Kernelized Complete Conditional Stein Discrepancy

Raghav Singhal  
New York University  
New York, USA  
rs4070@nyu.edu

Saad Lahlou  
New York University  
New York, USA  
msl596@nyu.edu

Rajesh Ranganath  
New York University  
New York, USA  
rajeshr@cims.nyu.edu

Abstract

Much of machine learning relies on comparing distributions with discrepancy measures. Stein’s method creates discrepancy measures between two distributions that require only the unnormalized density of one and samples from the other. Stein discrepancies can be combined with kernels to define the kernelized Stein discrepancies (ksds). While kernels make Stein discrepancies tractable, they pose several challenges in high dimensions. We introduce kernelized complete conditional Stein discrepancies (kcc-sds). Complete conditionals turn a multivariate distribution into multiple univariate distributions. We prove that kcc-sds detect convergence and non-convergence, and that they upper-bound ksd-s. We empirically show that kcc-sds detect non-convergence where ksd-s fail. Our experiments illustrate the difference between kcc-sds and ksd-s when comparing high-dimensional distributions and performing variational inference.

1 Introduction

Discrepancy measures that compare a distribution \( p \), known up to normalization, with a distribution \( q \), known via samples from it, can be used for finding good variational approximations (Ranganath et al., 2016), checking the quality of MCMC samplers (Gorham and Mackey, 2015, 2017), or goodness-of-fit testing (Liu et al., 2016). There are two difficulties with using traditional discrepancies like Wasserstein metrics or total variation distance for these tasks. First, \( p \) can be hard to sample, and second, computing these discrepancies requires an expensive maximization. These challenges lead to the following desiderata for a discrepancy \( D \) (Gorham and Mackey, 2015).

1. **Tractable** \( D \) uses samples from \( q \), evaluations of (unnormalized) \( p \), and has a closed form.

2. **Detect Convergence** If \( q_n \Rightarrow p \), then \( D(p, q_n) \rightarrow 0 \).

3. **Detect Non-Convergence** If \( D(p, q_n) \rightarrow 0 \), then that implies that \( q_n \Rightarrow p \).

These desiderata ensure that the discrepancy is non zero when \( p \) does not equal \( q \) and that it can be easily computed. To meet these desiderata, Chwialkowski et al., 2016; Oates et al., 2017; Gorham and Mackey, 2017 developed kernelized Stein discrepancies (ksds). Ksds measure the expectation of functions under \( q \) that have expectation zero under \( p \). These functions are constructed by applying Stein’s operator to a reproducing kernel Hilbert space.

In high dimensions most kernels evaluated on a pair of points are near zero. Thus, ksds in high dimensions can be near zero, making detecting differences between high dimensional distributions difficult. We develop kernelized complete conditional Stein discrepancies (kcc-sds). These discrepancies use complete conditionals: the distribution of one variable given the rest. The complete conditionals are univariate. Rather than using multivariate kernels, kcc-sds use multiple univariate kernels, making it easier to compare distributions in high dimensions.

A given Stein discrepancy relies on a supremum over a class of test functions called the Stein set. The kcc-sds differ from ksds in that kcc-sds’ Stein set consists of univariate functions.
Discrepancy
0
25
50
75
100
Dimension
2
6
10
Discrepancy
IMQ kernel
KCC-SD
KSD
Discrepancy
RBF kernel
KCC-SD
KSD
Figure 1: **KCC-SDs are more sensitive to differences than KSDs in high dimensions.** The figure compares \( p = \mathcal{N}(0, I_d) \) and \( q = \mathcal{N}(\mu, I_d) \), where \( \mu_1 = 5 \) and the rest of the components are zero and uses 1000 samples to compute the KCC-SD and the KSD with the RBF and IMQ kernel. KCC-SDs retain a better ability to tell \( p \) and \( q \) apart as the dimensions increase.

rather than multivariate functions. An immediate question is whether a Stein discrepancy with only univariate functions detects non-convergence. We prove under technical conditions that (1) KCC-SDs detect convergence and non-convergence, and (2) KCC-SDs are larger than KSDs for the same choice of kernel.

**Figure 1** compares KSDs and KCC-SDs with different kernels on each panel. The figure compares two Gaussian distributions, \( p = \mathcal{N}(0, I_d) \) and \( q = \mathcal{N}(\mu, I_d) \) where only one coordinate of \( \mu \) is non-zero, \( \mu_1 = 5 \). We then increase the dimension of the distribution and compare KCC-SD and KSD with both the IMQ and RBF kernels. We see that KCC-SDs retain their ability to distinguish distributions in high dimensions. We show that KCC-SDs can be used for variational inference and empirically compare distributions in cases where KSDs provably fail.

## 2 Stein Discrepancies

Stein’s method provides recipes for constructing expectation zero functions of distributions known up to normalization. For a distribution, \( p \), with a Lipschitz score function, we can create a Stein operator, \( \mathcal{A}_p(x) \), that acts on a test function \( f : \mathbb{R}^d \rightarrow \mathbb{R}^d \),

\[
\mathbb{E}_{p(x)}[(\mathcal{A}_p(x)f)(x)] = 0, \tag{1}
\]

where \( f \) is smooth and 1-Lipschitz \( L_1(p) \) function. This relation called *Stein’s identity* can be used to construct a discrepancy, where \( p \) is known only up to a normalization constant (Gorham and Mackey, 2015). Let \( \mathcal{H} \) be the Stein set, consisting of smooth Lipschitz functions satisfying a Neumann-type boundary condition:

\[
\mathbb{E}(q, \mathcal{A}_p, \mathcal{H}) = \sup_{f \in \mathcal{H}} \left| \mathbb{E}_{q(x)}[(\mathcal{A}_p(x)f)(x)] - \mathbb{E}_{p(x)}[(\mathcal{A}_p(x)f)(x)] \right| = \sup_{f \in \mathcal{H}} \left| \mathbb{E}_{q(x)}[\mathcal{A}_p(x)f(x)] \right|.
\]

Stein discrepancies can be computationally burdensome as the supremum lacks a closed form.

**Kernelized Stein Discrepancies.** To make the Stein discrepancy simpler to compute, Chwialkowski et al., 2016; Oates et al., 2017; Gorham and Mackey, 2017 combined the theory of reproducing kernels (Steinwart and Christmann, 2008) with the Stein discrepancy to introduce kernelized Stein discrepancies KSDs. KSDs restrict the Stein set to a reproducing kernel Hilbert space. Let \( k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \) be the kernel of a reproducing kernel Hilbert space (RKHS) \( \mathcal{K}_k \). The RKHS \( \mathcal{K}_k \) consists of functions \( g : \mathbb{R}^d \rightarrow \mathbb{R} \) with finite norm. Functions in the RKHS satisfy the reproducing property: \( g(x) = \langle g(\cdot), k(x, \cdot) \rangle_{\mathcal{K}_k} \) for all \( x \in \mathbb{R}^d \) and for all \( g \in \mathcal{K}_k \). KSDs are defined using a Stein set \( \mathcal{G}_k \): the set of vector-valued functions \( g = (g_1, \ldots, g_d) \) such that for all \( i, g_i \in \mathcal{K}_k \), and \( \sum_{i=1}^d \|g_i\|_{\mathcal{K}_k} \leq 1 \). KSDs have a closed-form.

