Stein Discrepancies

PART III ESSAY

ABSTRACT

Exact Bayesian inference is rarely feasible in practice, necessitating the use of approximate inference methods. One such method is Markov Chain Monte Carlo (MCMC), which outputs a discrete measure with the aim of approximating the posterior distribution. Increasingly, biased MCMC is being used to speed up mixing times at the expense of asymptotic exactness. It is desirable in this context to be able to quantify how well our discrete measure approximates the target posterior.

In this essay, we introduce the *Stein discrepancy*, which quantifies distance between a discrete measure and a target measure, whose density we need only know up to a normalising constant. We show that the Stein discrepancy can be *kernelised* to yield an analytic solution, and that a specific choice of kernel, namely the *IMQ kernel*, yields a kernelised Stein discrepancy (KSD) which can detect when a discrete measure converges to the target, and when it doesn't.

We also use the IMQ KSD to derive *Stein variational gradient descent* (SVGD), an expressive variational inference algorithm which is asymptotically exact. Finally, we highlight scalability issues present when calculating the IMQ KSD and using SVGD, and propose solutions for these.

Contents

Chapter 1

Introduction

1.1 Motivation

Many problems in statistics and machine learning require the approximation of a target probability measure P with another Q. Crucial to approximating P accurately is the ability to obtain a measure of "distance" between P and Q. A popular approach to measuring such distance is to consider an **integral probability metric** (IPM):

$$d_{\mathcal{H}}(Q, P) := \sup_{h \in \mathcal{H}} \left| \mathbb{E}_Q[h(\boldsymbol{X})] - \mathbb{E}_P[h(\boldsymbol{X})] \right|,$$

where $\mathcal{H}$ is some class of test functions.

Often, our ultimate goal is to approximate an integral under P of the form $\mathbb{E}_P[f(\boldsymbol{X})]$. With this in mind, given a sequence of approximating measures $(Q_n)_{n \geq 1}$, a natural requirement for $\mathcal{H}$ is that, as $n \to \infty$, $d_{\mathcal{H}}(Q_n, P) \to 0$ if and only if Q_n converges weakly to P, which we denote by $Q_n \Rightarrow P$. When this is the case, we say that $d_{\mathcal{H}}(\cdot, \cdot)$ **metrises** weak convergence.

One choice of $\mathcal{H}$ satisfying this requirement is $\mathcal{H} = \mathcal{W}_{\|\cdot\|_2} := \{h : \mathbb{R}^d \to \mathbb{R} \mid \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{|h(\boldsymbol{x}) - h(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|_2} \leq 1\}$, which corresponds to the **Wasserstein-1 metric**. Another choice is $\mathcal{H} = \mathrm{BL}_{\|\cdot\|_2} := \{h : \mathbb{R}^d \to \mathbb{R} : \sup_{\boldsymbol{x} \in \mathbb{R}^d} |h(\boldsymbol{x})| + \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{|h(\boldsymbol{x}) - h(\boldsymbol{y})|}{\|\boldsymbol{x} - \boldsymbol{y}\|_2} \leq 1\}$, which corresponds to the **bounded Lipschitz metric**.

These two metrics are known to metrise weak convergence. However, they both exhibit a common feature which inhibits calculation in many settings: they require the ability to integrate over both P and Q. In many cases, we are unable to integrate over the target P, and so it is of significant practical interest to find a choice of $\mathcal{H}$ which metrises weak convergence while only requiring integration over the approximating distribution Q.

Although not immediately obvious, such a choice of $\mathcal{H}$ does exist, as outlined by Gorham and Mackey [2015]. Their discovery of such an $\mathcal{H}$ was based on **Stein's method** [Stein, 1972], and proceeds as follows, for some class of sufficiently well-behaved target measures $P \in \mathcal{P}$ with differentiable densities p:

1. Identify an operator $\mathcal{T}_P$ which maps a set $\mathcal{G}$ of vector-valued functions $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$ to real-valued functions $\mathcal{T}_P \boldsymbol{g}$ such that

$$\mathbb{E}_P[(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{X})] = 0 \quad \text{for all } \boldsymbol{g} \in \mathcal{G}. \tag{1.1}$$

   Such an operator $\mathcal{T}_P$ is called a **Stein operator**; such a set $\mathcal{G}$ is called a *Stein set*. One can then define the **Stein discrepancy** as

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}) := \sup_{\boldsymbol{g} \in \mathcal{G}} \left| \mathbb{E}_Q[(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{X})] \right| = d_{\mathcal{T}_P \mathcal{G}}(Q, P).$$

2. Lower bound the Stein discrepancy by an IPM which is known to dominate weak convergence, so that $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}) \to 0$ implies $Q_n \Rightarrow P$ as $n \to \infty$. This allows us to detect when Q_n is not converging to P.

3. Upper bound the Stein discrepancy by an IPM known to be dominated by weak convergence, ensuring that $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}) \to 0$ whenever $Q_n \Rightarrow P$ as $n \to \infty$, allowing us to detect when Q_n is converging to P.

The most popular choice for the Stein operator is the **Langevin Stein operator**

$$(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{x}) := \frac{1}{p(\boldsymbol{x})} \nabla_{\boldsymbol{x}} \cdot (p(\boldsymbol{x}) \boldsymbol{g}(\boldsymbol{x})) = \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{s}_p(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{g}(\boldsymbol{x}),$$

where $\boldsymbol{s}_p(\boldsymbol{x}) := \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})$ is the score function. To see that this satisfies (1.1), we integrate by parts:

$$\mathbb{E}_P[\boldsymbol{g}(\boldsymbol{X})^\top \boldsymbol{s}_p(\boldsymbol{X})] = \int_{\mathbb{R}^d} \boldsymbol{g}(\boldsymbol{x})^\top \nabla_{\boldsymbol{x}} p(\boldsymbol{x}) d\boldsymbol{x} = -\int_{\mathbb{R}^d} \nabla_{\boldsymbol{x}} \cdot \boldsymbol{g}(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = -\mathbb{E}_P[\nabla_{\boldsymbol{x}} \cdot \boldsymbol{g}(\boldsymbol{x})],$$

assuming appropriate boundary conditions which will hold for any reasonable class of test functions $\mathcal{G}$.

An additional appealing property of the Langevin Stein operator is that it depends on P only through the score $\boldsymbol{s}_p$, and so it is only required for us to know the density p up to a multiplicative constant. This is particularly useful in the context of Bayesian inference, where P is a posterior distribution and, for example, Q_n is the discrete measure outputted by some Markov Chain Monte Carlo (MCMC) algorithm.

1.2 Kernelised Stein discrepancies

Computing the Stein discrepancy involves taking the supremum over a complicated set of functions. While Gorham and Mackey [2015] proposed methods for solving such an optimisation problem, an alternative, proposed by Chwialkowski et al. [2016] and Gorham and Mackey [2017], is to *kernelise* the Stein discrepancy, which allows us to calculate the supremum analytically.

Let $\mathcal{H}_k$ be a reproducing kernel Hilbert space (RKHS), with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$. For $d \in \mathbb{N}$, we can associate with the product space $\mathcal{H}_k^d$ the inner product $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathcal{H}_k^d} := \sum_{i=1}^d \langle u_i, v_i \rangle_{\mathcal{H}_k}$ for $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{H}_k^d$. We then define the **kernel Stein set** to be the unit ball with respect to the norm induced by this inner product:

$$\mathcal{G}_k := \left\{ \boldsymbol{g} \in \mathcal{H}_k^d \,:\, \|\boldsymbol{g}\|_{\mathcal{H}_k^d} \leq 1 \right\}. \tag{1.2}$$

Since the RKHS $\mathcal{H}_k$ is the completion of the linear span $\{\sum_{i=1}^n \alpha_i k(\cdot, \boldsymbol{x}_i) : n \in \mathbb{N},\ \alpha_i \in \mathbb{R},\ \boldsymbol{x}_i \in \mathbb{R}^d\}$, we have $\boldsymbol{g} \in \mathcal{G}_k \iff -\boldsymbol{g} \in \mathcal{G}_k$. Thus, since $\mathcal{T}_P(-\boldsymbol{g}) = -\mathcal{T}_P(\boldsymbol{g})$, we can remove the modulus in the Stein discrepancy. Furthermore, the Langevin Stein operator can be decomposed as $\mathcal{T}_P \boldsymbol{g}(\cdot) = \sum_{i=1}^d \mathcal{T}_P^{(i)} g_i(\cdot)$, where $\mathcal{T}_P^{(i)} h(\boldsymbol{x}) = h(\boldsymbol{x}) \nabla_{x_i} \log p(\boldsymbol{x}) + \nabla_{x_i} h(\boldsymbol{x})$, so

$$\begin{aligned}
\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_k) = \sup_{\boldsymbol{g} \in \mathcal{G}_k} \sum_{i=1}^d \mathbb{E}_Q[(\mathcal{T}_P^{(i)} g_i)(\boldsymbol{X})] &= \sup_{\boldsymbol{g} \in \mathcal{G}_k} \sum_{i=1}^d \mathbb{E}_Q[\mathcal{T}_P^{(i)} \langle g_i, k(\boldsymbol{X}, \cdot) \rangle_{\mathcal{H}_k}] \\
&= \sup_{\boldsymbol{g} \in \mathcal{G}_k} \sum_{i=1}^d \left\langle g_i, \mathbb{E}_Q \left[\mathcal{T}_P^{(i)} k(\boldsymbol{X}, \cdot) \right] \right\rangle_{\mathcal{H}_k} = \sup_{\boldsymbol{g} \in \mathcal{G}_k} \langle \boldsymbol{g}, \mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot) \rangle_{\mathcal{H}_k^d},
\end{aligned}$$

where

$$\boldsymbol{\xi}_p(\boldsymbol{x}, \boldsymbol{y}) := \boldsymbol{s}_p(\boldsymbol{x}) k(\boldsymbol{x}, \boldsymbol{y}) + \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{y}). \tag{1.3}$$

By the Cauchy-Schwarz inequality on the Cartesian product Hilbert space $\mathcal{H}_k^d$, we have

$$\langle \boldsymbol{g}, \mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot) \rangle_{\mathcal{H}_k^d} \leq \|\boldsymbol{g}\|_{\mathcal{H}_k^d} \|\mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot)\|_{\mathcal{H}_k^d},$$

and so for $\boldsymbol{g} \in \mathcal{G}_k$,

$$\langle \boldsymbol{g}, \mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot) \rangle_{\mathcal{H}_k^d} \leq \| \mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot) \|_{\mathcal{H}_k^d}. \tag{1.4}$$

Therefore if $\boldsymbol{g}(\cdot) := \mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot) / \| \mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot) \|_{\mathcal{H}_k^d}$ is in the kernel Stein set $\mathcal{G}_k$, then this will achieve the upper bound in (1.4) and thus attain the supremum defined in the Stein discrepancy. This choice of $\boldsymbol{g}$ does indeed belong to $\mathcal{G}_k$: it has unit norm, and $\xi_{p,i}(\boldsymbol{x}, \cdot) = k(\boldsymbol{x}, \cdot) \nabla_{x_i} \log p(\boldsymbol{x}) + \nabla_{x_i} k(\boldsymbol{x}, \cdot)$ is a linear combination of $k(\boldsymbol{x}, \cdot)$, which belongs to $\mathcal{H}_k$ by definition, and $\nabla_{x_i} k(\boldsymbol{x}, \cdot)$, which is also in $\mathcal{H}_k$, as we now justify. Take $\frac{1}{h_n}(k(\boldsymbol{x} + h_n \boldsymbol{e}_i, \cdot) - k(\boldsymbol{x}, \cdot)) \in \mathcal{H}_k$ for all n, where $\boldsymbol{e}_i$ is the i^{th} standard basis vector and $(h_n)_{n \geq 1}$ is such that $h_n \to 0$ as $n \to \infty$. Taking $n \to \infty$ gives $\nabla_{x_i} k(\boldsymbol{x}, \cdot) \in \mathcal{H}_k$. Then we have

$$\mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_k) = \left\langle \frac{\mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot)}{\| \mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot) \|_{\mathcal{H}_k^d}}, \mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot) \right\rangle_{\mathcal{H}_k^d} = \| \mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot) \|_{\mathcal{H}_k^d}.$$

Observe further that

$$\begin{aligned} \langle \xi_{p,i}(\boldsymbol{x}, \cdot), \xi_{p,i}(\boldsymbol{y}, \cdot) \rangle_{\mathcal{H}_k} &= \langle s_{p,i}(\boldsymbol{x}) k(\boldsymbol{x}, \cdot) + \nabla_{x_i} k(\boldsymbol{x}, \cdot),\ s_{p,i}(\boldsymbol{y}) k(\boldsymbol{y}, \cdot) + \nabla_{y_i} k(\boldsymbol{y}, \cdot) \rangle_{\mathcal{H}_k} \\ &= s_{p,i}(\boldsymbol{x}) s_{p,i}(\boldsymbol{y}) k(\boldsymbol{x}, \boldsymbol{y}) + s_{p,i}(\boldsymbol{x}) \nabla_{y_i} k(\boldsymbol{x}, \boldsymbol{y}) + s_{p,i}(\boldsymbol{y}) \nabla_{x_i} k(\boldsymbol{x}, \boldsymbol{y}) + \nabla_{x_i} \nabla_{y_i} k(\boldsymbol{x}, \boldsymbol{y}), \end{aligned}$$

so, defining the **Stein reproducing kernel**

$$\begin{aligned} k_p(\boldsymbol{x}, \boldsymbol{y}) &:= \langle \boldsymbol{\xi}_p(\boldsymbol{x}, \cdot), \boldsymbol{\xi}_p(\boldsymbol{y}, \cdot) \rangle_{\mathcal{H}_k^d} \\ &= \boldsymbol{s}_p(\boldsymbol{x})^\top \boldsymbol{s}_p(\boldsymbol{y}) k(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{s}_p(\boldsymbol{x})^\top \nabla_{\boldsymbol{y}} k(\boldsymbol{x}, \boldsymbol{y}) + \boldsymbol{s}_p(\boldsymbol{y})^\top \nabla_{\boldsymbol{x}} k(\boldsymbol{x}, \boldsymbol{y}) + \nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{y}} k(\boldsymbol{x}, \boldsymbol{y}), \end{aligned}$$

we get the squared **kernelised Stein discrepancy** (KSD)

$$\begin{aligned} \mathcal{S}^2(Q, \mathcal{T}_P, \mathcal{G}_k) &= \langle \mathbb{E}_{\boldsymbol{X} \sim Q}[\boldsymbol{\xi}_p(\boldsymbol{X}, \cdot)], \mathbb{E}_{\boldsymbol{X}' \sim Q}[\boldsymbol{\xi}_p(\boldsymbol{X}', \cdot)] \rangle_{\mathcal{H}_k^d} \\ &= \mathbb{E}_{\boldsymbol{X}, \boldsymbol{X}' \sim Q}[k_p(\boldsymbol{X}, \boldsymbol{X}')], \end{aligned}$$

where $\boldsymbol{X}, \boldsymbol{X}' \overset{\text{i.i.d.}}{\sim} Q$. When $Q(\cdot) = \frac{1}{n} \sum_{i=1}^n \delta_{\boldsymbol{x}_i}(\cdot)$, we have

$$\mathcal{S}^2(Q, \mathcal{T}_P, \mathcal{G}_k) = \frac{1}{n^2} \sum_{i,j=1}^n k_p(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

1.3 Outline of chapters

In Chapter 2, we prove that a particular choice of kernel, namely the IMQ kernel, with suitable parameters, yields a KSD that metrises weak convergence for a large class $\mathcal{P}$ of target measures, following the approach of Gorham and Mackey [2017].

In Chapter 3, we demonstrate how the formulation of the KSD can be used to develop a flexible variational inference algorithm known as **Stein variational gradient descent** (SVGD) [Liu and Wang, 2016]. We will prove that it is asymptotically exact using results from Liu [2017] and Gorham et al. [2020], the latter of which receives greater focus in Chapter 4, where we address scalability issues associated with KSDs. In this chapter, we also discuss an approach by Huggins and Mackey [2018] to alleviating the cost of calculating the KSD, which grows quadratically in the number of samples.

In Chapter 5, we put this theory into practice, demonstrating the practicality of the various Stein discrepancies we have introduced, as well as the SVGD algorithm, drawing on experiments carried out in Gorham and Mackey [2017] and Gorham et al. [2020]. We also focus on the scalability issues presented in Chapter 4.

Chapter 2

Assessing Convergence with Kernelised Stein Discrepancies

The goal of this chapter is to find a kernel $k(\cdot, \cdot)$ whose corresponding KSD metrises weak convergence for a large class $\mathcal{P}$ of target measures P with differentiable densities p. We will be interested in so-called *distantly dissipative* target distributions.

2.1 Definition (Distant dissipativity)**.** *Let $r > 0$. We say that a distribution P is* **distantly dissipative** *if*

$$\kappa_0 := \liminf_{r \to \infty} \kappa(r) > 0, \tag{2.1}$$

where

$$\kappa(r) := \inf_{\boldsymbol{x}, \boldsymbol{y}: \|\boldsymbol{x} - \boldsymbol{y}\|_2 = r} \left\{ -2 \frac{\langle \boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_p(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle}{\|\boldsymbol{x} - \boldsymbol{y}\|_2^2} \right\}.$$

We let $\mathcal{P}$ be the class of distantly dissipative distributions with Lipschitz score function $\boldsymbol{s}_p$. As an example, $\mathcal{P}$ includes all distributions which are strongly log-concave outside of a compact set, including Bayesian linear, logistic and Huber regression posteriors with Gaussian priors.

As mentioned above, in order to show that the KSD metrises weak convergence, we aim to both upper and lower bound it by IPMs known to metrise weak convergence. We first consider the easier – and more general – direction, where we obtain an upper bound on the KSD for a variety of kernels $k(\cdot, \cdot)$.

2.1 Detecting convergence: upper bounding the KSD

A KSD detects convergence if $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$ whenever $Q_n \Rightarrow P$. To aid our discussion, we define the following quantities. For any function $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^{d'}$, set

$$M_0(\boldsymbol{g}) := \sup_{\boldsymbol{x} \in \mathbb{R}^d} \|\boldsymbol{g}(\boldsymbol{x})\|_2, \quad M_1(\boldsymbol{g}) := \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{\|\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{g}(\boldsymbol{y})\|_2}{\|\boldsymbol{x} - \boldsymbol{y}\|_2}, \quad M_2(\boldsymbol{g}) := \sup_{\boldsymbol{x} \neq \boldsymbol{y}} \frac{\|\nabla_{\boldsymbol{x}} \boldsymbol{g}(\boldsymbol{x}) - \nabla_{\boldsymbol{y}} \boldsymbol{g}(\boldsymbol{y})\|_{\text{op}}}{\|\boldsymbol{x} - \boldsymbol{y}\|_2}. \tag{2.2}$$

We first prove the following lemma.

2.2 Lemma. *Let $\boldsymbol{Z} \sim P$ and $\boldsymbol{X} \sim \mu$. Suppose that $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$ is such that $\max\{M_0(\boldsymbol{g}), M_1(\boldsymbol{g}), M_2(\boldsymbol{g})\} < \infty$. If $\boldsymbol{s}_p$ is Lipschitz with $\mathbb{E}_P[\|\boldsymbol{s}_p(\boldsymbol{Z})\|_2^2] < \infty$, then*

$$|\mathbb{E}_\mu[(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{X})]| \leq (M_0(\boldsymbol{g}) M_1(\boldsymbol{s}_p) + M_2(\boldsymbol{g}) d) W_2(\mu, P) + \sqrt{2 M_0(\boldsymbol{g}) M_1(\boldsymbol{g}) \mathbb{E}_P[\|\boldsymbol{s}_p(\boldsymbol{Z})\|_2^2] W_2(\mu, P)},$$

where $W_p(\cdot, \cdot)$ denotes the Wasserstein-p distance.

