, we get:

$$\begin{aligned}&\partial_{\alpha^i}\int_{\mathcal{X}}\int_{\mathcal{X}} p_\alpha(x)\partial_{\theta^j}\delta_{p_\theta,p_\alpha}(x)^\top K(x,y)\delta_{p_\theta,p_\alpha}(y)p_\alpha(y)\mathrm{d}x\mathrm{d}y\Big|_{\alpha=\theta}\\ &= \partial_{\alpha^i}\int_{\mathcal{X}}\int_{\mathcal{X}} p_\alpha(x)\big(\partial_{\theta^j}m_\theta^\top(\nabla\log p_\theta - \nabla\log p_\alpha) + m_\theta^\top\partial_{\theta^j}\nabla\log p_\theta\big)^\top K(x,y)\delta_{p_\theta,p_\alpha}(y)p_\alpha(y)\mathrm{d}x\mathrm{d}y\Big|_{\alpha=\theta}\\ &= \int_{\mathcal{X}}\int_{\mathcal{X}} p_\alpha(x)\big(m_\theta^\top\partial_{\theta^j}\nabla\log p_\theta\big)^\top K(x,y)\partial_{\alpha^i}\delta_{p_\theta,p_\alpha}(y)p_\alpha(y)\mathrm{d}x\mathrm{d}y\Big|_{\alpha=\theta}\\ &= -\int_{\mathcal{X}}\int_{\mathcal{X}} p_\alpha(x)\big(m_\theta^\top\partial_{\theta^j}\nabla\log p_\theta\big)^\top K(x,y)\big(m_\theta^\top\partial_{\alpha^i}\nabla\log p_\alpha\big)(y)p_\alpha(y)\mathrm{d}x\mathrm{d}y\Big|_{\alpha=\theta}\\ &= -\int_{\mathcal{X}}\int_{\mathcal{X}}\big(m_\theta^\top\partial_{\theta^j}\nabla\log p_\theta\big)^\top(x)K(x,y)\big(m_\theta^\top\partial_{\theta^i}\nabla\log p_\theta\big)(y)\mathrm{d}\mathbb{P}_\theta(x)\mathrm{d}\mathbb{P}_\theta(y).\end{aligned}$$

Similarly, we also get:

$$
\begin{aligned}
&\partial_{\alpha^i} \int_{\mathcal{X}} \int_{\mathcal{X}} p_\alpha(x)\delta_{p_\theta,p_\alpha}(x)^\top K(x,y)\partial_{\theta^j}\delta_{p_\theta,p_\alpha}(y)p_\alpha(y)\mathrm{d}x\mathrm{d}y\big|_{\alpha=\theta} \\
&= -\int_{\mathcal{X}} \int_{\mathcal{X}} \big(m_\theta^\top \partial_{\theta^i}\nabla \log p_\theta\big)^\top(x)K(x,y)\big(m_\theta^\top \partial_{\theta^j}\nabla \log p_\theta\big)(y)\mathrm{d}\mathbb{P}_\theta(x)\mathrm{d}\mathbb{P}_\theta(y) \\
&= -\int_{\mathcal{X}} \int_{\mathcal{X}} \big(m_\theta^\top \partial_{\theta^i}\nabla \log p_\theta\big)^\top(y)K(y,x)\big(m_\theta^\top \partial_{\theta^j}\nabla \log p_\theta\big)(x)\mathrm{d}\mathbb{P}_\theta(y)\mathrm{d}\mathbb{P}_\theta(x) \\
&= -\int_{\mathcal{X}} \int_{\mathcal{X}} \big(m_\theta^\top \partial_{\theta^i}\nabla \log p_\theta\big)^\top(y)K(x,y)^\top\big(m_\theta^\top \partial_{\theta^j}\nabla \log p_\theta\big)(x)\mathrm{d}\mathbb{P}_\theta(y)\mathrm{d}\mathbb{P}_\theta(x) \\
&= -\int_{\mathcal{X}} \int_{\mathcal{X}} \big(m_\theta^\top \partial_{\theta^j}\nabla \log p_\theta\big)(x)^\top K(x,y)\big(m_\theta^\top \partial_{\theta^i}\nabla \log p_\theta\big)(y)\mathrm{d}\mathbb{P}_\theta(y)\mathrm{d}\mathbb{P}_\theta(x).
\end{aligned}
$$

Hence, we conclude that

$$
\tfrac{1}{2}\partial_{\alpha^i}\partial_{\theta^j}\,\mathrm{DKSD}_{K,m}(\mathbb{P}_\alpha,\mathbb{P}_\theta)^2 = -\int_{\mathcal{X}} \int_{\mathcal{X}} \big(m_\theta^\top \partial_{\theta^j}\nabla \log p_\theta\big)(x)^\top K(x,y)\big(m_\theta^\top \partial_{\theta^i}\nabla \log p_\theta\big)(y)\mathrm{d}\mathbb{P}_\theta(y)\mathrm{d}\mathbb{P}_\theta(x)
$$

The information tensor is positive semi-definite. Indeed writing $V_\theta(y) \equiv m_\theta^\top(y)\nabla_y\langle v, \nabla_\theta \log p_\theta\rangle$:

$$
\begin{aligned}
\langle v, g(\theta)v\rangle &= v^i g_{ij}(\theta)v^j \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} \big(m_\theta^\top(x)\nabla_x\langle v, \nabla_\theta \log p_\theta\rangle\big)^\top K(x,y)\big(m_\theta^\top(y)\nabla_y\langle v, \nabla_\theta \log p_\theta\rangle\big)\mathrm{d}\mathbb{P}_\theta(x)\mathrm{d}\mathbb{P}_\theta(y) \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} \big\langle m_\theta^\top(x)\nabla_x\langle v, \nabla_\theta \log p_\theta\rangle, K(x,y)m_\theta^\top(y)\nabla_y\langle v, \nabla_\theta \log p_\theta\rangle\big\rangle\mathrm{d}\mathbb{P}_\theta(x)\mathrm{d}\mathbb{P}_\theta(y) \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}} \langle V_\theta(x), K(x,y)V_\theta(y)\rangle\mathrm{d}\mathbb{P}_\theta(x)\mathrm{d}\mathbb{P}_\theta(y) \geq 0
\end{aligned}
$$

since K is IPD. ■

C.2 Information Semi-Metric of Diffusion Score Matching

A similar calculation allows us to derive the metric tensor for DSM. The proposition below generalises [35], who derived the metric tensor for SM.

Proposition 7 (Information Tensor DSM). *The information tensor defined by DSM is positive semi-definite and has components*

$$
g_{\mathrm{DSM}}(\theta)_{ij} = \int_{\mathcal{X}} \big\langle m^\top \nabla\partial_{\theta^i}\log p_\theta, m^\top\nabla\partial_{\theta^j}\log p_\theta\big\rangle \mathrm{d}\mathbb{P}_\theta.
$$

Proof The information metric is given by $g(\theta)_{ij} = -\frac{1}{2}\frac{\partial^2}{\partial\alpha^i\partial\theta^j}\mathrm{DSM}(p_\alpha\|p_\theta)|_{\alpha=\theta}$. Recall

$$
\mathrm{DSM}(p_\alpha\|p_\theta) = \int_{\mathcal{X}} \big\|m^\top(\nabla\log p_\theta - \nabla\log p_\alpha)\big\|_2^2 p_\alpha \mathrm{d}x.
$$

Moreover

$$
\begin{aligned}
\tfrac{1}{2}\partial_{\alpha^i}\partial_{\theta^j}\,\mathrm{DSM}(p_\alpha\|p_\theta)\big|_{\alpha=\theta} &= \tfrac{1}{2}\partial_{\alpha^i}\partial_{\theta^j}\int_{\mathcal{X}} \big\|m^\top(\nabla\log p_\theta - \nabla\log p_\alpha)\big\|_2^2 p_\alpha \mathrm{d}x\big|_{\alpha=\theta} \\
&= \partial_{\alpha^i}\int_{\mathcal{X}} \big(m^\top(\nabla\log p_\theta - \nabla\log p_\alpha)\big)\cdot\big(m^\top\partial_{\theta^j}\nabla\log p_\theta\big)p_\alpha\mathrm{d}x\big|_{\alpha=\theta} \\
&= \int_{\mathcal{X}} \big(m^\top(\nabla\log p_\theta - \nabla\log p_\alpha)\big)\cdot\big(m^\top\partial_{\theta^j}\nabla\log p_\theta\big)\partial_{\alpha^i}p_\alpha\mathrm{d}x\big|_{\alpha=\theta} \\
&\quad - \int_{\mathcal{X}} \big(m^\top\partial_{\alpha^i}\nabla\log p_\alpha\big)\cdot\big(m^\top\partial_{\theta^j}\nabla\log p_\theta\big)p_\alpha\mathrm{d}x\big|_{\alpha=\theta} \\
&= -\int_{\mathcal{X}} \big(m^\top\partial_{\theta^i}\nabla\log p_\theta\big)\cdot\big(m^\top\partial_{\theta^j}\nabla\log p_\theta\big)\mathrm{d}\mathbb{P}_\theta.
\end{aligned}
$$

Finally g is semi-positive definite,

$$
\begin{aligned}
\langle v, g(\theta)v\rangle = v^i g_{ij}(\theta)v^j &= \int_{\mathcal{X}} v^i m_{rs}^\top\partial_{x^s}\partial_{\theta^i}\log p_\theta m_{rl}^\top\partial_{x^l}\partial_{\theta^j}\log p_\theta v^j\mathrm{d}\mathbb{P}_\theta \\
&= \int_{\mathcal{X}} m_{rs}^\top\partial_{x^s}\langle v, \nabla_\theta\log p_\theta\rangle m_{rl}^\top\partial_{x^l}\langle v, \nabla_\theta\log p_\theta\rangle\mathrm{d}\mathbb{P}_\theta \\
&= \int_{\mathcal{X}} \big\langle m^\top\nabla_x\langle v, \nabla_\theta\log p_\theta\rangle, m^\top\nabla_x\langle v, \nabla_\theta\log p_\theta\rangle\big\rangle\mathrm{d}\mathbb{P}_\theta \\
&= \int_{\mathcal{X}} \|m^\top\nabla_x\langle v, \nabla_\theta\log p_\theta\rangle\|^2\mathrm{d}\mathbb{P}_\theta \geq 0
\end{aligned}
$$

■

D Proofs of Consistency and Asymptotic Normality

In this appendix, we prove several results concerning the consistency and asymptotic normality of DKSD and DSM estimators.