**Proposition 1** (Gorham and Mackey, 2017) Suppose \( k \in \mathbb{C}^{1,1}(\mathbb{R}^d, \mathbb{R}^d) \), then for all \( j \in \{1, \ldots, d\} \) define the function,

\[
k^j_{0}(x, y) = \mathcal{A}^j_{p(x)}\mathcal{A}^j_{p(y)}k(x, y), \tag{2}
\]
where, for instance if $A_p$ is the Langevin-Stein operator, then for any $f: \mathbb{R}^d \to \mathbb{R}^d$
\[
(A_p f)(x) = \sum_{j=1}^{d} (A^j_{p(x)}) f(x) = \sum_{j=1}^{d} f(x)^T \nabla x_j \log p(x) + \nabla x_j f(x)
\]

Now, if $\sum_{j=1}^{d} \mathbb{E}_{q(x) \times q(x')}[k^j_{cc}(x, x')^{1/2}] < \infty$, then kernelized Stein discrepancy has a closed form, $S(q, A_p, G_k) = \|w\|_2$ where $w^2 = \mathbb{E}_{q(x) \times q(y)}[k^j_{cc}(x, y)]$, and $x, y \sim_d q$.

This theorem shows that when the Stein set is chosen via an RKHS, the Stein discrepancy can be computed in closed form. When the distribution $p$ lies in the class of distantly dissipative \(^1\) distributions, $\mathcal{P}(\mathbb{R}^d)$, KSDS provably detect convergence and non-convergence for $d \leq 3$, for kernels like the radial basis function or the inverse multi-quadratic (IMQ) (Gorham and Mackey, 2017). In $d > 3$, the KSD with thin tailed kernels like the RBF do not detect non-convergence. But the KSD with the IMQ kernel with $\beta \in (0, 1)$ does detect non-convergence. However all of these kernels shrink as the $\|\cdot\|_2$ grows, which mean their associated KSD become less sensitive in higher dimensions (see Figure 1).

3 Complete Conditional Stein Discrepancy

Complete conditionals are univariate conditional distributions, $p(x_i \mid x_{-i})$, where $x_{-i} = \{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d\}$. Complete conditional distributions are the basis for many inference procedures including the Gibbs sampler (Geman and Geman, 1984) and coordinate ascent variational inference (Ghahramani and Beal, 2001).

We construct cc-sds and their kernelized versions, KCC-SDS, and show that KCC-SDS satisfy the desiderata. For a broad family of kernels, we show that KCC-SDS upper bound traditional KSDS. We focus on the Langevin-Stein operator (Mira et al., 2013; Gorham and Mackey, 2015; Oates et al., 2017),
\[
(A_p f)(x) = f(x)^T \nabla x \log p(x) + \nabla x : f(x).
\]

The analysis done here can be applied other operators based on the gradient of the log probability.

**Definition.** Using complete conditionals, we define a new operator that can be used to compare distributions in arbitrary dimensions. For any $f: \mathbb{R}^d \to \mathbb{R}^d$ with univariate component functions, $f_i: \mathbb{R} \to \mathbb{R}$, we can apply the complete conditional factorization,
\[
(A_p f)(x) = \sum_{i=1}^{d} (A^i_{p(x_i \mid x_{-i})} f_i)(x) = \sum_{i=1}^{d} (A^i_{p(x)} f_i)(x)
\]
\[
= \sum_{i=1}^{d} f_i(x_i) \nabla x_i \log p(x_i \mid x_{-i}) + \nabla x_i f_i(x_i) = f(x)^T \nabla x \log p(x) + \nabla x : f(x),
\]

where $\nabla x_i \log p(x_i \mid x_{-i}) = \nabla x_i \log p(x)$. Note that although the test functions $f_i$ are univariate, the Stein operator applied to each component function, $h_i = A^i_{p(x_i \mid x_{-i})} f_i$, is a scalar valued function of multiple variables, $h_i: \mathbb{R}^d \to \mathbb{R}$.

This factorization yields the same operator as the Langevin-Stein operator in Equation (3). A two variable example for cc-sds is in Appendix A. The key difference is that the Stein set for cc-sd consists of univariate component functions. Formally, we define the function space $\mathcal{C} = \{ f: \mathbb{R}^d \to \mathbb{R}^d \mid f(x) = (f_1(x_1), \ldots, f_d(x_d)) \}$. Then the complete conditional Stein discrepancy $S(q, A_p, \mathcal{C})$

\(^1\)A distribution $p$ is distantly dissipative if $\kappa_0 = \lim_{r \to 0} \kappa(r) < \infty$, where
\[
\kappa(r) = \inf \left\{ \frac{\kappa^p(x \mid x')}{\|x - y\|^2}: \|x - y\|_2 = r \right\} < \infty.
\]

Common examples include finite Gaussian mixtures with the same variance, and strongly log-concave distributions.
is defined as the Stein discrepancy restricted to the function set consisting of univariate component functions,
\[ S(q, A_p, C) = \sup_{f \in C} \mathbb{E}_{q(x)}[|A_p(x)f(x)|]. \]
CC-SDs do not require the complete conditionals for \( p \) or \( q \). Like the original Stein discrepancy, the suprema in CC-SDs can be hard to compute. Instead, we introduce their kernelized form, the kernelized complete conditional Stein discrepancy (KCC-SD).

4 Kernelized Complete Conditional Stein Discrepancy

Similar to the construction of KSDs from the Stein discrepancy, we meld the theory of reproducing kernels with complete conditional Stein discrepancies to obtain KCC-SDs. In this section we show that KCC-SDs satisfy all three desiderata: (1) a closed and tractable form, (2) detection of convergence, and (3) detection of non-convergence. We also show KCC-SDs upper bound KSDs, and the difference between the two increases as the dimension of the distribution increases.

**KCC-SDs admit a closed form.** Let \( k^i : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \) be a reproducing kernel for the reproducing kernel Hilbert space \( \mathcal{K}_k^i \), then the Stein set \( \mathcal{C}_k \) for KCC-SDs consists of functions \( f(x) = (f_1(x_1), \ldots, f_d(x_d)) \) where each \( f_i \) is a univariate function and \( f_i \in \mathcal{K}_k^i \) with \( \sum_i \|f_i\|_{\mathcal{K}_k^i} \leq 1 \). Note that it is possible for the kernel to change with each dimension, but for simplicity we focus on a single kernel for all dimensions and drop the index on the kernel. We now show that KCC-SDs can be computed in closed form.

**Theorem 1** Let \( k \in \mathbb{C}^{1,1}(\mathbb{R}, \mathbb{R}) \) then for all \( j \in \{1, \ldots, d\} \) define the complete conditional Stein kernel, \( k_{cc} \) as follows:
\[
k_{cc}^j(x, y) = A_p^j A_q^j \kappa(x, y)
= s^j_p(x) s^j_q(y) k(x_j, y_j) + s^j_p(x) \nabla y_j k(x_j, y_j) \\
+ s^j_q(y) \nabla x_j k(x_j, y_j) + \nabla x_j \nabla y_j k(x_j, y_j),
\]
Now, if \( \sum_{j=1}^d \mathbb{E}_{q(x) \times q(x')}[k_{cc}^j(x, x')^{1/2}] < \infty \), then \( S(q, A_p, C_k) = \|w\|_2 \), where \( w_j^2 = \mathbb{E}_{q(x) \times q(x')}[k_{cc}^j(x, x')], \) and \( x, x' \iidsim q \).

The proof is in Appendix D. Note that the closed form for KCC-SDs is the same as KSDs but the kernels are now univariate rather than multivariate.

**KCC-SDs detect convergence.** KCC-SDs can be upper bounded with the Wasserstein distance (W2). This shows that if \( q_n \Rightarrow p \) as \( n \to \infty \), then KCC-SDs go to zero, satisfying desideratum 2.

**Proposition 2** Suppose \( k \in \mathbb{C}^{2,2}_b(\mathbb{R}, \mathbb{R}) \) and \( \nabla_x \log p \) is Lipschitz with \( \mathbb{E}_{p(x)}[\|\nabla \log p(x)\|_2^2] < \infty \), if \( q_n \Rightarrow p \), then \( S(q_n, A_p, C_k) \to 0 \).