Proof. By Jensen's inequality, we have $\mathbb{E}_P[\|\boldsymbol{s}_p(\boldsymbol{Z})\|_2] \leq \sqrt{\mathbb{E}_P[\|\boldsymbol{s}_p(\boldsymbol{Z})\|_2^2]} < \infty$, which is sufficient for us to be able to integrate by parts and obtain $\mathbb{E}_P[(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{Z})] = 0$. Then, using the triangle inequality, Jensen's inequality and the Fenchel-Young inequality for dual norms,

$$\begin{aligned}
|\mathbb{E}_\mu[(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{X})]| &= |\mathbb{E}_P[(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{Z})] - \mathbb{E}_\mu[(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{X})]| \\
&= |\mathbb{E}[\langle \boldsymbol{s}_p(\boldsymbol{Z}), \boldsymbol{g}(\boldsymbol{Z}) - \boldsymbol{g}(\boldsymbol{X})\rangle + \langle \boldsymbol{s}_p(\boldsymbol{Z}) - \boldsymbol{s}_p(\boldsymbol{X}), \boldsymbol{g}(\boldsymbol{X})\rangle + \nabla_{\boldsymbol{Z}} \cdot \boldsymbol{g}(\boldsymbol{Z}) - \nabla_{\boldsymbol{X}} \cdot \boldsymbol{g}(\boldsymbol{X})]| \\
&\leq \mathbb{E}[|\langle \boldsymbol{s}_p(\boldsymbol{Z}), \boldsymbol{g}(\boldsymbol{Z}) - \boldsymbol{g}(\boldsymbol{X})\rangle|] + (M_0(\boldsymbol{g}) M_1(\boldsymbol{s}_p) + M_2(\boldsymbol{g}) d)\mathbb{E}[\|\boldsymbol{X} - \boldsymbol{Z}\|_2].
\end{aligned}$$

By the Cauchy-Schwarz inequality and the fact that $\min\{a, b\} \leq \sqrt{ab}$ for $a, b \geq 0$, the first term can be upper bounded as follows:

$$\begin{aligned}
\mathbb{E}[|\langle \boldsymbol{s}_p(\boldsymbol{Z}), \boldsymbol{g}(\boldsymbol{Z}) - \boldsymbol{g}(\boldsymbol{X})\rangle|] &\leq \mathbb{E}[\min\{2M_0(\boldsymbol{g}), M_1(\boldsymbol{g})\|\boldsymbol{X} - \boldsymbol{Z}\|_2\} \, \|\boldsymbol{s}_p(\boldsymbol{Z})\|_2] \\
&\leq (2M_0(\boldsymbol{g}) M_1(\boldsymbol{g}))^{1/2} \mathbb{E}\left[\|\boldsymbol{X} - \boldsymbol{Z}\|_2^{1/2} \|\boldsymbol{s}_p(\boldsymbol{Z})\|_2\right] \\
&\leq \sqrt{2M_0(\boldsymbol{g}) M_1(\boldsymbol{g}) \mathbb{E}[\|\boldsymbol{X} - \boldsymbol{Z}\|_2] \mathbb{E}_P[\|\boldsymbol{s}_p(\boldsymbol{Z})\|_2^2]}.
\end{aligned}$$

Taking the infimum of these bounds over all couplings of $\boldsymbol{X} \sim \mu$ and $\boldsymbol{Z} \sim P$ yields the desired result. □

2.3 Theorem (KSDs detect convergence)**.** *Let k be a kernel such that $\nabla_{\boldsymbol{x}}^\ell \nabla_{\boldsymbol{y}}^\ell k(\boldsymbol{x}, \boldsymbol{y})$ is continuous and uniformly bounded for $\ell = 0, 1, 2$. If $\boldsymbol{s}_p$ is Lipschitz with $\mathbb{E}_P[\|\boldsymbol{s}_p(\boldsymbol{Z})\|_2^2] < \infty$, then $\mathcal{S}(Q_n, \mathcal{T}_p, \mathcal{G}_k) \to 0$ whenever $Q_n \Rightarrow P$.*

Proof. For any function $f : \mathbb{R}^d \to \mathbb{R}$ for which the appropriate derivatives exist, define

$$D^\alpha f(\boldsymbol{x}) := \frac{\partial^{|\alpha|}}{(dx_1)^{\alpha_1} \dots (dx_d)^{\alpha_d}} f(\boldsymbol{x}),$$

where $|\alpha| = \sum_{j=1}^d \alpha_j$. Take any $\boldsymbol{g} \in \mathcal{G}_k$, and choose any $\alpha \in \mathbb{N}^d$ such that $|\alpha| \leq 2$. Then by Cauchy-Schwarz we have

$$\sup_{\boldsymbol{x} \in \mathbb{R}^d} |D^\alpha g_j(\boldsymbol{x})| = \sup_{\boldsymbol{x} \in \mathbb{R}^d} |D^\alpha \langle g_j, k(\boldsymbol{x}, \cdot)\rangle_{\mathcal{H}_k}| \leq \sup_{\boldsymbol{x} \in \mathbb{R}^d} \|g_j\|_{\mathcal{H}_k} \|D^\alpha k(\boldsymbol{x}, \cdot)\|_{\mathcal{H}_k} = \|g_j\|_{\mathcal{H}_k} \sup_{\boldsymbol{x} \in \mathbb{R}^d} (D_{\boldsymbol{x}}^\alpha D_{\boldsymbol{y}}^\alpha k(\boldsymbol{x}, \boldsymbol{x}))^{1/2}.$$

Since $\sum_{j=1}^d \|g_j\|_{\mathcal{H}_k}^2 \leq 1$ for all $\boldsymbol{g} \in \mathcal{G}_k$ and $D_{\boldsymbol{x}}^\alpha D_{\boldsymbol{y}}^\alpha k(\boldsymbol{x}, \boldsymbol{x})$ is uniformly bounded in $\boldsymbol{x}$ for all $|\alpha| \leq 2$, the entries of $\boldsymbol{g}(\boldsymbol{x})$, $\nabla_{\boldsymbol{x}} \boldsymbol{g}(\boldsymbol{x})$ and $\nabla_{\boldsymbol{x}}^2 \boldsymbol{g}(\boldsymbol{x})$ are uniformly bounded in $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{g} \in \mathcal{G}_k$. So there exists $\lambda > 0$ such that $\sup_{\boldsymbol{g} \in \mathcal{G}_k} \max\{M_0(\boldsymbol{g}), M_1(\boldsymbol{g}), M_2(\boldsymbol{g})\} \leq \lambda < \infty$. The result then follows from Lemma 2.2 as

$$\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k) \leq \lambda \left((M_1(\boldsymbol{s}_p) + d) W_2(\mu, P) + \sqrt{2\mathbb{E}_P[\|\boldsymbol{s}_p(\boldsymbol{Z})\|_2^2] W_2(\mu, P)} \right). \qquad \square$$

This more general result shows in particular that the IMQ KSD detects convergence for targets $P \in \mathcal{P}$ with $\mathbb{E}_P[\|\boldsymbol{s}_p(\boldsymbol{Z})\|_2^2] < \infty$.

2.2 Detecting non-convergence: lower bounding the KSD

A KSD detects non-convergence if $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$ implies $Q_n \Rightarrow P$. The proof of this direction is eminently more involved. While Gorham and Mackey [2017] showed that KSDs based on many different popular kernels detect non-convergence in the univariate case, many of these fail to replicate this property in higher dimensions. This failure lies in their inability to enforce *uniform tightness* (defined below).

Fortunately, the **IMQ kernel** $k(\boldsymbol{x}, \boldsymbol{y}) = (c^2 + \|\boldsymbol{x} - \boldsymbol{y}\|_2^2)^\beta$ with $c > 0$ and $\beta \in (-1, 0)$ has sufficiently heavy tails for its corresponding KSD to enforce such tightness, as we show in this section. We break the proof down into two parts: we first prove a lower bound for a general KSD in terms of the *tightness rate* (also defined below), and then show that the IMQ KSD automatically enforces tightness.

2.2.1 Tightness lower bound

As promised, we start with two definitions.

2.4 Definition (Uniform tightness)**.** *A sequence of probability measures $(\mu_n)_{n\geq 1}$ is said to be* **uniformly tight** *if for every $\epsilon > 0$, there exists $R(\epsilon) < \infty$ such that*

$$\limsup_n \mu_n(\|\boldsymbol{X}\|_2 > R(\epsilon)) \leq \epsilon.$$

The uniform tightness condition essentially states that no mass in the sequence $(\mu_n)_{n\geq 1}$ escapes to infinity. When k decays more rapidly than $\boldsymbol{s}_p$ grows, the KSD ignores excess mass in the tails and hence can be driven to zero by a non-tight sequence of increasingly diffuse probability measures.

2.5 Definition (Tightness rate)**.** *Let μ be a probability measure on $\mathbb{R}^d$ and $\epsilon > 0$. The* **tightness rate** *is defined as*

$$R(\mu, \epsilon) := \inf\{r \geq 0 : \mu(\|\boldsymbol{X}\|_2 > r) \leq \epsilon\}. \tag{2.3}$$

The main theorem of this section, which lower bounds the KSD in terms of the tightness rate, is stated as follows.

2.6 Theorem (KSD tightness lower bound)**.** *Suppose $P \in \mathcal{P}$ and let μ be a probability measure with tightness rate $R(\mu, \epsilon)$ defined in (2.3). Suppose further that $k(\boldsymbol{x}, \boldsymbol{y}) = \Psi(\boldsymbol{x} - \boldsymbol{y})$ with $\Psi \in C^2$ and $F(t) := \sup_{\|\boldsymbol{\omega}\|_\infty \leq t} \hat{\Psi}(\boldsymbol{\omega})^{-1}$ finite for all $t > 0$, where $\hat{\Psi}$ denotes the generalised Fourier transform of Ψ. Then there exists a constant $\mathcal{M}_P$ such that, for all $\rho, \epsilon, \delta > 0$,*

$$\begin{aligned} d_{BL_{\|\cdot\|_2}}(\mu, P) \leq & \rho\sqrt{d}(1 + M_1(\boldsymbol{s}_p)\mathcal{M}_P) + \epsilon + \min\{\epsilon, 1\}\left(2 + \epsilon + \left(M_1(\boldsymbol{s}_p)\rho\sqrt{d} + \tfrac{d\delta^{-1}\theta_{d-1}}{\theta_d}\right)\mathcal{M}_P\right) \\ & + (2\pi)^{-d/4} V_d^{1/2} \mathcal{M}_P (R(\mu, \epsilon) + 2\delta)^{d/2} F\left(\tfrac{12d\log 2}{\pi}(c_{\rho,\delta} + M_1(\boldsymbol{s}_p)\mathcal{M}_P)\epsilon^{-1}\right)^{1/2} \mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k), \end{aligned}$$

where $\theta_d := d\int_0^1 \exp\left\{-1/(1-r^2)\right\} r^{d-1} dr$ for $d > 0$, $\theta_0 := e^{-1}$, V_d is the volume of the unit Euclidean ball in dimension d, and

$$c_{\rho,\delta} := 1 + M_1(\boldsymbol{s}_p)\mathcal{M}_P(1 + d) + \left\{2 + (M_1(\boldsymbol{s}_p)\rho + 1/\rho)\sqrt{d}\mathcal{M}_P\right\} \tfrac{d\delta^{-1}\theta_{d-1}}{\theta_d} + \delta^{-2}\tfrac{22}{\theta_d}\mathcal{M}_P.$$

In order to prove this, we first prove two lemmas.

2.2.1.1 Stein approximations with finite RKHS norm

2.7 Lemma. *Suppose that $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d$ is bounded and belongs to $L^2(\mathbb{R}^d) \cap C^1(\mathbb{R}^d)$. Suppose further that $h = \mathcal{T}_P \boldsymbol{g}$ and $\boldsymbol{s}_p$ are Lipschitz, and that $k(\boldsymbol{x}, \boldsymbol{y}) = \Psi(\boldsymbol{x} - \boldsymbol{y})$ for some $\Psi \in C^2$ with generalised Fourier transform $\hat{\Psi}$. Then, for every $\epsilon > 0$, we can find $\boldsymbol{g}_\epsilon \in \mathcal{H}_k^d$ such that $\sup_{\boldsymbol{x}\in\mathbb{R}^d} |(\mathcal{T}_P \boldsymbol{g}_\epsilon)(\boldsymbol{x}) - (\mathcal{T}_P \boldsymbol{g})(\boldsymbol{x})| \leq \epsilon$ and*

$$\|\boldsymbol{g}_\epsilon\|_{\mathcal{H}_k^d} \leq (2\pi)^{-d/4} F\left(\tfrac{12d\log 2}{\pi}(M_1(h) + M_1(\boldsymbol{s}_p)M_0(\boldsymbol{g}))\epsilon^{-1}\right)^{1/2} \|\boldsymbol{g}\|_{L^2},$$

where $F(t) := \sup_{\|\boldsymbol{\omega}\|_\infty \leq t} \hat{\Psi}(\boldsymbol{\omega})^{-1}$.

Proof. First, define the function $S : \mathbb{R}^d \to \mathbb{R}$ by $S(\boldsymbol{x}) := \prod_{j=1}^d \frac{\sin x_j}{x_j}$. Then $S \in L^2$ and $\int_{\mathbb{R}^d} \|\boldsymbol{x}\|_2 S(\boldsymbol{x})^4 d\boldsymbol{x} < \infty$. If we also define the density function $\rho(\boldsymbol{x}) := Z^{-1} S(\boldsymbol{x})^4$, where $Z := \int_{\mathbb{R}^d} S(\boldsymbol{x})^4 d\boldsymbol{x} = (2\pi/3)^d$ is the normalisation constant, one can then check that $\hat{\rho}(\boldsymbol{\omega})^2 \leq (2\pi)^{-d} \mathbb{1}_{\{\|\boldsymbol{\omega}\|_\infty \leq 4\}}$.

Let $\boldsymbol{Y} \sim \rho$. For each $\delta > 0$, define $\rho_\delta(\boldsymbol{x}) := \delta^{-d}\rho(\boldsymbol{x}/\delta)$, and for any function $\boldsymbol{f}$, set $\boldsymbol{f}_\delta(\boldsymbol{x}) := \mathbb{E}[\boldsymbol{f}(\boldsymbol{x} - \delta\boldsymbol{Y})]$. Since $h = \mathcal{T}_P \boldsymbol{g}$ is assumed to be Lipschitz, this implies that $|h_\delta(\boldsymbol{x}) - h(\boldsymbol{x})| \leq \delta M_1(h)\mathbb{E}_\rho[\|\boldsymbol{Y}\|_2]$ for all $\boldsymbol{x} \in \mathbb{R}^d$.

Now, since

$$(\mathcal{T}_P \boldsymbol{g}_\delta)(\boldsymbol{x}) = \mathbb{E}_\rho[\langle \boldsymbol{s}_p(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{x} - \delta\boldsymbol{Y})\rangle] + \mathbb{E}[\nabla_{\boldsymbol{x}} \cdot \boldsymbol{g}(\boldsymbol{x} - \delta\boldsymbol{Y})]$$

and

$$h_\delta(\boldsymbol{x}) = \mathbb{E}_\rho[\langle \boldsymbol{s}_p(\boldsymbol{x} - \delta\boldsymbol{Y}), \boldsymbol{g}(\boldsymbol{x} - \delta\boldsymbol{Y})\rangle] + \mathbb{E}[\nabla_{\boldsymbol{x}} \cdot \boldsymbol{g}(\boldsymbol{x} - \delta\boldsymbol{Y})]$$

for all $\boldsymbol{x} \in \mathbb{R}^d$, we deduce that

$$\begin{aligned}
|(\mathcal{T}_P \boldsymbol{g}_\delta)(\boldsymbol{x}) - h_\delta(\boldsymbol{x})| &= |\mathbb{E}_\rho[\langle \boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_p(\boldsymbol{x} - \delta\boldsymbol{Y}), \boldsymbol{g}(\boldsymbol{x} - \delta\boldsymbol{Y})\rangle]| \\
&\leq \mathbb{E}_\rho[\|\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_p(\boldsymbol{x} - \delta\boldsymbol{Y})\|_2 \|\boldsymbol{g}(\boldsymbol{x} - \delta\boldsymbol{Y})\|_2] \\
&\leq M_0(\boldsymbol{g}) M_1(\boldsymbol{s}_p) \delta \mathbb{E}_\rho[\|\boldsymbol{Y}\|_2],
\end{aligned}$$

by Cauchy-Schwarz, Jensen's inequality and the fact that $\boldsymbol{s}_p$ is Lipschitz. Thus, for any $\delta > 0$,

$$|(\mathcal{T}_P \boldsymbol{g}_\delta)(\boldsymbol{x}) - (\mathcal{T}_P \boldsymbol{g})(\boldsymbol{x})| \leq |(\mathcal{T}_P \boldsymbol{g}_\delta)(\boldsymbol{x}) - h_\delta(\boldsymbol{x})| + |h_\delta(\boldsymbol{x}) - h(\boldsymbol{x})| \leq \delta(M_1(h) + M_1(\boldsymbol{s}_p) M_0(\boldsymbol{g})) \mathbb{E}_\rho[\|\boldsymbol{Y}\|_2], \quad (2.4)$$

by the triangle inequality. Letting $\tilde{\epsilon} = \epsilon / ((M_1(h) + M_1(\boldsymbol{s}_p) M_0(\boldsymbol{g})) \mathbb{E}_\rho[\|\boldsymbol{Y}\|_2])$, we have $M_0(\mathcal{T}_P \boldsymbol{g}_{\tilde{\epsilon}} - \mathcal{T}_P \boldsymbol{g}) \leq \epsilon$.

By Theorem 10.12 of Wendland [2004], $\mathcal{H}_k^d$ is precisely the subset of $L^2(\mathbb{R}^d) \cap C^1(\mathbb{R}^d)$ whose elements $\boldsymbol{f}$ have finite RKHS norm:

$$\|\boldsymbol{f}\|_{\mathcal{H}_k^d}^2 = \frac{1}{(2\pi)^{d/2}} \sum_{i=1}^d \int_{\mathbb{R}^d} \frac{|\hat{f}_i(\boldsymbol{\omega})|^2}{\hat{\Psi}(\boldsymbol{\omega})} d\boldsymbol{\omega} < \infty.$$

It is therefore enough to prove the bound on the RKHS norm of $\boldsymbol{g}_\delta$. First note that, by the convolution theorem, $\|\boldsymbol{g}_\delta\|_{\mathcal{H}_k^d}^2$ is equal to

$$\begin{aligned}
\frac{1}{(2\pi)^{d/2}} \sum_{i=1}^d \int_{\mathbb{R}^d} \frac{|\hat{g}_{\delta,i}(\boldsymbol{\omega})|^2}{\hat{\Psi}(\boldsymbol{\omega})} d\boldsymbol{\omega} &= (2\pi)^{d/2} \sum_{i=1}^d \int_{\mathbb{R}^d} \frac{|\hat{g}_i(\boldsymbol{\omega})|^2 \hat{\rho}_\delta(\boldsymbol{\omega})^2}{\hat{\Psi}(\boldsymbol{\omega})} d\boldsymbol{\omega} \\
&\leq (2\pi)^{-d/2} \left\{ \sup_{\|\boldsymbol{\omega}\|_\infty \leq 4\delta^{-1}} \hat{\Psi}(\boldsymbol{\omega})^{-1} \right\} \sum_{i=1}^d \int_{\mathbb{R}^d} |\hat{g}_i(\boldsymbol{\omega})|^2 d\boldsymbol{\omega},
\end{aligned}$$

where the inequality follows from the fact that $\hat{\rho}_\delta(\boldsymbol{\omega}) = \hat{\rho}(\delta\boldsymbol{\omega})$. Therefore, overloading the definition $\|\boldsymbol{f}\|_{L^2} := \sum_{i=1}^d \|f_i\|_{L^2}$, we have

$$\|\boldsymbol{g}_\delta\|_{\mathcal{H}_k^d} \leq (2\pi)^{-d/4} F (4\delta^{-1})^{1/2} \|\boldsymbol{g}\|_{L^2},$$

since $\|\boldsymbol{g}\|_{L^2} = \|\hat{\boldsymbol{g}}\|_{L^2}$ by Plancherel's theorem. The result follows from computing $\int_{\mathbb{R}} \sin^4(x)/x^4 dx = 2\pi/3$ and observing that

$$\int_{\mathbb{R}^d} \|\boldsymbol{x}\|_2 \prod_{j=1}^d \frac{\sin^4 x_j}{x_j^4} d\boldsymbol{x} \leq \int_{\mathbb{R}^d} \|\boldsymbol{x}\|_1 \prod_{j=1}^d \frac{\sin^4 x_j}{x_j^4} d\boldsymbol{x} = \sum_{j=1}^d \int_{\mathbb{R}^d} \frac{\sin^4 x_j}{|x_j|^3} \prod_{k \neq j} \frac{\sin^4 x_k}{x_k^4} d\boldsymbol{x} = 2d(\log 2) \left(\frac{2\pi}{3}\right)^{d-1},$$

which in turn implies that $\mathbb{E}_\rho[\|\boldsymbol{Y}\|_2] \leq \frac{3d \log 2}{\pi}$. □

2.2.1.2 Smoothed indicator function

2.8 Lemma. *For any compact set $K \subseteq \mathbb{R}^d$ and $\delta > 0$, define the* **set inflation** *$K^{2\delta} := \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{y}\|_2 \leq 2\delta$ for all $y \in K\}$. Then there exists a function $v_{K,\delta} : \mathbb{R}^d \to [0,1]$ such that*

1. *$v_{K,\delta}(\boldsymbol{x}) = 1$ for all $\boldsymbol{x} \in K$ and $v_{K,\delta}(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \notin K^{2\delta}$;*
2. *$\|\nabla_{\boldsymbol{x}} v_{K,\delta}(\boldsymbol{x})\|_2 \leq \frac{d\delta^{-1}\theta_{d-1}}{\theta_d} \mathbb{1}_{\{\boldsymbol{x} \in K^{2\delta} \setminus K\}}$;*
3. *$\|\nabla_{\boldsymbol{x}}^2 v_{K,\delta}(\boldsymbol{x})\|_{op} \leq \delta^{-2} \frac{22}{\theta_d} \mathbb{1}_{\{\boldsymbol{x} \in K^{2\delta} \setminus K\}}$.*

Proof. First, define the standard normalised *bump function*

$$\psi(\boldsymbol{x}) := J_d^{-1} \exp\left\{-1/(1-\|\boldsymbol{x}\|_2^2)\right\} \mathbb{1}_{\{\|\boldsymbol{x}\|_2<1\}},$$

where $J_d = \int_{\mathcal{B}(\mathbf{0},1)} \exp\left\{-1/(1-\|\boldsymbol{x}\|_2^2)\right\} d\boldsymbol{x} = \theta_d V_d$ is the normalising constant.