D.1 Diffusion Kernel Stein Discrepancies

Given the Stein kernel (3) we want to estimate $\theta^* \equiv \text{argmin}_{\theta\in\Theta} \text{DKSD}_{K,m}(\mathbb{Q},\mathbb{P}_\theta)^2 = \text{argmin}_{\theta\in\Theta} \int_{\mathcal{X}}\int_{\mathcal{X}} k^0_\theta(x,y)\mathbb{Q}(\mathrm{d}x)\mathbb{Q}(\mathrm{d}y)$ using a sequence of estimators $\hat{\theta}^{\text{DKSD}}_n \in \text{argmin}_{\theta\in\Theta}\widehat{\text{DKSD}}_{K,m}(\mathbb{Q},\mathbb{P}_\theta)^2$ that minimise the U-statistic approximation (4). We will assume we are in the specified setting $\mathbb{Q}=\mathbb{P}_{\theta^*}\in\mathcal{P}_\Theta$. In the misspecified setting it is necessary to further assume the existence of a unique minimiser.

D.1.1 Strong Consistency

We first prove a general strong consistency result based on an equicontinuity assumption:

Lemma 2. *Let $\mathcal{X}=\mathbb{R}^d$. Suppose $\{\theta\mapsto k^0_\theta(x,y)\},\{\theta\mapsto\mathbb{Q}_z k^0_\theta(x,z)\}$ are equicontinuous on any compact subset $C\subset\Theta$ for x,y in a sequence of sets whose union has full $\mathbb{Q}$-measure, and $\|s_{p_\theta}(x)\|\leq f_1(x)$, $\|\nabla_x\cdot m_\theta(x)\|\leq f_2(x)$, $\|\nabla_x\cdot(m_\theta(x)K(x,y))\|\leq f_3(x,y)$, $|\text{Tr}[m(y)\nabla_y\nabla_x\cdot(m(x)K)]|\leq f_4(x,y)$ hold on C, where $f_1(x)\sqrt{K(x,x)_{ii}}\in L^1(\mathbb{Q})$, and $f_4, f_3f_2, f_1f_3\in L^1(\mathbb{Q}\otimes\mathbb{Q})$. Assume further that $\theta\mapsto\mathbb{P}_\theta$ is injective. Then we have a unique minimiser θ^*, and if either Θ is compact, or $\theta^*\in int(\Theta)$ and Θ and $\theta\mapsto\widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n,\mathbb{P}_\theta)^2$ are convex, then $\hat{\theta}^{\text{DKSD}}_n$ is strongly consistent.*

Proof

Note $\text{DKSD}_{K,m}(\mathbb{Q},\mathbb{P}_\theta)^2=0$ iff $\mathbb{P}_\theta=\mathbb{P}_{\theta^*}$ by Proposition 4, which implies $\theta=\theta^*$ since $\theta\mapsto\mathbb{P}_\theta$ is injective. Thus we have a unique minimiser at θ^*.

Suppose first Θ is compact and take $C=\Theta$. Note

$$\begin{aligned}|k^0(x,y)|\leq&|\langle s_p(x),Ks_p(y)\rangle|+|\langle\nabla_y\cdot(m(y)K),s_p(x)\rangle|+|\langle\nabla_x\cdot(m(x)K),s_p(y)\rangle|\\&+|\langle\nabla_x\cdot(m(x)K),\nabla_y\cdot m\rangle|+|\text{Tr}[m(y)\nabla_y\nabla_x\cdot(m(x)K)]|\\&\leq|\langle s_p(x),Ks_p(y)\rangle|+f_3(y,x)f_1(x)+f_3(x,y)f_1(y)+f_3(x,y)f_2(y)+f_4(x,y),\end{aligned}$$

From the reproducing property $f(x)=\langle f,K(\cdot,x)v\rangle_{\mathcal{H}^d}$, for any $f\in\mathcal{H}^d$, $v\in\mathbb{R}^d$. Using $K(y,x)=K(x,y)^\top$ we have $K(\cdot,x)_{,i}=K(x,\cdot)_{i,}$, where $K(\cdot,x)_{,i}$ and $K(x,\cdot)_{i,}$ denote the i^{th} column and row respectively, which implies that $K(x,\cdot)_{i,},K(\cdot,x)_{,i}\in\mathcal{H}^d$ and $f(x)_i=\langle f,K(\cdot,x)_{,i}\rangle_{\mathcal{H}^d}$. Choosing $f=K(\cdot,y)_{,j}$ implies

$$\begin{aligned}K(x,y)_{ij}=\langle K(\cdot,y)_{,j},K(\cdot,x)_{,i}\rangle_{\mathcal{H}^d}&\leq\|K(\cdot,y)_{,j}\|_{\mathcal{H}^d}\|K(\cdot,x)_{,i}\|_{\mathcal{H}^d}\\&=\sqrt{\langle K(\cdot,y)_{,j},K(\cdot,y)_{,j}\rangle_{\mathcal{H}^d}}\sqrt{\langle K(\cdot,x)_{,i},K(\cdot,x)_{,i}\rangle_{\mathcal{H}^d}}\\&=\sqrt{K(y,y)_{jj}}\sqrt{K(x,x)_{ii}}.\end{aligned}$$

It follows that

$$\begin{aligned}\langle s_p(x),Ks_p(y)\rangle=(s_p)_i(x)K(x,y)_{ij}(s_p)_j(y)&\leq(s_p)_i(x)\sqrt{K(x,x)_{ii}}\sqrt{K(y,y)_{jj}}(s_p)_j(y)\\&\leq\|s_p(x)\|_\infty\sqrt{K(x,x)_{ii}}\sqrt{K(y,y)_{jj}}\|s_p(y)\|_\infty\\&\leq Cf_1(x)\sqrt{K(x,x)_{ii}}\sqrt{K(y,y)_{jj}}f_1(y),\end{aligned}$$

where the constant $C>0$ arises from the norm-equivalence of $\|s_p(y)\|$ and $\|s_p(y)\|_\infty$. Hence k^0 is integrable. Thus by theorem 1 [61],

$$\sup_\theta\left|\widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n,\mathbb{P}_\theta)^2-\text{DKSD}_{K,m}(\mathbb{Q},\mathbb{P}_\theta)^2\right|\xrightarrow{a.s.}0$$

and $\theta\mapsto\text{DKSD}_{K,m}(\mathbb{Q},\mathbb{P}_\theta)^2$ are continuous. By theorem 2.1 [48] then $\hat{\theta}^{\text{DKSD}}_n\xrightarrow{a.s.}\theta^*$.

On the other hand, if Θ is convex we follow a similar strategy to the proof of theorem 2.7 [48]. Since $\theta^*\in\text{int}(\Theta)$, we can find a $\epsilon>0$ for which $C=\overline{B}(\theta^*,2\epsilon)\subset\Theta$ is a closed ball containing θ^* (which is compact since $\Theta\subset\mathbb{R}^m$). Using the compact case, we know any sequence of estimators $\tilde{\theta}^{\text{DKSD}}_n\in\text{argmin}_{\theta\in C}\widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n,\mathbb{P}_\theta)^2$ is strongly consistent for θ^*. In particular, there exists N_0 a.s. s.t. for $n>N_0$, $\|\tilde{\theta}^{\text{DKSD}}_n-\theta^*\|<\epsilon$. If $\theta\notin C$, there exists $\lambda\in[0,1)$ s.t. $\lambda\tilde{\theta}^{\text{DKSD}}_n+(1-\lambda)\theta$ lies on the boundary of the closed ball C. Using convexity and the fact $\tilde{\theta}^{\text{DKSD}}_n$ is a minimiser over C, $\widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n,\mathbb{P}_{\tilde{\theta}^{\text{DKSD}}_n})^2\leq\widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n,\mathbb{P}_{\lambda\tilde{\theta}^{\text{DKSD}}_n+(1-\lambda)\theta})^2\leq\lambda\widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n,\mathbb{P}_{\tilde{\theta}^{\text{DKSD}}_n})^2+(1-\lambda)\widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n,\mathbb{P}_\theta)^2$

which implies $\widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_{\tilde{\theta}_n^{\text{DKSD}}})^2 \leq \widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2$ and $\tilde{\theta}_n^{\text{DKSD}}$ is the global minimum of $\theta \mapsto \widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2$ for $n > N_0$. ■

When k^0 is Fréchet differentiable on Θ equicontinuity can be obtained using the Mean value theorem, which simplifies the assumptions under which strong consistency holds.

