Random Feature Stein Discrepancies

Jonathan H. Huggins*  
Department of Biostatistics, Harvard  
jhuggins@mit.edu  
Lester Mackey*  
Microsoft Research New England  
lmackey@microsoft.com

Abstract

Computable Stein discrepancies have been deployed for a variety of applications, ranging from sampler selection in posterior inference to approximate Bayesian inference to goodness-of-fit testing. Existing convergence-determining Stein discrepancies admit strong theoretical guarantees but suffer from a computational cost that grows quadratically in the sample size. While linear-time Stein discrepancies have been proposed for goodness-of-fit testing, they exhibit avoidable degradations in testing power—even when power is explicitly optimized. To address these shortcomings, we introduce feature Stein discrepancies (ΦSDs), a new family of quality measures that can be cheaply approximated using importance sampling. We show how to construct ΦSDs that provably determine the convergence of a sample to its target and develop high-accuracy approximations—random ΦSDs (RΦSDs)—which are computable in near-linear time. In our experiments with sampler selection for approximate posterior inference and goodness-of-fit testing, RΦSDs perform as well or better than quadratic-time KSDs while being orders of magnitude faster to compute.

1 Introduction

Motivated by the intractable integration problems arising from Bayesian inference and maximum likelihood estimation [9], Gorham and Mackey [10] introduced the notion of a Stein discrepancy as a quality measure that can potentially be computed even when direct integration under the distribution of interest is unavailable. Two classes of computable Stein discrepancies—the graph Stein discrepancy [10, 12] and the kernel Stein discrepancy (KSD) [6, 11, 19, 21]—have since been developed to assess and tune Markov chain Monte Carlo samplers, test goodness-of-fit, train generative adversarial networks and variational autoencoders, and more [6, 10–12, 16–19, 27]. However, in practice, the cost of these Stein discrepancies grows quadratically in the size of the sample being evaluated, limiting scalability. Jitkrittum et al. [16] introduced a special form of KSD termed the finite-set Stein discrepancy (FSSD) to test goodness-of-fit in linear time. However, even after an optimization-based preprocessing step to improve power, the proposed FSSD experiences a unnecessary degradation of power relative to quadratic-time tests in higher dimensions.

To address the distinct shortcomings of existing linear- and quadratic-time Stein discrepancies, we introduce a new class of Stein discrepancies we call feature Stein discrepancies (ΦSDs). We show how to construct ΦSDs that provably determine the convergence of a sample to its target, thus making them attractive for goodness-of-fit testing, measuring sample quality, and other applications. We then introduce a fast importance sampling-based approximation we call random ΦSDs (RΦSDs). We provide conditions under which, with an appropriate choice of proposal distribution, an RΦSD is close in relative error to the ΦSD with high probability. Using an RΦSD, we show how, for any $\gamma > 0$, we can compute $O_P(N^{-1/2})$-precision estimates of an ΦSD in $O(N^{1+\gamma})$ (near-linear) time when the ΦSD precision is $\Omega(N^{-1/2})$. Additionally, to enable applications to goodness-of-fit testing,

*Contributed equally
we (1) show how to construct RΦSDs that can distinguish between arbitrary distributions and (2) describe the asymptotic null distribution when sample points are generated i.i.d. from an unknown distribution. In our experiments with biased Markov chain Monte Carlo (MCMC) hyperparameter selection and fast goodness-of-fit testing, we obtain high-quality results—which are comparable to or better than those produced by quadratic-time KSDs—using only ten features and requiring orders-of-magnitude less computation.

**Notation** For measures μ₁, μ₂ on $\mathbb{R}^D$ and functions $f : \mathbb{R}^D \to \mathbb{C}$, $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{C}$, we let $\mu_1(f) := \int f(x)\mu_1(dx)$, $(\mu_k(k)) := \int k(x,x')\mu_1(dx')$, and $(\mu_1 \times \mu_2)(k) := \int \int k(x_1,x_2)\mu_1(dx_1)\mu_2(dx_2)$. We denote the generalized Fourier transform of $f$ by $\hat{f}$ or $\mathcal{F}(f)$ and its inverse by $\mathcal{F}^{-1}(f)$. For $r \geq 1$, let $L^r := \{ f : \|f\|_{L^r} := (\int |f(x)|^r \, dx)^{1/r} < \infty \}$ and $C^n$ denote the space of $n$-times continuously differentiable functions. We let $\overset{D}{\to}$ and $\overset{p}{\to}$ denote convergence in distribution and in probability, respectively. We let $\overline{a}$ denote the complex conjugate of $a$. For $D \in \mathbb{N}$, define $[D] := \{1, \ldots, D\}$. The symbol $\gtrsim$ indicates greater than up to a universal constant.

## 2 Feature Stein discrepancies

When exact integration under a target distribution $P$ is infeasible, one often appeals to a discrete measure $Q_N = \frac{1}{N} \sum_{n=1}^{N} \delta_{x_n}$ to approximate expectations, where the sample points $x_1, \ldots, x_N \in \mathbb{R}^D$ are generated from a Markov chain or quadrature rule. The aim in sample quality measurement is to quantify how well $Q_N$ approximates the target in a manner that (a) recognizes when a sample sequence is converging to the target, (b) highlights when a sample sequence is not converging to the target, and (c) is computationally efficient. It is natural to frame this comparison in terms of an integral probability metric (IPM) [20], $d_H(Q_N, P) := \sup_{h \in H} |Q_N(h) - P(h)|$, measuring the maximum discrepancy between target and sample expectations over a class of test functions. However, when generic integration under $P$ is intractable, standard IPMs like the 1-Wasserstein distance and Dudley metric may not be efficiently computable.

To address this need, Gorham and Mackey [10] introduced the Stein discrepancy framework for generating IPM-type quality measures with no explicit integration under $P$. For any approximating probability measure $\mu$, each Stein discrepancy takes the form

$$d_{T,G}(\mu, P) = \sup_{g \in G} |\mu(Tg)| \quad \text{where} \quad \forall g \in G, P(Tg) = 0.$$

Here, $T$ is an operator that generates mean-zero functions under $P$, and $G$ is the Stein set of functions on which $T$ operates. For concreteness, we will assume that $P$ has $C^1$ density $p$ with support $\mathbb{R}^d$ and restrict our attention to the popular Langevin Stein operator [10, 21] defined by $Tg := \sum_{d=1}^{D} T_d g_d$ for $(T_d g_d)(x) := p(x)^{-1} \partial_d p(x) g_d(x)$ and $g : \mathbb{R}^D \to \mathbb{R}^D$. To date, two classes of computable Stein discrepancies with strong convergence-determining guarantees have been identified. The graph Stein discrepancies [10, 12] impose smoothness constraints on the functions $g$ and are computed by solving a linear program, while the kernel Stein discrepancies [6, 11, 19, 21] define $G$ as the unit ball of a reproducing kernel Hilbert space and are computed in closed-form. Both classes, however, suffer from a computational cost that grows quadratically in the number of sample points. Our aim is to develop alternative discrepancy measures that retain the theoretical and practical benefits of existing Stein discrepancies at a greatly reduced computational cost.

Our strategy is to identify a family of convergence-determining discrepancy measures that can be accurately and inexpensively approximated with random sampling. To this end, we define a new domain for the Stein operator centered around a feature function $\Phi : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{C}$ which, for some $r \in [1, \infty)$ and all $x, z \in \mathbb{R}^D$, satisfies $\Phi(x, \cdot) \in L^r$ and $\Phi(\cdot, z) \in C^1$:

$$G_{\Phi, r} := \left\{ g : \mathbb{R}^D \to \mathbb{R} \mid g_d(x) = \int \Phi(x, z) f_d(z) \, dz \quad \text{with} \quad \sum_{d=1}^{D} \|f_d\|_{L^r}^2 \leq 1 \right\}.$$

When combined with the Langevin Stein operator $T$, this feature Stein set gives rise to a feature Stein discrepancy (ΦSD) with an appealing explicit form $(\sum_{d=1}^{D} \|\mu(T_d \Phi)\|_{L^r}^2)^{1/2}$:

$$\PhiSD_{g, r}(\mu, P) := \sup_{g \in G_{\Phi, r}} |\mu(Tg)|^2 = \sup_{g \in G_{\Phi, r}} \left[ \sum_{d=1}^{D} \|\mu(T_d g_d)\|^2 \right]^{1/2}.$$
Ψ(Φ(x, z) = \int \text{d}z.) This follows from the definition \cite{6, 11, 19, 21} of the Stein discrepancy (KSD) with base reproducing kernel \( k(x, y) = \int \Phi(x, z)\Phi(y, z) \text{d}z. \) However, we will see in Sections 3 and 5 that there are significant differences between practical and theoretical benefits of using KSDs with \( r \neq 2 \). Namely, we will be able to approximate \( \Phi_{\text{SD}}^r_D \) with \( r \neq 2 \) more effectively with a smaller sampling budget. If \( \Phi(x, z) = e^{-i(z-x)} (\Psi(z)^{1/2} \text{ and } \nu \sim \Psi \text{ for } \Psi \in L^2, \) then \( \Phi_{\text{SD}}^r_D \) is the random Fourier feature (RFF) approximation \cite{22} to KSDs with \( k(x, y) = \Psi(x - y). \) Chwialkowski et al. \cite{5, Prop. 1} showed that the RFF approximation can be a desirable choice of discrepancy measure, as there exist uncountably many pairs of distinct distributions that, with high probability, cannot be distinguished by the RFF approximation. Following Chwialkowski et al. \cite{5} and Jitkrittum et al. \cite{16}, we show how to select \( \Phi \) and \( \nu \) to avoid this property in Section 4. The random finite set Stein discrepancy \cite{FSSD-rand, 16} with proposal \( \nu \) is an \( R\Phi_{\text{SD}}^{r,D} \) with \( \Phi(x, z) = f(x, z)\nu(z)^{1/2} \) for \( f \) a real analytic and \( C_0 \)-universal \cite{3, Def. 4.1} reproducing kernel. In Section 3.1, we will see that features \( \Phi \) of a different form give rise to strong convergence-determining properties.