Letting $\boldsymbol{W} \sim \psi$, define $v_{K,\delta}(\boldsymbol{x}) := \mathbb{E}[\mathbb{1}_{\{\boldsymbol{x}+\delta\boldsymbol{W}\in K^\delta\}}]$ as the smoothed approximation of $\boldsymbol{x} \mapsto \mathbb{1}_{\{\boldsymbol{x}\in K\}}$. The support $\mathrm{supp}(\boldsymbol{W})$ of $\boldsymbol{W}$ is exactly $\mathcal{B}(\mathbf{0},1)$, so we can immediately conclude property 1.

It is also clear that $\mathrm{supp}(\nabla v_{K,\delta}) \subseteq K^{2\delta}\backslash K$. Therefore, to prove property 2, we need only consider $\boldsymbol{x} \in K^{2\delta}\backslash K$. Since

$$\nabla_{\boldsymbol{x}} v_{K,\delta}(\boldsymbol{x}) = \frac{1}{\delta^{d+1}} \int_{\mathcal{B}(\boldsymbol{x},\delta)} \nabla_{\boldsymbol{x}} \psi\left(\tfrac{\boldsymbol{x}-\boldsymbol{y}}{\delta}\right) \mathbb{1}_{\{\boldsymbol{y}\in K^\delta\}} d\boldsymbol{y}$$

by the Leibniz rule, setting $K^\delta_{\boldsymbol{x}} := \delta^{-1}(K^\delta - \boldsymbol{x})$ and using Jensen's inequality, we have

$$\begin{aligned}\|\nabla_{\boldsymbol{x}} v_{K,\delta}(\boldsymbol{x})\|_2 &\le \frac{1}{\delta^{d+1}} \int_{\mathcal{B}(\boldsymbol{x},\delta)\cap K^\delta} \|\nabla\psi(\tfrac{\boldsymbol{x}-\boldsymbol{y}}{\delta})\|_2 d\boldsymbol{y} \\ &= \delta^{-1} \int_{\mathcal{B}(\mathbf{0},1)\cap K^\delta_{\boldsymbol{x}}} \|\nabla_{\boldsymbol{z}}\psi(\boldsymbol{z})\|_2 d\boldsymbol{z} \le \delta^{-1} \int_{\mathcal{B}(\mathbf{0},1)} \|\nabla_{\boldsymbol{z}}\psi(\boldsymbol{z})\|_2 d\boldsymbol{z},\end{aligned}$$

where we used the substitution $\boldsymbol{z} := (\boldsymbol{x}-\boldsymbol{y})/\delta$. By substituting $r = \|\boldsymbol{z}\|_2$ and integrating by parts, we have

$$\begin{aligned}\int_{\mathcal{B}(\mathbf{0},1)} \|\nabla_{\boldsymbol{z}}\psi(\boldsymbol{z})\|_2 d\boldsymbol{z} &= J_d^{-1} \int_0^1 \tfrac{2r}{(1-r^2)^2} \exp\left\{-\tfrac{1}{1-r^2}\right\} (dV_d r^{d-1}) dr \\ &= \frac{d}{\theta_d}\left(\left[-r^{d-1}\exp\left\{-\tfrac{1}{1-r^2}\right\}\right]\Big|_{r=0}^{r=1} + \int_0^1 (d-1) r^{d-2} \exp\left\{-\tfrac{1}{1-r^2}\right\} dr\right) \\ &= \tfrac{d}{\theta_d}[e^{-1}\mathbb{1}_{\{d=1\}} + \mathbb{1}_{\{d>1\}}\theta_{d-1}] = \tfrac{d\theta_{d-1}}{\theta_d},\end{aligned}$$

yielding property 2.

Finally, for property 3, first observe that we again have $\mathrm{supp}(\nabla^2 v_{K,\delta}) \subseteq K^{2\delta} \setminus K$, so we only need to check for $\boldsymbol{x} \in K^{2\delta} \setminus K$. Analogously to above, we have

$$\|\nabla^2 v_{K,\delta}(\boldsymbol{x})\|_{op} \le \frac{1}{\delta^{d+2}} \int_{\mathcal{B}(\boldsymbol{x},\delta)\cap K^\delta} \|\nabla^2\psi\left(\tfrac{\boldsymbol{x}-\boldsymbol{y}}{\delta}\right)\|_{op} d\boldsymbol{y} \le \delta^{-2} M_1(\nabla\psi) \int_{\mathcal{B}(\mathbf{0},1)\cap K^\delta_{\boldsymbol{x}}} d\boldsymbol{z} \le \delta^{-2} M_1(\nabla\psi) V_d.$$

It is simple to show that

$$\nabla^2\psi(\boldsymbol{x}) = J_d^{-1} \frac{\exp\{-1/(1-\|\boldsymbol{x}\|_2^2)\}}{(1-\|\boldsymbol{x}\|_2^2)^2} \cdot \left[\frac{4}{(1-\|\boldsymbol{x}\|_2^2)^2}\boldsymbol{x}\boldsymbol{x}^\top - \frac{8}{1-\|\boldsymbol{x}\|_2^2}\boldsymbol{x}\boldsymbol{x}^\top - 2I_d\right] \mathbb{1}_{\{\|\boldsymbol{x}\|_2<1\}},$$

so by the triangle inequality we have

$$\|\nabla^2\psi(\boldsymbol{x})\|_{op} \le J_d^{-1} e^{-1/(1-r^2)} \left(\frac{4r^2}{(1-r^2)^4} + \frac{8r^2}{(1-r^2)^3} + \frac{2}{(1-r^2)^2}\right) \mathbb{1}_{\{r<1\}},$$

where $r = \|\boldsymbol{x}\|_2$. One can then observe by plotting the graph of the expression on the right that $M_1(\nabla\psi) \le 22 J_d^{-1}$ and so $M_1(\nabla v_{K,\delta}) \le \delta^{-2}\frac{22}{\theta_d}$, as desired. □

2.2.1.3 Proof of Theorem 2.6

We will now prove the tightness lower bound of Theorem 2.6. To do so, we first note that by Theorem 5 and Section 4.2 of Gorham et al. [2018], for any $h \in \mathrm{BL}_{\|\cdot\|_2}$ there exists $\boldsymbol{g} \in C^1$ which solves the Stein equation $\mathcal{T}_P\boldsymbol{g} = h - \mathbb{E}_P[h(\boldsymbol{Z})]$ and satisfies $M_0(\boldsymbol{g}) \le \mathcal{M}_P$, where $\mathcal{M}_P$ is a constant independent of h and $\boldsymbol{g}$.

Proof of Theorem 2.6. Let $h \in \mathrm{BL}_{\|\cdot\|_2}$, and choose $\boldsymbol{g} \in C^1$ which solves the Stein equation and satisfies $M_0(\boldsymbol{g}) \le \mathcal{M}_P$. Next, we will show that we can approximate $\mathcal{T}_P\boldsymbol{g}$ arbitrarily well by a function in a scaled copy of $\mathcal{T}_P\mathcal{G}_k$.

Smoothing g by convolution. Fix any $\rho > 0$, and define $\boldsymbol{g}_\rho(\boldsymbol{x}) := \mathbb{E}[\boldsymbol{g}(\boldsymbol{x} - \rho\boldsymbol{U})]$, where $\boldsymbol{U} \sim \mathcal{N}(\boldsymbol{0}, I_d)$. We have $M_0(\boldsymbol{g}_\rho) \le M_0(\boldsymbol{g}) \le \mathcal{M}_P$ and, integrating by parts, we get

$$M_1(\boldsymbol{g}_\rho) = \sup_{\boldsymbol{x}\in\mathbb{R}^d} \|\mathbb{E}[\nabla_{\boldsymbol{x}} \boldsymbol{g}(\boldsymbol{x} - \rho\boldsymbol{U})]\|_2 = \sup_{\boldsymbol{x}\in\mathbb{R}^d} \|\mathbb{E}[\boldsymbol{g}(\boldsymbol{x} - \rho\boldsymbol{U})\boldsymbol{U}]\|_2/\rho \le M_0(\boldsymbol{g})\mathbb{E}[\|\boldsymbol{U}\|_2]/\rho \le \mathcal{M}_P\sqrt{d}/\rho.$$

The same argument as in (2.4) gives

$$M_0(\mathcal{T}_P\boldsymbol{g}_\rho - \mathcal{T}_P\boldsymbol{g}) \le \rho(M_1(\mathcal{T}_P\boldsymbol{g}) + M_1(\boldsymbol{s}_p)\mathcal{M}_P)\mathbb{E}[\|\boldsymbol{U}\|_2] \le \rho\sqrt{d}(1 + M_1(\boldsymbol{s}_p)\mathcal{M}_P).$$

Finally, we show that $M_0(\mathcal{T}_P\boldsymbol{g}_\rho)$ and $M_1(\mathcal{T}_P\boldsymbol{g}_\rho)$ are bounded uniformly in $\boldsymbol{g}$. Setting $h_\rho(\boldsymbol{x}) := \mathbb{E}[h(\boldsymbol{x} - \rho\boldsymbol{U})]$, note that $(\mathcal{T}_P\boldsymbol{g}_\rho)(\boldsymbol{x}) = h_\rho(\boldsymbol{x}) - \mathbb{E}_{\boldsymbol{Z}\sim P}[h(\boldsymbol{Z})] + \mathbb{E}_{\boldsymbol{U}\sim\mathcal{N}(\boldsymbol{0},I_d)}[\langle \boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_p(\boldsymbol{x} - \rho\boldsymbol{U}), \boldsymbol{g}(\boldsymbol{x} - \rho\boldsymbol{U})\rangle]$. Thus,

$$M_0(\mathcal{T}_P\boldsymbol{g}_\rho) \le M_0(h_\rho - \mathbb{E}_P[h_\rho(\boldsymbol{Z})]) + M_1(\boldsymbol{s}_p)\rho\mathbb{E}[\|\boldsymbol{U}\|_2]M_0(\boldsymbol{g}_\rho) \le 2 + M_1(\boldsymbol{s}_p)\rho\sqrt{d}\mathcal{M}_P,$$

where we have used the fact that $h_\rho \in \mathrm{BL}_{\|\cdot\|_2}$. Furthermore, integrating by parts gives

$$\begin{aligned}
M_1(\mathcal{T}_P\boldsymbol{g}_\rho) &\le M_1(h_\rho - \mathbb{E}_P[h(\boldsymbol{Z})]) + M_1(\mathcal{T}_P\boldsymbol{g}_\rho - h_\rho + \mathbb{E}_P[h(\boldsymbol{Z})]) \\
&\le 1 + M_1(\boldsymbol{s}_p)M_0(\boldsymbol{g}_\rho) + \sup_{\boldsymbol{x}\in\mathbb{R}^d} \|\mathbb{E}[\nabla_{\boldsymbol{U}}\langle \boldsymbol{s}_p(\boldsymbol{x} - \rho\boldsymbol{U}) - \boldsymbol{s}_p(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{x} - \rho\boldsymbol{U})\rangle]/\rho\|_2 \\
&\le 1 + M_1(\boldsymbol{s}_p)\mathcal{M}_P + \sup_{\boldsymbol{x}\in\mathbb{R}^d} \|\mathbb{E}[\boldsymbol{U}\langle \boldsymbol{s}_p(\boldsymbol{x} - \rho\boldsymbol{U}) - \boldsymbol{s}_p(\boldsymbol{x}), \boldsymbol{g}(\boldsymbol{x} - \rho\boldsymbol{U})\rangle]/\rho\|_2 \\
&\le 1 + M_1(\boldsymbol{s}_p)\mathcal{M}_P + \mathbb{E}[\|\boldsymbol{U}\|_2^2]M_1(\boldsymbol{s}_p)M_0(\boldsymbol{g}_\rho) \le 1 + (1 + d)M_1(\boldsymbol{s}_p)\mathcal{M}_P.
\end{aligned}$$

Truncating $\boldsymbol{g}_\rho$. Let $\epsilon, \delta > 0$, and let $K = \mathcal{B}(\boldsymbol{0}, R(\mu, \epsilon))$ in Lemma 2.8. It then follows that we can set $\boldsymbol{g}_{K,\delta}(\boldsymbol{x}) := \boldsymbol{g}_\rho(\boldsymbol{x})v_{K,\delta}(\boldsymbol{x})$ as a smooth, truncated approximation to $\boldsymbol{g}_\rho$. We have

$$\begin{aligned}
(\mathcal{T}_P\boldsymbol{g}_\rho)(\boldsymbol{x}) - (\mathcal{T}_P\boldsymbol{g}_{K,\delta})(\boldsymbol{x}) &= (1 - v_{K,\delta}(\boldsymbol{x}))\left\{\boldsymbol{s}_p(\boldsymbol{x})^\top\boldsymbol{g}_\rho(\boldsymbol{x}) + \nabla_{\boldsymbol{x}}\cdot\boldsymbol{g}_\rho(\boldsymbol{x})\right\} + \nabla_{\boldsymbol{x}}v_{K,\delta}(\boldsymbol{x})^\top\boldsymbol{g}_\rho(\boldsymbol{x}) \\
&= (1 - v_{K,\delta}(\boldsymbol{x}))(\mathcal{T}_P\boldsymbol{g}_\rho)(\boldsymbol{x}) + \nabla_{\boldsymbol{x}}v_{K,\delta}(\boldsymbol{x})^\top\boldsymbol{g}_\rho(\boldsymbol{x}),
\end{aligned}$$

so properties 1 and 2 of Lemma 2.8 imply that $(\mathcal{T}_P\boldsymbol{g}_\rho)(\boldsymbol{x}) = (\mathcal{T}_P\boldsymbol{g}_{K,\delta})(\boldsymbol{x})$ for $\boldsymbol{x} \in K$, $(\mathcal{T}_P\boldsymbol{g}_{K,\delta})(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \notin K^{2\delta}$, and

$$\begin{aligned}
|(\mathcal{T}_P\boldsymbol{g}_\rho)(\boldsymbol{x}) - (\mathcal{T}_P\boldsymbol{g}_{K,\delta})(\boldsymbol{x})| &\le |(\mathcal{T}_P\boldsymbol{g}_\rho)(\boldsymbol{x})| + \|\nabla_{\boldsymbol{x}}v_{K,\delta}(\boldsymbol{x})\|_2\|\boldsymbol{g}_\rho(\boldsymbol{x})\|_2 \\
&\le |(\mathcal{T}_P\boldsymbol{g}_\rho)(\boldsymbol{x})| + \tfrac{d\delta^{-1}\theta_{d-1}}{\theta_d}\|\boldsymbol{g}_\rho(\boldsymbol{x})\|_2 \le 2 + (M_1(\boldsymbol{s}_p)\rho\sqrt{d} + \tfrac{d\delta^{-1}\theta_{d-1}}{\theta_d})\mathcal{M}_P
\end{aligned}$$

for $\boldsymbol{x} \in K^{2\delta} \setminus K$ by Cauchy-Schwarz. Furthermore,

$$\begin{aligned}
M_1(\mathcal{T}_P\boldsymbol{g}_{K,\delta}) &\le M_1((\mathcal{T}_P\boldsymbol{g}_\rho)v_{K,\delta}) + M_1(\nabla v_{K,\delta}{}^\top\boldsymbol{g}_\rho) \\
&\le M_1(\mathcal{T}_P\boldsymbol{g}_\rho)M_0(v_{K,\delta}) + M_0(\mathcal{T}_P\boldsymbol{g}_\rho)M_1(v_{K,\delta}) + M_1(\nabla v_{K,\delta})M_0(\boldsymbol{g}_\rho) + M_0(\nabla v_{K,\delta})M_1(\boldsymbol{g}_\rho) \\
&\le 1 + M_1(\boldsymbol{s}_p)\mathcal{M}_P(1 + d) + (2 + M_1(\boldsymbol{s}_p)\rho\sqrt{d}\mathcal{M}_P)\tfrac{d\delta^{-1}\theta_{d-1}}{\theta_d} + \delta^{-2}\tfrac{22}{\theta_d}\mathcal{M}_P + \tfrac{d\delta^{-1}\theta_{d-1}}{\theta_d}\mathcal{M}_P\sqrt{d}/\rho \\
&=: c_{\rho,\delta}.
\end{aligned}$$

By Lemma 2.7, there is a function $\tilde{\boldsymbol{g}}_\epsilon \in \mathcal{H}_k^d$ such that $|(\mathcal{T}_P\tilde{\boldsymbol{g}}_\epsilon)(\boldsymbol{x}) - (\mathcal{T}_P\boldsymbol{g}_{K,\delta})(\boldsymbol{x})| \le \epsilon$ and

$$\|\tilde{\boldsymbol{g}}_\epsilon\|_{\mathcal{H}_k^d} \le (2\pi)^{-d/4}F\left(\tfrac{12d\log 2}{\pi}(c_{\rho,\delta} + M_1(\boldsymbol{s}_p)\mathcal{M}_P\epsilon^{-1})\right)^{1/2}\|\boldsymbol{g}_{K,\delta}\|_{L^2}. \tag{2.5}$$

Then, since $\mathcal{T}_P\boldsymbol{g}_{K,\delta}$ and $\mathcal{T}_P\boldsymbol{g}_\rho$ are identical on K, we have $|(\mathcal{T}_P\tilde{\boldsymbol{g}}_\epsilon)(\boldsymbol{x}) - (\mathcal{T}_P\boldsymbol{g}_\rho)(\boldsymbol{x})| \le \epsilon$ for all $\boldsymbol{x} \in K$. For $\boldsymbol{x} \notin K$, the triangle inequality gives

$$\begin{aligned}
|(\mathcal{T}_P\tilde{\boldsymbol{g}})(\boldsymbol{x}) - (\mathcal{T}_P\boldsymbol{g}_\rho)(\boldsymbol{x})| &\le |(\mathcal{T}_P\tilde{\boldsymbol{g}}_\epsilon)(\boldsymbol{x}) - (\mathcal{T}_P\boldsymbol{g}_{K,\delta})(\boldsymbol{x})| + |(\mathcal{T}_P\boldsymbol{g}_{K,\delta})(\boldsymbol{x}) - (\mathcal{T}_P\boldsymbol{g}_\rho)(\boldsymbol{x})| \\
&\le 2 + \epsilon + (M_1(\boldsymbol{s}_p)\rho\sqrt{d} + \tfrac{d\delta^{-1}\theta_{d-1}}{\theta_d})\mathcal{M}_P.
\end{aligned}$$

We can upper bound $\|\boldsymbol{g}_{K,\delta}\|_{L^2}$ in (2.5) as $\|\boldsymbol{g}_{K,\delta}\|_{L^2} \leq \mathrm{Vol}(K^{2\delta})^{1/2} M_0(\boldsymbol{g}_\rho) \leq \mathrm{Vol}(K^{2\delta})^{1/2} \mathcal{M}_P$.