Theorem 2 (Strong Consistency DKSD). *Let $\mathcal{X} = \mathbb{R}^d$, $\Theta \subset \mathbb{R}^m$. Suppose that K is bounded with bounded derivatives up to order 2, that $k^0(x, y)$ is continuously-differentiable on an $\mathbb{R}^m$-open neighbourhood of Θ, and that for any compact subset $C \subset \Theta$ there exist functions f_1, f_2, g_1, g_2 such that*

1. *$\|m^\top(x)\nabla \log p_\theta(x)\| \leq f_1(x)$, where $f_1 \in L^1(\mathbb{Q})$ and continuous.*
2. *$\|\nabla_\theta\big(m(x)^\top \nabla \log p_\theta(x)\big)\| \leq g_1(x)$, where $g_1 \in L^1(\mathbb{Q})$ is continuous.*
3. *$\|m(x)\| + \|\nabla_x m(x)\| \leq f_2(x)$ where $f_2 \in L^1(\mathbb{Q})$ and continuous.*
4. *$\|\nabla_\theta m(x)\| + \|\nabla_\theta \nabla_x m(x)\| \leq g_2(x)$ where $g_2 \in L^1(\mathbb{Q})$ is continuous.*

Assume further that $\theta \mapsto \mathbb{P}_\theta$ is injective. Then we have a unique minimiser θ^*, and if either Θ is compact, or $\theta^* \in int(\Theta)$ and Θ and $\theta \mapsto \widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2$ are convex, then $\hat{\theta}_n^{\text{DKSD}}$ is strongly consistent.

Proof Let $\|K\| + \|\nabla_x K\| + \|\nabla_x \nabla_y K\| \leq K_\infty$. Note $\|\nabla_y \cdot (m(y)K)\| \leq 2f_2(y)K_\infty$ and $|\text{Tr}[m(y)\nabla_y \nabla_x \cdot (m(x)K)]| \leq 2f_2(y)f_2(x)K_\infty$ so

$$|k_\theta^0(x, y)| \leq f_1(x)K_\infty f_1(y) + 2f_2(x)K_\infty f_1(y) + 2f_2(y)K_\infty f_1(x) + 3K_\infty f_2(x)f_2(y)$$

which is symmetric and integrable by assumption. Let S_m, $m = 1, 2, \ldots$ be an increasing sequence of closed balls in $\mathbb{R}^d$, such that $\cup_{m=1}^\infty S_m = \mathbb{R}^d$. Moreover,

$$\begin{aligned}
\|\nabla_\theta \langle s_p(x), K s_p(y)\rangle\| &\leq g_1(x)f_1(y)K_\infty + g_1(y)f_1(x)K_\infty \\
\|\nabla_\theta \langle \nabla_y \cdot (m(y)K), s_p(x)\rangle\| &\leq 2K_\infty g_2(y)f_1(x) + 2f_2(y)g_1(x)K_\infty \\
\|\nabla_\theta \langle \nabla_x \cdot (m(x)K), \nabla_y \cdot m\rangle\| &\leq 2K_\infty g_2(x)f_2(y) + 2K_\infty f_2(x)g_2(y) \\
\|\nabla_\theta \text{Tr}[m(y)\nabla_y \nabla_x \cdot (m(x)K)]\| &\leq 2K_\infty g_2(y)f_2(x) + 2K_\infty f_2(y)g_2(x)
\end{aligned}$$

thus $\|\nabla_\theta k_\theta^0(x, y)\|$ is bounded above by a continuous integrable symmetric function, $(x, y) \mapsto s(x, y)$, which attains a maximum on the compact spaces $S_m \times S_m$. By the MVT applied on the $\mathbb{R}^m$-open neighbourhood of Θ, $|k_\theta^0(x, y) - k_\alpha^0(x, y)| \leq \|\nabla_\theta k_\theta^0(x, y)\|\|\theta - \alpha\| \leq s(x, y)\|\theta - \alpha\| \leq \max_{x,y \in S_m} s(x, y)\|\theta - \alpha\|$, and $k_\theta^0(x, y)$ is equicontinuous in $\theta \in C$ for $x, y \in S_m$. Similarly, since s is integrable, $|\int_{\mathcal{X}} k_\theta^0(x, y)\mathbb{Q}(dy) - \int_{\mathcal{X}} k_\alpha^0(x, z)\mathbb{Q}(dz)| \leq \|\nabla_\theta \int_{\mathcal{X}} k_\theta^0(x, z)\text{d}\mathbb{Q}(z)\|\|\theta - \alpha\| \leq \int_{\mathcal{X}} \|\nabla_\theta k_\theta^0(x, z)\|\text{d}\mathbb{Q}(z)\|\theta - \alpha\| \leq \max_{x \in S_m} \mathbb{Q}_z s(x, z)\|\theta - \alpha\| \leq$ is equicontinuous in $\theta \in C$ for $x \in S_m$. The rest follows as in the previous proposition. ■

D.1.2 Asymptotic Normality

Theorem 10. *Let $\mathcal{X}$ and Θ be open subsets of $\mathbb{R}^d$ and $\mathbb{R}^m$ respectively. Let K be a bounded kernel with bounded derivatives up to order 2 and suppose that $\hat{\theta}_n^{\text{DKSD}} \xrightarrow{p} \theta^*$ and that there exists a compact neighbourhood $\mathcal{N} \subset \Theta$ of θ^* such that $\theta \to \widehat{\text{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2$ is twice continuously $\mathbb{R}^m$-differentiable in $\mathcal{N}$ and for $\theta \in \mathcal{N}$,*

1. *$\|m^\top(x)\nabla \log p_\theta(x)\| + \|\nabla_\theta\big(m(x)^\top \nabla \log p_\theta(x)\big)\| \leq f_1(x)$, where $f_1 \in L^2(\mathbb{Q})$ and continuous.*
2. *$\|m(x)\| + \|\nabla_x m(x)\| + \|\nabla_\theta m(x)\| + \|\nabla_\theta \nabla_x m(x)\| \leq f_2(x)$ where $f_2 \in L^2(\mathbb{Q})$ and continuous.*
3. *$\|\nabla_\theta \nabla_\theta\big(m(x)^\top \nabla \log p_\theta(x)\big)\| + \|\nabla_\theta \nabla_\theta \nabla_\theta\big(m(x)^\top \nabla \log p_\theta(x)\big)\| \leq g_1(x)$, where $g_1 \in L^1(\mathbb{Q})$ and continuous.*

4. $\|\nabla_\theta\nabla_\theta m(x)\|+\|\nabla_\theta\nabla_\theta\nabla_x m(x)\|+\|\nabla_\theta\nabla_\theta\nabla_\theta m(x)\|+\|\nabla_\theta\nabla_\theta\nabla_\theta\nabla_x m(x)\| \le g_2(x)$ *where* $g_2 \in L^1(\mathbb{Q})$ *and continuous.*

Suppose also that the information tensor g is invertible at θ^*. Then :

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

where

$$\Sigma = \int_{\mathcal{X}} \mathrm{d}\mathbb{Q}(x)\big(\int_{\mathcal{X}} \mathrm{d}\mathbb{Q}(y)\nabla_{\theta^*}k_\theta^0(x,y)\big) \otimes \big(\int_{\mathcal{X}} \mathrm{d}\mathbb{Q}(z)\nabla_{\theta^*}k_\theta^0(x,z)\big).$$

Proof Note that $\nabla_\theta\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n,\mathbb{P}_\theta)^2 = \frac{1}{N(N-1)}\sum_{i\neq j}\nabla_\theta k_\theta^0(X_i,X_j)$. Let $\mu(\theta) \equiv \mathbb{Q}\otimes\mathbb{Q}[\nabla_\theta k_\theta^0]$. Assumptions 1 and 2 imply that $\mathbb{Q}\otimes\mathbb{Q}[\|\nabla_\theta k_\theta^0\|^2] < \infty$. By [27, Theorem 7.1] it follows that

$$\sqrt{n}\Big(\nabla_\theta\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n,\mathbb{P}_\theta)^2 - \mu(\theta)\Big) \xrightarrow{d} \mathcal{N}(0, 4\Sigma(\theta))$$

where

$$\begin{aligned}\Sigma &= \mathbb{Q}\big[\mathbb{Q}_2\big[\nabla_\theta k_\theta^0 - \mu(\theta)\big]\otimes\mathbb{Q}_2\big[\nabla_\theta k_\theta^0 - \mu(\theta)\big]\big]\\ &= \int_{\mathcal{X}}\big(\int_{\mathcal{X}}\nabla_\theta k_\theta^0(x,y)\mathrm{d}\mathbb{Q}(y) - \mu(\theta)\big)\otimes\big(\int_{\mathcal{X}}\nabla_\theta k_\theta^0(x,z)\mathrm{d}\mathbb{Q}(z) - \mu(\theta)\big)\mathrm{d}\mathbb{Q}(x)\end{aligned}$$

Note that $\mu(\theta^*) = \mathbb{Q}\otimes\mathbb{Q}[\nabla_\theta k_\theta^0|_{\theta^*}] = \nabla_\theta\big(\mathbb{Q}\otimes\mathbb{Q}[k_\theta^0]\big)|_{\theta=\theta^*}$, and if $\mathbb{Q}\otimes\mathbb{Q}[k_\theta^0]$ is differentiable around θ^*, then the first order optimality condition implies $\mu(\theta^*) = 0$.