The proof follows from Gorham and Mackey, 2017 and is in Appendix E and this proposition applies to kernels like the RBF, IMQ and Matern kernels.

**KCC-SD detects non-convergence.** In this section, we show that KCC-SDs detect non-convergence by showing that when the KCC-SD converges to zero, the Fisher divergence converges to zero. The Fisher divergence measures the error between the score function of two distributions. It is defined as
\[
\mathcal{F}(q, p) = \mathbb{E}_{q(x)} \left[ \| \nabla_x \log p(x) - \nabla_x \log q(x) \|^2_2 \right].
\]
The following lemma shows that if \( \nabla_x \log p(x), \nabla_x \log q(x) \) are Lipschitz and \( p(x), q(x) > 0 \) for all \( x \in \mathbb{R}^d \), then when the Fisher divergence between two distributions goes to zero, the two distributions are equal in distribution.

**Lemma 1** Suppose \( \nabla_x \log p \) and \( \nabla_x \log q \) are Lipschitz with \( \mathbb{E}_{q(x)} \left[ \| \nabla_x \log p(x) \|_2^2 \right] < \infty \), and
\[
\mathbb{E}_{q(x)} \left[ \| \nabla_x \log q(x) \|_2^2 \right] < \infty \text{ If } \mathcal{F}(q, p) = 0, \text{ then } q \overset{d}{=} p.
We use this lemma to show KCC-SD going to zero implies equality in distribution.

**Theorem 2** Suppose \( k \in C^2_b(\mathbb{R}, \mathbb{R}) \) is integrally strictly positive definite, and \( \nabla_x \log p \) and \( \nabla_x \log q \) are Lipschitz with \( E_{q(x)} \left[ \| \nabla_x \log p(x) \|_2^2 \right] < \infty \), and \( E_{q(x)} \left[ \| \nabla_x \log q(x) \|_2^2 \right] < \infty \) and suppose \( p(x), q(x) > 0 \) for all \( x \in \mathbb{R}^d \). If \( S(q_n, A_p, C_k) \to 0 \), then \( F(q_n, p) \to 0 \).

The proof is in Appendix F. Unlike a full theory of weak convergence, the proof for Theorem 2 requires the score function of \( q \). We empirically show that KCC-SD detects non-convergence for distributions that do not have score functions even where the KSD fails to detect non-convergence.

For \( d > 3 \), Gorham and Mackey, 2017 show that KSDs fail to detect non-convergence for commonly used kernels like the RBF. When kernels decay faster than the score function grows, KSDs ignore the tails. This problem gets worse in higher dimensions for the RBF kernel and Matern kernel. For instance, if \( x, y \sim \mathcal{N}(0, I_d) \) then \( E[\| x - y \|_2^2] = 2d \), which causes the RBF kernel to decay rapidly in high dimensions, leading to a low discrepancy value even if the distributions are different.

In Figure 2 we compare a non-tight sequence \( q_n \) to a Gaussian target \( p = \mathcal{N}(0, I_d) \) from Gorham and Mackey, 2017. For each \( n \), let \( q_n \) be the empirical distribution over points \( \{x_i\}_{i=1}^n \) where \( \| x_i \|_2 \leq 2n^{1/d} \log n \) and \( \| x_i - x_j \|_2 \geq 2 \log n \) for all \( i, j \). For a kernel like the RBF, this will cause the kernel to decay as we increase the sample size, as \( k(x_i, x_j) = e^{-\beta \| x_i - x_j \|_2^2} \leq e^{-4\beta \log n^2} = n^{-4\beta \log n} \). This sequence of \( q_n \) does not have a score function. Unlike the KSD, Figure 2 shows that the KCC-SD with both the RBF and IMQ kernels is able to detect non-convergence.

Even when KSDs detect convergence in high dimensions, they can be too small to be of practical use, thereby making them poor assessments of sample quality. Figure 1 depicts this problem for two Gaussian distributions, \( p \) is the standard Gaussian and \( q \) is a Gaussian with the mean of one dimension set to 5. The plots show how KSDs decrease with increasing dimension. After dimension 10, the KSD has becomes very small for the RBF kernel, and even if we use IMQ kernel, which detects non-convergence, the KSD still becomes smaller.

**KCC-SDs upper bounds KSDs.** In this section we show that KCC-SDs are upper bounds on KSDs. The difference between the discrepancies grows as the dimensionality increases.

Suppose that the KSD and the KCC-SD have the same type of kernel with the same kernel parameters. We show that the KCC-SD is a upper bound of the KSD, given that the kernel satisfies the following conditions:

**C1** \( k(x_j, y_j) - k(x, y) \geq 0 \) for all \( x, y \in \mathbb{R}^d \) and for all \( 1 \leq j \leq d \).

**C2** Define the univariate kernel difference as \( k_d(x_j, y_j; x_{-j}, y_{-j}) = k(x_j, y_j) - k(x, y) \), where we fix \( x_{-j}, y_{-j} \). Then \( k_d \) is an integrally strictly positive definite kernel.

In Appendix G, we show that both the RBF and the IMQ kernels satisfy these conditions. The proofs follow from Schoenberg connection between monotone and positive definite functions (Fasshauer, 2003).
Figure 3: KCC-SDs converge to zero with i.i.d samples from \( p \). Here we compute both discrepancies with \( n = 1000 \) samples in \( d = 10 \), and \( p = \frac{1}{2}N(-10, \Sigma_1) + \frac{1}{2}N(10, \Sigma_2) \). We compute the discrepancies using the RBF and IMQ kernels.

**Theorem 3** Suppose \( k \) satisfies conditions C1 and C2 and \( \nabla_x \log p, \nabla_x \log q \) are Lipschitz with 
\[
E_q(\| \nabla_x \log p(x) \|^2), E_q(\| \nabla_x \log q(x) \|^2) < \infty \quad \text{and} \quad S(q, A_p, C_k), S(q, A_p, G_k) < \infty.
\]
Then the KSD and the KCC-SD satisfy the relation
\[
S(q, A_p, G_k) \leq S(q, A_p, C_k).
\] (4)

The proof is in Appendix G. The diagram below shows the relations between the discrepancies. By construction the Stein set for cc-sd is a subset of the Stein set for the Stein discrepancy. This means the Stein discrepancy dominates the cc-sd. Kernelization shrinks Stein sets, so kernelized variants are dominated by their corresponding non-kernelized variants. **Theorem 3** shows that KCC-SDs dominate KSDs.

\[
\begin{array}{c}
\text{SD} \\
\rightarrow \text{CC-SD} \\
\geq \text{KSD} \\
\leq \text{KCC-SD}
\end{array}
\]

5 Experiments

We developed the kernelized complete conditional Stein discrepancies. Here we empirically study their use for performing sample quality checks and variational inference. We detail the variational inference algorithms in Appendix B. We study two kernels: IMQ and RBF. For the IMQ kernel, \( k(x, y) = (c + \|x - y\|^2_2)^{-\beta} \), we take \( c = 1 \) and \( \beta = \frac{1}{2} \) and for the RBF kernel, \( k(x, y) = e^{-\beta \|x - y\|^2_2} \) we choose \( \beta = \frac{1}{2} \).

5.1 Distribution Tests

Figure 1 shows the effect of increasing dimension when comparing two Gaussian distributions which only differ in one coordinate of the mean. KSDs decrease as the dimension increases, unlike KCC-SDs. **Figure 2** with the RBF kernel shows empirically that KCC-SD detects non-convergence where KCC does not. Despite showing that KCC-SDs upper bound KSDs, in **Figure 3** we show that both KCC-SD and KSD converge to zero at a similar rate when \( p = q \) for a mixture of Gaussians where each component has different non-diagonal covariance matrices.