Our choice of K ensures that $\mu(\mathbb{1}_{\{\boldsymbol{X} \notin K\}}) \leq \min\{\epsilon, 1\}$, so

$$
\begin{aligned}
&|\mathbb{E}_\mu[h(\boldsymbol{X})] - \mathbb{E}_P[h(\boldsymbol{Z})]| = |\mathbb{E}_\mu[(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{X})]| \\
&\leq |\mathbb{E}_\mu[(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{X}) - (\mathcal{T}_P \boldsymbol{g}_\rho)(\boldsymbol{X})]| + |\mathbb{E}_\mu[(\mathcal{T}_P \boldsymbol{g}_\rho)(\boldsymbol{X}) - (\mathcal{T}_P \tilde{\boldsymbol{g}}_\epsilon)(\boldsymbol{X})]| + |\mathbb{E}_\mu[(\mathcal{T}_P \tilde{\boldsymbol{g}}_\epsilon)(\boldsymbol{X})]| \\
&\leq M_0(\mathcal{T}_P \boldsymbol{g} - \mathcal{T}_P \boldsymbol{g}_\rho) + |\mathbb{E}_\mu[(\mathcal{T}_P \boldsymbol{g}_\rho)(\boldsymbol{X}) - (\mathcal{T}_P \tilde{\boldsymbol{g}}_\epsilon)(\boldsymbol{X}) \mathbb{1}_{\{\boldsymbol{X} \in K\}}]| \\
&\qquad\qquad + |\mathbb{E}_\mu[(\mathcal{T}_P \boldsymbol{g}_\rho)(\boldsymbol{X}) - (\mathcal{T}_P \tilde{\boldsymbol{g}}_\epsilon)(\boldsymbol{X}) \mathbb{1}_{\{\boldsymbol{X} \notin K\}}]| + |\mathbb{E}_\mu[(\mathcal{T}_P \tilde{\boldsymbol{g}}_\epsilon)(\boldsymbol{X})]| \\
&\leq \rho\sqrt{d}(1 + M_1(\boldsymbol{s}_p)\mathcal{M}_P) + \epsilon + \min\{\epsilon, 1\}(2 + \epsilon + (M_1(\boldsymbol{s}_p)\rho\sqrt{d} + \tfrac{d\delta^{-1}\theta_{d-1}}{\theta_d})\mathcal{M}_P) + \|\tilde{\boldsymbol{g}}_\epsilon\|_{\mathcal{H}_k^d} \mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k) \\
&\leq \rho\sqrt{d}(1 + M_1(\boldsymbol{s}_p)\mathcal{M}_P) + \epsilon + \min\{\epsilon, 1\}(2 + \epsilon + (M_1(\boldsymbol{s}_p)\rho\sqrt{d} + \tfrac{d\delta^{-1}\theta_{d-1}}{\theta_d})\mathcal{M}_P) \\
&\qquad + (2\pi)^{-d/4} F\left(\tfrac{12d\log 2}{\pi}(c_{\rho,\delta} + M_1(\boldsymbol{s}_p)\mathcal{M}_P \epsilon^{-1})\right)^{1/2} \mathrm{Vol}(\mathcal{B}(\mathbf{0}, R(\mu, \epsilon) + 2\delta))^{1/2} \mathcal{M}_P \mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k)
\end{aligned}
$$

The result follows from substituting in $\mathrm{Vol}(\mathcal{B}(\mathbf{0}, r)) = V_d r^d$ and then taking the supremum over all $h \in \mathrm{BL}_{\|\cdot\|_2}$. □

2.2.2 IMQ KSD detects non-convergence

Our next theorem shows that KSDs based on IMQ kernels automatically enforce tightness, and hence detect non-convergence, whenever $\beta \in (-1, 0)$.

2.9 Theorem (IMQ KSD detects non-convergence). *Suppose $P \in \mathcal{P}$ and $k(\boldsymbol{x}, \boldsymbol{y}) = (c^2 + \|\boldsymbol{x} - \boldsymbol{y}\|_2^2)^\beta$ for some $c > 0$ and $\beta \in (-1, 0)$. If $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$, then $Q_n \Rightarrow P$ as $n \to \infty$.*

In order to prove this we first prove two lemmas, and then Theorem 2.12, from which Theorem 2.9 quickly follows.

2.2.2.1 Generalised multiquadric Stein sets yield coercive functions

Set $\Psi_{c,\beta}(\boldsymbol{x}) := (c^2 + \|\boldsymbol{x}\|_2^2)^\beta$, where $c > 0$ and $\beta \in (-1, 0)$. In this section, we will show that we can find functions in $\mathcal{T}_P \mathcal{G}_k$ which grow sufficiently rapidly as $\|\boldsymbol{x}\|_2 \to \infty$. To do so we use the following results from Wendland [2004]:

- $\Psi_{c,\beta}$ has generalised Fourier transform

$$
\hat{\Psi}_{c,\beta}(\boldsymbol{\omega}) = \frac{2^{1+\beta}}{\Gamma(-\beta)} \left(\frac{\|\boldsymbol{\omega}\|_2}{c}\right)^{-\beta - d/2} K_{\beta + d/2}(c\|\boldsymbol{\omega}\|_2), \tag{2.6}
$$

  where $K_v(z)$ is the *modified Bessel function of the third kind*.

- If we further define

$$
\tau_v := \begin{cases} \sqrt{\pi/2} & \text{if } |v| \geq 1/2, \\ \frac{\sqrt{\pi} 3^{|v| - 1/2}}{2^{|v|+1}\Gamma(|v| + 1/2)} & \text{if } |v| < 1/2, \end{cases}
$$

  then we have the following bounds on $K_v(z)$ for $v \in \mathbb{R}$ and $z > 0$:

  B1. $K_v(z) \geq \tau_v \, e^{-z}/\sqrt{z}$ for $z \geq 1$;

  B2. $K_v(z) \geq e^{-1} \, \tau_v \, z^{-|v|}$ for $z \leq 1$;

  B3. $K_v(z) \leq \sqrt{2\pi/z} \, e^{-z + v^2/(2z)}$ for $z > 0$;

  B4. $K_v(z) \leq 2^{|v|-1} \, \Gamma(|v|) \, z^{-|v|}$ for $v \neq 0, z > 0$.

2.10 Lemma. *Let $P \in \mathcal{P}$ and $k(\boldsymbol{x}, \boldsymbol{y}) = \Psi_{c,\beta}(\boldsymbol{x} - \boldsymbol{y})$. Then for any $\alpha \in (0, \frac{1}{2}(\beta+1))$ and $a > c/2$, there exist functions $\mathring{\boldsymbol{g}} \in \mathcal{G}$ and $\mathcal{D}(a, c, \alpha, \beta, d)$ such that $\mathcal{T}_P \mathring{\boldsymbol{g}}$ is bounded below by*

$$\zeta(a, c, \alpha, \beta, d) := -\frac{\mathcal{D}(a, c, \alpha, \beta, d)^{1/2}}{2\alpha} \left[\frac{M_1(\boldsymbol{s}_p) R_0^2 + \|\boldsymbol{s}_p(\boldsymbol{0})\|_2 R_0 + d}{a^{2(1-\alpha)}} \right],$$

where $R_0 := \inf\{r > 0 : \omega(r') \geq 0 \; \forall r' \geq r\}$. Furthermore,

$$\liminf \|\boldsymbol{x}\|_2^{-2\alpha} (\mathcal{T}_P \mathring{\boldsymbol{g}})(\boldsymbol{x}) \geq \frac{\alpha}{\mathcal{D}(a, c, \alpha, \beta, d)^{1/2}} \omega_0 \quad \text{as } \|\boldsymbol{x}\|_2 \to \infty. \tag{2.7}$$

Proof. Define the function $\boldsymbol{g} : \mathcal{X} \to \mathbb{R}^d$ via

$$g_j(\boldsymbol{x}) := \nabla_{x_j} \Psi_{a,\alpha}(\boldsymbol{x}) = 2\alpha x_j (a^2 + \|\boldsymbol{x}\|_2^2)^{\alpha - 1}, \;\; j = 1, \ldots, d.$$

Recall that $\boldsymbol{g} \in \mathcal{H}_k^d$ if and only if $\|g_j\|_{\mathcal{H}_k} = \left\| \hat{g}_j / \sqrt{\hat{\Psi}_{c,\beta}} \right\|_{L^2} < \infty$ for each $j = 1, \ldots, d$. Noting that $\hat{g}_j(\boldsymbol{\omega}) = (i\omega_j)\hat{\Psi}_{c,\beta}(\boldsymbol{\omega})$, we have

$$\begin{aligned}
\|\boldsymbol{g}\|_{\mathcal{H}_k^d}^2 = \sum_{j=1}^d \|g_j\|_{\mathcal{H}_k}^2 &= \sum_{j=1}^d (2\pi)^{-d/2} \int_{\mathbb{R}^d} \hat{g}_j(\boldsymbol{\omega}) \overline{\hat{g}_j(\boldsymbol{\omega})} / \hat{\Psi}_{c,\beta}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\
&= (2\pi)^{-d/2} \sum_{j=1}^d \int_{\mathbb{R}^d} \frac{2^{2(1+\alpha)}/\Gamma(-\alpha)^2}{2^{1+\beta}/\Gamma(-\beta)} \frac{a^{2\alpha+d}}{c^{\beta+d/2}} \omega_j^2 \|\boldsymbol{\omega}\|_2^{\beta-2\alpha-d/2} \frac{K_{\alpha+d/2}(a\|\boldsymbol{\omega}\|_2)^2}{K_{\beta+d/2}(c\|\boldsymbol{\omega}\|_2)} d\boldsymbol{\omega} \\
&= \underbrace{(2\pi)^{-d/2} \frac{2^{2(1+\alpha)}/\Gamma(-\alpha)^2}{2^{1+\beta}/\Gamma(-\beta)} \frac{a^{2\alpha+d}}{c^{\beta+d/2}}}_{=: \mathcal{C}_0(a,c,\alpha,\beta,d)} \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_2^{\beta-2\alpha-d/2+2} \frac{K_{\alpha+d/2}(a\|\boldsymbol{\omega}\|_2)^2}{K_{\beta+d/2}(c\|\boldsymbol{\omega}\|_2)} d\boldsymbol{\omega}.
\end{aligned}$$

We decompose the integral into two parts: one over the unit ball $\mathcal{B}(\boldsymbol{0}, 1)$ and the other over its complement $\mathcal{B}(\boldsymbol{0}, 1)^c$. Using bounds B2 and B4, and the substitution $r = \|\boldsymbol{\omega}\|_2$, we have

$$\begin{aligned}
\int_{\mathcal{B}(\boldsymbol{0},1)} \|\boldsymbol{\omega}\|_2^{\beta-2\alpha-d/2+2} \frac{K_{\alpha+d/2}(a\|\boldsymbol{\omega}\|_2)^2}{K_{\beta+d/2}(c\|\boldsymbol{\omega}\|_2)} d\boldsymbol{\omega} &\leq \int_{\mathcal{B}(\boldsymbol{0},1)} \|\boldsymbol{\omega}\|_2^{\beta-2\alpha-d/2+2} \frac{2^{2\alpha+d-2}\Gamma(\alpha+d/2)^2 (a\|\boldsymbol{\omega}\|_2)^{-2\alpha-d}}{e^{-1}\tau_{\beta+d/2}\|c\boldsymbol{\omega}\|_2^{-\beta-d/2}} d\boldsymbol{\omega} \\
&= 2^{2\alpha+d-2}\Gamma(\alpha+d/2)^2 \frac{e}{\tau_{\beta+d/2}} \frac{c^{\beta+d/2}}{a^{2\alpha+d}} \int_{\mathcal{B}(\boldsymbol{0},1)} \|\boldsymbol{\omega}\|_2^{2\beta-4\alpha-d+2} e^{c\|\boldsymbol{\omega}\|_2} d\boldsymbol{\omega} \\
&= \underbrace{dV_d 2^{2\alpha+d-2}\Gamma(\alpha+d/2)^2 \frac{e}{\tau_{\beta+d/2}} \frac{c^{\beta+d/2}}{a^{2\alpha+d}}}_{=: \mathcal{C}_1(a,c,\alpha,\beta,d)} \int_0^1 r^{2\beta-4\alpha+1} e^{cr} dr,
\end{aligned}$$

where V_d is the volume of the d-dimensional unit ball. By our choice of α, we can bound the integral as

$$\int_0^1 r^{2\beta-4\alpha+1} e^{cr} dr \leq e^c \int_0^1 r^{2\beta-4\alpha+1} dr = \frac{e^c}{2\beta - 4\alpha + 2}.$$

Using bounds B1 and B3, we have

$$\begin{aligned}
\int_{\mathcal{B}(\boldsymbol{0},1)^c} \|\boldsymbol{\omega}\|_2^{\beta-2\alpha-d/2+2} \frac{K_{\alpha+d/2}(a\|\boldsymbol{\omega}\|_2)^2}{K_{\beta+d/2}(c\|\boldsymbol{\omega}\|_2)} d\boldsymbol{\omega} &\leq \int_{\mathcal{B}(\boldsymbol{0},1)^c} \|\boldsymbol{\omega}\|_2^{\beta-2\alpha-d/2+2} \frac{2\pi/(a\|\boldsymbol{\omega}\|_2) \cdot e^{-2a\|\boldsymbol{\omega}\|_2 + (\alpha+d/2)^2/(a\|\boldsymbol{\omega}\|_2)}}{\tau_{\beta+d/2} e^{-c\|\boldsymbol{\omega}\|_2} / \sqrt{c\|\boldsymbol{\omega}\|_2}} d\boldsymbol{\omega} \\
&\leq \frac{2\pi\sqrt{c}}{a\tau_{\beta+d/2}} \int_{\mathcal{B}(\boldsymbol{0},1)^c} \|\boldsymbol{\omega}\|_2^{\beta-2\alpha-d/2+3/2} e^{(c-2a)\|\boldsymbol{\omega}\|_2 + (\alpha+d/2)^2/(a\|\boldsymbol{\omega}\|_2)} d\boldsymbol{\omega} \\
&= dV_d \frac{2\pi\sqrt{c}}{a\tau_{\beta+d/2}} \int_1^\infty r^{\beta-2\alpha+d/2+1/2} e^{(c-2a)r + (\alpha+d/2)^2/(ar)} dr.
\end{aligned}$$

Recalling that $c - 2a < 0$, and using the substitution $t = (2a - c)r$, the integral above can be upper bounded as follows:

$$\int_1^\infty r^{\beta-2\alpha+d/2+1/2} e^{(c-2\alpha)r+(\alpha+d/2)^2/(ar)} dr \le e^{(c-2a)+(\alpha+d/2)^2/a} \int_1^\infty r^{\beta-2\alpha+d/2+1/2} e^{(c-2a)r} dr$$
$$= e^{(c-2a)+(\alpha+d/2)^2/a} (2a-c)^{-\beta+2\alpha-d/2-3/2} \Gamma(\beta - 2\alpha + d/2 + 3/2, 2a - c) =: \mathcal{C}_2(a, c, \alpha, \beta, d),$$

where $\Gamma(s, x) := \int_x^\infty t^{s-1} e^{-t} dt$ is the *upper incomplete gamma function*. Thus, $\boldsymbol{g} \in \mathcal{H}_k^d$ with norm upper bounded by $\mathcal{D}(a, b, \alpha, \beta, d)^{1/2}$, where

$$\mathcal{D}(a, b, \alpha, \beta, d) := \mathcal{C}_0(a, b, \alpha, \beta, d) \left(\mathcal{C}_1(a, c, \alpha, \beta, d) \frac{e^c}{2\beta - 4\alpha + 2} + \mathcal{C}_2(a, c, \alpha, \beta, d) \frac{2\pi d V_d \sqrt{c}}{a \tau_{\beta+d/2}} \right).$$

Now set $\mathring{\boldsymbol{g}} := -\mathcal{D}(a, c, \alpha, \beta)^{-1/2} \boldsymbol{g} \in \mathcal{G}_k$. By our definition of $\boldsymbol{g}$, we have

$$\frac{\mathcal{D}(a, c, \alpha, \beta, d)^{1/2}}{2\alpha} (\mathcal{T}_P \mathring{\boldsymbol{g}})(\boldsymbol{x}) = -\frac{\boldsymbol{s}_p(\boldsymbol{x})^\top \boldsymbol{x}}{(a^2 + \|\boldsymbol{x}\|_2^2)^{1-\alpha}} - \frac{d}{(a^2 + \|\boldsymbol{x}\|_2^2)^{1-\alpha}} + \frac{2(1-\alpha)\|\boldsymbol{x}\|_2^2}{(a^2 + \|\boldsymbol{x}\|_2^2)^{2-\alpha}} \tag{2.8}$$
$$\ge -\frac{\boldsymbol{s}_p(\boldsymbol{x})^\top \boldsymbol{x}}{(a^2 + \|\boldsymbol{x}\|_2^2)^{1-\alpha}} - \frac{d}{a^{2(1-\alpha)}}. \tag{2.9}$$

By the distant dissipativity of P, we have $\limsup_{\|\boldsymbol{x}\|_2 \to \infty} \frac{1}{\|\boldsymbol{x}\|_2^2} \boldsymbol{s}_p(\boldsymbol{x})^\top \boldsymbol{x} \le -\frac{1}{2}\omega_0$, with $\omega_0 > 0$ as defined in (2.1). Thus the first term in (2.8) has a growth rate of at least $\omega_0 \|\boldsymbol{x}\|_2^{2\alpha}/2$ as $\|\boldsymbol{x}\|_2 \to \infty$. From this we deduce (2.7).

Indeed, by our choice of R_0 and the distant dissipativity of P, we have $-\boldsymbol{s}_p(\boldsymbol{x})^\top \boldsymbol{x} \ge 0$ for $\boldsymbol{x} \notin \mathcal{B}(\boldsymbol{0}, R_0)$. Finally, using the Cauchy-Schwarz inequality,

$$|\boldsymbol{s}_p(\boldsymbol{x})^\top \boldsymbol{x}| \le |(\boldsymbol{s}_p(\boldsymbol{x}) - \boldsymbol{s}_p(\boldsymbol{0}))^\top \boldsymbol{x}| + |\boldsymbol{s}_p(\boldsymbol{0})^\top \boldsymbol{x}| \le M_1(\boldsymbol{s}_p)\|\boldsymbol{x}\|_2^2 + \|\boldsymbol{s}_p(\boldsymbol{0})\|_2 \|\boldsymbol{x}\|_2.$$

Then for any $\boldsymbol{x} \in \mathcal{B}(\boldsymbol{0}, R_0)$, we have $-\boldsymbol{s}_p(\boldsymbol{x})^\top \boldsymbol{x} \ge -M_1(\boldsymbol{s}_p)R_0^2 - \|\boldsymbol{s}_p(\boldsymbol{0})\|_2 R_0$. Applying these lower bounds to (2.9) yields the uniform lower bound $\zeta(a, c, \alpha, \beta, d)$. □

2.2.2.2 Coercive functions yield tightness

2.11 Lemma. *Suppose $\boldsymbol{g} \in \mathcal{G}_k$ is such that $\mathcal{T}_P \boldsymbol{g}$ is bounded below by $\zeta \in \mathbb{R}$ and $\liminf_{\|\boldsymbol{x}\|_2 \to \infty} \|\boldsymbol{x}\|_2^{-u} (\mathcal{T}_P \boldsymbol{g})(\boldsymbol{x}) > \eta$ for some $\eta, u > 0$. Then for any probability measure μ, and ϵ small enough, we have*

$$R(\mu, \epsilon) \le \left[\frac{1}{\epsilon\eta} (\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k) - \zeta) \right]^{1/u}. \tag{2.10}$$

Thus, provided $\limsup_{m \to \infty} \mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_k)$ is finite, $(\mu_m)_{m \ge 1}$ is uniformly tight.