Consider now $\nabla_\theta\nabla_\theta\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\},\mathbb{P}_\theta)^2 = \frac{1}{n(n-1)}\sum_{i\neq j}\nabla_\theta\nabla_\theta k_\theta^0(X_i,X_j)$. Note

$$\begin{aligned}\|\nabla_\theta\nabla_\theta\nabla_\theta\langle s_p(x), Ks_p(y)\rangle\| &\lesssim g_1(x)K_\infty f_1(y) + f_1(x)K_\infty g_1(y)\\ \|\nabla_\theta\nabla_\theta\nabla_\theta\langle\nabla_y\cdot(m(y)K), s_p(x)\rangle\| &\lesssim g_2(y)K_\infty f_1(x) + f_2(y)K_\infty g_1(x)\\ \|\nabla_\theta\nabla_\theta\nabla_\theta\langle\nabla_x\cdot(m(x)K),\nabla_y\cdot m\rangle\| &\lesssim f_2(y)K_\infty g_2(x) + g_2(y)K_\infty f_2(x)\\ \|\nabla_\theta\nabla_\theta\nabla_\theta\mathrm{Tr}[m(y)\nabla_y\nabla_x\cdot(m(x)K)] &\lesssim g_2(y)K_\infty f_2(x) + f_2(y)K_\infty g_2(x)\end{aligned}$$

Hence by Assumptions 1-4 $\|\nabla_\theta\nabla_\theta\nabla_\theta k_\theta^0\|$ is bounded above by a continuous integrable symmetric function and we can apply the MVT to show equicontinuity as in the proof of Proposition 2. Moreover the conditions of [61, Theorem 1] hold for the components of $\nabla_\theta\nabla_\theta k_\theta^0$, so that $\sup_{\theta\in\mathcal{N}}\left|\frac{1}{n(n-1)}\sum_{i\neq j}\partial_{\theta^a}\partial_{\theta^b}k_\theta^0(X_i,X_j) - \mathbb{Q}\otimes\mathbb{Q}\partial_{\theta^a}\partial_{\theta^b}k_\theta^0\right| \xrightarrow{a.s.} 0$ as $n\to\infty$, for all a and b.

Finally we observe that $\mathbb{Q}\otimes\mathbb{Q}\partial_{\theta^a}\partial_{\theta^b}k_\theta^0\big|_{\theta=\theta^*} = g_{ab}(\theta^*)$, where g is the information metric associated with $\mathrm{DKSD}_{K,m}$. Indeed using $\delta_{p,q} = 0$ if $p = q$

$$\begin{aligned}&\mathbb{Q}\otimes\mathbb{Q}\partial_{\theta^a}\partial_{\theta^b}k_\theta^0\big|_{\theta=\theta^*}\\ &= \partial_{\theta^a}\partial_{\theta^b}\int_{\mathcal{X}}\int_{\mathcal{X}}p_{\theta^*}(x)\delta_{p_\theta,p_{\theta^*}}(x)^\top K(x,y)\delta_{p_\theta,p_{\theta^*}}(y)p_{\theta^*}(y)\mathrm{d}x\mathrm{d}y\big|_{\theta=\theta^*}\\ &= \partial_{\theta^a}\int_{\mathcal{X}}\int_{\mathcal{X}}p_{\theta^*}(x)\partial_{\theta^b}\delta_{p_\theta,p_{\theta^*}}(x)^\top K(x,y)\delta_{p_\theta,p_{\theta^*}}(y)p_{\theta^*}(y)\mathrm{d}x\mathrm{d}y\big|_{\theta=\theta^*}\\ &+ \partial_{\theta^a}\int_{\mathcal{X}}\int_{\mathcal{X}}p_{\theta^*}(x)\delta_{p_\theta,p_{\theta^*}}(x)^\top K(x,y)\partial_{\theta^b}\delta_{p_\theta,p_{\theta^*}}(y)p_{\theta^*}(y)\mathrm{d}x\mathrm{d}y\big|_{\theta=\theta^*}\\ &= \int_{\mathcal{X}}\int_{\mathcal{X}}p_{\theta^*}(x)\partial_{\theta^b}\delta_{p_\theta,p_{\theta^*}}(x)^\top K(x,y)\partial_{\theta^a}\delta_{p_\theta,p_{\theta^*}}(y)p_{\theta^*}(y)\mathrm{d}x\mathrm{d}y\big|_{\theta=\theta^*}\\ &+ \int_{\mathcal{X}}\int_{\mathcal{X}}p_{\theta^*}(x)\partial_{\theta^a}\delta_{p_\theta,p_{\theta^*}}(x)^\top K(x,y)\partial_{\theta^b}\delta_{p_\theta,p_{\theta^*}}(y)p_{\theta^*}(y)\mathrm{d}x\mathrm{d}y\big|_{\theta=\theta^*}\\ &= 2\int_{\mathcal{X}}\int_{\mathcal{X}}\big(m_{\theta^*}^\top(x)\nabla_x\partial_{\theta^{j*}}\log p_{\theta^*}\big)^\top K(x,y)\big(m_{\theta^*}^\top(y)\nabla_y\partial_{\theta^{i*}}\log p_{\theta^*}\big)\mathrm{d}\mathbb{P}_{\theta^*}(x)\mathrm{d}\mathbb{P}_{\theta^*}(y),\end{aligned}$$

so $\mathbb{Q}\otimes\mathbb{Q}\partial_{\theta^a}\partial_{\theta^b}k_\theta^0\big|_{\theta=\theta^*} = g_{ab}(\theta^*)$. The conditions of [48, Theorem 3.1] hold, from which the advertised result follows. ■

D.2 Diffusion Score Matching

Recall that the DSM is given by:

$$\mathrm{DSM}(\mathbb{Q}\|\mathbb{P}_\theta) = \int_{\mathcal{X}}\Big(\big\|m^\top\nabla_x\log p_\theta\big\|_2^2 + \|m^\top\nabla\log q\|_2^2 + 2\nabla\cdot\big(mm^\top\nabla\log p_\theta\big)\Big)\mathrm{d}\mathbb{Q}$$

and we wish to estimate

$$\theta^* = \mathrm{argmin}_{\theta\in\Theta} \int_{\mathcal{X}} \Big(\big\|m^\top \nabla_x \log p_\theta\big\|_2^2 + 2\nabla\cdot\big(mm^\top \nabla \log p_\theta\big)\Big) \mathrm{d}\mathbb{Q} \equiv \mathrm{argmin}_{\theta\in\Theta} \int_{\mathcal{X}} F_\theta \mathrm{d}\mathbb{Q}$$

with a sequence of M-estimators $\hat{\theta}_n^{\mathrm{DSM}} = \mathrm{argmin}_{\theta\in\Theta} \frac{1}{n}\sum_i^n F_\theta(X_i)$. Recall also we have

$$F_\theta(x) = \big\|m^\top \nabla_x \log p_\theta\big\|_2^2 + 2\big\langle \nabla\cdot(mm^\top), \nabla \log p_\theta\big\rangle + 2\mathrm{Tr}\big[mm^\top \nabla^2 \log p_\theta\big].$$

We will have a unique minimiser θ^* whenever the map $\theta \mapsto \mathbb{P}_\theta$ is injective.

D.2.1 Weak Consistency of DSM

Theorem 11. *Suppose $\mathcal{X}$ be open subset of $\mathbb{R}^d$, and $\Theta \subset \mathbb{R}^m$. Suppose $\log p_\theta(\cdot)$ is $C^2(\mathcal{X})$ and $m \in C^1(\mathcal{X})$, and $\|\nabla_x \log p_\theta(x)\| \le f_1(x)$. Suppose also that $\|\nabla_x\nabla_x \log p_\theta(x)| \le f_2(x)$ on any compact set $C \subset \Theta$, where $\|m^\top\| f_1 \in L^2(\mathbb{Q})$, $\|\nabla\cdot(mm^\top)\| f_1 \in L^1(\mathbb{Q})$, $\|mm^\top\|_\infty f_2 \in L^1(\mathbb{Q})$. If either Θ is compact, or Θ and $\theta\mapsto F_\theta$ are convex and $\theta^* \in int(\Theta)$, then $\hat{\theta}_n^{\mathrm{DSM}}$ is weakly consistent for θ^*.*

Proof By assumption $\theta \mapsto F_\theta(x)$ is continuous. Suppose Θ is compact, taking $C = \Theta$, note

$$\begin{aligned}
|F_\theta| &= \Big| \big\|m^\top \nabla_x \log p_\theta\big\|_2^2 + 2\nabla\cdot\big(mm^\top\nabla \log p_\theta\big)\Big| \\
&= \Big| \big\|m^\top \nabla_x \log p_\theta\big\|_2^2 + 2\big(\nabla\cdot(mm^\top)\cdot\nabla \log p_\theta + \mathrm{Tr}\big[mm^\top\nabla^2 \log p_\theta\big]\big)\Big| \\
&\lesssim \|m^\top\|^2 f_1^2 + 2\|\nabla\cdot(mm^\top)\| f_1 + 2\|mm^\top\|_\infty f_2
\end{aligned}$$

which is integrable, so the conditions of Lemma 2.4 [48] are satisfied so $\theta \mapsto \mathbb{Q}F_\theta$ is continuous, and $\sup_\Theta |\frac{1}{n}\sum_i^n F_\theta(X_i) - \mathbb{Q}F_\theta| \xrightarrow{p} 0$, and thus from theorem 2.1 [48] $\hat{\theta}_n^{\mathrm{DSM}} \xrightarrow{p} \theta^*$. If Θ is convex, note that the sum of convex functions is convex, so $\theta \mapsto \frac{1}{n}\sum_i^n F_\theta(X_i)$ is convex, and we can follow a derivation analogous to the one in Theorem 4. ■

D.2.2 Asymptotic Normality of DSM

Theorem 12. *Suppose $\mathcal{X}, \Theta$ be open subsets of $\mathbb{R}^d$ and $\mathbb{R}^m$ respectively. If **(i)** $\hat{\theta}_n^{\mathrm{DSM}} \xrightarrow{p} \theta^*$, **(ii)** $\theta \mapsto \log p_\theta(x)$ is twice continuously differentiable on a closed ball $\bar{B}(\epsilon, \theta^*) \subset \Theta$, and*