In Appendix C we conduct more tests to study the rate of convergence to zero when both distributions are the same. We also compare two Gaussian distributions with increasing distance between their means, there we see that KCC-SD is more sensitive to changes than KSD in high dimensions.

5.2 Sample Quality Checks

Here we show that KCC-SD can be used for sample quality checks.
Figure 4: kcc-sd is a provable method to compute sample quality and does not assume asymptotic exactness of the samples, unlike standard methods like Effective Sample Size, here $\epsilon$ is the stepsizes for SGLD. We use the RBF kernel to compute the kcc-sd value.

Selecting Sampler Hyperparameters. Stochastic gradient Langevin dynamics (SGLD) is a biased MCMC sampler based on adding noise to the standard stochastic gradient optimization method (Welling and Teh, 2011). Since this method makes use of subsampling, it has allowed MCMC to scale to large datasets and large models. In this experiment we do posterior inference for a two-layer neural network, with a sigmoid activation function, for a regression task. We used the yacht hydrodynamics dataset (Gerritsma et al., 1981) from the UCI dataset repository.

Since biased methods trade sampling efficiency for asymptotic exactness, standard MCMC diagnostics are not applicable as they do not account for asymptotic bias. We use kcc-sd to assess sample quality from biased MCMC samplers. Selecting the stepsize $\epsilon$ is an important task to ensure the samples are approximately from the posterior (Welling and Teh, 2011). When $\epsilon$ is too small, then SGLD is not exploring the space enough and there is high autocorrelation between the samples. However, when $\epsilon$ is too big, then SGLD has higher bias and is unstable.

For $\epsilon \in [10^{-8}, 10^{-3}]$ we generate 5 independent chains with minibatch 32. Each chain consists of 10,000 samples with a burnin phase of 50,000 samples. We compare kcc-sd to inverse effective sample size. Effective sample size relies on asymptotic exactness of the samples, which is violated by stochastic gradient Langevin dynamics. Figure 4 compares these two metrics. While $\epsilon = 10^{-4}$ has the lowest kcc-sd value, the effective sample size measure is maximized by the value $\epsilon = 10^{-2}$.

5.3 Stein Variational Gradient Descent

We compare SVGD to the complete conditional Stein variational gradient descent (cc-SVGD) algorithm by training a Bayesian neural network and learning a multivariate Gaussian. We provide details for cc-SVGD in Appendix B.

Learning Multivariate Gaussians. We compare the performance of cc-SVGD and SVGD using the RBF kernel on learning a multivariate Gaussian target, $\mathcal{N} (\mu, \Sigma)$. We train both methods to learn a Gaussian target with diagonal covariance and non-diagonal covariance. We use a 100 particles initialized from $\mathcal{N} (0, I_d)$ and run both methods for 1000 iterations. We use gradient descent with a decreasing stepsize $\eta_t = \alpha (t+1)^{-1/2}$. Figure 5 displays the ksd between the target and the learnt distribution as the dimension increases. cc-SVGD has a lower KSD value, it learns a better approximation.

Bayesian Neural Network. We compare cc-SVGD and SVGD on Bayesian neural networks. We use a similar setting as (Liu and Wang, 2016), a neural network with one hidden layer, with a rectifier activation and 100 hidden units. We use 90% of the dataset as a training set and use the rest as the test set. The results are averaged over 5 trials. The minibatch size is 100 and the number of particles is 20 for both methods, we use the RBF kernel with $\beta = \frac{1}{2}$.

Table 1 shows that cc-SVGD performs better than SVGD and MAP on 3 out of 5 experiments. SVGD and MAP yield almost identical results. In these experiments, more accurate Bayesian inference seems to provide little advantage as MAP performs well.
Figure 5: (a) Average KSD with the RBF kernel over 5 trials of CC-SVGD and SVGD approximations with a correlated multivariate Gaussian target. We use a 100 particles and run for 1000 iterations using gradient descent. (b) Average KSD with the RBF kernel over 5 trials of CC-SVGD and SVGD with a multivariate Gaussian target with a diagonal covariance. We use a 100 particles and run it for a 1000 iterations using gradient descent.

Table 1: Benchmarks on Bayesian neural network. We show that in 3 out of the 5 tasks, CC-SVGD performs better than SVGD and MAP with the same network and hyperparameters.

| Dataset     | SVGD  | CC-SVGD | MAP  |
|-------------|-------|---------|------|
| Mean Test RMSE |       |         |      |
| Boston      | 2.565 | 2.404   | 2.567|
| Yacht       | 0.749 | 0.611   | 0.749|
| Protein     | 4.578 | 4.740   | 4.578|
| Concrete    | 5.297 | 5.609   | 5.297|
| Real Estate | 6.608 | 6.566   | 7.208|
| Mean Test Log-Likelihood |       |         |      |
| Boston      | -2.391| -2.369  | -2.392|
| Yacht       | -1.404| -1.167  | -1.404|
| Protein     | -2.944| -2.968  | -2.944|
| Concrete    | -3.052| -3.148  | -3.052|
| Real Estate | -3.378| -3.378  | -3.476|

6 Discussion

We developed kernelized complete conditional Stein discrepancies. We show that KCC-SDs are an upper bound on KSDs and can be tractably computed on samples given an unnormalized differentiable distribution. They lead to better sample quality measures and variational inference algorithms. The KSD with the RBF kernel is not able to detect non-convergence with non-tight sequences. However, we observe that the KCC-SD with RBF kernel not only detects non-convergence but also has a higher discrepancy than the KSD with IMQ. A proof that KCC-SD does or does not detect non-convergence for non-tight sequences with RBF kernels is a promising theoretical avenue of research. Empirically when distributions match, KCC-SD and KSD converge to zero at the same rate. While in Theorem 3, we show that the KCC-SD upper bounds the KSD, this means that the KCC-SD could provide a more powerful goodness-of-fit test (Liu et al., 2016). Like KSDs, KCC-SDs also suffer from a computational cost that grows quadratically with the number of samples. To address this, random feature Stein discrepancies (Huggins and Mackey, 2018) were developed. Applications of this method using KCC-SD are a promising avenue for research.

Acknowledgments

We would like to thank Jaan Altosaar, Mark Goldstein, Xintian Han, Aahlad Manas Puli, and Bharat Srikishan for their helpful feedback and comments.

References

Chwialkowski, K., Strathmann, H., and Gretton, A. (2016). A kernel test of goodness of fit.
A Construction of cc-sds

Consider a distribution \( p(x_1, x_2) \) and suppose we want to compare their complete conditionals, the univariate distributions \( p(x_1 \mid x_2) \) and \( q(x_1 \mid x_2) \). Using a similar analysis as Ranganath, 2018, we use the univariate Langevin-Stein operator to compare these complete conditionals. Let \( h \) be a univariate function. The Langevin-Stein operator applied to \( h \) yields

\[
(A_{p(x_1 \mid x_2)} h)(x) = h(x_1) \nabla_{x_1} \log p(x_1, x_2) + \nabla_{x_2} h(x_1) = h(x_1) \nabla_{x_1} \log p(x_1, x_2) + \nabla_{x_1} h(x_1).
\]

The equality uses the fact that the score function of the conditional distribution is the score function of the joint, \( \nabla_{x_1} \log p(x_1, x_2) = \nabla_{x_1} \log p(x_1, x_2) \). Note that although the test function \( h \) is a univariate function, the operator applied to \( h \), \( g_h = A_{p(x_1 \mid x_2)} h \), is a scalar-valued function of multiple variables, \( g_h : \mathbb{R}^2 \to \mathbb{R} \).
If \( p(x_1 \mid x_2) = q(x_1 \mid x_2) \) for all inputs \( x_1 \), Stein’s identity applies and
\[
E_{q(x_1 \mid x_2)}[(A_{p(x_1 \mid x_2)}h)(x)] = 0.
\]