Proof. Define $\gamma(r) := \inf\{(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{x}) - \zeta \mid \|\boldsymbol{x}\|_2 \ge r\}$. This is non-negative and non-decreasing, so by Markov's inequality we have

$$\mu(\|\boldsymbol{X}\|_2 \ge r) \le \frac{\mathbb{E}_\mu[\gamma(\|\boldsymbol{X}\|_2)]}{\gamma(r)} \le \frac{\mathbb{E}_\mu[(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{X}) - \zeta]}{\gamma(r)}$$

for any measure μ. It is therefore clear that if $\epsilon \ge (\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k) - \zeta)/\gamma(r_\epsilon)$, then $\mu(\|\boldsymbol{X}\|_2 \ge r_\epsilon) \le \epsilon$. The assumption of the lemma implies that $\gamma(r) \ge \eta r^u$ for large enough r. Thus, if we suppose that

$$r_\epsilon \ge \left[\frac{1}{\epsilon\eta} (\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k) - \zeta) \right]^{1/u},$$

where ϵ is small enough to yield $\gamma(r_\epsilon) \ge \eta r^u$, then

$$\epsilon \ge \frac{\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k) - \zeta}{\eta r_\epsilon^u} \ge \frac{\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k) - \zeta}{\gamma(r_\epsilon)},$$

and so we have $\mu(\|\boldsymbol{X}\|_2 \ge r_\epsilon) \le \epsilon$, and (2.10) follows from the definition of $R(\mu, \epsilon)$. Furthermore, if $\limsup_{m \to \infty} \mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_k)$ is finite, then so is $\limsup_{m \to \infty} R(\mu_m, \varepsilon)$, so $(\mu_m)_{m \ge 1}$ is uniformly tight. □

2.2.2.3 Proof of Theorem 2.9

We are now ready to prove Theorem 2.9, which follows quickly from the following bound.

2.12 Theorem (IMQ KSD lower bound). *Let $P \in \mathcal{P}$ and let $k(\boldsymbol{x}, \boldsymbol{y}) = (c^2 + \|\boldsymbol{x} - \boldsymbol{y}\|_2^2)^\beta$ for some $c > 0$ and $\beta \in (-1, 0)$. Let $\alpha \in (0, \frac{1}{2}(\beta + 1))$ and $a > c/2$. Then there exists $\epsilon_0 > 0$ and a constant $\mathcal{M}_P$ such that, for all μ,*

$$d_{BL_{\|\cdot\|_2}}(\mu, P) \leq \inf_{\epsilon \in [0, \epsilon_0), \delta, \rho > 0} \rho\sqrt{d}(1 + M_1(\boldsymbol{s}_p)\mathcal{M}_P) + \left(3 + \epsilon + (M_1(\boldsymbol{s}_p)\rho\sqrt{d} + \tfrac{d\delta^{-1}\theta_{d-1}}{\theta_d})\mathcal{M}_P\right)\epsilon + (2\pi)^{-d/4}\mathcal{M}_P V_d^{1/2} \times$$
$$\left[\left(\frac{\mathcal{D}(a, c, \alpha, \beta, d)^{1/2}(\mathcal{S}(\mu, \mathcal{T}, \mathcal{G}_k) - \zeta(a, c, \alpha, \beta, d))}{\alpha\omega_0\epsilon}\right)^{1/\alpha} + 2\delta\right]^{d/2} \sqrt{F_{IMQ}\left(\tfrac{12d\log 2}{\pi}(c_{\rho,\delta} + M_1(\boldsymbol{s}_p)\mathcal{M}_P)\epsilon^{-1}\right)}\mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k),$$

where

$$F_{IMQ}(t) := \frac{\Gamma(-\beta)}{2^{1+\beta}}\left(\frac{\sqrt{d}}{c}\right)^{\beta+d/2} \frac{t^{\beta+d/2}}{K_{\beta+d/2}(c\sqrt{d}t)}. \tag{2.11}$$

Proof. For any $\alpha \in (0, \frac{1}{2}(\beta+1))$ and $a > c/2$, we can apply Lemma 2.10 to find $\mathring{\boldsymbol{g}} \in \mathcal{G}_k$ such that $\mathcal{T}_P\mathring{\boldsymbol{g}}$ is bounded below by a constant $\zeta(a, c, \alpha, \beta, d)$ and grows at a rate of $\|\boldsymbol{x}\|_2^{2\alpha}$ as $\|\boldsymbol{x}\|_2 \to \infty$. We can plug the upper bound on $R(\mu, \epsilon)$ given by Lemma 2.11 into the bound given by Theorem 2.6. It is clear from (2.6) that $\hat{\Psi}_{c,\beta}$ is monotonically decreasing in $\|\boldsymbol{\omega}\|_2$, so by noting that $\|\boldsymbol{\omega}\|_\infty \leq t$ implies $\|\boldsymbol{\omega}\|_2 \leq \sqrt{d}t$, we can also substitute (2.11) into the bound from Theorem 2.6. Finally, as indicated by (2.7), we can take $\eta \to \frac{\alpha}{\mathcal{D}(a,c,\alpha,\beta,d)^{1/2}}\omega_0$, which yields the desired result. □

Finally, we prove Theorem 2.9.

Proof of Theorem 2.9. Recall that $F_{\text{IMQ}}(t)$ is finite for all $t > 0$, so if $\mathcal{S}(\mu_m, \mathcal{T}_P, \mathcal{G}_k) \xrightarrow{m\to\infty} 0$, then by Theorem 2.12 we have

$$\limsup_{m\to\infty} d_{\text{BL}_{\|\cdot\|_2}}(\mu_m, P) \leq \rho\sqrt{d}(1 + M_1(\boldsymbol{s}_p)\mathcal{M}_P) + (3 + \epsilon + (M_1(\boldsymbol{s}_p)\rho\sqrt{d} + \tfrac{\delta^{-1}d\theta_{d-1}}{\theta_d})\mathcal{M}_P)\epsilon.$$

Taking $\epsilon, \rho \to 0$ then yields $d_{\text{BL}_{\|\cdot\|_2}}(\mu_m, P) \xrightarrow{m\to\infty} 0$. The result follows since $d_{\text{BL}_{\|\cdot\|_2}}(\cdot, \cdot)$ metrises weak convergence. □

The result of all of this is that we finally have a KSD that metrises weak convergence over a large class of probability distributions. In the next chapter, we explore how the formulation of the KSD, along with the theory introduced in this chapter, can be used to formulate, and prove the convergence of, an expressive variational inference algorithm.

Chapter 3

Stein Variational Gradient Descent

3.1 SVGD algorithm

Variational inference approximates the target density p by minimising the *Kullback-Leibler* (KL) *divergence*

$$\mathrm{KL}(q \,\|\, p) := \mathbb{E}_Q[\log q(\boldsymbol{X})] - \mathbb{E}_Q[\log p(\boldsymbol{X})]$$

over $Q \in \mathcal{Q}$ for some family of distributions $\mathcal{Q}$. Usually, $\mathcal{Q}$ is taken to be some parametric family $\mathcal{Q} = \{q_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. However, such families are often too restrictive to return an accurate approximation to the target, and have to be carefully constructed depending on the specific modelling problem at hand. A more expressive family $\mathcal{Q}$ would lead to a more general-purpose algorithm; the tradeoff, however, is that the corresponding optimisation of $\mathrm{KL}(q \,\|\, p)$ is more difficult to solve for such $\mathcal{Q}$.

Fortunately, the KSD allows for an expressive yet tractable choice of $\mathcal{Q}$. Liu and Wang [2016] proposed considering the following transformation $\boldsymbol{T} : \mathcal{X} \to \mathcal{X}$ defined as a small perturbation of the identity map:

$$\boldsymbol{T}(\boldsymbol{x}) = \boldsymbol{x} + \epsilon \boldsymbol{\phi}(\boldsymbol{x}). \tag{3.1}$$

An intuitive approach to minimising $\mathrm{KL}(q \,\|\, p)$ is to perform gradient descent by iteratively applying transformations of the form (3.1) to $\boldsymbol{x}$, each time choosing $\boldsymbol{\phi}$ so as to maximise the gradient $\nabla_\epsilon \mathrm{KL}(q_{[\boldsymbol{T}]} \,\|\, p)|_{\epsilon=0}$, where $q_{[\boldsymbol{T}]}$ is as defined below.

3.1 Proposition. *Consider the transformation $\boldsymbol{T}$ defined in (3.1), and let $q_{[\boldsymbol{T}]}(\boldsymbol{z})$ be the density of $\boldsymbol{z} := \boldsymbol{T}(\boldsymbol{x})$, where $\boldsymbol{x} \sim q(\boldsymbol{x})$. We have*

$$\nabla_\epsilon \mathrm{KL}(q_{[\boldsymbol{T}]} \,\|\, p)|_{\epsilon=0} = -\mathbb{E}_Q[\boldsymbol{s}_p(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{\phi}(\boldsymbol{x})].$$

Proof. By the change of variables formula,

$$q_{[\boldsymbol{T}]}(\boldsymbol{z}) = q(\boldsymbol{T}^{-1}(\boldsymbol{z})) \cdot |\det(\nabla_{\boldsymbol{z}} \boldsymbol{T}^{-1}(\boldsymbol{z}))|.$$

Letting $\tilde{\boldsymbol{x}} \sim p(\tilde{\boldsymbol{x}})$ and setting $\tilde{\boldsymbol{z}} := \boldsymbol{T}^{-1}(\tilde{\boldsymbol{x}})$,

$$p_{[\boldsymbol{T}^{-1}]}(\tilde{\boldsymbol{z}}) = p(\boldsymbol{T}(\tilde{\boldsymbol{z}})) \cdot |\det(\nabla_{\tilde{\boldsymbol{z}}} \boldsymbol{T}(\tilde{\boldsymbol{z}}))|.$$

Then

$$\mathrm{KL}(q_{[\boldsymbol{T}]} \,\|\, p) = \mathbb{E}_{\boldsymbol{z} \sim q_{[\boldsymbol{T}]}(\boldsymbol{z})}\left[\log \frac{q_{[\boldsymbol{T}]}(\boldsymbol{z})}{p(\boldsymbol{z})}\right] = \mathbb{E}_{\boldsymbol{x} \sim q(\boldsymbol{x})}\left[\log \frac{q(\boldsymbol{x})}{p_{[\boldsymbol{T}^{-1}]}(\boldsymbol{x})}\right] = \mathrm{KL}(q \,\|\, p_{[\boldsymbol{T}^{-1}]}), \tag{3.2}$$

and so

$$\begin{aligned}\nabla_\epsilon \mathrm{KL}(q_{[\boldsymbol{T}]} \,\|\, p) &= -\mathbb{E}_{\boldsymbol{x} \sim q(\boldsymbol{x})}[\nabla_\epsilon \log p_{[\boldsymbol{T}^{-1}]}(\boldsymbol{x})] \\ &= -\mathbb{E}_{\boldsymbol{x} \sim q(\boldsymbol{x})}[\boldsymbol{s}_p(\boldsymbol{T}(\boldsymbol{x}))^\top \nabla_\epsilon \boldsymbol{T}(\boldsymbol{x}) + \mathrm{Trace}((\nabla_{\boldsymbol{x}} \boldsymbol{T}(\boldsymbol{x}))^{-1} \cdot \nabla_\epsilon \nabla_{\boldsymbol{x}} \boldsymbol{T}(\boldsymbol{x}))].\end{aligned}$$

Evaluating at $\epsilon = 0$ yields the desired result. □

This result allows us to find the steepest descent direction $\tilde{\phi}_{q,p}$ over $\phi \in \mathcal{G}_k$, when $\mathcal{G}_k$ is the unit ball of some product Hilbert space.

3.2 Corollary. *Let $\mathcal{G}_k$ be the kernel Stein set (1.2). Over all $\phi \in \mathcal{G}_k$, the direction of steepest descent is given by*

$$\tilde{\phi}_{q,p}(\cdot) = \frac{\mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot)}{\|\mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot)\|_{\mathcal{H}_k^d}}. \tag{3.3}$$

Proof. Observe that

$$\sup_{\boldsymbol{\phi} \in \mathcal{G}_k} -\nabla_\epsilon \mathrm{KL}(q_{[\boldsymbol{T}]} \,\|\, p)|_{\epsilon=0} = \sup_{\boldsymbol{\phi} \in \mathcal{G}_k} \mathbb{E}_Q[\boldsymbol{s}_p(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \cdot \boldsymbol{\phi}(\boldsymbol{x})] = \mathcal{S}(Q, \mathcal{T}_P, \mathcal{G}_k).$$

We know that the supremum in the KSD is achieved by

$$\tilde{\boldsymbol{\phi}}_{q,p}(\cdot) = \frac{\mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot)}{\|\mathbb{E}_Q \boldsymbol{\xi}_p(\boldsymbol{X}, \cdot)\|_{\mathcal{H}_k^d}},$$

so this is our direction of steepest descent over the unit ball $\mathcal{G}_k$ of $\mathcal{H}_k^d$. □

These results motivate Algorithm 3.1, known as **Stein variational gradient descent** (SVGD). The normalising factor in (3.3) can be absorbed into the learning rates $(\epsilon_\ell)_{\ell=0}^L$.

Algorithm 3.1: Stein Variational Gradient Descent (SVGD)

Input: The score function $\boldsymbol{s}_p(\boldsymbol{x})$, initial particles $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^n$, kernel k, number of iterations L and step sizes $\{\epsilon_\ell\}_{\ell=0}^{L-1}$;
for $\ell = 0, \ldots, L-1$ **do**
    $\boldsymbol{x}_i^{(\ell+1)} \leftarrow \boldsymbol{x}_i^{(\ell)} + \epsilon_\ell \tilde{\boldsymbol{\phi}}_{\mu_n^{(\ell)},p}(\boldsymbol{x}_i^{(\ell)}), \quad i = 1, \ldots, n$, where
    $\tilde{\boldsymbol{\phi}}_{\mu_n^{(\ell)},p}(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^n \left\{ \boldsymbol{s}_p(\boldsymbol{x}_j^{(\ell)}) k(\boldsymbol{x}_j^{(\ell)}, \boldsymbol{x}) + \nabla_{\boldsymbol{x}_j^{(\ell)}} k(\boldsymbol{x}_j^{(\ell)}, \boldsymbol{x}) \right\}.$

3.2 Convergence of SVGD

Let $\mu_n^{(\ell)}(\cdot) := \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_i^{(\ell)}}(\cdot)$ be the discrete probability measure of the n particles at the ℓ^{th} time step of the SVGD algorithm. In this section, we will show that $\mu_n^{(\ell)} \Rightarrow P$ as $n, \ell \to \infty$ when k is the IMQ kernel and $P \in \mathcal{P}$. Thus, SVGD has the attractive property that, given sufficient computational resources, it can approximate the target arbitrarily well, which is typically not the case with standard parametric variational inference.

First, let's introduce some notation. We denote the optimal transformation between two probability measures ν and π by $\boldsymbol{T}^{\nu,\pi,\epsilon}(\boldsymbol{x}) := \boldsymbol{x} + \epsilon \tilde{\boldsymbol{\phi}}_{\nu,\pi}(\boldsymbol{x})$. We then define a mapping $\Phi_{\pi,\epsilon}$ on $\mathcal{P}$ by $\Phi_{\pi,\epsilon} : \mu \mapsto \boldsymbol{T}_\#^{\mu,\pi,\epsilon}\mu$, where $\boldsymbol{T}_\#^{\mu,\pi,\epsilon}$ denotes the *pushforward measure* of the transformation $\boldsymbol{T}^{\mu,\pi,\epsilon}$.

We then have $\mu_n^{(\ell+1)} = \Phi_{\pi,\epsilon_\ell}(\mu_n^{(\ell)})$. Taking the infinite particle limit $n \to \infty$, and given some initial probability measure $\mu_\infty^{(0)}$, we can define another sequence of probability measures $(\mu_\infty^{(\ell)})_{\ell \geq 0}$ through the updates

$$\mu_\infty^{(\ell+1)} := \Phi_{\pi,\epsilon_\ell}(\mu_\infty^{(\ell)}).$$

3.2.1 Convergence to large sample limit

We first show that for each time step ℓ, the finite particle update converges to the infinite particle update as the number of particles goes to infinity.

3.3 Theorem. *Consider the SVGD algorithm as outlined above, and suppose that $W_1(\mu_n^{(0)}, \mu_\infty^{(0)}) \xrightarrow{n\to\infty} 0$. Suppose further that there exist constants $C_1, C_2 > 0$ such that*

$$M_1(\boldsymbol{\xi}_p(\boldsymbol{x}, \cdot)) \leq C_1(1 + \|\boldsymbol{x}\|_2) \quad \textit{and} \quad M_1(\boldsymbol{\xi}_p(\cdot, \boldsymbol{y})) \leq C_2(1 + \|\boldsymbol{y}\|_2), \tag{3.4}$$

where $\boldsymbol{\xi}_p$ is as defined in (1.3) and M_1 is as defined in (2.2). Then $W_1(\mu_n^{(\ell)}, \mu_\infty^{(\ell)}) \to 0$ as $n \to \infty$ for each $\ell \in \{1, 2, \dots\}$.

Proof. We prove this by induction on the time step ℓ. The base case $\ell = 0$ is satisfied by assumption. Now suppose $W_1(\mu_n^{(\ell)}, \mu_\infty^{(\ell)}) \xrightarrow{n\to\infty} 0$ for some $\ell \geq 0$. Then there exists $C' > 0$ such that

$$\sup_{n\geq 1}\{1 + \epsilon_\ell C_1(1 + \mu_n^{(\ell)}(\|\cdot\|_2)) + \epsilon_\ell C_2(1 + \mu_\infty^{(\ell)}(\|\cdot\|_2))\} \leq C'.$$

The assumptions (3.4) and Lemma 3.4 below imply that

$$\begin{aligned} W_1(\mu_n^{(\ell+1)}, \mu_\infty^{(\ell+1)}) &= W_1(\Phi_{p,\epsilon_\ell}(\mu_n^{(\ell)}), \Phi_{p,\epsilon_\ell}(\mu_\infty^{(\ell)})) \\ &\leq W_1(\mu_n^{(\ell)}, \mu_\infty^{(\ell)})(1 + \epsilon_\ell C_1(1 + \mu_n^{(\ell)}(\|\cdot\|_2)) + \epsilon_\ell C_2(1 + \mu_\infty^{(\ell)}(\|\cdot\|_2))) \\ &\leq C' W_1(\mu_n^{(\ell)}, \mu_\infty^{(\ell)}) \xrightarrow{n\to\infty} 0, \end{aligned}$$

which completes the proof. □

It thus remains to prove the lemma used in the proof above.

3.4 Lemma. *Suppose that, for some constants $\tilde{C}_1, \tilde{C}_2 > 0$, we have*

$$\sup_{\boldsymbol{y}\in\mathcal{X}} \|\nabla_{\boldsymbol{y}} \boldsymbol{\xi}_p(\boldsymbol{x}, \boldsymbol{y})\|_{op} \leq \tilde{C}_1(1 + \|\boldsymbol{x}\|_2), \quad \textit{and} \quad \sup_{\boldsymbol{x}\in\mathcal{X}} \|\nabla_{\boldsymbol{x}} \boldsymbol{\xi}_p(\boldsymbol{x}, \boldsymbol{y})\|_{op} \leq \tilde{C}_2(1 + \|\boldsymbol{y}\|_2).$$

Then, for any $\epsilon > 0$ and probability measures μ, ν,

$$W_1(\Phi_{p,\epsilon}(\mu), \Phi_{p,\epsilon}(\nu)) \leq W_1(\mu, \nu)(1 + \epsilon\tilde{C}_1(1 + \mu(\|\cdot\|_2)) + \epsilon\tilde{C}_2(1 + \nu(\|\cdot\|_2))).$$

Proof. Let $(\boldsymbol{X}', \boldsymbol{Z}')$ be an optimal 1-Wasserstein coupling of (μ, ν). Then

$$\begin{aligned} \|\boldsymbol{T}^{\mu,p,\epsilon}(\boldsymbol{x}) - \boldsymbol{T}^{\mu,p,\epsilon}(\boldsymbol{x})\|_2 &\leq \|\boldsymbol{x} - \boldsymbol{z}\|_2 + \epsilon\|\mathbb{E}[\boldsymbol{\xi}_p(\boldsymbol{X}', \boldsymbol{x}) - \boldsymbol{\xi}_p(\boldsymbol{X}', \boldsymbol{z})]\|_2 + \epsilon\|\mathbb{E}[\boldsymbol{\xi}_p(\boldsymbol{X}', \boldsymbol{z}) - \boldsymbol{\xi}_p(\boldsymbol{Z}', \boldsymbol{z})]\|_2 \\ &\leq \|\boldsymbol{x} - \boldsymbol{z}\|_2(1 + \epsilon\tilde{C}_1(1 + \mathbb{E}[\|\boldsymbol{X}'\|_2])) + \epsilon\tilde{C}_2\mathbb{E}[\|\boldsymbol{X}' - \boldsymbol{Z}'\|_2](1 + \|\boldsymbol{z}\|_2) \\ &= \|\boldsymbol{x} - \boldsymbol{z}\|_2(1 + \epsilon\tilde{C}_1(1 + \mu(\|\cdot\|_2))) + \epsilon\tilde{C}_2 W_1(\mu, \nu)(1 + \|\boldsymbol{z}\|_2). \end{aligned}$$

By definition of $\boldsymbol{X}'$ and $\boldsymbol{Z}'$, we have $\boldsymbol{T}^{\mu,\nu,\epsilon}(\boldsymbol{X}') \sim \Phi_{p,\epsilon}(\mu)$ and $\boldsymbol{T}^{\mu,\nu,\epsilon}(\boldsymbol{Z}') \sim \Phi_{p,\epsilon}(\nu)$. Hence,

$$\begin{aligned} W_1(\Phi_{p,\epsilon}(\mu), \Phi_{p,\epsilon}(\nu)) &\leq \mathbb{E}[\|\boldsymbol{T}^{\mu,p,\epsilon}(\boldsymbol{X}') - \boldsymbol{T}^{\nu,p,\epsilon}(\boldsymbol{Z}')\|_2] \\ &\leq \mathbb{E}[\|\boldsymbol{X}' - \boldsymbol{Z}'\|_2](1 + \epsilon\tilde{C}_1(1 + \mu(\|\cdot\|_2))) + \epsilon\tilde{C}_2 W_1(\mu, \nu)(1 + \mathbb{E}[\|\boldsymbol{Z}'\|_2]) \\ &= W_1(\mu, \nu)(1 + \epsilon\tilde{C}_1(1 + \mu(\|\cdot\|_2)) + \epsilon\tilde{C}_2(1 + \nu(\|\cdot\|_2))), \end{aligned}$$

as desired. □

Crucially, the growth assumptions on the kernel k and the target P in the statement of Theorem 3.3 are satisfied by the IMQ kernel, as well as other kernels, and by target measures P with Lipschitz score function $\boldsymbol{s}_p$.