- ***(iii)*** *$\|mm^\top\| + \|\nabla_x\cdot(mm^\top)\| \le f_1(x)$, and $\|\nabla_x \log p\| + \|\nabla_{\theta^*}\nabla_x \log p\| + \|\nabla_{\theta^*}\nabla_x\nabla_x \log p\| \le f_2(x)$, with $f_1 f_2^2 \in L^2(\mathbb{Q})$*
- ***(iv)*** *for $\theta \in \bar{B}(\epsilon, \theta^*)$ $\|\nabla_\theta\nabla_x \log p\|^2 + \|\nabla_x \log p\|\|\nabla_\theta\nabla_\theta\nabla_x \log p\| + \|\nabla_\theta\nabla_\theta\nabla_x \log p\| + \|\nabla_\theta\nabla_\theta\nabla_x\nabla_x \log p\| \le g_1(x)$, and $f_1 g_2 \in L^1(\mathbb{Q})$,*

*and **(v)** and the information tensor is invertible at θ^*. Then*

$$\sqrt{n}\Big(\hat{\theta}_n^{\mathrm{DSM}} - \theta^*\Big) \xrightarrow{d} \mathcal{N}\big(0, g^{-1}(\theta^*)\mathbb{Q}[\nabla_{\theta^*}F_\theta \otimes \nabla_{\theta^*}F_\theta] g^{-1}(\theta^*)\big)$$

Proof From **(ii)** $\theta \mapsto F_\theta$ is twice continuously differentiable on a ball $B(\epsilon, \theta^*) \subset \Theta$. Note $\nabla_\theta \frac{1}{N}\sum_i^N F_\theta(X_i) = \frac{1}{N}\sum_i^N \nabla_\theta F_\theta(X_i)$, then by **(i,ii)** and the first order optimality condition $\mathbb{Q}[\nabla_{\theta^*}F_\theta(X_i)] = \nabla_{\theta^*}\mathbb{Q}[F_\theta(X_i)] = 0$. Note

$$\begin{aligned}
\|\nabla_{\theta^*}F_\theta(x)\| &\lesssim \|mm^\top\|\|\nabla_x \log p\|\|\nabla_{\theta^*}\nabla_x \log p\| + \|\nabla_x\cdot(mm^\top)\|\|\nabla_{\theta^*}\nabla_x \log p\| \\
&\quad + \|mm^\top\|\|\nabla_{\theta^*}\nabla_x\nabla_x \log p\| \\
&\lesssim f_1(x) f_2(x)[f_2(x) + 2].
\end{aligned}$$

Hence from **(iii)** $\nabla_{\theta^*}F_\theta \in L^2(\mathbb{Q})$, so by the CLT

$$\sqrt{n}\nabla_{\theta^*}\frac{1}{n}\sum_i^n F_\theta(X_i) \xrightarrow{d} \mathcal{N}(0, \mathbb{Q}[\nabla_{\theta^*}F_\theta \otimes \nabla_{\theta^*}F_\theta]).$$

From **(i)**, $\theta \mapsto \nabla_\theta \nabla_\theta F_\theta(x)$ is continuous on $\overline{B}(\epsilon, \theta^*)$. Moreover

$$\begin{aligned}\|\nabla_\theta \nabla_\theta F_\theta(x)\| \lesssim \|mm^\top\| \big(\|\nabla_\theta \nabla_x \log p\|^2 + \|\nabla_x \log p\| \|\nabla_\theta \nabla_\theta \nabla_x \log p\| \big) \\ + \|\nabla \cdot (mm^\top)\| \|\nabla_\theta \nabla_\theta \nabla_x \log p\| + \|mm^\top\| \|\nabla_\theta \nabla_\theta \nabla_x \nabla_x \log p\| \\ \lesssim f_1(x) g_1(x)\end{aligned}$$

thus from **(iv)**, the assumptions of Lemma 2.4 [48] applied to $\overline{B}(\epsilon, \theta^*)$ hold, and $\sup_{\overline{B}(\epsilon,\theta^*)} \big| \frac{1}{n} \sum_i^n \partial_{\theta^a} \partial_{\theta^b} F_\theta|_{\theta^*}(X_i) - \mathbb{Q} \partial_{\theta^a} \partial_{\theta^b} F_\theta|_{\theta^*} \big| \xrightarrow{p} 0$. As in Theorem 5 $\mathbb{Q} \partial_{\theta^a} \partial_{\theta^b} F_\theta|_{\theta^*} = g_{ab}(\theta^*)$ is the information tensor, which is continuous at θ^* by Lemma 2.4. The result follows by theorem 3.1 [48]. ■

D.3 Strong Consistency and Central Limit Theorems for Exponential Families

Let $\mathcal{X}$ be an open subset of $\mathbb{R}^d$, $\Theta \subset \mathbb{R}^m$. Consider the case when the density p lies in an exponential family, i.e. $p_\theta(x) \propto \exp(\langle \theta, T(x) \rangle_{\mathbb{R}^m} - c(\theta)) \exp(b(x))$, where $\theta \in \mathbb{R}^m$ and sufficient statistic $T = (T_1, \ldots, T_m) : \mathcal{X} \to \mathbb{R}^m$. Then $\nabla T \in \Gamma(\mathcal{X}, \mathbb{R}^{m \times d})$ and $\nabla_x \log p_\theta = \nabla_x b + \theta \cdot \nabla_x T$, $\nabla_\theta \nabla_x \log p_\theta = \nabla_x T^\top$.

D.3.1 Strong Consistency of the Minimum Diffusion Kernel Stein Discrepancy Estimator

We consider a RKHS $\mathcal{H}^d$ of functions $f : \mathcal{X} \to \mathbb{R}^d$ with matrix kernel K. Recall the Stein kernel is

$$\begin{aligned}k^0 &= \nabla_x \log p \cdot m(x) K m(y)^\top \nabla_y \log p + \nabla_x \cdot (m(x)K) \cdot \nabla_y \cdot m + \mathrm{Tr}[m(y) \nabla_y \nabla_x \cdot (m(x)K)] \\ &\quad + \nabla_y \cdot (m(y)K) \cdot m(x)^\top \nabla_x \log p + \nabla_x \cdot (m(x)K) \cdot m(y)^\top \nabla_y \log p\end{aligned}$$

Given a (i.i.d.) sample $X_i \sim \mathbb{Q}$, we can define an estimator using the U-statistic

$$\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2 = \tfrac{2}{n(n-1)} \textstyle\sum_{1 \leq i < j \leq n} k^0(X_i, X_j).$$

For the case where the density p lies in an exponential family, then $k^0 = \theta^\top A \theta + v^\top \theta + c$ where $A \in \Gamma(\mathcal{X} \times \mathcal{X}, \mathbb{R}^{m \times m}), v \in \Gamma(\mathcal{X} \times \mathcal{X}, \mathbb{R}^m)$ are given by (we set $\phi \equiv m^\top \nabla T^\top \in \Gamma(\mathcal{X}, \mathbb{R}^{d \times m})$)

$$\begin{aligned}A &= \phi(x)^\top K(x,y) \phi(y) \\ v^\top &= \nabla_y b \cdot m(y) K(y,x) \phi(x) + \nabla_x b \cdot m(x) K(x,y) \phi(y) \\ &\quad + \nabla_x \cdot (m(x)K) \cdot \phi(y) + \nabla_y \cdot (m(y)K) \cdot \phi(x) \\ c &= \nabla_x b \cdot m(x) K(x,y) m(y)^\top \nabla_y b + \nabla_x \cdot (m(x)K) \cdot \nabla_y \cdot m + \mathrm{Tr}[m(y) \nabla_y \nabla_x \cdot (m(x)K)] \\ &\quad + \nabla_y \cdot (m(y)K) \cdot m(x)^\top \nabla_x b + \nabla_x \cdot (m(x)K) \cdot m(y)^\top \nabla_y b\end{aligned}$$

Lemma 3. *Suppose K is IPD, that ∇T has linearly independent rows, that m is invertible, and $\|\phi\|_{L^1(\mathbb{Q})} < \infty$. Then the matrix $\int_\mathcal{X} A \mathbb{Q} \otimes \mathbb{Q}$ is symmetric positive definite.*

Proof The matrix $B = \int_\mathcal{X} A \mathbb{Q} \otimes \mathbb{Q}$ is symmetric

$$\begin{aligned}(\textstyle\int_\mathcal{X} A \mathbb{Q} \otimes \mathbb{Q})^\top &= \textstyle\int_\mathcal{X} A(x,y)^\top \mathbb{Q}(\mathrm{d}x) \otimes \mathbb{Q}(\mathrm{d}y) = \int_\mathcal{X} \nabla_y T m(y) K(x,y)^\top m(x)^\top \nabla_x T^\top \mathbb{Q}(\mathrm{d}x) \otimes \mathbb{Q}(\mathrm{d}y) \\ &= \textstyle\int_\mathcal{X} \nabla_y T m(y) K(y,x) m(x)^\top \nabla_x T^\top \mathbb{Q}(\mathrm{d}y) \otimes \mathbb{Q}(\mathrm{d}x) = \int_\mathcal{X} A \mathbb{Q} \otimes \mathbb{Q}.\end{aligned}$$

Moreover, set $\phi \equiv m^\top \nabla T^\top$, so $A(x,y) = \phi(x)^\top K(x,y) \phi(y)$. If $v \neq 0$, then $u \equiv \phi v \neq 0$ as $\nabla T^\top$ has full column rank (i.e., the vectors $\{\nabla T_i\}$ are linearly independent) and m is invertible, and $\|\phi v\|_{L^1(\mathbb{Q})} = \int_\mathcal{X} \|\phi(x) v\|_1 \mathrm{d}x \leq \|v\|_1 \int_\mathcal{X} \|\phi(x)\|_1 \mathrm{d}x < \infty$ implies $\mathrm{d}\mu_i \equiv u_i \mathrm{d}\mathbb{Q}$ is a finite signed Borel measure for each i. Clearly

$$\begin{aligned}v^\top (\textstyle\int_\mathcal{X} A \mathbb{Q} \otimes \mathbb{Q}) v &= \textstyle\int_\mathcal{X} u(x)^\top K(x,y) u(y) \mathbb{Q}(\mathrm{d}x) \mathbb{Q}(\mathrm{d}y) \\ &= \textstyle\int_\mathcal{X} K(x,y)_{ij} u_i(x) u_j(y) \mathbb{Q}(\mathrm{d}x) \mathbb{Q}(\mathrm{d}y) \\ &= \textstyle\int_\mathcal{X} K(x,y)_{ij} \mu_i(\mathrm{d}x) \mu_j(\mathrm{d}y) \geq 0.\end{aligned}$$