Now, two distributions match only if their complete conditionals match. This means we can combine the complete conditional Stein operators to compare multivariate distributions. For any \( f(x) = (f_1(x_1), f_2(x_2)) \), we can compare the distributions \( p, q \) as follows:
\[
E_{q(x)}[(A_{p(x)}f)(x)] = E_{q(x)}\left[ E_{q(x_1 \mid x_2)_1}[(A^1_{p(x_1 \mid x_2)}f_1)(x_1)] + E_{q(x_2 \mid x_1)_2}[(A^2_{p(x_2 \mid x_1)}f_2)(x_2)] \right]
= E_{q(x)}\left[ A^1_{p(x_1 \mid x_2)}f_1 + A^2_{p(x_2 \mid x_1)}f_2(x) \right],
\]
where we use the fact that \( E_{q(x)}E_{q(x_1 \mid x_2)}h(x) = E_{q(x_1, x_2)}h(x) \). Observe that the discrepancy can be computed without the use of complete conditionals of \( p \) or \( q \). Hence, the complete conditional factorization suggests the use of test functions with univariate component functions.

**B Variational Inference Using Stein Discrepancies**

Variational inference casts Bayesian inference as an optimization problem. This is typically formulated as minimizing the KL divergence between the posterior and variational family, \( q_\lambda \), over \( \lambda \). Operator variational inference (OPVI) (Ranganath et al., 2016) uses Stein discrepancies as objectives for variational inference. Stein variational gradient descent (SVGD) (Liu and Wang, 2016) uses Stein’s method to iteratively transform a set of particles to match the posterior. We describe how to use kcc-sds in OPVI and SVGD yielding black box variational inference algorithms (Ranganath et al., 2014).

**B.1 Operator Variational Inference**

OPVI suggests the use of a neural network to learn the optimal test function, \( f^* \). This increases the difficulty of optimization. We introduce the use of kcc-sds and ksd as objective functions in operator variational inference. This removes the need to estimate an optimal test function.

Given a parametric model family, \( q_\lambda(z) \) and data model \( p(x, z) \), kernelized OPVI solves the following optimization problem:
\[
\lambda^* = \inf_{\lambda} \sup_{f \in C_k} \left[ E_{p(z \mid x)}[(A_{p(z \mid x)}f)(z)] - E_{q_\lambda(z)}[(A_{q_\lambda(z)}f)(z)] \right]^2
= \inf_{\lambda} \sup_{f \in C_k} \left[ E_{q_\lambda(z)}[(A_{q_\lambda(z)}f)(z)] \right]^2
= \inf_{\lambda} \mathbb{S}(q_\lambda, A_{p(z \mid x)}, C_k)^2.
\]

Unbiased estimation of \( \mathbb{S}(q_\lambda, A_{p(z \mid x)}, C_k)^2 \) only requires the evaluation of the model score function, \( \nabla_z \log p(x, z) \) and samples from \( q_\lambda \). As the only requirement for the variational approximation is sampling (and differentiability for gradients), this allows flexibility to choose variational families where a tractable density is not available. Such distributions are called variational programs and were studied in (Ranganath et al., 2016).

**B.2 Stein Variational Gradient Descent**

SVGD, a particle based variational inference algorithm, is based on creating a set \( Q \) of distributions, which consists of distributions which are obtained by taking smooth and invertible transformations \( T \), of a reference distribution \( q \). The resulting \( q_T \in Q \) is defined as
\[
q_T(z) = q(T^{-1}(z)) \cdot |\det(\nabla_z T^{-1}(z))|
\]
where \( T^{-1} \) and \( \nabla_z T^{-1}(z) \) denote the inverse and the Jacobian matrix of the inverse.
Now, suppose we choose the family of transformations, $T$, to be small perturbations of the identity map of the form $T(x) = x + \epsilon \phi(x)$, where $\phi(x)$ is a smooth function belonging to a suitable function family. The Stein operator is equivalent to the derivative of the KL divergence (Liu and Wang, 2016). We note that the Stein operator in the KL divergence derivative are built from the matrix Stein operator. However, the derivative uses the trace of the matrix Stein operator which is equal to the Stein operator we study.

**Theorem 4** (Liu and Wang, 2016) Let $T(x) = x + \epsilon \phi(x)$ and $q_T(z)$ be the density of $z = T(x)$ where $x \sim q(x)$. Then
\[\nabla_x \text{KL}(q_T || p)_{|\epsilon = 0} = -E_{q(x)}[\text{Tr}(\mathcal{A}_p(x) \phi)(x)],\]
where $(\mathcal{A}_p(x) \phi)(x) = \nabla_x \log p(x) \phi(x)^T + \nabla_x \phi(x)$ is the matrix Stein operator.

The following lemma identifies the maximal perturbation direction $\phi^*_p,q$ that gives the steepest decrease in the KL divergence:

**Lemma 2** (Liu and Wang, 2016) Assume the conditions in Theorem 4, consider all the perturbation directions $\phi$ in the ball $\mathcal{B} = \{ \phi \in \mathcal{G}_k : ||\phi||_{\mathcal{G}_k} \leq S(q, \mathcal{A}_p, \mathcal{G}_k) \}$ of function space $\mathcal{G}_k$. The direction of steepest descent that maximizes the negative gradient in Equation (6) is given by
\[\phi^*_p,q(\cdot) = E_{q(x)}[k(x, \cdot) \nabla_x \log p(x) + \nabla_x k(x, \cdot)],\]
which implies that $\nabla_x \text{KL}(q_T | p)_{|\epsilon = 0} = -S(q, \mathcal{A}_p, \mathcal{G}_k)^2$.

Here, we propose the use of the kcc-sd Stein set, $\mathcal{C}_k$, as the perturbation family for SVGD rather than using the KS SD Stein set for the perturbation family. The optimal function for SVGD using the kcc-sd Stein set, $\psi^*_p,q : \mathbb{R}^d \rightarrow \mathbb{R}^d$, is defined as
\[\psi^*_p,q(\cdot) = E_{q(x)}[k(x, \cdot) \nabla_x \log p(x) + \nabla_x k(x, \cdot)]\]
where $(k(x, y) \otimes \nabla_x \log p(x))_i = k(x^i, y^i) \nabla_{x^i} \log p(x)$, with $x^i$ denoting the $i$-th dimension and $(\nabla_x k(x, y))_i = \nabla_{x^i} k(x^i, y^i)$. Here, we state a similar lemma to Lemma 2 to show that if the perturbation functions belong to the kcc-sd Stein set, then there is a closed form optimal perturbation function.

**Lemma 3** (cc-svgd) Assume the conditions in Theorem 4, consider all the perturbation directions $\phi$ in the ball $\mathcal{B} = \{ \phi \in \mathcal{C}_k : ||\phi||_{\mathcal{C}_k} \leq S(q, \mathcal{A}_p, \mathcal{C}_k) \}$ of function space $\mathcal{C}_k$. The direction of steepest descent that maximizes the negative gradient in Equation (6) is given by
\[\psi^*_p,q(\cdot) = E_{q(x)}[k(x, \cdot) \nabla_x \log p(x) + \nabla_x k(x, \cdot)]\]
which implies that $\nabla_x \text{KL}(q_T | p)_{|\epsilon = 0} = -S(q, \mathcal{A}_p, \mathcal{C}_k)^2$.

We refer to SVGD using kcc-sd updates as complete conditional Stein variational gradient descent (cc-svgd).