3.2.2 Convergence of large sample limit

We have shown that $\mu_n^{(\ell)} \Rightarrow \mu_\infty^{(\ell)}$ as $n \to \infty$ for all ℓ. Our goal now is to show that $\mu_\infty^{(\ell)} \Rightarrow P$ as $\ell \to \infty$, and hence that $\hat{\mu}_n^{(\ell)} \Rightarrow P$ as $n, \ell \to \infty$. In order to do so, we first prove the following lemma.

3.5 Lemma. *Let $B \in \mathbb{R}^{r\times r}$ and let $0 < \epsilon < \frac{1}{\rho(B+B^\top)}$, where $\rho(\cdot)$ denotes the spectral radius. Then $I_r + \epsilon(B + B^\top)$ is positive definite and*

$$\log|\det(I_r + \epsilon B)| \geq \epsilon\mathrm{Trace}(B) - \epsilon^2 \frac{\|B\|_F^2}{1 - \epsilon\rho(B + B^\top)},$$

where $\|\cdot\|_F$ is the Frobenius norm. Thus, if we take $\epsilon \leq \frac{1}{2\rho(B+B^\top)}$, then

$$\log|\det(I_r + \epsilon B)| \geq \epsilon\mathrm{Trace}(B) - 2\epsilon^2\|B\|_F^2.$$

Proof. First note that when $\epsilon < 1/\rho(B + B^\top)$, we have

$$\rho(I_r + \epsilon(B + B^\top)) \geq 1 - \epsilon\rho(B + B^\top) > 0,$$

so $I_r + \epsilon(B + B^\top)$ is positive definite. Now,

$$\begin{aligned}\log|\det(I_r + \epsilon B)| = \frac{1}{2}\log\det\{(I_r + \epsilon B)(I_r + \epsilon B)^\top\} &= \frac{1}{2}\log\det\{I_r + \epsilon(B + B^\top) + \epsilon^2 BB^\top\} \\ &\geq \frac{1}{2}\log\det\{I_r + \epsilon(B + B^\top)\},\end{aligned}$$

since $I_r + \epsilon(B + B^\top)$ and $\epsilon^2 BB^\top$ are both positive semi-definite.

Let $\{\lambda_i\}$ be the eigenvalues of $B + B^\top$. Then, setting $f_i(s) := \log(1 + s\epsilon\lambda_i)$ for $s \in [0, 1]$, we have

$$\log(1 + \epsilon\lambda_i) = f_i(1) = \int_0^1 \frac{\epsilon\lambda_i}{1 + s\epsilon\lambda_i} ds$$

by Taylor's theorem, so

$$\begin{aligned}\log\det(I_r + \epsilon(B + B^\top)) - \epsilon\text{Trace}(B + B^\top) &= \sum_i \{\log(1 + \epsilon\lambda_i) - \epsilon\lambda_i\} \\ &= \sum_i \left\{\int_0^1 \frac{\epsilon\lambda_i}{1 + s\epsilon\lambda_i} ds - \epsilon\lambda_i\right\} \\ &= -\sum_i \int_0^1 \frac{s\epsilon^2\lambda_i^2}{1 + s\epsilon\lambda_i} ds \\ &\geq -\sum_i \frac{\epsilon^2\lambda_i^2}{1 - s\epsilon\max|\lambda_i|} \int_0^1 sds \\ &\geq \frac{\epsilon^2}{2(1 - \epsilon\rho(B + B^\top))} \sum_i \lambda_i^2 = -\frac{\epsilon^2}{2} \frac{\|B + B^\top\|_F^2}{1 - \epsilon\rho(B + B^\top)}.\end{aligned}$$

Hence,

$$\begin{aligned}\log|\det(I_r + \epsilon B)| &\geq \frac{1}{2}\log\det\{I_r + \epsilon(B + B^\top)\} \\ &\geq \frac{\epsilon}{2}\text{Trace}(B + B^\top) - \frac{\epsilon^2}{4}\frac{\|B + B^\top\|_F^2}{1 - \epsilon\rho(B + B^\top)} \\ &\geq \epsilon\text{Trace}(B) - \epsilon^2\frac{\|B\|_F^2}{1 - \epsilon\rho(B + B^\top)},\end{aligned}$$

since $\|B + B^\top\|_F \leq \|B\|_F + \|B^\top\|_F = 2\|B\|_F$. □

We now prove that $\mu_\infty^{(\ell)} \Rightarrow P$ as $\ell \to \infty$.

3.6 Theorem (Convergence of SVGD). *Suppose that* $\text{KL}(\mu_\infty^{(0)}\|P) < \infty$ *and that the step sizes* ϵ_ℓ *are at most*

$$\tilde{\epsilon}_\ell := \left(2\sup_{\boldsymbol{x}\in\mathbb{R}^d} \rho\left(\nabla\tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p} + \nabla\tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}^\top\right)\right)^{-1},$$

where $\rho(\cdot)$ *denotes the spectral norm. Set* $R := c^{2\beta}\left(M_1(\boldsymbol{s}_p)/2 + 4d/c^2\right)$. *Then*

$$\text{KL}(\mu_\infty^{(\ell+1)}\|P) - \text{KL}(\mu_\infty^{(\ell)}\|P) \leq -\epsilon_\ell(1 - \epsilon_\ell R)\,\mathcal{S}^2(\mu_\infty^{(\ell)}, \mathcal{T}_P, \mathcal{G}_k).$$

Thus, if we take $\epsilon_\ell \propto \mathcal{S}(\mu_\infty^{(\ell)}, \mathcal{T}_P, \mathcal{G}_k)^K$ *for any* $K > 0$, *then* $\hat{\mu}_\infty^{(\ell)} \Rightarrow P$ *as* $\ell \to \infty$.

Proof. Using the argument in (3.2) and the fact that the inverse of $\boldsymbol{T}^{\mu,\nu,\epsilon}$ is $\boldsymbol{T}^{\mu,\nu,-\epsilon}$, we have

$$\begin{aligned}\text{KL}(\mu_\infty^{(\ell+1)}\|P) - \text{KL}(\mu_\infty^{(\ell)}\|P) &= \text{KL}(\boldsymbol{T}^{\mu_\infty^{(\ell)},P,\epsilon_\ell}\mu_\infty^{(\ell)}\|P) - \text{KL}(\mu_\infty^{(\ell)}\|P) \\ &= \text{KL}(\mu_\infty^{(\ell)}\|\boldsymbol{T}_\#^{\mu_\infty^{(\ell)},P,-\epsilon_\ell}P) - \text{KL}(\mu_\infty^{(\ell)}\|P) \\ &= -\mathbb{E}_{\boldsymbol{x}\sim\mu_\infty^{(\ell)}}\Big[\underbrace{\log|\det(\nabla_{\boldsymbol{x}}\boldsymbol{T}^{\mu_\infty^{(\ell)},P,\varepsilon_\ell}(\boldsymbol{x}))|}_{(1)} + \underbrace{\log p(\boldsymbol{T}^{\mu_\infty^{(\ell)},P,\varepsilon_\ell}(\boldsymbol{x})) - \log p(\boldsymbol{x})}_{(2)}\Big], \quad (3.5)\end{aligned}$$

where the final equality follows from applying the change of variables formula for the transformation $\boldsymbol{T}^{\mu_\infty^{(\ell)},P,-\epsilon_\ell}$. For term (1), take $B := \nabla_{\boldsymbol{x}} \tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x})$ in Lemma 3.5 and $\epsilon < 1/(2\rho(B + B^\top))$. Then

$$\log|\det(\nabla_{\boldsymbol{x}} \boldsymbol{T}^{\mu_\infty^{(\ell)},P,\varepsilon_\ell}(\boldsymbol{x}))| \geq \epsilon \text{Trace}(\nabla_{\boldsymbol{x}} \tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x})) - 2\epsilon^2 \|\nabla_{\boldsymbol{x}} \tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x})\|_F^2.$$

For term (2), by defining $\boldsymbol{x}_s := \boldsymbol{T}^{\mu_\infty^{(\ell)},P,s\epsilon_\ell}(\boldsymbol{x})$ for $s \in [0,1]$ and applying Taylor's theorem, we get

$$\begin{aligned}
\log p(\boldsymbol{x}) - \log p(\boldsymbol{T}^{\mu_\infty^{(\ell)},P,\epsilon_\ell}(\boldsymbol{x})) &= -\int_0^1 \nabla_s \log p(\boldsymbol{x}_s) ds \\
&= -\int_0^1 \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_s)^\top (\epsilon \boldsymbol{\phi}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x})) ds \\
&= -\epsilon \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x})^\top \tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x}) - \int_0^1 (\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}_s) - \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}))^\top (\epsilon \tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x})) ds \\
&\leq -\epsilon \boldsymbol{s}_p(\boldsymbol{x})^\top \tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x}) + \epsilon^2 M_1(\boldsymbol{s}_p) \cdot \|\tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x})\|_2^2 \int_0^1 ds \\
&= -\epsilon \boldsymbol{s}_p(\boldsymbol{x})^\top \tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x}) + \frac{\epsilon^2}{2} M_1(\boldsymbol{s}_p) \cdot \|\tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x})\|_2^2.
\end{aligned}$$

Thus (3.5) can be upper bounded as follows:

$$\begin{aligned}
\text{KL}(\mu_\infty^{(\ell+1)} \| P) - \text{KL}(\mu_\infty^{(\ell)} \| P) &\leq -\epsilon \left(\boldsymbol{s}_p(\boldsymbol{x})^\top \tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x}) + \nabla_{\boldsymbol{x}} \cdot \tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x}) \right) + \Delta \\
&= -\epsilon \mathcal{S}^2(\mu_\infty^{(\ell)}, \mathcal{T}_P, \mathcal{G}_k) + \Delta,
\end{aligned}$$

by definition of $\tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x})$, where

$$\Delta := \epsilon^2 \mathbb{E}_{\mu_\infty^{(\ell)}} \left[\frac{1}{2} M_1(\boldsymbol{s}_p) \cdot \|\tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x})\|_2^2 + 2\|\nabla_{\boldsymbol{x}} \tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x})\|_F^2 \right].$$

We now want to bound the terms in Δ. Letting $\tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p} = (\phi_1, \ldots, \phi_d)^\top$, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\|\tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x})\|_2^2 = \sum_{i=1}^d \phi_i(\boldsymbol{x})^2 &= \sum_{i=1}^n (\langle k(\boldsymbol{x}, \cdot), \phi_i(\cdot) \rangle_{\mathcal{H}_k})^2 \\
&\leq \sum_{i=1}^d \|k(\boldsymbol{x}, \cdot)\|_{\mathcal{H}_k}^2 \cdot \|\phi_i\|_{\mathcal{H}_k}^2 \\
&= k(\boldsymbol{x}, \boldsymbol{x}) \cdot \|\tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}\|_{\mathcal{H}_k^d}^2 = c^{2\beta} \mathcal{S}^2(\mu_\infty^{(\ell)}, \mathcal{T}_P, \mathcal{G}_k),
\end{aligned}$$

for the IMQ kernel k. Furthermore,

$$\begin{aligned}
\|\nabla_{\boldsymbol{x}} \tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}(\boldsymbol{x})\|_F^2 = \sum_{i,j=1}^d \partial_{x_j} \phi_i(\boldsymbol{x})^2 &= \sum_{i,j=1}^d (\langle \partial_{x_j} k(\boldsymbol{x}, \cdot), \phi_i(\cdot) \rangle_{\mathcal{H}_k})^2 \\
&\leq \sum_{i,j=1}^d \|\partial_{x_j} k(\boldsymbol{x}, \cdot)\|_{\mathcal{H}_k}^2 \cdot \|\phi_i\|_{\mathcal{H}_k}^2 \\
&= \nabla_{\boldsymbol{x}} \cdot \nabla_{\boldsymbol{x}'} k(\boldsymbol{x}, \boldsymbol{x}')\Big|_{\boldsymbol{x}=\boldsymbol{x}'} \sum_{i=1}^d \|\phi_i\|_{\mathcal{H}_k}^2 \\
&= -2d\beta c^{2(\beta-1)} \|\tilde{\boldsymbol{\phi}}_{\mu_\infty^{(\ell)},p}\|_{\mathcal{H}_k^d}^2 \leq 2dc^{2(\beta-1)} \mathcal{S}^2(\mu_\infty^{(\ell)}, \mathcal{T}_P, \mathcal{G}_k)
\end{aligned}$$

for $\beta \in (-1, 0)$. Substituting these bounds into Δ, we conclude that

$$\text{KL}(\mu_\infty^{(\ell+1)} \| P) - \text{KL}(\mu_\infty^{(\ell)} \| P) \leq -\epsilon(1 - \epsilon R) \mathcal{S}^2(\mu_\infty^{(\ell)}, \mathcal{T}_P, \mathcal{G}_k). \qquad \square$$

Combining the results of Theorems 3.3 and 3.6, we have that $\mu_n^{(\ell)} \Rightarrow P$ as $n, \ell \to \infty$ with sufficiently small step sizes, and hence that SVGD can approximate the target arbitrarily well. This is a very attractive property to have in an approximate inference algorithm which is not true of standard parametric variational inference approaches; however, it is an asymptotic result, and so depends deeply on its scalability. This is our focus in Chapter 4.

Chapter 4

Scalable Stein Discrepancies

Despite its satisfying theoretical results, the IMQ KSD presents issues in terms of its applicability to large-scale problems. For example, consider the case where the target is the posterior

$$p(\boldsymbol{x}|\mathcal{D}) = p_0(\boldsymbol{x}) \prod_{r=1}^{R} p(\mathcal{D}_r|\boldsymbol{x}),$$

where $\boldsymbol{x} \in \mathbb{R}^d$ and $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_R\}$ is a (potentially very large) set of training data. Assuming each likelihood $p(\mathcal{D}_r|\boldsymbol{x})$ can be evaluated cheaply enough, and letting $Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\boldsymbol{x}_i}(\cdot)$, the computational cost of calculating $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k)$ is $\mathcal{O}(n^2 dR)$. In the context of SVGD, at each update we must compute the optimal direction, at cost $\mathcal{O}(ndR)$, n times, thus also giving a complexity of $\mathcal{O}(n^2 dR)$ per update. This setting presents us with several limitations:

1. The most immediately obvious scalability issue is that the computational cost is quadratic in n, the number of samples. In Section 4.1, we introduce a *random feature Stein discrepancy* [Huggins and Mackey, 2018] which is inspired by the IMQ KSD and is computable in near-linear time, as a remedy for this problem.

2. Secondly, the ability to perform reliable inference in higher dimensions is dependent on having sufficient data to do so. Consequently, although a computation time linear in d is in itself acceptable, R will typically have to scale with d, which in turn impedes both the evaluation of the KSD *and* the application of SVGD to large-scale Bayesian inference. We discuss *stochastic Stein discrepancies* [Gorham et al., 2020] as a solution to this issue in Section 4.2.

There is another potential scalability issue, which we will not discuss here. It is not one of computational complexity, but instead one of degrading performance in higher dimensions:

3. Reddi et al. [2014] demonstrated that many methods based on evaluating kernels and distances in high dimensions exhibit rapidly deteriorating performance. In this sense, the IMQ KSD presents a double threat: it involves evaluating the IMQ kernel on high-dimensional inputs, and it involves computing high-dimensional inner products. A recent proposal for remedying this problem has been to *slice* the KSD, i.e. to project both the input and the score function onto one dimensional "slices" [Gong et al., 2021]. It is not clear, however, as to whether or not sliced KSDs retain the convergence-monitoring properties of the regular IMQ KSD.

4.1 Random feature Stein discrepancies

We now let $\Phi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{C}$ be a feature function such that $\Phi(\boldsymbol{x}, \cdot) \in L^r$ and $\Phi(\cdot, \boldsymbol{z}) \in C^1$ for all $\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^d$ and some $r \in [1, \infty)$. Rather than taking the Stein set to be the unit ball of some product Hilbert space, we

instead let

$$\mathcal{G}_{\Phi,r} := \left\{ \boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^d \,\middle|\, g_i(\boldsymbol{x}) = \int \Phi(\boldsymbol{x}, \boldsymbol{z}) \overline{f_i(\boldsymbol{z})} d\boldsymbol{z} \;\text{ with }\; \sum_{i=1}^d \|f_i\|_{L^s}^2 \leq 1, \;\text{ where } s = \tfrac{r}{r-1} \right\}.$$

We can now define the (squared) **feature Stein discrepancy** (FSD) which, analogously to the KSD, has an analytic solution:

$$\begin{aligned}
\mathrm{F}_{\Phi,r}\mathcal{S}^2(\mu, P) &:= \sup_{\boldsymbol{g} \in \mathcal{G}_{\Phi,r}} |\mathbb{E}_\mu[(\mathcal{T}_P \boldsymbol{g})(\boldsymbol{X})]|^2 = \sup_{\boldsymbol{g} \in \mathcal{G}_{\Phi,r}} \left| \sum_{i=1}^d \mathbb{E}_\mu[(\mathcal{T}_P^{(i)} g_d)(\boldsymbol{X})] \right|^2 \\
&= \sup_{\boldsymbol{f} \,:\, v_d = \|f_i\|_{L^s},\, \|\boldsymbol{v}\|_2 \leq 1} \left| \sum_{i=1}^d \int \mathbb{E}_\mu[(\mathcal{T}_P^{(i)} \Phi)(\boldsymbol{X}, \boldsymbol{z})] \overline{f_i(\boldsymbol{z})} d\boldsymbol{z} \right|^2 \\
&= \sup_{\boldsymbol{v} : \|\boldsymbol{v}\|_2 \leq 1} \left| \sum_{i=1}^d \|\mathbb{E}_\mu[(\mathcal{T}_P^{(i)} \Phi)(\boldsymbol{X}, \cdot)]\|_{L^r} v_d \right|^2 = \sum_{i=1}^d \|\mathbb{E}_\mu[(\mathcal{T}_P^{(i)} \Phi)(\boldsymbol{X}, \cdot)]\|_{L^r}^2.
\end{aligned}$$

To reduce the cost of large n, we can approximate $\mathrm{F}_{\Phi,r}\mathcal{S}^2(\mu, P)$ using an importance sample of size M:

$$\mathrm{RF}_{\Phi,r,\nu,M}\mathcal{S}^2(\mu, P) := \sum_{i=1}^d \left(\frac{1}{M} \sum_{j=1}^M \nu(\boldsymbol{Z}_j)^{-1} |\mathbb{E}_{\boldsymbol{X} \sim \mu}[(\mathcal{T}_P^{(i)} \Phi)(\boldsymbol{X}, \boldsymbol{Z}_j)]|^r \right)^{2/r}, \quad \text{where } \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M \overset{\text{i.i.d.}}{\sim} \nu.$$

We call this the (squared) **random feature Stein discrepancy** (RFSD). In what follows, we intend to give an outline of the theory which justifies using the RFSD to monitor the convergence of a discrete measure Q_n to a target P.