3 Selecting a Random Feature Stein Discrepancy

In this section, we provide guidance for selecting the components of an R\( \Phi_{\text{SD}} \) to achieve our theoretical and computational goals. We first discuss the choice of the feature function \( \Phi \) and order \( r \) and then turn our attention to the proposal distribution \( \nu. \) Finally, we detail two practical choices of R\( \Phi_{\text{SD}} \) that will be used in our experiments. To ease notation, we will present theoretical guarantees in terms of the sample measure \( Q_N \), but all results continue to hold if any approximating probability measure \( \mu \) is substituted for \( Q_N \).

3.1 Selecting a feature function \( \Phi \)

A principal concern in selecting a feature function is ensuring that the \( \Phi_{\text{SD}} \) detects non-convergence—that is, \( Q_N \Rightarrow P \) whenever \( \Phi_{\text{SD}}^{r,D}(Q_N, P) \to 0. \) To ensure this, we will construct \( \Phi_{\text{SD}} \)s that upper bound a reference KSD known to detect non-convergence. This is enabled by the following inequality proved in Appendix A.

**Proposition 3.1 (KSD–\( \Phi_{\text{SD}} \) inequality).** If \( k(x, y) = \int \mathcal{F}(\Phi(x, \cdot))(\omega) \mathcal{F}(\Phi(y, \cdot))(\omega) \rho(\omega) \text{d}\omega, \) \( r \in [1, 2], \) and \( \rho \in L^2 \) for \( t = r/(2 - r), \) then

\[
\text{KSD}^2(Q_N, P) \leq \|\rho\|_{L^r} \Phi_{\text{SD}}^{r,D}(Q_N, P).
\]

(2)

Our strategy is to first pick a KSD that detects non-convergence and then choose \( \Phi \) and \( r \) such that (2) applies. Unfortunately, KSDs based on many common base kernels, like the Gaussian and Matérn, fail to detect non-convergence when \( D > 2 \) \cite{11, Thm. 6}. A notable exception is the KSD with inverse multiquadric (IMQ) base kernel.
Example 3.1 (IMQ kernel). The IMQ kernel is given by $\psi_{c,\beta}^{\text{IMQ}}(x - y) := (c^2 + \|x - y\|^2_2)^{\beta}$, where $c > 0$ and $\beta < 0$. Gorham and Mackey [11, Thm. 8] proved that when $\beta \in (-1, 0)$, KSDs with an IMQ base kernel determine weak convergence on $\mathbb{R}^D$ whenever $P \in \mathcal{P}$, the set of distantly dissipative distributions for which $\nabla \log p$ is Lipschitz.\(^2\)

Let $m_N := \mathbb{E}_{X \sim Q_N}[X]$ denote the mean of $Q_N$. We would like to consider a broader class of base kernels, the form of which we summarize in the following assumption:

**Assumption A.** The base kernel has the form $k(x, y) = A_N(x)\psi(x - y)A_N(y)$ for $\psi \in C^2$, $A \in C^1$, and $A_N(x) := A(x - m_N)$, where $A > 0$ and $\nabla \log A$ is bounded and Lipschitz.

The IMQ kernel falls within the class defined by Assumption A (let $A = 1$ and $\Psi = \psi_{c,\beta}^{\text{IMQ}}$). On the other hand, our next result, proved in Appendix B, shows that **tilted base kernels** with $A$ increasing sufficiently quickly also control convergence.

**Theorem 3.2** (Tilted KSDs detect non-convergence). Suppose that $P \in \mathcal{P}$, Assumption A holds, $1/A \in L^2$, and $H(u) := \sup_{\omega \in \mathbb{R}^D} e^{-\|\omega\|^2_2/(2a^2)} / \Psi(\omega)$ is finite for all $u > 0$. Then for any sequence of probability measures $\mu_N \Rightarrow P$ if $\mu_N \Rightarrow P$ then $\mu_N \Rightarrow P$.

**Example 3.2** (Tilted hyperbolic secant kernel). The hyperbolic secant (sech) function is $\text{sech}(u) := 2/(e^u + e^{-u})$. For $x \in \mathbb{R}^D$ and $a > 0$, define the sech kernel $\psi_a^{\text{sech}}(x) := \prod_{d=1}^D \text{sech}(\sqrt{2/a} x_d)$. Since $\psi_a^{\text{sech}}(\omega) = \psi_{1/a}^{\text{sech}}(\omega)/a^D$, KSD$_k$ from Theorem 3.2 detects non-convergence when $\Psi = \psi_a^{\text{sech}}$ and $A^{-1} \in L^2$. Valid tilting functions include $A(x) = \prod_{d=1}^D e^{(1 + \|x\|^2_d)/4a^2}$ for any $c > 0$ and $F(x) = (c^2 + \|x\|^2_2)^b$ for any $b > D/4$ (to ensure $A^{-1} \in L^2$).

With our appropriate reference KSDs in hand, we will now design upper bounding $\Phi$-SDs. To accomplish this we will have $\Phi$ mimic the form of the base kernels in Assumption A:

**Assumption B.** Assumption A holds and $\Phi(x, z) = A_N(x)F(x - z)$, where $F \in C^1$ is positive, and there exists a norm $\|\cdot\|$ and constants $s, C > 0$ such that

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

In addition, there exist a constant $\xi \in (0, 1]$ and continuous, non-increasing function $f$ such that $\xi f(\|x\|) \leq F(x) \leq f(\|x\|)$.

Assumption B requires a minimal amount of regularity from $F$, essentially that $F$ is sufficiently smooth and behave as if it is a function only of the norm of its argument. A conceptually straightforward choice would be to set $F = \mathcal{F}^{-1}(\hat{\Psi}^{1/2})$—that is, to be the square root kernel of $\Psi$. We would then have that $\Psi(x - y) = \int F(x - z)F(y - z) \, dz$, so in particular $\Phi$-$\text{SD}_{k,2} = \text{KSD}_k$. Since the exact square-root kernel of a base kernel can be difficult to compute in practice, we require only that $F$ be a suitable approximation to the square root kernel of $\Psi$:

**Assumption C.** Assumption B holds, and there exists a **smoothness parameter** $\bar{\lambda} \in (1/2, 1]$ such that if $\lambda \in (1/2, \bar{\lambda})$, then $F/\psi^{\lambda/2} \in L^2$.

Requiring that $F/\psi^{\lambda/2} \in L^2$ is equivalent to requiring that $F$ belongs to the reproducing kernel Hilbert space $K_\lambda$ induced by the kernel $\mathcal{F}^{-1}(\hat{\Psi}^{\lambda})$. The smoothness of the functions in $K_\lambda$ increases as $\lambda$ increases. Hence $\bar{\lambda}$ quantifies the smoothness of $F$ relative to $\Psi$.

Finally, we would like an assurance that the $\Phi$-$\text{SD}$ detects convergence—that is, $\Phi$-$\text{SD}_{k,2}(Q_N, P)$ converges to $P$ whenever $Q_N$ converges to $P$ in a suitable metric. The following result, proved in Appendix C, provides such a guarantee for both the $\Phi$-$\text{SD}$ and the R$\Phi$-$\text{SD}$.

**Proposition 3.3.** Suppose Assumption B holds with $F \in L^1$, $1/A$ bounded, $x \mapsto x/A(x)$ Lipschitz, and $\mathbb{E}_P[A(Z)||Z||^2_2 < \infty$. If the tilted Wasserstein distance

$$W_{\lambda_N}(Q_N, P) := \sup_{h \in H} |Q_N(A_N h) - P(A_N h)| (H := \{h : \|\nabla h\|_2 \leq 1, \forall x \in \mathbb{R}^D\})$$

\(^2\)We say $P$ satisfies distantly dissipativity [8, 12] if $\kappa_0 := \lim_{r \to \infty} \kappa(r) > 0$ for $\kappa(r) = \inf(-2(\nabla \log p(x) - \nabla \log p(y), x - y)/\|x - y\|^2_2 : \|x - y\|_2 = r)$.\]
converges to zero, then \( \PhiSD_{\Phi,r}(Q_N, P) \to 0 \) and \( R\PhiSD_{\Phi,r,\nu_N,M_N}(Q_N, P) \overset{L_p}{\to} 0 \) for any choices of \( r \in [1, 2, \nu_N, M_N \geq 1}. 

Remark 3.4. When \( A \) is constant, \( W_{A_N} \) is the familiar 1-Wasserstein distance.

### 3.2 Selecting an importance sampling distribution \( \nu \)

Our next goal is to select an R\( \Phi \)SD proposal distribution \( \nu \) for which the R\( \Phi \)SD is close to its reference \( \Phi \)SD even when the importance sample size \( M \) is small. Our strategy is to choose \( \nu \) so that the second moment of each R\( \Phi \)SD feature, \( w_d(Z, Q_N) := |(Q_N T_d \Phi)(Z)|^\nu / \nu(Z) \), is bounded by a power of its mean:

**Definition 3.5** ((\( C, \gamma \)) second moments). Fix a target distribution \( P \). For \( Z \sim \nu, d \in [D], \) and \( N \geq 1 \), let \( Y_{N,d} := w_d(Z, Q_N) \). If for some \( C > 0 \) and \( \gamma \in [0, 2] \) we have \( \mathbb{E}[Y_{N,d}] \leq C \mathbb{E}[Y_{N,d}]^{2-\gamma} \) for all \( d \in [D] \) and \( N \geq 1 \), then we say \((\Phi, r, \nu)\) yields \((C, \gamma)\) second moments for \( P \) and \( Q_N \).

The next proposition, proved in Appendix D, demonstrates the value of this second moment property.

**Proposition 3.6.** Suppose \((\Phi, r, \nu)\) yields \((C, \gamma)\) second moments for \( P \) and \( Q_N \). If \( M \geq 2C \mathbb{E}[Y_{N,d}]^{-\gamma} \log(D/\delta)/c^2 \) for all \( d \in [D] \), then, with probability at least \( 1 - \delta \),

\[
R\PhiSD_{\Phi,r,M}(Q_N, P) \geq (1 - \epsilon)^1/r \PhiSD_{\Phi,r}(Q_N, P).
\]

Under the further assumptions of Proposition 3.1, if the reference KSD\(_k\)(\( Q_N, P \)) \( \geq N^{-1/2} \), then a sample size \( M \geq N^{\gamma r/2}C ||\rho||_{L_1}^{\gamma r/2} \log(D/\delta)/c^2 \) suffices to have, with probability at least \( 1 - \delta \),

\[
||\rho||_{L_1}^{1/2} R\PhiSD_{\Phi,r,M}(Q_N, P) \geq (1 - \epsilon)^{1/r} \text{KSD}_k(Q_N, P).
\]

Notably, a smaller \( r \) leads to substantial gains in the sample complexity \( M = \Omega(N^{\gamma r/2}) \). For example, if \( r = 1 \), it suffices to choose \( M = \Omega(N^{1/2}) \) whenever the weight function \( w_d \) is bounded (so that \( \gamma = 1 \)); in contrast, existing analyses of random Fourier features \([15, 22, 25, 26, 30]\) require \( M = \Omega(N) \) to achieve the same error rates. We will ultimately show how to select \( \nu \) so that \( \gamma \) is arbitrarily close to 0. First, we provide simple conditions and a choice for \( \nu \) which guarantee \((C, 1)\) second moments.