Moreover since the kernel is IPD, if this equals zero then for all i: $0 = \mu_i(C) = u_i\mathbb{Q}(C) = \phi_{ij}v_j\mathbb{Q}(C)$ for all measurable sets C, which implies $\phi v = 0$ and thus $v = 0$. ■

Theorem 3. *Suppose K is IPD with bounded derivative up to order 2, that ∇T has linearly independent rows, and m is invertible. Suppose $\|\phi\|, \|\nabla_x b\|\|m\|, \|\nabla_x m\| + \|m\| \in L^1(\mathbb{Q})$. The minimiser $\hat{\theta}_n^{\mathrm{DKSD}}$ of $\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)$ exists eventually, and converges almost surely to the minimiser θ^* of $\mathrm{DKSD}_{K,m}(\mathbb{Q}, \mathbb{P}_\theta)$.*

Proof

Let $X_i : \Omega \to \mathcal{X} \subset \mathbb{R}^d$ be independent $\mathbb{Q}$-distributed random vectors. The U-statistic $A_n \equiv \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} A(X_i, X_j)$ is symmetric semi-definite. Since $\int_{\mathcal{X}} \|A\| \mathrm{d}\mathbb{Q} \otimes \mathbb{Q} < \infty$, by theorem 1 [28] the components of A_n converge to the components of B almost surely, and since the matrix inverse is a continuous map, by the continuous mapping theorem the components of A_n^{-1} (the inverse exists eventually) converge almost surely to B^{-1}. Hence the minimiser of $\widehat{\mathrm{DKSD}}_{K,m}(\{X_i\}_{i=1}^n, \mathbb{P}_\theta)^2 = \theta^\top A_n \theta + v_n^\top \theta + c$ where $v_n \equiv \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} v(X_i, X_j)$ exists eventually.

$$
\begin{aligned}
|A(x,y)| &\lesssim K_\infty \|\phi(x)\| \|\phi(y)\| \\
\|v\| &\lesssim K_\infty \|\nabla_y b\| \|m(y)\| \|\phi(x)\| + K_\infty \|\nabla_x b\| \|m(x)\| \|\phi(y)\| \\
&\quad + (\|\nabla_x m\| + \|m(x)\|) K_\infty \|\phi(y)\| + (\|\nabla_y m\| + \|m(y)\|) K_\infty \|\phi(x)\| \\
|c| &\lesssim K_\infty \|\nabla_x b\| \|m(x)\| \|m(y)\| \|\nabla_y b\| + K_\infty (\|\nabla_x m\| + \|m(x)\|) \|\nabla_y m\| + \\
&\quad + K_\infty \|m(y)\| (1 + \|m(x)\| + \|\nabla_x m\|) \\
&\quad + K_\infty (\|\nabla_y m\| + \|m(y)\|) \|\nabla_x m\| \|\nabla_x b\| + K_\infty (\|\nabla_x m\| + \|m(x)\|) \|\nabla_y m\| \|\nabla_y b\|
\end{aligned}
$$

and it follows from the integrability assumptions that $\mathbb{Q} \otimes \mathbb{Q} |k_\theta^0| < \infty$. Since the product and sum of random variables that converge a.s. converge a.s., we have that $\hat{\theta}_n^{\mathrm{DKSD}} \to \theta^*$ a.s.,

$$
\hat{\theta}_n^{\mathrm{DKSD}} = -\tfrac{1}{2} A_n^{-1} v_n \xrightarrow{a.s.} -\tfrac{1}{2} B^{-1} v = \theta^*.
$$

■

D.3.2 Asymptotic Normality of the Kernel Stein Estimator

We now consider the distribution of $\sqrt{n}(\hat{\theta}_n^{\mathrm{DKSD}} - \theta^*)$. Recall that $A \in \Gamma(\mathcal{X}, \mathbb{R}^{m \times m}), v \in \Gamma(\mathcal{X}, \mathbb{R}^m)$, and for n large enough A_n^{-1} exists a.s., and $\hat{\theta}_n^{\mathrm{DKSD}} = -\frac{1}{2} A_n^{-1} v_n$.

Theorem 4. *Suppose $\|\phi\|, \|\nabla_x b\|\|m\|, \|\nabla_x m\| + \|m\| \in L^2(\mathbb{Q})$. Then the* DKSD *estimator is asymptotically normal.*

Proof From the integrability assumptions, it follows that $v, A \in L^2(\mathbb{Q} \otimes \mathbb{Q})$, and since $\mathcal{X}$ has finite $\mathbb{Q} \otimes \mathbb{Q}$-measure, $v, A \in L^1(\mathbb{Q} \otimes \mathbb{Q})$. Assume first that $m = 1$. Hence the tuple $U_n \equiv (v_n, A_n) : \Omega \to \mathbb{R}^2$, with $\mathbb{E}[U_n] = (\int_{\mathcal{X}} v \mathbb{Q} \otimes \mathbb{Q}, \int_{\mathcal{X}} A \mathbb{Q} \otimes \mathbb{Q}) \equiv (U_1, U_2)$, is asymptotically normal

$$
\sqrt{n}(U_n - \mathbb{E}[U_n]) \xrightarrow{d} \mathcal{N}(0, 4\Sigma)
$$

where, setting $v^0 = v - U_1$ and $A^0 = A - U_2$

$$
\begin{aligned}
\Sigma &= \mathbb{E}\big[\big(\textstyle\int_{\mathcal{X}} v^0(X,y) \mathrm{d}\mathbb{Q}(y), \int_{\mathcal{X}} A^0(X,y) \mathrm{d}\mathbb{Q}(y)\big) \otimes \big(\int_{\mathcal{X}} v^0(X,y) \mathrm{d}\mathbb{Q}(y), \int_{\mathcal{X}} A^0(X,y) \mathrm{d}\mathbb{Q}(y)\big)\big] \\
&= \begin{pmatrix} \int_{\mathcal{X}} v^0(x,y) \mathrm{d}\mathbb{Q}(y) \int_{\mathcal{X}} v^0(x,z) \mathrm{d}\mathbb{Q}(z) \mathrm{d}\mathbb{Q}(x) & \int_{\mathcal{X}} v^0(x,y) \mathrm{d}\mathbb{Q}(y) \int_{\mathcal{X}} A^0(x,z) \mathrm{d}\mathbb{Q}(z) \mathrm{d}\mathbb{Q}(x) \\ \int_{\mathcal{X}} v^0(x,y) \mathrm{d}\mathbb{Q}(y) \int_{\mathcal{X}} v^0(x,z) \mathrm{d}\mathbb{Q}(z) \mathrm{d}\mathbb{Q}(x) & \int_{\mathcal{X}} A^0(x,y) \mathrm{d}\mathbb{Q}(y) \int_{\mathcal{X}} A^0(x,z) \mathrm{d}\mathbb{Q}(z) \mathrm{d}\mathbb{Q}(x) \end{pmatrix}
\end{aligned}
$$

Since $\hat{\theta}_n^{\mathrm{DKSD}} = g(U_n)$, $\theta^* = g(U)$ where $g(x,y) \equiv -\frac{1}{2} x/y$, we can apply the delta method which states

$$
\sqrt{n}(\hat{\theta}_n^{\mathrm{DKSD}} - \theta^*) = \sqrt{n}(g(U_n) - g(U)) \xrightarrow{d} \mathcal{N}\big(0, 4\nabla g(U) \Sigma \nabla g(U)^\top\big)
$$

and $\nabla g(U) = \big(-1/2U_2, U_1/2U_2^2\big)$. Now let m be arbitrary. Since $A \in L^2(\mathbb{Q})$ then setting $A^0 \equiv A - \int_{\mathcal{X}} A\mathbb{Q} \otimes \mathbb{Q}$ we find

$$\sqrt{n}(A_n - \mathbb{E}[A_n]) \xrightarrow{d} \mathcal{N}(0, 4\Sigma_1), \qquad \Sigma_1 \equiv \int_{\mathcal{X}} \big[\int_{\mathcal{X}} A^0(x,y)\mathrm{d}\mathbb{Q}(y) \otimes \int_{\mathcal{X}} A^0(x,y)\mathrm{d}\mathbb{Q}(y)\big]\mathrm{d}\mathbb{Q}(x)$$

and similarly , with $v^0 \equiv v - \int v\mathrm{d}\mathbb{Q} \otimes \mathrm{d}\mathbb{Q}$

$$\sqrt{n}(v_n - \mathbb{E}[v_n]) \xrightarrow{d} \mathcal{N}(0, 4\Sigma_2), \qquad \Sigma_2 \equiv \int_{\mathcal{X}} \big[\int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y) \otimes \int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y)\big]\mathrm{d}\mathbb{Q}(x).$$

and

$$\sqrt{n}((v_n, A_n) - \mathbb{E}[(v_n, A_n)]) \xrightarrow{d} \mathcal{N}(0, 4\Sigma)$$

where

$$\Sigma = \int_{\mathcal{X}} \big[\big(\int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y), \int_{\mathcal{X}} A^0(x,y)\mathrm{d}\mathbb{Q}(y)\big) \otimes \big(\int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y), \int_{\mathcal{X}} A^0(x,y)\mathrm{d}\mathbb{Q}(y)\big)\big]\mathrm{d}\mathbb{Q}(x).$$