4.1 Proposition (KSD-FSD inequality)**.** *Let $k(\boldsymbol{x}, \boldsymbol{y}) = \int \mathcal{F}(\Phi(\boldsymbol{x}, \cdot))(\boldsymbol{\omega}) \overline{\mathcal{F}(\Phi(\boldsymbol{y}, \cdot))(\boldsymbol{\omega})} \rho(\boldsymbol{\omega}) d\boldsymbol{\omega}$ be a kernel, where $\mathcal{F}$ denotes the generalised Fourier transform, and let $r \in [1, 2]$. Setting $t := r/(2-r)$, if $\rho \in L^t$, then we have*

$$\mathcal{S}^2(Q_n, \mathcal{T}_P, \mathcal{G}_k) \leq \|\rho\|_{L^t} \, \mathrm{F}_{\Psi_{c,\beta},r}\mathcal{S}^2(Q_n, P).$$

Proof. Let $s = r/(r-1)$. Using Hölder's inequality and the Babenko-Beckner inequality, we have

$$\begin{aligned}
\mathcal{S}^2(Q_n, \mathcal{T}_P, \mathcal{G}_k) &= \sum_{i=1}^d \int |\mathcal{F}(\mathbb{E}_{\boldsymbol{X} \sim Q_n}[(\mathcal{T}_P^{(i)} \Phi)(\boldsymbol{X}, \cdot)])(\boldsymbol{\omega})|^2 \rho(\boldsymbol{\omega}) d\boldsymbol{\omega} \leq \|\rho\|_{L^t} \sum_{i=1}^d \|\mathcal{F}(\mathbb{E}_{\boldsymbol{X} \sim Q_n}[(\mathcal{T}_P^{(i)} \Phi)(\boldsymbol{X}, \cdot)])\|_{L^s}^2 \\
&\leq (r^{1/r}/s^{1/s})^d \|\rho\|_{L^t} \sum_{i=1}^d \|\mathbb{E}_{\boldsymbol{X} \sim Q_n}[(\mathcal{T}_P^{(i)} \Phi)(\boldsymbol{X}, \cdot)]\|_{L^r}^2 = (r^{1/r}/s^{1/s})^d \|\rho\|_{L^t} \mathrm{F}_{\Phi,r}\mathcal{S}^2(Q_n, P),
\end{aligned}$$

since $(r^{1/r} s^{1/s})^{d/2} \leq 1$ for $s = r/(r-1)$. □

As a consequence of this proposition, if we choose k to be the IMQ kernel with $c > 0$ and $\beta \in (-1, 0)$, then any Φ satisfying the condition in Proposition 4.1 will yield a FSD which detects non-convergence. Since $k(\boldsymbol{x}, \boldsymbol{y}) = \Psi_{c,\beta}(\boldsymbol{x} - \boldsymbol{y})$, it makes sense to take $\Phi(\boldsymbol{x}, \boldsymbol{y}) = F(\boldsymbol{x} - \boldsymbol{y})$ for some F which satisfies the following regularity conditions.

4.2 Assumption. *The function $F \in C^1$ is positive, and there exists a norm $\|\cdot\|$ and constants $s, C \geq 0$ such that*

$$|\partial_{x_i} \log F(\boldsymbol{x})| \leq C(1 + \|\boldsymbol{x}\|^s), \quad \lim_{\|\boldsymbol{x}\| \to \infty} (1 + \|\boldsymbol{x}\|^s) F(\boldsymbol{x}) = 0, \quad \text{and } F(\boldsymbol{x} - \boldsymbol{z}) \leq C F(\boldsymbol{z}) / F(\boldsymbol{x}).$$

Furthermore, there is a constant $\underline{c} \in (0, 1]$ and a continuous, non-increasing function f such that $\underline{c} f(\|\boldsymbol{x}\|) \leq F(\boldsymbol{x}) \leq f(\|\boldsymbol{x}\|)$.

How do we choose F? An intuitive choice would be $F = \mathcal{F}^{-1}(\hat{\Psi}_{c,\beta}^{1/2})$, so that $\Psi_{c,\beta}(\boldsymbol{x} - \boldsymbol{y}) = \int F(\boldsymbol{x} - \boldsymbol{z}) F(\boldsymbol{y} - \boldsymbol{z}) d\boldsymbol{z}$, and in particular, $F_{\Phi,2}\mathcal{S}(\mu, P) = \mathcal{S}(\mu, \mathcal{T}_P, \mathcal{G}_k)$. However, this choice of F is difficult to compute in practice, so we instead only require F to be a good approximation to $\mathcal{F}^{-1}(\hat{\Psi}_{c,\beta})$.

4.3 Assumption. *Assumption 4.2 holds, and there exists a* smoothness parameter $\bar{\lambda} \in (1/2, 1]$ *such that* $\hat{F}/\hat{\Psi}_{c,\beta}^{\lambda/2} \in L^2$ *for all* $\lambda \in (1/2, \bar{\lambda})$.

The following proposition shows that certain RFSDs detect weak convergence, in probability.

4.4 Proposition. *Suppose that Assumption 4.2 holds with* $F \in L^r$ *and that* $\mathbb{E}_{\boldsymbol{X}\sim P}[\|\boldsymbol{X}\|_2^2] < \infty$. *If* $W_1(Q_n, P) \to 0$ *as* $n \to \infty$, *then* $\mathrm{F}_{\Phi,r}\mathcal{S}(Q_n, P) \to 0$ *and* $\mathrm{RF}_{\Phi,r,\nu_n,M_n}\mathcal{S}(Q_n, P) \xrightarrow{P} 0$ *for any* $r \in [1, 2]$, ν_n *and* M_n.

The aim when choosing ν is to ensure that the second moment of each *RFSD feature*

$$w_i(\boldsymbol{Z}, Q_n) := |\mathbb{E}_{\boldsymbol{X}\sim Q_n}[(\mathcal{T}_P^{(i)}\Phi)(\boldsymbol{X}, \boldsymbol{Z})]|^r/\nu(\boldsymbol{Z})$$

is bounded by a power of its mean, as outlined by the following definition.

4.5 Definition ((C,γ) second moments)**.** *We say that* (Φ, r, ν) *yields* (C, γ) *second moments for* P *and* Q_n *if* $\mathbb{E}[w_i(\boldsymbol{Z}, Q_n)^2] \leq C\mathbb{E}[w_i(\boldsymbol{Z}, Q_n)]^{2-\gamma}$ *for all* $i = 1, \ldots, d$ *and* $n \geq 1$.

4.6 Proposition. *Suppose that* (Φ, r, ν) *yields* (C, γ) *second moments for* P *and* Q_n. *Under the assumptions of Proposition 4.1, if the reference KSD* $\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \gtrsim n^{-1/2}$, *and* $M \gtrsim n^{\gamma r/2} C\|\rho\|_{L^t}^{\gamma r/2} \log(d/\delta)/\epsilon^2$, *then with probability at least* $1 - \delta$,

$$\|\rho\|_{L^t}^{1/2} RF_{\Phi,r,\nu,M}\mathcal{S}(Q_n, P) \geq (1 - \epsilon)^{1/r}\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k).$$

The following final assumption ensures in particular that Proposition 4.1, and hence Proposition 4.6, applies.

4.7 Assumption. *Assumption 4.2 holds,* $\omega_1^2\hat{\Psi}_{c,\beta}^{1/2}(\boldsymbol{\omega}) \in L^1$, *and* $\hat{\Psi}_{c,\beta}/\hat{F}^2 \in L^t$, *where* $t = r/(2 - r)$.

Huggins and Mackey [2018] proceed to show that, if Assumptions 4.2, 4.3 and 4.7 hold, and there exists $\mathcal{C} > 0$ such that

$$\mathbb{E}_{\boldsymbol{X}\sim Q_n}[(1 + \|\boldsymbol{X}\| + \|\boldsymbol{X} - \mathbb{E}_{\boldsymbol{X}'\sim Q_n}[\boldsymbol{X}']\|^s)/F(\boldsymbol{X} - \mathbb{E}_{\boldsymbol{X}'\sim Q_n}[\boldsymbol{X}'])] \leq \mathcal{C} \quad \text{for all } n \geq 1,$$

then there exists $b \in [0, 1)$ such that for any $\xi \in (0, 1 - b)$, $c > 0$ and $\alpha > 2(1 - \bar{\lambda})$, if $\nu(\boldsymbol{z}) \geq c\Psi_{c,\beta}(\boldsymbol{z} - \mathbb{E}_{\boldsymbol{X}\sim Q_n}[\boldsymbol{X}])^{\xi r}$, then there exists a constant $C_\alpha > 0$ such that (Φ, r, ν) yields $(C_\alpha, \gamma_\alpha)$ second moments for P and Q_n, where $\gamma_\alpha := \alpha + (2 - \alpha)\xi/(2 - b - \xi)$. Thus, by increasing the smoothness $\bar{\lambda}$ of F and decreasing the overdispersion parameter ξ of ν, we can make γ arbitrarily close to 0, and hence yield RFSDs which are computable in arbitrarily-close-to-linear time.

A particular case, which they show to satisfy the necessary assumptions, is choosing $\bar{\lambda}$ and $\underline{\xi} \in (0, 1/2)$ – the minimum ξ that we can choose when constructing ν – and then taking $F = \Psi_{c',\beta'}$, where $c' = \bar{\lambda}c/2$, $\beta' \in [-d/(2\underline{\xi}), -\beta/(2\underline{\xi}) - d/(2\underline{\xi}))$, $r = -d/(2\beta'\underline{\xi})$, $\xi \in (\underline{\xi}, 1)$ and $\nu(\boldsymbol{z}) \propto \Psi_{c',\beta'}(\boldsymbol{z} - \mathbb{E}_{\boldsymbol{X}\sim Q_n}[\boldsymbol{X}])^{\xi r}$. This is known as the L^r IMQ RFSD. In particular, taking $\beta' = -d/(2\underline{\xi})$ yields $r = 1$. The importance sampling distribution ν in this case is easy to sample from since it is a multivariate t-distribution.

4.2 Stochastic Stein discrepancies

Consider the case where our target is the posterior

$$p(\boldsymbol{x}|\mathcal{D}) \propto p_0(\boldsymbol{x}) \prod_{r=1}^{R} p(\mathcal{D}_r|\boldsymbol{x}).$$

We can decompose the Langevin Stein operator as $\mathcal{T}_P = \sum_{r=1}^{R} \mathcal{T}_P^{(r)}$, where

$$(\mathcal{T}_P^{(r)}\boldsymbol{g})(\boldsymbol{x}) = \boldsymbol{s}_{p_r}(\boldsymbol{x})^\top \boldsymbol{g}(\boldsymbol{x}) + \tfrac{1}{R}\nabla_{\boldsymbol{x}} \cdot \boldsymbol{g}(\boldsymbol{x}), \quad \text{and } p_r(\boldsymbol{x}) := p_0(\boldsymbol{x})^{1/R} p(\mathcal{D}_r|\boldsymbol{x}).$$

(Note that this is different to the decomposition that we have used previously.) The idea behind reducing the cost induced by large R is to randomly subsample the base operators $\mathcal{T}_P^{(r)}$. To this end, for the discrete measure $Q_n(\cdot) := \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_i}(\cdot)$, we define the **stochastic Stein discrepancy** as

$$\text{stoch}_m\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}) := \sup_{\boldsymbol{g}\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^n \frac{R}{m}(\mathcal{T}_{\sigma_i}\boldsymbol{g})(\boldsymbol{x}_i) \right|,$$

where m is our chosen batch size and the $\sigma_i \subseteq \{1,\dots,R\}$ are subsets of size m sampled uniformly at random (without replacement). By using arguments analogous to those used in Section 1.2, one can show that taking the Stein set to be the unit ball of a product Hilbert space, i.e. $\mathcal{G} = \mathcal{G}_k$, again yields an analytic solution to the optimisation over $\boldsymbol{g} \in \mathcal{G}$:

$$\text{stoch}_m\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) = \frac{1}{n^2}\sum_{i,j=1}^n k_p^{\sigma_i,\sigma_j}(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

where

$$k_p^{\sigma,\tilde{\sigma}}(\boldsymbol{x},\boldsymbol{y}) := \tfrac{R^2}{m^2}\boldsymbol{s}_{p_\sigma}(\boldsymbol{x})^\top \boldsymbol{s}_{p_{\tilde{\sigma}}}(\boldsymbol{y})k(\boldsymbol{x},\boldsymbol{y}) + \tfrac{R}{m}\boldsymbol{s}_{p_\sigma}(\boldsymbol{x})^\top \nabla_{\boldsymbol{y}}k(\boldsymbol{x},\boldsymbol{y}) + \tfrac{R}{m}\boldsymbol{s}_{p_{\tilde{\sigma}}}(\boldsymbol{y})^\top \nabla_{\boldsymbol{x}}k(\boldsymbol{x},\boldsymbol{y}) + \nabla_{\boldsymbol{x}}\cdot\nabla_{\boldsymbol{y}}k(\boldsymbol{x},\boldsymbol{y}).$$

The stochastic IMQ KSD metrises weak convergence in a general sense, as outlined by the following theorems, whose proofs are given in Gorham et al. [2020].

4.8 Theorem (Stochastic IMQ KSD detects convergence). *If $P \in \mathcal{P}$, then $\text{stoch}_m\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \to 0$ whenever $Q_n \Rightarrow P$.*

4.9 Theorem (Stochastic IMQ KSD detects non-convergence). *Let $P \in \mathcal{P}$, and suppose that $\sup_{\boldsymbol{x}\in\mathcal{X}} \frac{\|\log p_\sigma(\boldsymbol{x})\|_2}{1+\|\boldsymbol{x}\|_2} < \infty$ for all $\sigma \in \binom{[R]}{m}$. If $Q_n \not\Rightarrow P$ as $n \to \infty$, then $\text{stoch}_m\mathcal{S}(Q_n, \mathcal{T}_P, \mathcal{G}_k) \not\to 0$.*

4.2.1 Stochastic SVGD

As mentioned above, using SVGD to approximate the posterior $p(\boldsymbol{x}|\mathcal{D})$ requires $\mathcal{O}(n^2dR)$ operations per update. The stochastic KSD presents us with an alternative algorithm, **stochastic SVGD**, which requires $\mathcal{O}(n^2dm)$ operations per update, with $m \ll R$. Rather than using the optimal direction

$$\tilde{\boldsymbol{\phi}}_{\mu_n^{(\ell)},p}(\cdot) = \frac{1}{n}\sum_{j=1}^n \left\{ \boldsymbol{s}_p(\boldsymbol{x}_j^{(\ell)})k(\boldsymbol{x}_j^{(\ell)},\cdot) + \nabla_{\boldsymbol{x}_j^{(\ell)}}k(\boldsymbol{x}_j^{(\ell)},\cdot) \right\},$$

as in Algorithm 3.1, we instead use the stochastic direction

$$\tilde{\boldsymbol{\phi}}^{\text{stoch}}_{\hat{\mu}_{n,m}^{(\ell)},p}(\cdot) = \frac{1}{n}\sum_{j=1}^n \left\{ \frac{R}{m}\boldsymbol{s}_{p_{\sigma_j}}(\boldsymbol{x}_j^{(\ell)})k(\boldsymbol{x}_j^{(\ell)},\cdot) + \nabla_{\boldsymbol{x}_j^{(\ell)}}k(\boldsymbol{x}_j^{(\ell)},\cdot) \right\},$$

resulting in Algorithm 4.1. For n particles and batch size m, we denote the discrete measure at time ℓ by $\hat{\mu}_{n,m}^{(\ell)}$, and the stochSVGD update by

$$\hat{\mu}_{n,m}^{(\ell+1)} = \Phi_{\epsilon_\ell,p}^m(\hat{\mu}_{n,m}^{(\ell)}).$$

Algorithm 4.1: Stochastic Stein Variational Gradient Descent (stochSVGD)

Input: The score function $\boldsymbol{s}_p(\boldsymbol{x})$, initial particles $\{\boldsymbol{x}_i^{(0)}\}_{i=1}^n$, kernel k, batch size m, number of iterations L and step sizes $\{\epsilon_\ell\}_{\ell=0}^{L-1}$;
for $\ell = 0,\dots,L-1$ **do**
    For $i = 1,\dots,n$, sample an independent mini-batch σ_i of size m from $\{1,\dots,R\}$;
    $\boldsymbol{x}_i^{(\ell+1)} \leftarrow \boldsymbol{x}_i^{(\ell)} + \epsilon_\ell \tilde{\boldsymbol{\phi}}^{\text{stoch}}_{\hat{\mu}_{n,m}^{(\ell)},p}(\boldsymbol{x}_i^{(\ell)}), \quad i = 1,\dots,n$, where
        $\tilde{\boldsymbol{\phi}}^{\text{stoch}}_{\hat{\mu}_{n,m}^{(\ell)},p}(\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^n \left\{ \frac{R}{m}\boldsymbol{s}_{p_{\sigma_i}}(\boldsymbol{x}_j^{(\ell)})k(\boldsymbol{x}_j^{(\ell)},\boldsymbol{x}) + \nabla_{\boldsymbol{x}_j^{(\ell)}}k(\boldsymbol{x}_j^{(\ell)},\boldsymbol{x}) \right\}.$

We will now prove that this new algorithm converges almost surely as $n, \ell \to \infty$. Note that the infinite particle sequence $\{\mu_\infty^{(\ell)}\}_{\ell \geq 0}$ is the same as before, so we only need to show that $\hat{\mu}_{n,m}^{(\ell)} \Rightarrow \mu_\infty^{(\ell)}$ almost surely as $n \to \infty$, for each ℓ.

4.10 Theorem (Convergence of stochSVGD). *Consider the stochSVGD algorithm as outlined in Algorithm 4.1, and suppose that* $W_1(\hat{\mu}_{n,m}^{(0)}, \mu_\infty^{(0)}) \xrightarrow{n\to\infty} 0$. *Suppose that condition (3.4) holds from the original proof, along with the additional conditions that, for some* $C_0 > 0$,

$$\max_{\sigma \in \binom{[R]}{m}} \sup_{\boldsymbol{y} \in \mathbb{R}^d} \|\boldsymbol{s}_{p_\sigma}(\boldsymbol{x}) k(\boldsymbol{x}, \boldsymbol{y})\|_2 \leq C_0(1 + \|\boldsymbol{x}\|_2), \tag{4.1}$$

$$\max_{\sigma \in \binom{[R]}{m}} \sup_{\boldsymbol{y} \in \mathbb{R}^d} \|\nabla_{\boldsymbol{x}}(\boldsymbol{s}_{p_\sigma}(\boldsymbol{x}) k(\boldsymbol{x}, \boldsymbol{y}))\|_{op} \quad \textit{is bounded on compact sets } J, \tag{4.2}$$

where $\binom{[R]}{m}$ *is the set of all subsets of* R *of size* m. *Then* $W_1(\hat{\mu}_{n,m}^{(\ell)}, \mu_\infty^{(\ell)}) \xrightarrow{a.s.} 0$ *as* $n \to \infty$.