**Theorem 3** shows that $S(q, \mathcal{A}_p, \mathcal{C}_k) \geq S(q, \mathcal{A}_p, \mathcal{G}_k)$. This implies that the perturbation function provided by kcc-sd, decreases the KL-divergence more than the perturbation function provided by KS D. If $T(x) = x + \epsilon \psi^*_p,q(x)$, then
\[\nabla_x \text{KL}(q_T || p)_{|\epsilon = 0} = -S(q, \mathcal{A}_p, \mathcal{C}_k)^2 \leq -S(q, \mathcal{A}_p, \mathcal{G}_k)^2 ,\]
which shows that the kcc-sd perturbation function points in a steeper direction of descent than the KS D perturbation function. Note, that this is only a locally optimal step. Given particles $\{x_i\}_{i=1}^n$, we calculate the svgd update for a particle $x_j$ as
\[\phi^*_p,q(x_j) = \frac{1}{n} \sum_{i=1}^n k(x_i, x_j) \nabla x_i \log p(x_i) + \nabla x_j k(x_i, x_j) .\]
We can see that if the particles are far apart then $k(x_i, x_j) = e^{-\beta ||x_i - x_j||^2}$ gets small. This means that SVGD reduces to performing MAP updates for these particles, $\phi^*_p,q(x_j) = \frac{1}{p} \nabla x_j \log p(x_j)$. We demonstrate this phenomenon with the two layer Bayesian neural network, by calculating the Frobenius norm between the matrix of MAP updates and the SVGD updates.
Algorithm 1 Complete Conditional Stein Variational Gradient Descent

Input: Model \( \log p(x) \), initialize particles \( \{x_i^0\}_{i=1}^n \) from \( q_0 \)

while not converged do

\( x_i^{t+1} = x_i^t + \epsilon_t \psi^{\ast}_{p,q}(x_i^t) \) where \( \hat{\psi}^{\ast}_{p,q}(x) = \frac{1}{n} \sum_{j=1}^n k_{cc}(x_j, x) \nabla_{x_j} \log p(x_j) + \nabla_{x_j} k_{cc}(x_j, x) \)

end while

Bayesian Neural Network. Consider a Bayesian neural network with one hidden layer with 50 hidden units with a ReLU \((x) = \max(0, x)\) activation, and we use the Boston housing dataset for our experiments.

Here we compare SVGD and CC-SVGD to MAP. As shown in Figure 6, SVGD reduces to MAP in higher dimensions, unlike CC-SVGD which increases as the dimension grows. We use a standard Gaussian to initialize the weight matrices, this causes the particles to be far apart and thus reduces the SVGD update to MAP updates, with no interaction between particles without tricks like the median heuristic.

![Figure 6: SVGD updates reduce to MAP updates in high dimensions.](image)

We compare the SVGD update and the kcc-sd update to the MAP update on the Boston housing dataset. We increase the dimension of the hidden layer and show that SVGD decreases and CC-SVGD increases as the dimension increases. We use twenty particles and plot the difference between the updates after 10 iterations.

B.3 Operator Variational Inference

In this section, we use kcc-sd as an objective for variational inference.

Bayesian Linear Regression. Consider the Bayesian regression problem, \( y = x^T w + b \), where \( x, w \in \mathbb{R}^d \). We model the data as \( p(y | x, w) = \mathcal{N}(y | x^T w, \sigma^2_y I) \) with a normal prior on \( w \). We perform posterior approximation using the variational family, \( q_\lambda(w) = \mathcal{N}(w | \mu_w, \sigma^2_w I) \), where the variational parameters are \( \mu_w \) and \( \sigma_w \).

We run OPVI with kcc-sd and KSD using the IMQ kernel for different dimensions, with a fixed number of data points, \( x_i, y_i \), and use \( n_w = 100 \) latent samples, used to calculate the Stein discrepancies with the AdaGrad optimizer. The dataset was generated by randomly picking \( w \). We observed that OPVI with kernelized discrepancies requires manual tuning of the optimizer. In Table 2 we list the \( L_2 \) norm between the learned posterior mean and the true mean after 20 iterations.

| Dimension | KSD MSE | KCC-SD MSE |
|-----------|---------|------------|
| \( d = 3 \) | 0.34 | 0.05 |
| \( d = 5 \) | 0.38 | 0.19 |
| \( d = 20 \) | 4.39 | 1.26 |
| \( d = 50 \) | 6.31 | 3.25 |

Table 2: Operator Variational Inference
We increase the mean of \( C \) where
\[
\mu_j = \langle A_{f_0} \rangle_{K_k} f_0(x_j) = f_0(x_j) \nabla_{x_j} \log p(x) + \nabla_{x_j} f_0(x_j), \quad \text{for } f_0 : \mathbb{R} \rightarrow \mathbb{R}
\]
Let \( \Psi_k : \mathbb{R} \rightarrow \mathcal{K}_k \) be the 1-dimensional canonical feature map, defined by
\[
\Psi_{k,x_j}(z_j) = k(x_j, z_j).
\]
Now, since \( k \in C^1([0,1] \times [0,1]) \) follows (Gorham and Mackey, 2017) and using Corollary 3.46 from (Steinwart and Christmann, 2008), we show that for all \( f \in \mathcal{C}_k \)
\[
\mathcal{A}_{p(x)} f(x) = \sum_{j=1}^d (\mathcal{A}_{p(x)} f_j)(x) = \sum_{j=1}^d \mathcal{A}_{p(x)} f_j(x) = \sum_{j=1}^d \langle f_j, \Psi_{k,x_j} \rangle_{\mathcal{K}_k} = \sum_{j=1}^d \langle f_j, A_{p(x)} \Psi_{k,x_j} \rangle_{\mathcal{K}_k},
\]
where \( \mathcal{C}_k = \{ f : \mathbb{R}^d \rightarrow \mathbb{R}^d \mid f_j \in \mathcal{K}_k, \text{ for all } j \leq d \} \).

Then using Lemma 4.34 of (Steinwart and Christmann, 2008), gives for all \( x, y \in \mathbb{R}^d \).
\[
\langle A_{p(x)} \Psi_{k,x_j}, A_{p(y)} \Psi_{k,y_j} \rangle = \langle s^p(x) \Psi_{k,x_j} + \nabla_{x_j} \Psi_{k,x_j}, s^p(y) \Psi_{k,y_j} + \nabla_{y_j} \Psi_{k,y_j} \rangle_{\mathcal{K}_k}
\]
\[
= s^p_j(x) s^p_j(y) k(x_j, y_j) + s^p_j(x) \nabla_{y_j} k(x_j, y_j) + s^p_j(y) \nabla_{x_j} k(x_j, y_j) + \nabla_{x_j} \nabla_{y_j} k(x_j, y_j)
\]
\[
= k^d_{\text{loc}}(x, y).
\]

Now, assuming that Equation (10) is integrable with respect to \( q \), we have that \( A_{p(x)} \Psi_k \) is Bochner \( q \)-integrable for each \( j \). In other words, we can interchange expectation and the inner product. This
Figure 8: **KCC-SDs converge to zero with i.i.d samples from** $p$. Here we compute KCC-SD and KSD for $p = \mathcal{N}(0, I_d)$, with $n = 1000$ samples in $d = 10$. We can observe that both the KSD and the KCC-SD converge to zero at the same rate.