Let $\mathcal{D} \equiv \mathbb{R}^m \times \mathbb{R}^{m\times m}$, which we equip with coordinates $z_{ijk} = (x_i, y_{jk})$. Consider the function $g : \mathcal{D} \to \mathbb{R}^m$, $(x,y) \mapsto -\frac{1}{2}y^{-1}x$, so $g(v_n, A_n) = \theta_n^{\mathrm{DKSD}}$. Note $\Sigma \in \mathcal{D} \times \mathcal{D}$ and $\nabla g : \mathcal{D} \to \mathrm{End}(\mathcal{D}, \mathbb{R}^m) \cong \mathbb{R}^m \times \mathcal{D}$, so that $\nabla g(U)\Sigma\nabla g(U)^\top \in \mathbb{R}^{m\times m}$. First consider the matrix inversion $h(y) = y^{-1}$, so $\nabla h(y) \in \mathbb{R}^{(m\times m)\times(m\times m)}$, and $\nabla h(y)_{(ij)(kr)} = \partial_{y^{kr}} h_{ij}$. Since $h(y)_{ij}y_{jl} = \delta_{il}$ we have $0 = \partial_{kr}(h(y)_{ij}y_{jl}) = \partial_{kr}(h(y)_{ij})y_{jl} + h(y)_{ij}\delta_{jk}\delta_{rl} = \partial_{kr}(h(y)_{ij})y_{jl} + h(y)_{ik}\delta_{rl}$ and

$$\nabla h(y)_{(is)(kr)} = \partial_{kr}(h(y)_{ij})y_{jl}h(y)_{ls} = -h_{ik}\delta_{rl}h(y)_{ls} = -h(y)_{ik}h(y)_{rs}$$

and clearly $f : x \mapsto x$, then $\nabla f(x) = 1_{m\times m}$. Moreover

$$\partial_{y^{ab}} g_i(z) = \partial_{y^{ab}}(h(y)_{ij}f(x)_j) = \partial_{y^{ab}}(h(y)_{ij})x_j = -h(y)_{ia}h(y)_{bj}x_j, \quad \partial_{x^l} g_i(z) = h(y)_{il}$$

Then

$$\begin{aligned}(\nabla g(z)\Sigma)_{ir} &= \partial_v g_i \Sigma_{vr} = g_{i,x^l}\Sigma_{x^l r} + g_{i,y^{ab}}\Sigma_{y^{ab}r} = h(y)_{il}\Sigma_{x^l r} + \partial_{y^{ab}}(h(y)_{is})x_s\Sigma_{y^{ab}r} \\ &= h(y)_{il}\Sigma_{x^l r} - h(y)_{ia}h(y)_{bs}x_s\Sigma_{y^{ab}r},\end{aligned}$$

so

$$\begin{aligned}(\nabla g(z)\Sigma\nabla g(z)^\top)_{ic} &= (\nabla g(z)\Sigma)_{ir}(\nabla g(z))_{cr} = (\nabla g(z)\Sigma)_{ir}\partial_r g_c \\ &= h(y)_{il}\Sigma_{x^l r}\partial_r g_c - h(y)_{ia}h(y)_{bs}x_s\Sigma_{y^{ab}r}\partial_r g_c\end{aligned}$$

with

$$\begin{aligned}h(y)_{il}\Sigma_{x^l r}\partial_r g_c &= h(y)_{il}\Sigma_{x^l x^b}\partial_{x^b} g_c + h(y)_{il}\Sigma_{x^l y^{as}}\partial_{y^{as}} g_c \\ &= h(y)_{il}\Sigma_{x^l x^b}h(y)_{cb} - h(y)_{il}\Sigma_{x^l y^{as}}h_{ca}(y)h(y)_{sj}x_j\end{aligned}$$

and

$$\begin{aligned}-h(y)_{ia}h(y)_{bs}x_s\Sigma_{y^{ab}r}\partial_r g_c &= -h(y)_{ia}h(y)_{bs}x_s\big(\Sigma_{y^{ab}x^k}\partial_{x^k} g_c + \Sigma_{y^{ab}y^{ld}}\partial_{y^{ld}} g_c\big) \\ &= -h(y)_{ia}h(y)_{bs}x_s\big(\Sigma_{y^{ab}x^k}h(y)_{ck} - \Sigma_{y^{ab}y^{ld}}h(y)_{cl}h(y)_{dj}x_j\big).\end{aligned}$$

Note we have

$$\begin{aligned}\Sigma_{xx} &= \int_{\mathcal{X}}\int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y) \otimes \int_{\mathcal{X}} v^0(x,z)\mathrm{d}\mathbb{Q}(z)\mathrm{d}\mathbb{Q}(x) \equiv \int_{\mathcal{X}} T(x) \otimes T(x)\mathrm{d}\mathbb{Q}(x) \\ \Sigma_{xy} &= \int_{\mathcal{X}}\int_{\mathcal{X}} v^0(x,y)\mathrm{d}\mathbb{Q}(y) \otimes \int_{\mathcal{X}} A^0(x,z)\mathrm{d}\mathbb{Q}(z)\mathrm{d}\mathbb{Q}(x) \equiv \int_{\mathcal{X}} T(x) \otimes L(x)\mathrm{d}\mathbb{Q}(x) \\ \Sigma_{yy} &= \int_{\mathcal{X}}\int_{\mathcal{X}} A^0(x,y)\mathrm{d}\mathbb{Q}(y) \otimes \int_{\mathcal{X}} A^0(x,z)\mathrm{d}\mathbb{Q}(z)\mathrm{d}\mathbb{Q}(x) \equiv \int_{\mathcal{X}} L(x) \otimes L(x)\mathrm{d}\mathbb{Q}(x)\end{aligned}$$

then

$$\begin{aligned}4\nabla g(U_1,U_2)\Sigma\nabla g(U_1,U_2)^\top &= \int_{\mathcal{X}}(U_2^{-1}T) \otimes (TU_2^{-1})\mathrm{d}\mathbb{Q} \\ &\quad - 2\int_{\mathcal{X}}\big(U_2^{-1}LU_2^{-1}U_1\big) \otimes \big(TU_2^{-1}\big)\mathrm{d}\mathbb{Q} \\ &\quad + \int_{\mathcal{X}}\big(U_2^{-1}LU_2^{-1}U_1\big) \otimes \big(U_2^{-1}LU_2^{-1}U_1\big)\mathrm{d}\mathbb{Q}\end{aligned}$$

■

D.3.3 Diffusion Score Matching Asymptotics

Consider the loss function

$$L(x,\theta) = \big\langle \nabla \log p_\theta, mm^\top \nabla \log p_\theta \big\rangle + 2\big(\nabla \cdot (mm^\top) \cdot \nabla \log p_\theta + \mathrm{Tr}\big[mm^\top \nabla^2 \log p_\theta\big]\big).$$

For the exponential family $L(x,\theta) = \theta^\top A\theta + v^\top\theta + c$, where (we set $S = mm^\top$)

$$\begin{aligned} A &= \nabla T S \nabla T^\top \\ v^\top &= 2\nabla b \cdot S\nabla T^\top + 2\nabla \cdot S \cdot \nabla T^\top + 2\mathrm{Tr}\big[S\nabla^2 T_i\big]e_i \\ c &= \nabla b \cdot S\nabla b + 2\nabla \cdot S \cdot \nabla b + 2\mathrm{Tr}[S\nabla\nabla b]. \end{aligned}$$

Theorem 13. *Suppose m is invertible and $\{\nabla T_i\}$ are linearly independent. Then if $A, v \in L^1(\mathbb{Q})$, $\hat{\theta}_n^{\mathrm{DSM}}$ eventually exists and is strongly consistent. If we also have $A, v \in L^2(\mathbb{Q})$, then $\hat{\theta}_n^{\mathrm{DSM}}$ is asymptotically normal.*

Proof Let $\mathcal{M} \equiv \int A \mathrm{d}\mathbb{Q}$, $H \equiv \int v \mathrm{d}\mathbb{Q}$. If $A = \nabla T mm^\top \nabla T^\top = \nabla Tm(\nabla Tm)^\top$ so $\mathrm{rank}(A) = \mathrm{rank}(\nabla Tm(\nabla Tm)^\top) = \mathrm{rank}(\nabla Tm) = \mathrm{rank}(\nabla T) = \mathrm{rank}(\nabla T^\top)$ if m is invertible. So if the vectors $\{\nabla T_i\}$ are linearly independent, then $\nabla T^\top$ has full column rank. Then A it is symmetric positive (strictly) definite and the minimum of $L(\theta) \equiv \int L(x,\theta)\mathrm{d}\mathbb{Q}(x)$ is $\theta^* = -\frac{1}{2}\mathcal{M}^{-1}H$ which for sufficiently large n can be estimated by the random variable $\hat{\theta}_n^{\mathrm{DSM}} \equiv -\frac{1}{2}\mathcal{M}_n^{-1}H_n$ which converges a.s. to θ.