**Proposition 3.7.** Assume that \( P \in \mathcal{P} \), Assumptions \( A \) and \( B \) hold with \( s = 0 \), and there exists a constant \( C' > 0 \) such that for all \( N \geq 1 \), \( Q_N[1 + \|\cdot\|]A_N \leq C' \). If \( \nu(z) \propto Q_N[1 + \|\cdot\|]\Phi(\cdot, z) \), then for any \( r \geq 1 \), \((\Phi, r, \nu)\) yields \((C, 1)\) second moments for \( P \) and \( Q_N \).

Proposition 3.7, which is proved in Appendix E, is based on showing that the weight function \( w_d(z, Q_N) \) is uniformly bounded. In order to obtain \((C, \gamma)\) moments for \( \gamma < 1 \), we will choose \( \nu \) such that \( w_d(z, Q_N) \) decays sufficiently quickly as \( \|z\| \to \infty \). We achieve this by choosing an over-dispersed \( \nu \)—that is, we choose \( \nu \) with heavy tails compared to \( F \). We also require two integrability conditions involving the Fourier transforms of \( \Phi \) and \( F \).

**Assumption D.** Assumptions \( A \) and \( B \) hold, \( \omega^2_0 \Phi^{1/2}(\omega) \in L^1 \), and for \( t = r/(2 - r) \), \( \hat{\Psi}/\hat{F}^2 \in L^t \).

The \( L^1 \) condition is an easily satisfied technical condition while the \( L^t \) condition ensures that the KSD-\( \Phi \)SD inequality (2) applies to our chosen \( \Phi \)SD.

**Theorem 3.8.** Assume that \( P \in \mathcal{P} \), Assumptions \( A \) to \( D \) hold, and there exists \( C > 0 \) such that,

\[
Q_N[1 + \|\cdot\| + \|\cdot - m_N\|^\alpha]A_N/F(\cdot - m_N) \leq C \quad \text{for all} \quad N \geq 1. \tag{3}
\]

Then there is a constant \( b \in [0, 1] \) such that the following holds. For any \( \xi \in (0, 1 - b), c > 0, \) and \( \alpha > 2(1 - \xi) \), if \( \nu(z) \geq c \Psi(z - m_N)^{\xi r} \), then there exists a constant \( C_\alpha > 0 \) such that \((\Phi, r, \nu)\) yields \((C_\alpha, \gamma_\alpha)\) second moments for \( P \) and \( Q_N \), where \( \gamma_\alpha := \alpha + (2 - \alpha)\xi/(2 - b - \xi) \).

Theorem 3.8 suggests a strategy for improving the importance sample growth rate \( \gamma \) of an R\( \Phi \)SD:

1. increase the smoothness \( \lambda \) of \( F \) and decrease the over-dispersion parameter \( \xi \) of \( \nu \).

---

1 Note that KSD\(_k\)(\( Q_N, P \)) = \( \Omega_P(N^{-1/2}) \) whenever the sample points \( x_1, \ldots, x_N \) are drawn i.i.d. from a distribution \( \mu \), since the scaled V-statistic N KSD\(_k^2\)(\( Q_N, P \)) diverges when \( \nu \neq P \) and converges in distribution to a non-zero limit when \( \nu = P \) [23, Thm. 32]. Moreover, working in a hypothesis testing framework of shrinking alternatives, Gretton et al. [13, Thm. 13] showed that KSD\(_k\)(\( Q_N, P \)) = \( \Theta(N^{-1/2}) \) was the smallest local departure distinguishable by an asymptotic KSD test.
3.3 Example RΦSDs

In our experiments, we will consider two RΦSDs that determine convergence by Propositions 3.1 and 3.3 and that yield \((C, \gamma)\) second moments for any \(\gamma \in (0, 1)\) using Theorem 3.8.

Example 3.3 \((L^2)\) tilted hyperbolic secant RΦSD. Mimicking the construction of the hyperbolic secant kernel in Example 3.2 and following the intuition that \(F\) should behave like the square root of \(\hat{\Psi}\), we choose \(F = \Psi_{20a}^{\text{sech}}\). As shown in Appendix I, if we choose \(r = 2\) and \(\nu(z) \propto \Psi_{40a}^{\text{sech}}(z - m_N)\) we can verify all the assumptions necessary for Theorem 3.8 to hold. Moreover, the theorem holds for any \(b > 0\) and hence any \(\xi \in (0, 1)\) may be chosen. Note that \(\nu\) can be sampled from efficiently using the inverse CDF method.

Example 3.4 \((L^2)\) IMQ RΦSD. We can also parallel the construction of the reference IMQ kernel \(k(x, y) = \hat{q}_{c,\beta}^\text{IMQ}(x - y)\) from Example 3.1, where \(c > 0\) and \(\beta \in [-D/2, 0)\). (Recall we have \(A = 1\) in Assumption A.) In order to construct a corresponding RΦSD we must choose the constant \(\tilde{\lambda} \in (1/2, 1)\) that will appear in Assumption C and \(\tilde{\xi} \in (0, 1/2)\), the minimum \(\xi\) we will be able to choose when constructing \(\nu\). We show in Appendix J that if we choose \(F = \hat{q}_{c',\beta'}^\text{IMQ}\), then Assumptions A to D hold when \(c' = \tilde{\lambda}c/2, \beta' \in [-D/(2\tilde{\xi}), -\beta/(2\tilde{\xi}) - D/(2\tilde{\xi})]\), \(r = -D/(2\beta'\tilde{\xi})\), \(\xi \in (\tilde{\xi}, 1)\), and \(\nu(z) \propto \Psi_{c',\beta'}^{\text{IMQ}}(z - m_N)^{\xi r}\). A particularly simple setting is given by \(\beta' = -D/(2\tilde{\xi})\), which yields \(r = 1\). Note that \(\nu\) can be sampled from efficiently since it is a multivariate \(t\)-distribution.

In the future it would be interesting to construct other RΦSDs. We can recommend the following fairly default procedure for choosing an RΦSD based on a reference KSD admitting the form in Assumption A. (1) Choose any \(\gamma > 0\), and set \(\alpha = \gamma/3, \lambda = 1 - \alpha/2, \text{and} \xi = 4\alpha/(2 + \alpha)\). These are the settings we will use in our experiments. It may be possible to initially skip this step and reason about general choices of \(\gamma, \xi, \text{and} \lambda\). (2) Pick any \(F\) that satisfies \(\hat{\Psi}/\hat{\Psi}^{2} \in L^2\) for some \(\lambda \in (1/2, \lambda)\) (that is, Assumption C holds) while also satisfying \(\hat{\Psi}/\hat{\Psi}^{2} \in L^t\) for some \(t \in [1, \infty]\). The selection of \(t\) induces a choice of \(r\) via Assumption D. A simple choice for \(F\) is \(\mathcal{F}^{-1}\hat{\Psi}_\lambda\). (3) Check if Assumption B holds (it usually does if \(F\) decays no faster than a Gaussian); if it does not, a slightly different choice of \(F\) should be made. (4) Choose \(\nu(z) \propto \Psi(z - m_N)^{\xi r}\).

4 Goodness-of-fit testing with RΦSDs

We now detail additional properties of RΦSDs relevant to testing goodness of fit. In goodness-of-fit testing, the sample points \((X_n)_{n=1}^N\) underlying \(Q_N\) are assumed to be drawn i.i.d. from a distribution \(\mu\), and we wish to use the test statistic \(F_{r,N} := \PhiSD_{\Phi,\nu,M}(Q_N; P)\) to determine whether the null hypothesis \(H_0: P = \mu\) or alternative hypothesis \(H_1: P \neq \mu\) holds. For this end, we will restrict our focus to real analytic \(\Phi\) and strictly positive analytic \(\nu\), as by Chwialkowski et al. [5, Prop. 2 and Lemmas 1-3], with probability \(1, \Phi = \mu \iff \PhiSD_{\Phi,\nu,M}(\mu, P) = 0\) when these properties hold. Thus, analytic RΦSDs do not suffer from the shortcoming of RFFs—which are unable to distinguish between infinitely many distributions with high probability [5].

It remains to estimate the distribution of the test statistic \(F_{r,N}\) under the null hypothesis and to verify that the power of a test based on this distribution approaches 1 as \(N \to \infty\). To state our result, we assume that \(M\) is fixed. Let \(\xi_{r,N,dm}(x) := (T_d\Phi)(x, Z_{N,m})/(M \nu(Z_{N,m}))^{1/r}\) for \(r \in [1, 2]\), where
(a) Step size selection using RΦSDs and quadratic-time KSD baseline. With $M \geq 10$, each quality measure selects a step size of $\varepsilon = .01$ or $.005$.

(b) SGLD sample points with equidensity contours of $p$ overlaid. The samples produced by SGLD with $\varepsilon = .01$ or $.005$ are noticeably better than those produced using smaller or large step sizes.

Figure 2: Hyperparameter selection for stochastic gradient Langevin dynamics (SGLD)

(a) Step size selection using RΦSDs and quadratic-time KSD baseline. With $M \geq 10$, each quality measure selects a step size of $\varepsilon = .01$ or $.005$.

(b) SGLD sample points with equidensity contours of $p$ overlaid. The samples produced by SGLD with $\varepsilon = .01$ or $.005$ are noticeably better than those produced using smaller or large step sizes.

Figure 2: Hyperparameter selection for stochastic gradient Langevin dynamics (SGLD)

Z_{N,m} \overset{\text{indep}}{\sim} \nu_N$, so that $\xi_{r,N}(x) \in \mathbb{R}^{D_M}$. The following result, proved in Appendix K, provides the basis for our testing guarantees.