Proof. We prove this by induction. The base case $\ell = 0$ is satisfied by assumption. Now, for any $\ell \geq 0$, assume that $\mathbb{P}(\mathcal{E}) = 1$, where $\mathcal{E}$ is the event that $W_1(\hat{\mu}_{n,m}^{(\ell)}, \mu_n^{(\ell)}) \to 0$ as $n \to \infty$. Then, since $W_1(\mu_n^{(\ell)}, \mu_\infty^{(\ell)}) \to 0$ as $n \to \infty$ by Theorem 3.3, we have $W_1(\hat{\mu}_{n,m}^{(\ell)}, \mu_\infty^{(\ell)}) \to 0$ as $n \to \infty$ on $\mathcal{E}$. Therefore, there exists a constant C' such that

$$\sup_{n \geq 1} 1 + \epsilon_\ell C_1(1 + \hat{\mu}_{n,m}^{(\ell)}(\|\cdot\|_2)) + \epsilon_\ell C_2(1 + \mu_\infty^{(\ell)}(\|\cdot\|_2)) \leq C'$$

on $\mathcal{E}$. By the triangle inequality,

$$\begin{aligned} W_1(\hat{\mu}_{n,m}^{(\ell+1)}, \mu_n^{(\ell+1)}) &= W_1(\Phi_{\epsilon_\ell,p}^m(\hat{\mu}_{n,m}^{(\ell)}), \Phi_{\epsilon_\ell,p}(\mu_n^{(\ell)})) \\ &\leq W_1(\Phi_{\epsilon_\ell,p}^m(\hat{\mu}_{n,m}^{(\ell)}), \Phi_{\epsilon_\ell,p}(\hat{\mu}_{n,m}^{(\ell)})) + W_1(\Phi_{\epsilon_\ell,p}(\hat{\mu}_{n,m}^{(\ell)}), \Phi_{\epsilon_\ell,p}(\mu_n^{(\ell)})). \end{aligned}$$

By the assumptions (3.4) and Lemma 3.4, on $\mathcal{E}$ we have

$$\begin{aligned} W_1(\Phi_{\epsilon_\ell,p}(\hat{\mu}_{n,m}^{(\ell)}), \Phi_{\epsilon_\ell,p}(\mu_n^{(\ell)})) &\leq W_1(\hat{\mu}_{n,m}^{(\ell)}, \mu_n^{(\ell)})(1 + \epsilon_\ell C_1(1 + \hat{\mu}_{n,m}^{(\ell)}(\|\cdot\|_2)) + \epsilon_\ell C_2(1 + \mu_\infty^{(\ell)}(\|\cdot\|_2))) \\ &\leq C' W_1(\hat{\mu}_{n,m}^{(\ell)}, \mu_n^{(\ell)}) \to 0 \quad \text{as } n \to \infty. \end{aligned}$$

Furthermore, using the growth assumptions (4.1) and (4.2), we can apply Lemma 4.11 below, which ensures that $W_1(\Phi_{\epsilon_\ell,p}^m(\hat{\mu}_{n,m}^{(\ell)}), \Phi_{\epsilon_\ell,p}(\hat{\mu}_{n,m}^{(\ell)})) \xrightarrow{\text{a.s.}} 0$ as $n \to \infty$. Hence, on $\mathcal{E}$, $W_1(\hat{\mu}_{n,m}^{(\ell+1)}, \mu_n^{(\ell+1)}) \to 0$ as $n \to \infty$, thus proving our claim. □

4.11 Lemma. *For any triangular array of points* $\{\boldsymbol{z}_i^n\}_{i \in \{1,\ldots,n\}, n \geq 1}$ *in* $\mathbb{R}^d$, *define the discrete probability measure* $\nu_n(\cdot) := \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{x}_i^n}(\cdot)$. *Suppose that* $\boldsymbol{s}_{p_\sigma}(\cdot)k(\cdot, \boldsymbol{y})$ *is continuous for each* $\boldsymbol{y} \in \mathbb{R}^d$ *and* $\sigma \in \binom{[R]}{m}$. *Set*

$$\begin{aligned} f_0(\boldsymbol{x}) &:= \sup_{\boldsymbol{y} \in \mathbb{R}^d, \sigma \in \binom{[R]}{m}} M_0(\boldsymbol{s}_{p_\sigma})|k(\boldsymbol{x}, \boldsymbol{y})|, \\ f_1(\boldsymbol{x}) &:= \sup_{\boldsymbol{y} \in \mathbb{R}^d, \sigma \in \binom{[R]}{m}} \|\nabla_{\boldsymbol{x}}(\boldsymbol{s}_p(\boldsymbol{x}) k(\boldsymbol{x}, \boldsymbol{y}))\|_{\text{op}}. \end{aligned}$$

If f_0 *is* ν_n*-uniformly integrable and* f_0, f_1 *are bounded on every compact set, then for any* $\epsilon > 0$,

$$W_1(\Phi_{\epsilon,p}^m(\nu_n), \Phi_{\epsilon,p}(\nu_n)) \xrightarrow{a.s.} 0$$

as $n \to \infty$.

This lemma is proved in Lemma 13 of Gorham et al. [2020]. It is again important to highlight that the additional growth conditions (4.1) and (4.2) are satisfied by the IMQ kernel and P with Lipschitz score function $\boldsymbol{s}_p$.

Chapter 5

Experiments

In this chapter, we demonstrate the validity of our results in practice. Section 5.1 compares the various Stein discrepancies that we have introduced and verifies that they behave as we expect, while also introducing an application of these discrepancies in the form of hyperparameter selection in a biased MCMC algorithm. Section 5.2 focuses on approximate inference, and in particular the application of SVGD and stochSVGD as variational inference algorithms. The code is available at `https://github.com/alex-hinds/stein_discrepancies`.

5.1 Measuring sample quality

In this section, we will compare the different Stein discrepancies we have considered in terms of their ability to monitor convergence in various settings.

5.1.1 Comparison to the Wasserstein-1 distance in one dimension

To check that our Stein discrepancies behave as we expect, we construct a very simple experiment influenced by Section 4.1 of Gorham and Mackey [2017]. We generate synthetic data $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_R\}$ from the $\mathcal{N}(\alpha \boldsymbol{e}_1, I_D)$ distribution

$$p(\boldsymbol{x}|\alpha) = \tfrac{1}{(2\pi)^{d/2}} e^{-\frac{1}{2}\|\boldsymbol{x}-\alpha \boldsymbol{e}_1\|_2^2},$$

where α is a parameter, taken to be 1 when generating the data. Assuming an improper uninformative prior over α, our posterior is

$$p(\alpha|\mathcal{D}) \propto \prod_{r=1}^{R} e^{-\frac{1}{2}\|\boldsymbol{x}_r-\alpha \boldsymbol{e}_1\|_2^2} \propto \exp\left\{-\tfrac{1}{2}\left(-2\alpha \sum_{r=1}^{R} x_{r1} + R\alpha^2\right)\right\} \sim \mathcal{N}\left(\frac{1}{R}\sum_{r=1}^{R} x_{r1}, \frac{1}{R}\right).$$

The purpose of this example is purely to demonstrate the efficacy of our Stein discrepancies to monitor convergence. Of course, in this example, we can perform exact inference and there is no need for approximate sampling. What's more, since we can sample efficiently from the posterior, we shouldn't expect any computational benefits from using the stochastic IMQ KSD or the L^1 IMQ RFSD over the standard IMQ KSD. Nevertheless, the fact that this posterior is one-dimensional and exact allows us to compare the IMQ KSD and its stochastic variant with the Wasserstein-1 distance.

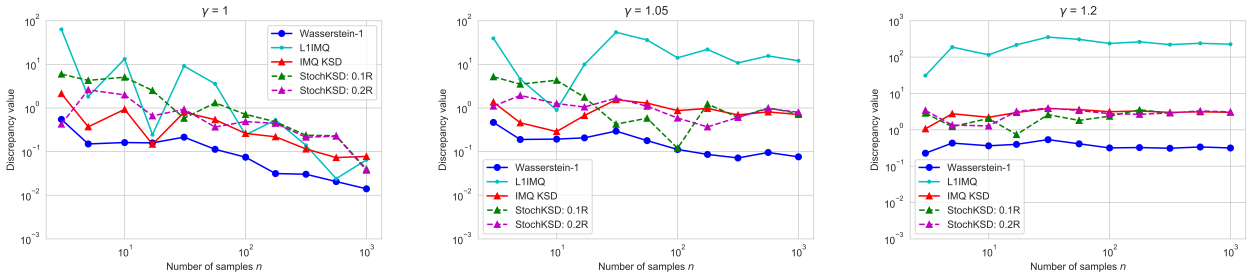
To make such a comparison, we generate sequences of i.i.d. random variables sampled from

$$\tilde{p}_\gamma(\alpha|\mathcal{D}) \sim \mathcal{N}\left(\gamma \times \frac{1}{R}\sum_{r=1}^{R} x_{r1}, \frac{1}{R}\right),$$

where $\gamma \in \mathbb{R}$. Clearly, when $\gamma = 1$, the sequence will converge weakly to the target $p(\alpha|\mathcal{D})$, but when $\gamma \neq 1$ it will not. We are interested in how effective our Stein discrepancies are in detecting small deviations of γ

from 1, particularly in comparison with the Wasserstein-1 distance.

We plot the values of the IMQ KSD, the Stochastic IMQ KSD with $m = 0.1R$ and $m = 0.2R$, the L^1 IMQ RFSD and the Wasserstein distance, for the sequence with three different values of γ and $R = 10$. We took the median over 10 evaluations for the RFSD and took $M = 100$.



(a) (b) (c)

Figure 5.1: Comparing KSDs to the Wasserstein-1 distance.

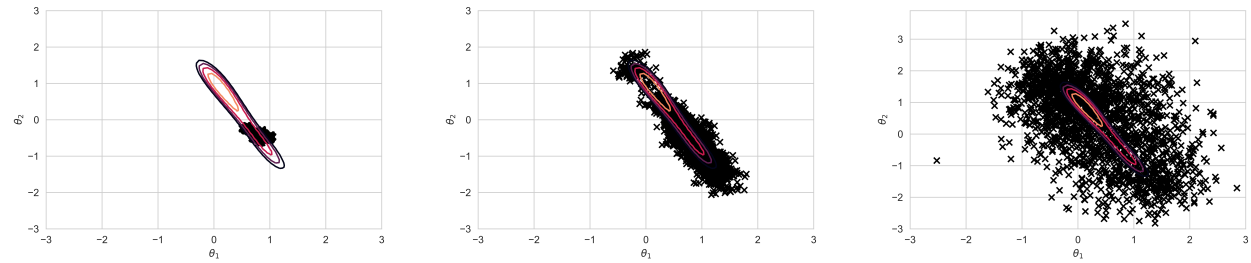
We see in Figure 5.1 that, as expected, all Stein discrepancies behave similarly to the Wasserstein-1 distance in detecting convergence in the case $\gamma = 1$. For the other values of γ tested, the Stein discrepancies detect non-convergence, and in fact seem to be more discriminative than the Wasserstein-1 distance.

5.1.2 Hyperparameter selection in Stochastic Gradient Langevin Dynamics

An MCMC algorithm which has gained huge popularity in recent years is **Stochastic Gradient Langevin Dynamics** (SGLD) [Welling and Teh, 2011]. Given $p(\boldsymbol{x}|\mathcal{D}) \propto p_0(\boldsymbol{x}) \prod_{r=1}^{R} p(\mathcal{D}_r|\boldsymbol{x})$, SGLD starts with initial particles $\boldsymbol{x}_0$ and makes the updates

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \frac{\epsilon_t}{2}\left(\nabla_{\boldsymbol{x}_t} \log p_0(\boldsymbol{x}_t) + \frac{R}{m}\sum_{r\in\sigma_t} \nabla_{\boldsymbol{x}_t} \log p(\mathcal{D}_r|\boldsymbol{x}_t)\right) + \boldsymbol{\eta}_t,$$

where σ_t is a subset of $\{1, \dots, R\}$ of size m sampled uniformly at random for each t, and $\boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{0}, \epsilon_t I_d)$.



(a) $\epsilon = 5 \times 10^{-5}$ (b) $\epsilon = 5 \times 10^{-3}$ (c) $\epsilon = 5 \times 10^{-2}$

Figure 5.2: Hyperparameter selection in SGLD.

Although convergence guarantees require that the step sizes $(\epsilon_t)_{t\geq 1}$ satisfy

$$\sum_{t=1}^{\infty} \epsilon_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \epsilon_t^2 < \infty,$$

it is common in practice to use a constant step size $\epsilon_t \equiv \epsilon$, which introduces bias. The choice of ϵ is important: if it is too small then the sequence will exhibit long mixing times, while if it is too large then the bias will lead

to a stationary distribution significantly different to the target.

As demonstrated in Gorham et al. [2020], we can calculate KSDs between the generated samples and the target for a range of ϵ. We consider the Gaussian mixture model (GMM) discussed in Welling and Teh [2011]:

$$\theta_1 \sim \mathcal{N}(0, \sigma_1^2), \quad \theta_2 \sim \mathcal{N}(0, \sigma_2^2),$$
$$x_r \sim \tfrac{1}{2}\mathcal{N}(\theta_1, \sigma_x^2) + \tfrac{1}{2}\mathcal{N}(\theta_1 + \theta_2, \sigma_x^2), \quad r = 1, \ldots, R,$$

where we take $\sigma_1^2 = 10$, $\sigma_2^2 = 1$ and $\sigma_x^2 = 2$. We generate the data x_r using $\theta_1 = 0$ and $\theta_2 = 1$. We see in Figure 5.2 that when ϵ is too small, the samples are underdispersed, and when ϵ is too large, the samples are overdispersed. The most acceptable value of ϵ of those plotted seems to be $\epsilon = 5 \times 10^{-3}$.

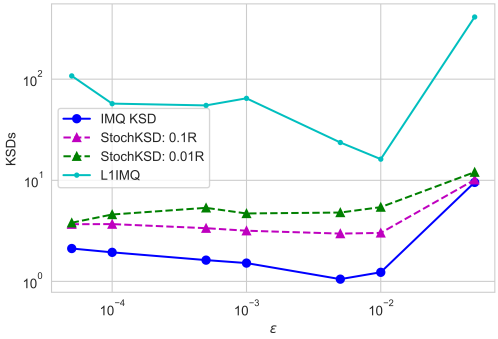
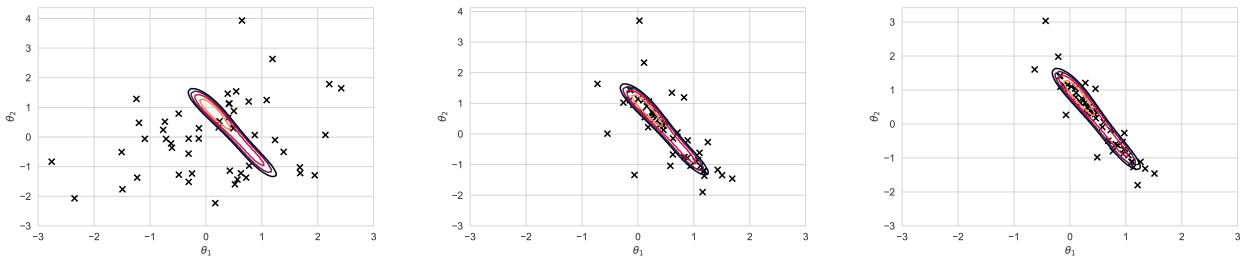


Figure 5.3: Various Stein discrepancies against the hyperparameter ϵ in SGLD.

Figure 5.3 shows the IMQ KSD, stochastic IMQ KSDs with $m = 0.01R$ and $m = 0.1R$, and the L^1 IMQ RFSD between SGLD samples and the target for various ϵ. For the L^1 IMQ RFSD, we take $M = 10$ and take the median over 5 evaluations. For the stochastic KSDs, we take the mean over 5 evaluations.

We see from Figure 5.3 that the KSD, RFSD and stochKSD with $m = 0.1R$ are minimised for ϵ somewhere between 10^{-3} and 10^{-2}, which agrees with Figure 5.2. The result is less clear for the stochastic KSD with $m = 0.01R$, which suggests that more evaluations of this discrepancy may be needed.



(a) 0 iterations (b) 100 iterations (c) 200 iterations

Figure 5.4: SVGD for the GMM posterior at multiple time steps.

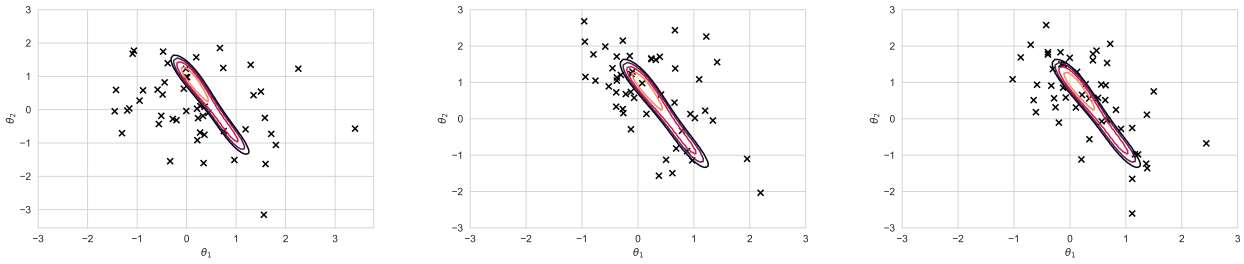
5.2 Particle-based variational inference

In this section, we assess the performance of SVGD and stochSVGD as approximate inference algorithms.

5.2.1 Convergence of SVGD

As a simple example, we return to the GMM example from Section 5.1.2. We plot the points outputted by SVGD and stochSVGD each with 50 particles at several stages. We use $R = 100$ likelihood terms and for the stochastic version, we use $m = 10$.

Figure 5.4 exemplifies the performance of SVGD. The particles appear to converge to the target very quickly. Figure 5.5 shows the stochSVGD particles converging, albeit requiring more iterations: this is to be expected since each stochSVGD iteration requires only one tenth of the likelihood evaluations required by full SVGD.



(a) 0 iterations (b) 100 iterations (c) 200 iterations

Figure 5.5: stochSVGD for the GMM posterior at multiple time steps.

Since we have samples from the SVGD algorithms and the score of the target, we can compute the IMQ KSD to monitor the convergence of the samples. Figure 5.6 shows the IMQ KSD at early time steps. SVGD repeatedly decreases the IMQ KSD, as to be expected by our theory, while the IMQ KSD for stochSVGD fluctuates due to the randomness of the subsampling procedure.

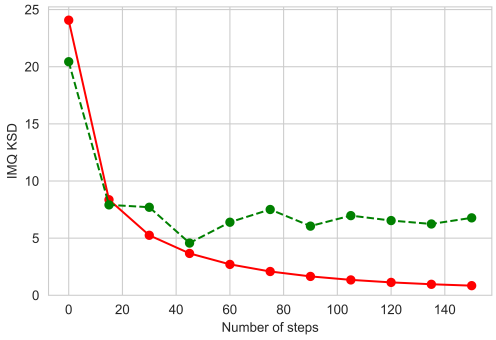


Figure 5.6: The IMQ KSD against the number of steps of SVGD (red) and stochSVGD (green).

We should expect that the IMQ KSD remains bounded away from zero for both algorithms, since we are trying to approximate a continuous distributions with a finite number of particles. The plots of the SVGD particles and the KSD plot both demonstrate that SVGD iteratively improves the particle approximation to the target GMM posterior.

Chapter 6

Conclusion

The recent popularity of biased MCMC algorithms has, among other applications, driven the need for a valid measure of distance between a discrete distribution of samples and a target distribution whose density is only known up to a multiplicative constant. An appropriate characterisation of 'validity' is the metrisation of weak convergence. We have shown that the IMQ KSD does indeed metrise weak convergence, while only requiring samples from one distribution and the score function of the other. This makes the IMQ KSD widely applicable, in particular for approximate inference.

Nevertheless, the scalability of the IMQ KSD is key. We have addressed issues concerning approximate inference on large datasets through the L^r IMQ RFSD and the stochastic IMQ KSD. We also introduced SVGD and stochSVGD, approximate inference algorithms in their own right, which are inspired by the KSD and stochastic KSD respectively, and are both (almost surely) asymptotically exact.

While we gave strong theoretical assurances for the IMQ KSD and its stochastic variant, future work could seek to obtain similar theoretical assurances for the sliced KSD introduced by Gong et al. [2021]. This would justify its use on high-dimensional data, and would help to combat the curse of dimensionality. It would also be of interest to obtain theoretical assurances for substituting a random feature version of the SVGD update direction in order to reduce the quadratic cost of SVGD in the number of particles.

Bibliography

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615, 2016.

A. Y. K. Foong, D. R. Burt, Y. Li, and R. E. Turner. On the expressiveness of approximate inference in bayesian neural networks, 2020.

W. Gong, Y. Li, and J. M. Hernez-Lobato. Sliced kernelized stein discrepancy, 2021.

J. Gorham and L. Mackey. Measuring sample quality with stein's method. *Advances in Neural Information Processing Systems*, 2015.

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1292–1301, 2017.

J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions, 2018.

J. Gorham, A. Raj, and L. Mackey. Stochastic stein discrepancies. In *Advances in Neural Information Processing Systems*, volume 33, pages 17931–17942, 2020.

J. Huggins and L. Mackey. Random feature stein discrepancies. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Q. Liu. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

S. J. Reddi, A. Ramdas, B. Ps, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions, 2014.

C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2. University of California Press. pp. 583–602*, 1972.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, 2011.

H. Wendland. Scattered data approximation. *Cambridge University Press*, 2004.