We consider the tuple $U_n \equiv (H_n, \mathcal{M}_n)$, so $\mathbb{E}[U_n] = (H, \mathcal{M})$. Since $A, v \in L^2(\mathbb{Q})$, then

$$\sqrt{n}(U_n - (H,\mathcal{M})) \xrightarrow{d} \mathcal{N}(0,\Gamma)$$

where, setting $v^0 = v - H$, $A^0 = A - \mathcal{M}$

$$\Gamma = \mathbb{E}\big[(v^0, A^0) \otimes (v^0, A^0)\big].$$

Let $\mathcal{D} \equiv \mathbb{R}^m \times \mathbb{R}^{m\times m}$, and consider $g : \mathcal{D} \to \mathbb{R}^m$, defined by $g(x,y) = -\frac{1}{2}y^{-1}x$. Using the Delta method

$$\sqrt{n}(\hat{\theta}_n^{\mathrm{DSM}} - \theta^*) \xrightarrow{d} \mathcal{N}\big(0, 4\nabla g(H,\mathcal{M})\Gamma\nabla g(H,\mathcal{M})^\top\big)$$

where, proceeding as in Appendix D.3.2, we find

$$\begin{aligned} 4\nabla g(H,\mathcal{M})\Gamma\nabla g(H,\mathcal{M})^\top &= \textstyle\int_{\mathcal{X}} (\mathcal{M}^{-1}v^0) \otimes (v^0\mathcal{M}^{-1})\mathrm{d}\mathbb{Q} \\ &\quad - 2\textstyle\int_{\mathcal{X}} \big(\mathcal{M}^{-1}A^0\mathcal{M}^{-1}H\big) \otimes \big(v^0\mathcal{M}^{-1}\big)\mathrm{d}\mathbb{Q} \\ &\quad + \textstyle\int_{\mathcal{X}} \big(\mathcal{M}^{-1}A^0\mathcal{M}^{-1}H\big) \otimes \big(\mathcal{M}^{-1}A^0\mathcal{M}^{-1}H\big)\mathrm{d}\mathbb{Q} \end{aligned}$$

■

E Robustness of Minimum Stein Discrepancy Estimators

In this section, we provide conditions on the Stein operator (and Stein class) to obtain robust estimators in the context of DKSD and DSM. In particular we prove Proposition 6 and derive the influence function of DSM.

E.1 Robustness of Diffusion Kernel Stein Discrepancy

Let $T : \mathcal{P}_\Theta \to \Theta$ with $T(\mathbb{P}) = \mathrm{argmin}_\Theta \, \mathrm{DKSD}_{K,m}(\mathbb{P}\|\mathbb{P}_\theta)$ be defined by $\mathrm{IF}(z,\mathbb{Q}) \equiv \lim_{t\to 0}(T(\mathbb{Q}+t(\delta_z - \mathbb{Q})) - T(\mathbb{Q}))/t$. Denote $\mathbb{Q}_t = \mathbb{Q} + t(\delta_z - \mathbb{Q})$, $\theta_t = T(\mathbb{Q}_t)$, $\theta_0 = T(\mathbb{Q})$. Note that by the first order optimality condition:

$$\nabla_\theta \textstyle\int_{\mathcal{X}}\int_{\mathcal{X}} k^0 \mathbb{Q}_t \otimes \mathbb{Q}_t|_{\theta_t} = \nabla_{\theta_t} \mathrm{DKSD}_{K,m}(\mathbb{Q}_t\|\mathbb{P}_\theta) = 0.$$

By the MVT, there exists $\bar{\theta}$ on the line joining θ_0 and θ_t for which

$$0 = \textstyle\int_{\mathcal{X}}\int_{\mathcal{X}} \nabla_\theta k^0|_{\theta_0} \mathbb{Q}_t \otimes \mathbb{Q}_t + \int_{\mathcal{X}}\int_{\mathcal{X}} \nabla_\theta\nabla_\theta k^0|_{\bar{\theta}} \mathbb{Q}_t \otimes \mathbb{Q}_t(\theta_t - \theta_0).$$

Expanding

$$\mathbb{Q}_t \otimes \mathbb{Q}_t \nabla_\theta k^0|_{\theta_0} = t^2(\delta_z - \mathbb{Q}) \otimes (\delta_z - \mathbb{Q}) \nabla_\theta k^0|_{\theta_0} + 2t\mathbb{Q}_y \nabla_\theta k^0|_{\theta_0}(z, y)$$

where we have used the optimality condition. On the other hand

$$\mathbb{Q}_t \otimes \mathbb{Q}_t \nabla_\theta \nabla_\theta k^0|_{\bar{\theta}} = (1-2t)\mathbb{Q} \otimes \mathbb{Q} \nabla_\theta \nabla_\theta k^0|_{\bar{\theta}} + t^2(\delta_z - \mathbb{Q}) \otimes (\delta_z - \mathbb{Q}) \nabla_\theta \nabla_\theta k^0|_{\bar{\theta}} + 2t\mathbb{Q}_y \nabla_\theta \nabla_\theta k^0|_{\bar{\theta}}(z, y).$$

Hence

$$\mathbb{Q}_y \nabla_\theta k^0|_{\theta_0}(z, y) = \tfrac{1}{2}\big((1-2t)\mathbb{Q} \otimes \mathbb{Q} \nabla_\theta \nabla_\theta k^0|_{\bar{\theta}} + 2t\mathbb{Q}_y \nabla_\theta \nabla_\theta k^0|_{\bar{\theta}}(z, y)\big)\tfrac{\theta_t - \theta_0}{t} + O(t),$$

and taking the limit $t \to 0$, $\bar{\theta} \to \theta_0$ and using a derivation as in the proof of Theorem 5

$$\mathbb{Q}_y \nabla_\theta k^0|_{\theta_0}(z, y) = \tfrac{1}{2}\int_{\mathcal{X}}\int_{\mathcal{X}} \nabla_\theta \nabla_\theta k^0|_{\theta_0} \mathrm{d}\mathbb{Q} \otimes \mathrm{d}\mathbb{Q}\, \mathrm{IF}(z, \mathbb{Q}) = g(\theta_0)\, \mathrm{IF}(z, \mathbb{Q})$$

hence the influence function is given by

$$\mathrm{IF}(z, \mathbb{Q}) = g(\theta_0)^{-1} \int_{\mathcal{X}} \nabla_\theta k^0|_{\theta_0}(z, y) \mathrm{d}\mathbb{Q}(y).$$

Suppose that the additional assumptions hold. We aim to show that $\mathrm{IF}(z, \mathbb{Q})$ is bounded independently of z. By assumption $\langle s_p(x), K(x, y)\nabla_\theta s_p(y)\rangle \to 0$ as $|y| \to \infty$ for $x \in \mathcal{X}$. Moreover $|\langle s_p(x), K(x, y)\nabla_\theta s_p(y)\rangle| \leq K_\infty \|s_p(x)\| \|\nabla_\theta s_p(y)\|$, which is $\mathbb{Q}$-integrable with respect to the x-variable, by assumption. By dominated convergence theorem, it follows that $\int \langle s_p(x), K(x, y)\nabla_\theta s_p(y)\rangle \mathrm{d}\mathbb{Q}(x) \to \infty$ as $|y| \to \infty$. Following a similar argument, and using the assumptions, a similar limit will hold for all terms in $\int \nabla_\theta k^0(z, y)\mathrm{d}\mathbb{Q}(y)$. Since the influence function is continuous in z it follows that $\sup_{z \in \mathcal{X}} \|\mathrm{IF}(z, \mathbb{P}_\theta)\| < \infty$.

E.2 Robustness of Diffusion Score Matching

The scoring rule $S : \mathcal{X} \times \mathcal{P}_{\mathcal{X}} \to \mathbb{R}$ of DSM is

$$S(x, \mathbb{P}_\theta) \equiv \tfrac{1}{2}\left\|m^\top \nabla_x \log p_\theta\right\|_2^2 + \nabla \cdot \left(mm^\top \nabla \log p_\theta\right)(x)$$

Indeed the proof of Theorem 9 we have

$$\int_{\mathcal{X}} \left\|m^\top \nabla \log q\right\|^2 \mathrm{d}\mathbb{Q} = -\int_{\mathcal{X}} \nabla \cdot \left(mm^\top \nabla \log q\right) \mathrm{d}\mathbb{Q}.$$

which implies $\mathbb{Q}S(\cdot, \mathbb{Q}) = -\frac{1}{2}\int_{\mathcal{X}} \left\|m^\top \nabla \log q\right\|^2 \mathrm{d}\mathbb{Q}$, so

$$\begin{aligned}\mathbb{Q}S(\cdot, \mathbb{P}_\theta) - \mathbb{Q}S(\cdot, \mathbb{Q}) &= \int_{\mathcal{X}} \left(\tfrac{1}{2}\left\|m^\top \nabla_x \log p_\theta\right\|_2^2 + \tfrac{1}{2}\left\|m^\top \nabla \log q\right\|^2 + \nabla \cdot \left(mm^\top \nabla \log p_\theta\right)\right) \mathrm{d}\mathbb{Q} \\ &= \mathrm{DSM}_m(\mathbb{Q}\|\mathbb{P}_\theta).\end{aligned}$$

The influence function is then $\mathrm{IF}(x, \mathbb{P}_\theta) = g_{\mathrm{DSM}}(\theta)^{-1} s(x, \theta)$, where

$$\begin{aligned}s(x, \theta) \equiv \nabla_\theta S(x, \theta) &= \tfrac{1}{2}\nabla_\theta \|m^\top \nabla_x \log p_\theta\|_2^2 + \nabla_\theta \nabla_x \cdot \left(mm^\top \nabla_x \log p_\theta\right) \\ &= \tfrac{1}{2}\nabla_\theta \|m^\top \nabla_x \log p_\theta\|_2^2 + \nabla_\theta\big(\big\langle \nabla_x \cdot (mm^\top), \nabla \log p_\theta \big\rangle + \mathrm{Tr}\big[mm^\top \nabla_x^2 \log p_\theta\big]\big) \\ &= \nabla_x \nabla_\theta \log p_\theta mm^\top \nabla_x \log p_\theta + (\nabla_x \nabla_\theta \log p_\theta)\nabla_x \cdot (mm^\top) + \mathrm{Tr}[mm^\top \nabla_x \nabla_x] \nabla_\theta \log p_\theta\end{aligned}$$

and where $g_{\mathrm{DSM}}(\theta) \equiv \mathbb{P}_\theta \nabla_\theta \nabla_\theta S(\cdot, \theta)$ is the information metric associated with DSM.