**Proposition 4.1** (Asymptotic distribution of RΦSD). Assume $\Sigma_{r,N} := \text{Cov}_{P}(\xi_{r,N})$ is finite for all $N$ and $\Sigma_r := \lim_{N \to \infty} \Sigma_{r,N}$ exists. Let $\zeta \sim \mathcal{N}(0, \Sigma_r)$. Then as $N \to \infty$: (1) under $H_0 : P = \mu$, $NF_{r,N} \overset{p}{\to} \sum_{d=1}^{D} (\sum_{m=1}^{M} |\zeta_{dm}|^r)^{2/r}$ and (2) under $H_1 : P \neq \mu$, $NF_{r,N} \overset{p}{\to} \infty$.

**Remark 4.2.** The condition $\Sigma_r := \lim_{N \to \infty} \Sigma_{r,N}$ holds if $\nu_N = \nu_0((\cdot \sim m_N))$ for a distribution $\nu_0$.

Our second asymptotic result provides a roadmap for using RΦSDs for hypothesis testing and is similar in spirit to Theorem 3 from Jitkrittum et al. [16]. In particular, it furnishes an asymptotic null distribution and establishes asymptotically full power.

**Theorem 4.3** (Goodness of fit testing with RΦSD). Let $\hat{\mu} := N^{-1} \sum_{n=1}^{N} \xi_{r,N}(X'_n)$ and $\hat{\Sigma} := N^{-1} \sum_{n=1}^{N} \xi_{r,N}(X'_n)\xi_{r,N}(X'_n)^T - \hat{\mu}\hat{\mu}^T$ with either $X'_n = X_n$ or $X'_n \overset{iid}{\sim} P$. Suppose for the test $NF_{r,N}$, the test threshold $\tau_\alpha$ is set to the $(1 - \alpha)$-quantile of the distribution of $\sum_{d=1}^{D} (\sum_{m=1}^{M} |\zeta_{dm}|^r)^{2/r}$, where $\zeta \sim \mathcal{N}(0, \hat{\Sigma})$. Then, under $H_0 : P = \mu$, asymptotically the false positive rate is $\alpha$. Under $H_1 : P \neq \mu$, the test power $\mathbb{P}_{H_1}(NF_{r,N} > \tau_\alpha) \to 1$ as $N \to \infty$.

5 Experiments

We now investigate the importance-sample and computational efficiency of our proposed RΦSDs and evaluate their benefits in MCMC hyperparameter selection and goodness-of-fit testing.\footnote{See https://bitbucket.org/jhhuggins/random-feature-stein-discrepancies for our code.} In our experiments, we considered the RΦSDs described in Examples 3.3 and 3.4: the tilted sech kernel using $r = 2$ and $A(x) = \prod_{d=1}^{D} e^{a_d \sqrt{1+x_d^2}}$ (L2 SechExp) and the inverse multiquadric kernel using $r = 1$ (L1 IMQ). We selected kernel parameters as follows. First we chose a target $\gamma$ and then selected $\lambda$, $\alpha$, and $\xi$ in accordance with the theory of Section 3 so that $(\Phi, r, \nu)$ yielded $(C_{\gamma, \gamma})$.
second moments. In particular, we chose $\alpha = \gamma/3$, $\bar{\lambda} = 1 - \alpha/2$, and $\xi = 4\alpha/(2 + \alpha)$. Except for the importance sample efficiency experiments, where we varied $\gamma$ explicitly, all experiments used $\gamma = 1/4$. Let $\text{med}_u$ denote the estimated median of the distance between data points under the $u$-norm, where the estimate is based on using a small subsample of the full dataset. For L2 SechExp, we took $a^{-1} = \sqrt{2\pi \text{med}_1}$, except in the sample quality experiments where we set $a^{-1} = \sqrt{2\pi}$. Finding hyperparameter settings for the L1 IMQ that were stable across dimension and appropriately controlled the size for goodness-of-fit testing required some care. However, we can offer some basic guidelines. We recommend choosing $\xi = \xi D/(D + df)$, which ensures $\nu$ has $df$ degrees of freedom. We specifically suggest using $df \in [0.5, 3]$ so that $\nu$ is heavy-tailed no matter the dimension. For most experiments we took $\beta = -1/2, c = 4 \text{med}_2$, and $df = 0.5$. The exceptions were in the sample quality experiments, where we set $c = 1$, and the restricted Boltzmann machine testing experiment, where we set $c = 10 \text{med}_2$ and $df = 2.5$. For goodness-of-fit testing, we expect appropriate choices for $c$ and $df$ will depend on the properties of the null distribution.

**Importance sample efficiency** To validate the importance sample efficiency theory from Sections 3.2 and 3.3, we calculated $P[R_{\Phi SD} > \Phi SD]/4$ as the importance sample size $M$ was increased. We considered choices of the parameters for L2 SechExp and L1 IMQ that produced $(C_\gamma, \gamma)$ second moments for varying choices of $\gamma$. The results, shown in Figs. 1a and 1b, indicate greater sample efficiency for L1 IMQ than L2 SechExp. L1 IMQ is also more robust to the choice of $\gamma$. Fig. 1c, which plots the values of $M$ necessary to achieve $\text{stdev}(R_{\Phi SD})/\Phi SD < 1/2$, corroborates the greater sample efficiency of L1 IMQ.

**Computational complexity** We compared the computational complexity of the RΦSDs (with $M = 10$) to that of the IMQ KSD. We generated datasets of dimension $D = 10$ with the sample size $N$ ranging from 500 to 5000. As seen in Fig. 3, even for moderate dataset sizes, the RΦSDs are computed orders of magnitude faster than the KSD. Other RΦSDs like FSSD and RFF obtain similar speed-ups; however, we will see the power benefits of the L1 IMQ and L2 SechExp RΦSDs below.

**Approximate MCMC hyperparameter selection** We follow the stochastic gradient Langevin dynamics [SGLD, 28] hyperparameter selection setup from Gorham and Mackey [10, Section 5.3]. SGLD with constant step size $\varepsilon$ is a biased MCMC algorithm that approximates the overdamped Langevin diffusion. No Metropolis-Hastings correction is used, and an unbiased estimate of the score function from a data subsample is calculated at each iteration. There is a bias-variance tradeoff in the choice of step size parameter: the stationary distribution of SGLD deviates more from its target as $\varepsilon$ grows, but as $\varepsilon$ gets smaller the mixing speed of SGLD decreases. Hence, an appropriate choice of $\varepsilon$ is critical for accurate posterior inference. We target the bimodal Gaussian mixture model (GMM) posterior of Welling and Teh [28] and compare the step size selection made by the two RΦSDs to that of IMQ KSD [11] when $N = 1000$. Fig. 2a shows that L1 IMQ and L2 SechExp agree with IMQ KSD (selecting $\varepsilon = .005$) even with just $M = 10$ importance samples. L1 IMQ continues to select $\varepsilon = .005$ while L2 SechExp settles on $\varepsilon = .01$, although the value for $\varepsilon = .005$ is only slightly larger. Fig. 2b compares the choices of $\varepsilon = .005$ and .01 to smaller and larger values of $\varepsilon$. The values of $M$ considered all represent substantial reductions in computation as the RΦSD replaces the $DN(N + 1)/2$ KSD kernel evaluations of the form $((T_d \otimes T_d)k)(x_n, x_m)$ with only $DNM$ feature function evaluations of the form $(T_d \Phi)(x_n, z_m)$.

**Goodness-of-fit testing** Finally, we investigated the performance of RΦSDs for goodness-of-fit testing. In our first two experiments we used a standard multivariate Gaussian $p(x) = \mathcal{N}(x | 0, I)$ as the null distribution while varying the dimension of the data. We explored the power of RΦSD-based tests compared to FSSD [16] (using the default settings in their code), RFF [22] (Gaussian and Cauchy kernels with bandwidth $= \text{med}_2$), and KSD-based tests [6, 11, 19] (Gaussian kernel with bandwidth
Figure 4: Quadratic-time KSD and linear-time RΦSD, FSSD, and RFF goodness-of-fit tests with $M = 10$ importance sample points (see Section 5 for more details). All experiments used $N = 1000$ except the multivariate $t$, which used $N = 2000$. (a) Size of tests for Gaussian null. (b, c, d) Power of tests. Both RΦSDs offer competitive performance.

6 Discussion and related work

In this paper, we have introduced feature Stein discrepancies, a family of computable Stein discrepancies that can be cheaply approximated using importance sampling. Our stochastic approximations, random feature Stein discrepancies (RΦSDs), combine the computational benefits of linear-time discrepancy measures with the convergence-determining properties of quadratic-time Stein discrepancies. We validated the benefits of RΦSDs on two applications where kernel Stein discrepancies have shown excellent performance: measuring sample quality and goodness-of-fit testing. Empirically, the L1 IMQ RΦSD performed particularly well: it outperformed existing linear-time KSD approximations in high dimensions and performed as well or better than the state-of-the-art quadratic-time KSDs.

RΦSDs could also be used as drop-in replacements for KSDs in applications to Monte Carlo variance reduction with control functionals [21], probabilistic inference using Stein variational gradient descent [18], and kernel quadrature [1, 2]. Moreover, the underlying principle used to generalize the KSD could also be used to develop fast alternatives to maximum mean discrepancies in two-sample testing applications [13]. Finally, while we focused on the Langevin Stein operator, our development is compatible with any Stein operator, including diffusion Stein operators [12].
Acknowledgments

Part of this work was done while JHH was a research intern at MSR New England.

References

[1] F. Bach. On the Equivalence between Kernel Quadrature Rules and Random Feature Expansions. *Journal of Machine Learning Research*, 18:1–38, 2017.

[2] F.-X. Briol, C. J. Oates, J. Cockayne, W. Y. Chen, and M. A. Girolami. On the Sampling Problem for Kernel Quadrature. In *International Conference on Machine Learning*, 2017.

[3] C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

[4] F. Chung and L. Lu. *Complex Graphs and Networks*, volume 107. American Mathematical Society, Providence, Rhode Island, 2006.

[5] K. Chwialkowski, A. Ramdas, D. Sejdinovic, and A. Gretton. Fast Two-Sample Testing with Analytic Representations of Probability Measures. In *Advances in Neural Information Processing Systems*, 2015.