Figure 9: **KCC-SDs converge to zero with i.i.d samples from** $p$. Here we compute KCC-SD and KSD, with $p = \mathcal{N}(0, \Sigma)$ where $\Sigma$ was generated randomly to have non-zero diagonal entries, both discrepancies were computed with $n = 1000$ samples and $d = 10$.

implies that

$$w_j^2 = \mathbb{E}[h_{cc}^j(x, y)] = \mathbb{E}[\langle A^j_{p(x)} \Psi_{k,x_j}, A^j_{p(y)} \Psi_{k,y_j} \rangle_{K_k}] = \langle \mathbb{E}[A^j_{p(x)} \Psi_{k,x_j}], \mathbb{E}[A^j_{p(y)} \Psi_{k,y_j}] \rangle_{K_k} = \left\| \mathbb{E}_{q(x)}[A^j_{p(x)} \Psi_{k,x_j}] \right\|_{K_k}^2,$$

where $x, y \sim q$.

Now using Equation (9) and Equation (11) and using Fenchel-Young inequality for the dual norm twice, we show that

$$S(q, A_p, C_k) = \sup_{f \in C_k} \mathbb{E}_{q(x)}[A_p(x)f(x)]$$

$$= \sup_{\|f\|_{K_k}} \sum_{j=1}^d \langle f_j, \mathbb{E}_{q(x)}[A^j_{p(x)} \Psi_{k,x_j}] \rangle_{K_k}$$

$$= \sup_{\|v\| \leq 1} \sum_{j=1}^d v_j \left\| \mathbb{E}_{q(x)}[A^j_{p(x)} \Psi_{k,x_j}] \right\|_{K_k}$$

$$= \sup_{\|v\| \leq 1} \sum_{j=1}^d v_j w_j = \|w\|.$$

(12)

Recall that the Stein set is defined as a product function space of univariate RKHS, limited to the unit ball $\|v\|$. We focus on $\|\cdot\|_2$ norm throughout, however it can be generalized to any norm. In other words, $C_k = \{f(x) = (f_1(x_1), \ldots, f_d(x_d)) \mid \|v\| \leq 1, \|f_j\|_{K_k} = v_j \}$. 

14
We prove Proposition 2 by using an important lemma from (Gorham and Mackey, 2017), which shows that:

**Lemma 4 (Gorham and Mackey, 2017)** Let $z \sim q$ and $x \sim q$, and if $s^p(x) = \nabla_x \log p(x)$ is Lipschitz and $E_{p(x)}[\|\nabla_x \log p(x)\|_2^2] < \infty$, then for any $f : \mathbb{R}^d \to \mathbb{R}$ with $M_0(f), M_2(f) < \infty$, we have the following upper bound

$$E_{p(x)}[A_p(x)f(x)] \leq C_1(f, s^p)W_2(p, q) + \sqrt{C_2(f, s^p)W_2(p, q)},$$

where $C_1(f, s^p) = (M_0(f)M_1(s^p) + M_2(f)d)$ and $C_2(f, s^p) = 2M_0(f)M_1(f)E_{p(x)}[\|\nabla_x \log p(x)\|_2^2]$ are constants which depend on $p$ and $f$. (Gorham and Mackey, 2017) also show that this lemma applies to kcc-sd since we have the same bound as the univariate ksd. We repeat their argument below.

First, we define the multi-index differential operator, $D^\alpha$, where $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$. For a differentiable function, $f : \mathbb{R}^d \to \mathbb{R}$, we have

$$D^\alpha f(x) = \frac{d^{|\alpha|}}{(dx_1)^{\alpha_1} \ldots (dx_d)^{\alpha_d}} f(x).$$

Now, let $g \in C_k$, then for any multi-index $\alpha \in \mathbb{N}^d$, with $|\alpha| \leq 2$. We can bound the derivative of each component function $D^\alpha g(x)$ by Cauchy-Schwartz and (Steinwart and Christmann, 2008) as

$$\sup_{x \in \mathbb{R}^d} |D^\alpha g_j(x_j)| = \sup_{x \in \mathbb{R}^d} |D^\alpha (g_j, k(x_j, \cdot))| \leq \sup_{x \in \mathbb{R}^d} \|g_j\|_{K_k} \|D^\alpha k(x_j, \cdot)\| \leq \|g_j\|_{K_k} \sup_{x \in \mathbb{R}^d} (D^\alpha x_i D^\alpha k)^{1/2},$$

and since $\|g_j\|_{K_k}$ is bounded for all $g \in C_k$ and as $D^\alpha_x D^\alpha_y k$ is bounded for all $|\alpha| \leq 2$. We have that $M_0(g)$, $M_1(g)$, $M_2(g)$ are all bounded. This means that Lemma 4 applies to kcc-sd, and hence kcc-sd detects convergence.

**F Detecting Non-Convergence**

We assume that $k(x, y)$, where $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, is an integrally strictly positive definite kernel. So that if $0 < \|f\|_2 < \infty$, then

$$\int \int f(xk(x, y)f(y)dx dy > 0. \tag{14}$$

Now, note that as $E_{q(x)}[A_{r(x)}f(x)] = 0$, we get the following form for the Stein discrepancy:

$$E_{q(x)}[A_{p(x)}f(x)] = E_{q(x)}[A_{p(x)}f(x) - A_{q(x)}f(x)]$$

$$= E_{q(x)}[(s_p(x) - s_q(x))^T f(x)], \tag{15}$$

where $s_j(x) = \nabla_x \log p(x)$ is the score function. Hence, the Complete Conditional Stein Discrepancy takes the following form