[6] K. Chwialkowski, H. Strathmann, and A. Gretton. A Kernel Test of Goodness of Fit. In *International Conference on Machine Learning*, 2016.

[7] DLMF. *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.1.2 of 2021-06-15. URL http://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[8] A. Eberle. Reflection couplings and contraction rates for diffusions. *Probability Theory and Related Fields*, 166(3-4):851–886, 2016.

[9] C. J. Geyer. Markov Chain Monte Carlo Maximum Likelihood. In *Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface*, pages 156–163, 1991.

[10] J. Gorham and L. Mackey. Measuring Sample Quality with Stein’s Method. In *Advances in Neural Information Processing Systems*, 2015.

[11] J. Gorham and L. Mackey. Measuring Sample Quality with Kernels. In *International Conference on Machine Learning*, 2017.

[12] J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring Sample Quality with Diffusions. *arXiv.org*, Nov. 2016, 1611.06972v3.

[13] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[14] R. Herb and P. Sally Jr. The Plancherel formula, the Plancherel theorem, and the Fourier transform of orbital integrals. In *Representation Theory and Mathematical Physics: Conference in Honor of Gregg Zuckerman's 60th Birthday, October 24–27, 2009, Yale University*, volume 557, page 1. American Mathematical Soc., 2011.

[15] J. Honorio and Y.-J. Li. The Error Probability of Random Fourier Features is Dimensionality Independent. *arXiv.org*, Oct. 2017, 1710.09953v1.

[16] W. Jitkrittum, W. Xu, Z. Szabó, K. Fukumizu, and A. Gretton. A Linear-Time Kernel Goodness-of-Fit Test. In *Advances in Neural Information Processing Systems*, 2017.

[17] Q. Liu and J. D. Lee. Black-box Importance Sampling. In *International Conference on Artificial Intelligence and Statistics*, 2017.

[18] Q. Liu and D. Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In *Advances in Neural Information Processing Systems*, 2016.
[19] Q. Liu, J. D. Lee, and M. I. Jordan. A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation. In International Conference on Machine Learning, 2016.

[20] A. Müller. Integral probability metrics and their generating classes of functions. Ann. Appl. Probab., 29(2):pp. 429–443, 1997.

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

[22] A. Rahimi and B. Recht. Random features for large-scale kernel machines. In Advances in Neural Information Processing Systems, 2007.

[23] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. The Annals of Statistics, 41(5):2263–2291, 2013.

[24] R. J. Serfling. Approximation Theorems of Mathematical Statistics. John Wiley & Sons, New York, 1980.

[25] B. K. Sriperumbudur and Z. Szabó. Optimal rates for random Fourier features. In Advances in Neural Information Processing Systems, pages 1144–1152, 2015.

[26] D. J. Sutherland and J. Schneider. On the error of random Fourier features. In Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence, pages 862–871, 2015.

[27] D. Wang and Q. Liu. Learning to Draw Samples - With Application to Amortized MLE for Generative Adversarial Learning. arXiv, stat.ML, 2016.

[28] M. Welling and Y. W. Teh. Bayesian Learning via Stochastic Gradient Langevin Dynamics. In International Conference on Machine Learning, 2011.

[29] H. Wendland. Scattered Data Approximation. Cambridge University Press, New York, NY, 2005.

[30] J. Zhao and D. Meng. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. Neural computation, 27(6):1345–1372, 2015.
Appendix for “Random Feature Stein Discrepancies”

A Proof of Proposition 3.1: KSD-ΦSD inequality

We apply the generalized Hölder’s inequality and the Babenko-Beckner inequality in turn to find
\[ \text{KSD}^2_2(Q_N, P) = \sum_{d=1}^D \int |\mathcal{F}(Q_N(T_\omega \Phi))| \Phi(\omega) \omega \mathrm{d}\omega \leq \|\mathcal{F}(Q_N(T_\omega \Phi))\|^2_{L^2} \cdot \sum_{d=1}^D \|Q_N(T_\omega \Phi)\|^2_{L^2} \cdot \Phi_{\Phi SD}^2(r, Q_N, P), \]
where
\[ t = \frac{c}{\sqrt{r}} \quad \text{and} \quad c_{r,d} := (r^{1/r}/s^{1/s})^{d/2} \leq 1 \quad \text{for} \quad s = r/(r-1). \]

B Proof of Theorem 3.2: Tilted KSDs detect non-convergence

For any vector-valued function \( f \), let \( M_1(f) = \sup_{x,y : \|x - y\|_2 \leq 1} \|f(x) - f(y)\|_2 \). The result will follow from the following theorem which provides an upper bound on the bounded Lipschitz metric \( d_{BL_{\|\cdot\|_2}}(\mu, P) \) in terms of the KSD and properties of \( A \) and \( \Psi \).

**Theorem B.1** (Tilted KSD lower bound). Suppose \( P \in \mathcal{P} \) and \( k(x, y) = A(x)\Psi(x - y)A(y) \) for \( \Psi \in C^2 \) and \( A \in C^1 \) with \( A > 0 \) and \( \nabla \log A \) bounded and Lipschitz. Then there exists a constant \( M_P \) such that, for all \( \epsilon > 0 \) and all probability measures \( \mu \),
\[ d_{BL_{\|\cdot\|_2}}(\mu, P) \leq \epsilon + C \text{KSD}_k(\mu, P), \]
where
\[ C := (2\pi)^{-d/4} \|1/A\|_L^2 \cdot M_P H(\|G\|_2 B(G)|1 + M_1(\log A) + M_P M_1(b + \nabla \log A))\epsilon^{-1})^{1/2}, \]
\[ H(t) := \sup_{\rho \in \mathbb{R}_+} e^{-\|\rho\|^2_2/(2t^2)} / \psi(\rho), \quad G \text{ is a standard Gaussian vector, and} \quad B(y) := \sup_{x \in \mathbb{R}^{d \times d}} A(x)/A(x + uy). \]

**Remarks** By bounding \( H \) and optimizing over \( \epsilon \), one can derive rates of convergence in \( d_{BL_{\|\cdot\|_2}} \).

Thm. 5 and Sec. 4.2 of Gorham et al. [12] provide an explicit value for the Stein factor \( M_P \).

Let \( A_\mu(x) = A(x - EX_{\sim \mu}[X]) \). Since \( \|1/A\|_L^2 = \|1/A_\mu\|_L^2 \), \( M_1(\log A_\mu) \leq M_1(\log A), M_1(\nabla \log A_\mu) \leq M_1(\nabla \log A) \), and \( \sup_{\rho \in \mathbb{R}_+} e^{-\|\rho\|^2_2/(2t^2)} / \psi(\rho) \), the exact conclusion of Theorem B.1 also holds when \( k(x, y) = A_\mu(x)\Psi(x - y)A_\mu(y) \). Moreover, since \( \log A \) is Lipschitz, \( B(y) \leq \epsilon \|\rho\|_L^2 \), \( \epsilon \|\rho\|_L^2 \), \( \mathbb{E}[\|G\|_2 B(G)] \) is finite. Now suppose \( \text{KSD}_k(\mu_N, P) \to 0 \) for a sequence of probability measures \( (\mu_N)_{N \geq 1} \). For any \( \epsilon > 0 \), \( \limsup_n d_{BL_{\|\cdot\|_2}}(\mu_N, P) \leq \epsilon \), since \( H(t) \) is finite for all \( t > 0 \). Hence, \( d_{BL_{\|\cdot\|_2}}(\mu_N, P) \to 0 \), and, as \( d_{BL_{\|\cdot\|_2}} \) metrizes weak convergence, \( \mu_N \Rightarrow P \).

B.1 Proof of Theorem B.1: Tilted KSD lower bound

Our proof parallels that of [11, Thm. 13]. Fix any \( h \in BL_{\|\cdot\|_2} \). Since \( A \in C^1 \) is positive, Thm. 5 and Sec. 4.2 of Gorham et al. [12] imply that there exists a \( g \in C^1 \) which solves the Stein equation \( T_P(Ag) = h - \mathbb{E}_P[h(Z)] \) and satisfies \( M_0(Ag) \leq M_P \) for \( M_P \) a constant independent of \( A, h, \) and \( g \). Since \( 1/A \in L^2 \), we have \( \|g\|_L^2 \leq M_P \|1/A\|_L^2 \).

Since \( \nabla \log A \) is bounded, \( A(x) \leq \exp(\|\rho\|_L^2) \) for some \( \gamma \). Moreover, any measure in \( \mathcal{P} \) is sub-Gaussian, so \( P \) has finite exponential moments. Hence, since \( A \) is also positive, we may define the tilted probability measure \( P_A \) with density proportional to \( Ap \). The identity \( T_{P}(Ag) = AT_{P_A} g \) implies that
\[ M_0(A\nabla T_{P_A} g) = M_0(A\nabla T_{P}(Ag) - T_{P}(Ag)\nabla \log A) \leq 1 + M_1(\log A). \]
Since \( b \) and \( \nabla \log A \) are Lipschitz, we may apply the following lemma, proved in Appendix B.2 to deduce that there is a function \( g_\epsilon \in K_{d_1}^c \) for \( k_1(x, y) := \Psi(x - y) \) such that \( |(T_{P}(Ag_\epsilon))(x) - (T_{P}(Ag))(x)| = A(x)|T_{P_A} g_\epsilon(x) - (T_{P_A} g)(x)| \leq \epsilon \) for all \( x \) with norm
\[ \|g_\epsilon\|_K_{d_1}^c \leq (2\pi)^{-d/4} H(\|G\|_2 B(G)|1 + M_1(\log A) + M_P M_1(b + \nabla \log A))\epsilon^{-1})^{1/2} \|1/A\|_L^2 M_P. \]
Lemma B.2 (Stein approximations with finite RKHS norm). Consider a function $A : \mathbb{R}^d \to \mathbb{R}$ satisfying $B(y) := \sup_{x \in \mathbb{R}^d, u \in [0, 1]} A(x) / A(x + uy)$). Suppose $g : \mathbb{R}^d \to \mathbb{R}^d$ is in $L^2 \cap C^1$. If $P$ has Lipschitz log density, and $k(x, y) = \Psi(x-y)$ for $\Psi \in C^2$ with generalized Fourier transform $\hat{\Psi}$, then for every $\epsilon \in (0, 1)$, there is a function $g_{\epsilon} : \mathbb{R}^d \to \mathbb{R}^d$ such that $|(T_{\epsilon} g_{\epsilon})(x) - (T_{\epsilon} g)(x)| \leq \epsilon / A(x)$ for all $x \in \mathbb{R}^d$ and

$$\|g_{\epsilon}\|_{K^d} \leq (2\pi)^{-d/2} H \left( \|G\|_2^2 \right) \left( M_0(A) / M_0(A) M_1(b) \right) \epsilon^{-1} \|\| \|_L^2,$$

where $H(t) := \sup_{x \in \mathbb{R}^d} e^{-\|x\|^2/(2t^2)} / \hat{\Psi}(x)$ and $G$ is a standard Gaussian vector.