$$E_{q(x)}[A_{p(x)}f(x)] = \sum_i E_{q(x)}[A_{i(x)}f^i(x)]$$

$$= \sum_i E_{q(x)}[(s^i_p(x) - s^i_q(x))^T f^i(x)]. \tag{16}$$

Note that this is different from the Langevin-Stein operator defined in Section 3.
Proof 1 (Detecting Non-Convergence) Define a function \( f^* : \mathbb{R} \to \mathbb{R} \),
\[
f^*(x_i) = \mathbb{E}_{q(y)}[A^i_{p(y)}k(x_i, y_i)] .
\]
Then using Equation (16), we obtain
\[
f^*(x_i) = \mathbb{E}_{q(y)}[(s^i_p(y) - s^i_q(y))k(x_i, y_i)] = \mathbb{E}_{q(y)}[(s^i_p(y_i | y_{-i}) - s^i_q(y_i | y_{-i}))k(x_i, y_i)] .
\]
Let \( b(x_i) \) be defined as follows:
\[
b(x_i) = \int_{x_{-i}} q(x_{-i}) \left( \nabla x_{-i} \log p(x_i | x_{-i}) - \nabla x_{-i} \log q(x_i | x_{-i}) \right) dx_{-i}
\]
\[
= \int_{x_{-i}} q(x_i, x_{-i}) \left( \nabla x_{-i} \log p(x_i, x_{-i}) - \nabla x_{-i} \log q(x_i, x_{-i}) \right) dx_{-i} .
\]
Hence, \( S(q, A_p, C_k) \) can be expressed as follows:
\[
S(q, A_p, C_k)^2 = \sum_i \mathbb{E}_{q(x)}[\mathbb{E}_{p(y)} f^*(x_i)]
\]
\[
= \sum_i \mathbb{E}_{q(x)} \mathbb{E}_{q(y)}[(s^i_p(x) - s^i_q(x))k(x_i, y_i)(s^i_p(y) - s^i_q(y))]
\]
\[
= \sum_i \int y_i \int x_i b(x_i) k(x_i, y_i) b(y_i) dx_i dy_i .
\]
Now, note that since \( k \) is integrally strictly positive definite, \( \|b(y_i)\|^2_2 = 0 \) if and only if
\[
\int_{y_i} \int_{x_i} b(x_i) k(x_i, y_i) b(y_i) dx_i dy_i = 0 ,
\]
so if \( S(p, q, C_k) = 0 \), then \( \|b(y_i)\|^2_2 = 0 \) for all \( i \leq d \). Note, that for fixed \( i \)
\[
\|b(y_i)\|^2_2 = \int y q(y_{-i})^2 q(y_i | y_{-i})^2 \left( s^i_p(y_i | y_{-i}) - s^i_q(y_i | y_{-i}) \right)^2 dy_{-i} = 0
\]
\[
= \mathbb{E}_{q(y)} \left[ q(y) \left( s^i_p(y_i | y_{-i}) - s^i_q(y_i | y_{-i}) \right)^2 \right]
\]
\[
= 0 .
\]
Finally, as \( q(y) > 0 \), we have that \( \left( \nabla y_i \log p(y_i | y_{-i}) - \nabla y_i \log q(y_i | y_{-i}) \right)^2 = 0 \) a.s. with respect to \( q(y) \). Therefore, as \( \nabla x_i \log p(x_i | x_{-i}) = \nabla x_i \log p(x) \), we can relate \( S(q, A_p, C_k) \) to the Fisher Score:
\[
0 = \sum_i \mathbb{E}_{q(y)} \left[ (\nabla y_i \log p(y_i | y_{-i}) - \nabla y_i \log q(y_i | y_{-i}))^2 \right]
\]
\[
= \sum_i \mathbb{E}_{q(y)} \left[ (\nabla y_i \log p(y) - \nabla y_i \log q(y))^2 \right]
\]
\[
= \mathbb{E}_{q(y)} \left[ \sum_i (\nabla y_i \log p(y) - \nabla y_i \log q(y))^2 \right]
\]
\[
= \mathbb{E}_{q(y)} \left[ ||\nabla y \log p(y) - \nabla y \log q(y)||_2^2 \right] = F(p, q) .
\]
Proof 2 (Fisher Score) Since \( p, q \) have full support on \( \mathbb{R}^d \), with \( p > 0, \forall x \in \mathbb{R}^d \). Then as \( ||\nabla x \log p(x) - \nabla x \log q(x)||_2^2 \geq 0 \) for all \( x \in \mathbb{R}^d \). If \( F(p, q) = \mathbb{E}_{q(x)} ||\nabla x \log p(x) - \nabla x \log q(x)||_2^2 \equiv 0, \) that implies that \( ||\nabla x \log p(x) - \nabla x \log q(x)||_2^2 = 0 \) almost surely with respect to \( q \). Which then implies that \( \log p(x) = \log q(x) + c \) almost surely (wrt \( q \)), which is equivalent to \( p(x) = q(x)e^c \). But as \( \int p(x)dx = e^c \int q(x)dx = 1 \), we get that \( c = 0 \).
G Relations

In this section we prove Theorem 3. We rely on the following conditions:

**C1** \( k(x_j, y_j) - k(x, y) \geq 0 \) for all \( x, y \in \mathbb{R}^d \)

**C2** \( k(x_j, y_j) - k(x, y) \) is a completely monotone function in \( (x_j - y_j)^2 \) with \( x_j, y_j \) fixed.

Then let \( k_d(x_j, y_j; x_j, y_j) = k(x_j, y_j) - k(x, y) \), then note that for the kbf, iMQ and Matern kernels, we observe that \( k_d = \Phi((x_j - y_j)^2) \), where \( \Phi : \mathbb{R}^+ \rightarrow \mathbb{R} \) is a completely monotone function.

**Lemma 5 (Schoenberg, 1938)** If a function \( \Phi : [0, \infty) \rightarrow \mathbb{R} \) is completely monotone but not constant, then \( \Phi(\|\cdot\|_2^2) \) is strictly positive definite and radial on \( \mathbb{R}^d \) for any \( d \).

Some common examples are the functions, \( \Phi(r) = e^{-\beta r} \), \( \Phi(r) = (r + c^2)^{-\beta} \).

Suppose \( k(x, y) = e^{-\|x-y\|_2^2} \), then note that \( k_d(x_j, y_j; x_j, y_j) = k(x_j, y_j) - k(x, y) \) is defined as

\[
k_d(x_j, y_j; x_j, y_j) = e^{-\|x_j-y_j\|_2^2} (1 - e^{-\|x_j-y_j\|_2^2}),
\]

since \( 1 - e^{-\|x_j-y_j\|_2^2} \) is fixed and positive, we note that \( k_d(x_j, y_j; x_j, y_j) = \Phi((x_j - y_j)^2) \), where \( C > 0 \) is a positive constant. This implies that \( \Phi(r) \) is completely monotone since

\[
(-1)^l \frac{d^l}{dr} \Phi(r) = C\beta^l e^{-\beta r} \geq 0.
\]

Also, suppose if \( k(x, y) = (c^2 + r)^{-\beta} \) then we get that

\[
k_d(x_j, y_j; x_j, y_j) = (c^2 + (x_j - y_j)^2)^{-\beta} - (c^2 + \|x-y\|_2^2)^{-\beta}.
\]

Next, note that \( \varphi(r) = (c^2 + r)^{-\beta} \) is a completely monotone function, since

\[
(-1)^l \frac{d^l}{dx} \varphi(r) = (-1)^l \beta^l (\beta + 1) \ldots (\beta + l - 1) (r + c^2)^{-\beta-l} \geq 0.
\]

As \( (c^2 + r)^{-\alpha} > (d^2 + r)^{-\alpha} \) if \( d > c \) for all \( \alpha > 0 \), we get that \( k_d(x_j, y_j; x_j, y_j) = \Phi((x_j - y_j)^2) \) is a completely monotone function.

**Lemma 6** Let \( k_d(x_j, y_j) = k(x_j, y_j) \) satisfy conditions C1 and C2 and let \( \nabla^2 \log p, \nabla^2 \log q \) be Lipschitz and \( E_{q(x)}[\|\nabla^2 \log p(x)\|^2], E_{q(x)}[\|\nabla^2 \log q(x)\|^2] < \infty \) and \( S(q, A_p, C_k), S(q, A_p, G_k) < \infty \). Then

\[
S(q, A_p, G_k) \leq S(q, A_p, C_k).
\]

**Proof 3 (Theorem 3)** Assuming the base kernel for KSD and kcc-sd satisfies the conditions C1 and C2 then following Appendix F KSD can be expressed as follows:

\[
S(q, A_p, G_k)^2 = \sum_{j=1}^{d} E_{q(x) \times q(y)} [(s_p^j(x) - s_q^j(x)) k(x, y) (s_p^j(y) - s_q^j(y))],
\]

and similarly, kcc-sd can be expressed as follows:

\[
S(q, A_p, C_k)^2 = \sum_{j=1}^{d} E_{q(x) \times q(y)} [(s_p^j(x) - s_q^j(x)) k(x_j, y_j) (s_p^j(y) - s_q^j(y))],
\]

then the difference between KCC-sd and KSD is:

\[
S(q, A_p, C_k)^2 - S(q, A_p, G_k)^2 = \sum_{j=1}^{d} E_{q(x) \times q(y)} [(s_p^j(x) - s_q^j(x)) k(x_j, y_j) - k(x, y) (s_p^j(y) - s_q^j(y))].
\]
Now, assuming $k_d(x_j, y_j) = k(x_j, y_j) - k(x, y)$ satisfies the conditions C1 and C2, we note that $k_d$ is integrally strictly positive semi-definite, implying that

$$
\int_{x_j} \int_{y_j} f(x_j) k_d(x_j, y_j; x_{-j}, y_{-j}) f(y_j) dx_j dy_j \geq 0,
$$

for all $x_{-j}, y_{-j}$ and for all univariate $\mathcal{L}^2 \mathbb{R}$ functions. Hence, $S(q, A_p, C_k) - S(q, A_p, G_k) \geq 0$, in other words KCC-SD upper bounds KSD.