Since $\|A g_{\epsilon}\|_{K^d} = \|g_{\epsilon}\|_{K^d}$, the triangle inequality and the definition of the KSD now yield

$$\left| \left| \mathbb{E}_{\mu}[h(X)] - \mathbb{E}_P[h(Z)] \right| \right| = \left| \left| \mathbb{E}_\mu[(T_{\epsilon} (Ag_{\epsilon}))(X)] \right| \right| \leq \|\| \| \|_L^2$$

The advertised conclusion follows by applying the bound (4) and taking the supremum over all $h \in BL_\mu$.

B.2 Proof of Lemma B.2: Stein approximations with finite RKHS norm

Assume $M_0(A) A \mathbb{V}T_{\epsilon} g) + M_0(A g) < \infty$, as otherwise the claim is vacuous. Our proof parallels that of Gorham and Mackey [11, Lem. 12]. Let $Y$ denote a standard Gaussian vector with density $\rho$. For each $\delta \in (0, 1)$, we define $\rho_\delta(x) = \delta^{-d/2} \rho(x / \delta)$, and for any function $f$ we write $f_\delta(x) \equiv \mathbb{E}[f(x + \delta Y)]$. Under our assumptions on $h = T_{\epsilon} g$ and $B$, the mean value theorem and Cauchy-Schwarz imply that for each $x \in \mathbb{R}^d$ there exists $u \in [0, 1]$ such that

$$h_\delta(x) - h(x) = \mathbb{E}_\rho[(h(x + \delta Y) - h(x))] \leq \delta M_0(A) \mathbb{V}T_{\epsilon} g) \mathbb{E}_\rho[\|Y\|_2 / A(x + \delta Y)] \leq \delta M_0(A) / M_0(A) M_1(b) \|\| \|_L^2.$$  

Now, for each $x \in \mathbb{R}^d$ and $\delta > 0$,

$$h_\delta(x) = \mathbb{E}_\rho[(b(x + \delta Y), g(x + \delta Y))] \leq \mathbb{E}_\rho[\|b(x + \delta Y)\|_2 / A(x + \delta Y)] \leq M_0(A) M_1(b) \|\| \|_L^2 / A(x + \delta Y)] \leq M_0(A) M_1(b) \|\| \|_L^2 / A(x + \delta Y)]$$

so, by Cauchy-Schwarz, the Lipschitzness of $b$, and our assumptions on $g$ and $B$,

$$|(T_{\epsilon} g)(x) - h_\delta(x)| \leq \mathbb{E}_\rho[\|b(x) - B(Y)\|_2 / A(x + \delta Y)] \leq M_0(A) M_1(b) \|\| \|_L^2 / A(x + \delta Y)] \leq \epsilon / A(x).$$

Thus, if we fix $\epsilon > 0$ and define $\bar{\epsilon} = \epsilon / \mathbb{E}_P[\|Y\|_2 B(Y)]$, then the triangle inequality implies

$$|(T_{\epsilon} g)(x) - h_\delta(x)| \leq \epsilon / A(x).$$

To conclude, we will bound $\|g_\delta\|_{K^d}$. By Wendland [29, Thm. 10.21],

$$\|g_\delta\|_{K^d}^2 = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \rho_\delta(\omega)^2 \Phi(\omega) \ d\omega = (2\pi)^{-d/2} \int_{\mathbb{R}^d} \rho_\delta(\omega)^2 \Phi(\omega) \ d\omega \leq \epsilon (2\pi)^{-d/2} \sup_{\omega \in \mathbb{R}^d} \rho_\delta(\omega)^2 \Phi(\omega) \ d\omega,$$

where we have used the Convolution Theorem [29, Thm. 5.16] and the identity $\rho_\delta(\omega) = \rho(\delta \omega)$. Finally, an application of Plancherel’s theorem [14, Thm. 1.1] gives $\|g_\delta\|_{K^d} \leq (2\pi)^{-d/4} F(\delta^{-1})^{1/2} \|g\|_{L^2}$.
C Proof of Proposition 3.3

We begin by establishing the $\Phi SD$ convergence claim. Define the target mean $m_P := \mathbb{E}_{Z \sim P}[Z]$. Since $\log A$ is Lipschitz and $A \geq 0$, $A_N \leq Ae^{m_N}$ and hence $P(A_N) < \infty$ and $\mathbb{E}_P[A_N(Z)\|Z\|^2_2] < \infty$ for all $N$ by our integrability assumptions on $P$.

Suppose $\mathcal{W}_{A_N}(Q_N, P) \to 0$, and, for any probability measure $\mu$ with $\mu(A_N) < \infty$, define the tilted probability measure $\mu_{A_N}$ via $d\mu_{A_N}(x) = d\mu(x)A_N(x)$. By the definition of $\mathcal{W}_{A_N}$, we have $|Q_N(A_N h) - P(A_N h)| \to 0$ for all $h \in \mathcal{H}$. In particular, since the constant function $h(x) = 1$ is in $\mathcal{H}$, we have $|Q_N(A_N) - P(A_N)| \to 0$. In addition, since the functions $f_N(x) = (x - m_N)/A_N(x)$ are uniformly Lipschitz in $N$, we have $m_N - m_P = Q_N(f_N) - P(f_N) \to 0$ and thus $A_N \to A_P$ for $A_P(x) := A(x - m_P) > 0$. Therefore, $P(A_N) \to P(A_P) > 0$, and, as $x/y$ is a continuous function of $(x, y)$ when $y > 0$, we have

$$Q_{N,A_N}(h) - P_{A_N}(h) = Q_N(A_N h)/Q_N(A_N) - P(A_N h)/P(A_N) \to 0$$

and hence the 1-Wasserstein distance $d_H(Q_{N,A_N}, P_{A_N}) \to 0$.

Now note that, for any $g \in \mathcal{G}_{\Phi/A_N,r}$,

$$Q_N(TA_N g) = Q_N(A_N T_{P_{A_N}} g) = Q_N(A_N)Q_{N,A_N}(T_{P_{A_N}} g) = ((Q_N(A_N) - P(A_N)) + P(A_N))Q_{N,A_N}(T_{P_{A_N}} g) \leq (\mathcal{W}_{A_N}(Q_N, P) + P(A_N))Q_{N,A_N}(T_{P_{A_N}} g)$$

where $T_{P_{A_N}}$ is the Langevin operator for the tilted measure $P_{A_N}$, defined by

$$(T_{P_{A_N}} g)(x) = \sum_{d=1}^D (p(x)A_N(x))^{-1} \partial_{z_d}(p(x)A_N(x)g_d(x)).$$

Taking a supremum over $g \in \mathcal{G}_{\Phi/A_N,r}$, we find

$$\Phi SD_{\Phi,r}(Q_N, P) \leq (\mathcal{W}_{A_N}(Q_N, P) + P(A_N)) \Phi SD_{\Phi/A_N,r}(Q_{N,A_N}, P_{A_N}).$$

Furthermore, since $\Phi(x, z)/A_N(x) = F(x - z)$, H"older’s inequality implies

$$\sup_{x \in \mathbb{R}^D} \|g(x)\|_\infty \leq \|F\|_{L_r},$$

$$\sup_{x \in \mathbb{R}^D, d \in [D]} \|\partial_{z_d}g(x)\|_\infty \leq \|\partial_{z_d}F\|_{L_r},$$

and

$$\sup_{x \in \mathbb{R}^D, d \in [D]} \|\partial_{z_d} \partial_{z_d'} g(x)\|_\infty \leq \|\partial_{z_d} \partial_{z_d'} F\|_{L_r}$$

for each $g \in \mathcal{G}_{\Phi/A_N,r}$. Since $\nabla \log p$ and $\nabla \log A_N$ are Lipschitz and $\mathbb{E}_P[A_N(Z)\|Z\|^2_2] < \infty$, we may therefore apply [11, Lem. 18] to discover that $\Phi SD_{\Phi/A_N,r}(Q_{N,A_N}, P_{A_N}) \to 0$ and hence $\Phi SD_{\Phi,r}(Q_N, P) \to 0$ whenever the 1-Wasserstein distance $d_H(Q_{N,A_N}, P_{A_N}) \to 0$.

To see that $R\Phi SD^2_{\Phi,r,v_N, M_N}(Q_N, P) \to 0$ whenever $\Phi SD^2_{\Phi,r}(Q_N, P) \to 0$, first note that since $r \in [1, 2]$, we may apply Jensen’s inequality to obtain

$$\mathbb{E}[R\Phi SD^2_{\Phi,r,v_N, M_N}(Q_N, P)] = \mathbb{E}\left[\sum_{d=1}^D \left(\frac{1}{M} \sum_{m=1}^M v_N(Z_m)^{-1} |Q_N(T_d\Phi)(Z_m)|^r\right)^2/r\right]$$

$$\leq \sum_{d=1}^D \mathbb{E}\left[\sum_{m=1}^M v_N(Z_m)^{-1} |Q_N(T_d\Phi)(Z_m)|^r\right]^{2/r}$$

$$= \Phi SD^2_{\Phi,r}(Q_N, P).$$

Hence, by Markov’s inequality, for any $\epsilon > 0$,

$$\mathbb{P}[R\Phi SD^2_{\Phi,r,v_N, M_N}(Q_N, P) > \epsilon] \leq \mathbb{E}[R\Phi SD^2_{\Phi,r,v_N, M_N}(Q_N, P)]/\epsilon \leq \Phi SD^2_{\Phi,r}(Q_N, P)/\epsilon \to 0,$$

yielding the result.
D  Proof of Proposition 3.6

To achieve the first conclusion, for each \( d \in [D] \), apply Corollary M.2 with \( \delta/D \) in place of \( \delta \) to the random variable

\[
\frac{1}{M} \sum_{m=1}^{M} w_d(Z_m, Q_N).
\]

The first claim follows by plugging in the high probability lower bounds from Corollary M.2 into \( \Phi_{SD}^r(Q_N, P) \) and using the union bound.

The equality \( \mathbb{E}[Y_d] = \Phi_{SD}^r(Q_N, P) \), the KSD-\( \Phi_{SD} \) inequality of Proposition 3.1 (\( \Phi_{SD}^r(Q_N, P) \geq KSD_{L^2}(Q_N, P) ||\rho||_{L^t}^{-r/2} \)), and the assumption \( KSD_{L^2}(Q_N, P) \geq N^{-1/2} \) imply that \( \mathbb{E}[Y_d] \geq N^{-r/2} ||\rho||_{L^t}^{-r/2} \). Plugging this estimate into the initial importance sample size requirement and applying the KSD-\( \Phi_{SD} \) inequality once more yield the second claim.

E  Proof of Proposition 3.7

It turns out that we obtain \((C, 1)\) moments whenever the weight functions \( w_d(z, Q_N) \) are bounded. Let \( Q(\Phi, \nu, C') := \{Q_N \mid \sup_{z,d} w_d(z, Q_N) < C'\} \).

**Proposition E.1.** For any \( C > 0 \), \((\Phi, r, \nu)\) yields \((C, 1)\) second moments for \( P \) and \( Q(\Phi, \nu, C') \).

**Proof** It follows from the definition of \( Q(\Phi, \nu, C) \) that

\[
\sup_{Q_N \in Q(\Phi, \nu, C)} \sup_{d,z} |(Q_N T_d \Phi)(z)|^r/\nu(z) \leq C.
\]

Hence for any \( Q_N \in Q(\Phi, \nu, C) \) and \( d \in [D] \), \( Y_d \leq C \) a.s. and thus

\[
\mathbb{E}[Y_d^2] \leq C'\mathbb{E}[Y_d].
\]

\(\square\)

Thus, to prove Proposition 3.7 it suffices to have uniform bound for \( w_d(z, Q_N) \) for all \( Q_N \in Q(C') \). Let \( \sigma(x) := 1 + ||x|| \) and fix some \( Q \in Q(C') \). Then \( \nu(z) = Q_N(\sigma \Phi(\cdot, z))/C(Q_N) \), where \( C(Q_N) := ||F||_{L^t} Q(\sigma A(\cdot, -m_N)) \leq ||F||_{L^t} C'. \) Moreover, for \( c, c' > 0 \) not depending on \( Q_N \),

\[
|(Q_N T_d \Phi)(z)|^r \leq Q_N(\|\partial_d \log p + \partial_d \log A(\cdot, -m_N) + \partial_d \log F(\cdot, z)\Phi(\cdot, z))\]

\[
\leq cQ_N(1 + \|\cdot\| + \|\cdot - m_N\|_a \Phi(\cdot, z))\]

\[
\leq c'(C')^{-1} Q_N(\sigma \Phi(\cdot, z)).
\]

Thus,

\[
w_d(z, Q_N) = \frac{|(Q_N T_d \Phi)(z)|^r}{\nu(z)} \leq \frac{C(Q)c'(C')^{-1} Q_N(\sigma \Phi(\cdot, z))}{Q_N(\sigma \Phi(\cdot, z))} \leq c'(C')^r ||F||_{L^t}.
\]

F  Technical Lemmas

**Lemma F.1.** If \( P \in \mathcal{P} \), Assumptions A to D hold, and (3) holds, then for any \( \lambda \in (1/2, \bar{\lambda}) \),

\[
|\langle Q N T_d \Phi \rangle(z)| \leq C_{\lambda, C} KSD_{k_d}^{2\lambda-1}.
\]

**Proof** Let \( \varsigma_d(\omega) := (1 + \omega d)^{-1} Q_N(\partial_d A(\cdot, -m_N)e^{-i\omega}) \). Applying Proposition H.1 with \( D = Q_N T_d A(\cdot, -m_N), h = F, \rho(\omega) = 1 + \omega d, \) and \( t = 1/2 \) yields

\[
\int |(Q_N T_d \Phi)(z)| \leq ||F||_{\Psi(\lambda)} \left((2\pi)^{-d/2} ||\varsigma_d||_{L^\infty} ||(1 + \partial_d)^{(1/4)} \Phi_{1/4}||_{L^2} \right)^{2-2\lambda} ||Q N T_d \Phi||_{\Phi}^{2\lambda-1}.
\]

The finiteness of \( ||F||_{\Psi(\lambda)} \) follows from Assumption C. Using \( P \in \mathcal{P} \), Assumption A, and (3) we have

\[
\varsigma_d(\omega) = (1 + \omega d)^{-1} Q_N(\partial_d \log p + \partial_d \log A(\cdot, -m_N) - i\omega d A(\cdot, -m_N))e^{-i\omega} \]

\[
\leq CQ_N(1 + \|\cdot\| A(\cdot, -m_N) \]

\[
\leq CC',
\]

15
so \(\|s_d\|_{L^\infty}\) is finite. The finiteness of \(\|(1 + \partial_d)\Psi^{(1/4)}\|_{L^2}\) follows from the Plancherel theorem and Assumption D. The result now follows upon noting that \(\|Q_N\nabla\Psi\|_\Psi = \text{KSD}_{k_d} .\)

**Lemma F.2.** If \(P \in \mathcal{P}, \text{Assumptions A and B hold, and (3) holds, then for some } b \in [0, 1), C_b > 0,
\[ |Q_N\nabla\Psi(z)| \leq C_b F(z - m_N)^{1-b} . \]

Moreover, \(b = 0\) if \(s = 0.\)

**Proof.** We have (with \(C\) a constant changing line to line)
\[ |Q_N\nabla\Psi(z)| \leq Q_N\nabla\Psi(\cdot, z) = Q_N(\partial_d \log p + \partial_d \log A(\cdot - m_N) + \partial_d \log F(\cdot - z)|A(\cdot - m_N)F(\cdot - z)) \leq CQ_N(1 + \|\| + \| - z\|^s(A(\cdot - m_N)F(\cdot - m_N)^{-1})F(z - m_N) \leq CQ_N(1 + \|\| + \| - m_N\|^s + \|z - m_N\|^s A(\cdot - m_N)F(\cdot - m_N)^{-1})F(z - m_N) \leq CC(1 + \|\| - m_N\|^s)F(z - m_N). \]

By assumption \((1 + \|\| - m_N\|^s)F(z) \to 0\) as \(\|\| \to \infty\), so for some \(C_b > 0\) and \(b \in [0, 1),\)
\[ (1 + \|\| - m_N\|^s) \leq C_b F(z)^{-b} . \]

**G.** Proof of Theorem 3.8: \((C, \gamma)\) second moment bounds for \(R\Phi SD\)

Take \(Q_N \in Q(\mathcal{C})\) fixed and let \(w_d(z) := w_d(z, Q_N)\). For a set \(S\) let \(\nu_S(S^c) := \int_{S^c \cap S^c} \nu(dz)\). Let \(K := \{x \in \mathbb{R}^D \|x - m_N\| \leq R\}.\) Recall that \(Z \sim \nu\) and \(Y_d = w_d(Z)\). We have
\[ E[Y_d^2] = E[w_d(Z)E] = E[w_d(Z)^2 1(Z \in K)] + E[w_d(Z)^2 1(Z \notin K)] \leq ||w_d||_{L^1(\nu)} ||w_d 1(\cdot \in K)||_{L^\infty(\nu)} + ||1(\cdot \notin K)||_{L^1(\nu)} ||w_d^2 1(\cdot \notin K)||_{L^\infty(\nu)} = ||Q_N\nabla\Psi||_{\mathcal{L}^1(\nu)} \sup_{z \in K} w_d(z) + 1(\Phi)^2 \sup_{z \notin K} w_d(z)^2 = E[Y_d] \sup_{z \in K} w_d(z) + 1(\Phi)^2 \sup_{z \notin K} w_d(z)^2 Without loss of generality we can take \(\nu(z) = \Psi(z - m_N)^{\xi_r}/\|\Psi^\xi\|_{L^1}\), since a different choice of \(\nu\) only affects constant factors. Applying Lemma F.1, Assumption D, and (2), we have
\[ \sup_{z \in K} w_d(z) \leq C_{\lambda, c} KSD_{k_d}^{(2\lambda - 1)} \sup_{z \in K} \nu(z)^{-1} \leq C_{\lambda, c} \Psi^{\xi_r} \|\|_{L^1} \sup_{z \in K} F(z - m_N)^{-\xi_r} KSD_{k_d}^{(2\lambda - 1)} \leq C_{\lambda, c} \xi_r^{\xi_r} \|\Psi^\xi\|_{L^1} \|1/\Phi^2\|_{L^1} f(R)^{-\xi_r} ||Q_N\nabla\Psi||_{L^r}^{(2\lambda - 1)} = C_{\lambda, c} (\Psi/\Phi)^{\xi_r} \|\|_{L^1} \|1/\Phi^2\|_{L^1} f(R)^{-\xi_r} ||Y_d||_{L^2}^{2\lambda - 1}. \]

Applying Lemma F.2 we have
\[ \sup_{z \in K^0} w_d(z)^2 \leq C_b^2 \sup_{z \in K^0} F(z - m_N)^{2(1-b)r}/\nu(z)^2 = C_b^2 \|\Psi^{\xi_r}\|_{L^1} \sup_{z \in K^0} F(z - m_N)^{2(1-b-\xi)r} = C_b^2 \|\Psi^{\xi_r}\|_{L^1} \sup_{z \in K^0} f(R)^{2(1-b-\xi)r}. \]

Thus, we have that
\[ E[Y_d^2] \leq C_{\lambda, c, r, \xi, \gamma} E[Y_d^{2\lambda} f(R)^{-\xi_r} + C_{b, \xi, r} f(R)^{2(1-b-\xi)r}. \]

As long as \(E[Y_d]^{2\lambda} \leq C_{b, \xi, r} f(0)^{2(1-b-\xi/2)r}/C_{\lambda, c, r, \xi, \gamma},\) since \(f\) is continuous and non-increasing to zero we can choose \(R\) such that \(f(R)^{2(1-b-\xi)r} = C_{\lambda, c, r, \xi} E[Y_d^{2\lambda}]/C_{b, \xi, r}\) and the result follows for
\[ E[Y_d^{2\lambda}] \leq C_{b, \xi, r} f(0)^{2(1-b-\xi/2)r}/C_{\lambda, c, r, \xi, \gamma}. \]

Otherwise, we can guarantee that \(E[Y_d^{2\lambda}] \leq C_a E[Y_d^{2\gamma\alpha}]\) be choosing \(C_a\) sufficiently large, since by assumption \(E[Y_d]^{\gamma\alpha}\) is uniformly bounded over \(Q_N \in Q(\mathcal{C}).\)
A uniform MMD-type bound

Let \( D \) denote a tempered distribution and \( \Psi \) a stationary kernel. Also, define \( \hat{D}(\omega) := D_x e^{-i(\omega, \hat{x})} \).

**Proposition H.1.** Let \( h \) be a symmetric function such that for some \( s \in (0, 1], h \in K_{\Psi(x)} \) and \( D_x h(x - \cdot) \in K_{\Psi(x)} \). Then

\[
|D_x h(x - \hat{x})| \leq \|h\|_{\Psi(x)} \left\| D_x \Psi^s(\hat{x} - \cdot) \right\|_{\Psi(x)}
\]

and for any \( t \in (0, s) \) any function \( \varphi(\omega) \),

\[
\left\| D_x \Psi^s(\hat{x} - \cdot) \right\|^{1-t}_{\Psi(x)} \leq \left( (2\pi)^{-d/2} \| \varphi^{-1} \hat{D} \|_{L^\infty} \| \varphi \hat{\Psi}^{s/2} \|_{L^2} \right)^{1-s} \left\| D_x \Psi(\hat{x} - \cdot) \right\|^{s-t}_{\Psi(x)}.
\]

Furthermore, if for some \( c > 0 \) and \( \varphi \in (0, s/2), \hat{h} \leq c \hat{\Psi}^r \), then

\[
\|h\|_{\Psi(x)} \leq \frac{c \| \Psi^{(r-s)/2} \|_{L^2}}{(2\pi)^{d/4}}.
\]

**Proof** The first inequality follows from an application of Cauchy-Schwartz:

\[
|D_x h(x - \hat{x})| = |\langle h(\cdot - z), D_x \Psi^s(\hat{x} - \cdot) \rangle_{\Psi(x)}| \\
\leq \|h(\cdot - z)\|_{\Psi(x)} \left\| D_x \Psi^s(\hat{x} - \cdot) \right\|_{\Psi(x)} \\
= \|h\|_{\Psi(x)} \left\| D_x \Psi^s(\hat{x} - \cdot) \right\|_{\Psi(x)}.
\]

For the first norm, we have

\[
\|h\|_{\Psi(x)}^2 = (2\pi)^{-d/2} \int \hat{h}^2(\omega) \frac{ds(\omega)}{\Psi^s(\omega)} d\omega \\
\leq c^2 (2\pi)^{-d/2} \int \hat{\phi}^{2r-s}(\omega) d\omega \\
= c^2 (2\pi)^{-d/2} \left\| \phi^{r-s/2} \right\|^2_{L^2}.
\]

Note that by the convolution theorem \( \mathcal{F}(D_x \Psi(\hat{x} - \cdot))(\omega) = \hat{D}(\omega) \hat{\Psi}^s(\omega) \). For the second norm, applying Jensen’s inequality and Hölder’s inequality yields

\[
\left\| D_x \Psi^s(\hat{x} - \cdot) \right\|_{\Psi(x)}^2 = (2\pi)^{-d/2} \int \hat{\Psi}(\omega) (\hat{D}(\omega)^2) \frac{ds(\omega)}{\Psi^s(\omega)} d\omega \\
= (2\pi)^{-d/2} \left( \int \hat{\Psi}^t \frac{ds(\omega)}{\hat{D}^2} \right) \int \frac{\hat{D}(\omega)^2}{\hat{\Psi}^t} \Psi(\omega)^{s-t} d\omega \\
\leq (2\pi)^{-d/2} \left( \int \hat{\Psi}^t \frac{ds(\omega)}{\hat{D}^2} \right) \int \frac{\hat{D}(\omega)^2}{\hat{\Psi}^t} \Psi(\omega)^{1-t} d\omega \\
= \left( (2\pi)^{-d/2} \int \hat{\Psi}^t \frac{ds(\omega)}{\hat{D}^2} \right)^\frac{s-t}{s-t} \left\| D_x \Psi(\hat{x} - \cdot) \right\|^{2\frac{s-t}{s-t}}_{\Psi(x)} \\
\leq \left( (2\pi)^{-d/2} \| \varphi^{-1} \hat{D} \|_{L^\infty} \int \varphi^{2r} \hat{\Psi}^t \right)^\frac{s-t}{s-t} \left\| D_x \Psi(\hat{x} - \cdot) \right\|^{2\frac{s-t}{s-t}}_{\Psi(x)} \\
= \left( (2\pi)^{-d/2} \| \varphi^{-1} \hat{D} \|_{L^\infty} \| \varphi \hat{\Psi}^{s/2} \|_{L^2} \right)^\frac{s-t}{s-t} \left\| D_x \Psi(\hat{x} - \cdot) \right\|^{2\frac{s-t}{s-t}}_{\Psi(x)}.
\]

\( \square \)
We verify each of the assumptions in turn. By construction or assumption each condition in Assumption A holds. Note in particular that $\Psi_{2a}^{\sech} \in C^\infty$. Since $e^{-a|x_d|} \leq \sech(ax_d) \leq 2e^{-a|x_d|}$, Assumption B holds with $\|\cdot\| = \|\cdot\|_1$, $f(R) = 2^d e^{-\sqrt{2\pi}aR}$, and $\ell = 2^{-d}$, and $s = 1$. In particular,

$$\partial_{x_d} \log \Psi_{2a}^{\sech}(x) = \sqrt{2\pi}e^{\sqrt{2\pi}ax_d} + \sum_{d' \neq d} D \log \sech(\sqrt{2\pi}ax_{d'})$$

$$\leq (\sqrt{2\pi}a)(1 + \sum_{d' \neq d} |x_{d'}|)$$

and using Proposition L.3 we have that

$$\hat{\Psi}_{a}^{\sech}(x - z) \leq e^{\sqrt{2\pi}|x||1} \hat{\Psi}_{a}^{\sech}(z) \leq 2^d \frac{\hat{\Psi}_{a}^{\sech}(z)}{\hat{\Psi}_{a}^{\sech}(x)}.$$

Assumption C holds with $\lambda = 1$ since for any $\lambda \in (0, 1)$, it follows from Proposition L.2 that

$$\tilde{f}_j / \Psi_{a}^{\sech} / \hat{\Phi}_{a}^{\sech} \leq (\Psi_{2a}^{\sech})^{1-\lambda} \in L^2.$$

The first part of Assumption D holds as well since by (6), $2^2 \hat{\Psi}_{a}^{\sech}(\omega) = a^{-D/2} \hat{\Psi}_{a}^{\sech}(\omega) \in L^2$.

Finally, to verify the second part of Assumption D, we first note that since $r = 2$, $t = \infty$. The assumption holds since by Proposition L.2, $\frac{\hat{\Psi}_{a}^{\sech}(\omega)}{\hat{\Psi}_{2a}^{\sech}(\omega)} \leq 1$.

We verify each of the assumptions in turn. By construction or assumption each condition in Assumption A holds. Note in particular that $\hat{\Psi}_{c,\beta}^{\IMQ} \in C^\infty$. Assumption B holds with $\|\cdot\| = \|\cdot\|_2$, $f(R) = ((c')^2 + R^2)^{\beta'}$, $\ell = 1$, and $s = 0$. In particular,

$$\partial_{x_d} \log \Psi_{c,\beta}^{\IMQ}(x) \leq \frac{2\beta'|x_d|}{(c')^2 + \|x\|_2^2} \leq -2\beta'$$

and

$$\Psi_{c,\beta}^{\IMQ}(x - z) = \left(\frac{(c')^2 + \|x - z\|_2^2}{(c')^2 + \|z\|_2^2}\right)^{-\beta'} \leq \left(\frac{(c')^2 + 2\|z\|_2^2 + 2\|x\|_2^2}{(c')^2 + \|z\|_2^2}\right)^{-\beta'} \leq \left(2 + \|x\|_2^2/(c')^2\right)^{-\beta'} = 2^{-\beta} \Psi_{c,\beta}^{\IMQ}(x)^{-1}.$$

By Wendland [29, Theorem 8.15], $\hat{\Psi}_{c,\beta}^{\IMQ}$ has generalized Fourier transform

$$\hat{\Psi}_{c,\beta}^{\IMQ}(\omega) = \frac{\Gamma(-\beta)}{\Gamma(1+\beta)} \frac{\omega}{c}^{1+\beta} \frac{||\omega||_2}{c}^{-\beta-D/2} K_{\beta+D/2}(c||\omega||_2),$$

where $K_c(z)$ is the modified Bessel function of the third kind. We write $a(\ell) \sim b(\ell)$ to denote asymptotic equivalence up to a constant: $\lim_{\|\omega\|_2 \to \infty} a(\ell)/b(\ell) = c$ for some $c \in (0, \infty)$. Asymptotically [7, eqs. 10.25.3, 10.30.2],

$$\hat{\Psi}_{c,\beta}^{\IMQ}(\omega) \sim \|\omega\|_2^{-\beta-D/2-1/2} e^{-c\|\omega\|_2}, \quad \|\omega\|_2 \to \infty \quad \text{and}$$

$$\hat{\Psi}_{c,\beta}^{\IMQ}(\omega) \sim \|\omega\|_2^{-(\beta+D/2)-|\beta+D/2|} \|\omega\|_2^{-2(\beta+D)+} \quad \|\omega\|_2 \to 0.$$
Assumption C holds since for any $\lambda \in (0, \lambda_c)$,
\[
\hat{\Psi}_{c',\beta}^{\text{IMQ}}(\hat{\Psi}_{c,\beta}^{\text{IMQ}})^{1/2} \sim \|\omega\|^2\left(\beta + D - 2\lambda/2 + (\beta + D - 2\lambda/2)\right)^{1/2} e^{(\beta - c)\lambda/2} \|\omega\|^2, \quad \|\omega\|^2 \to \infty \quad \text{and}
\sim \|\omega\|^2\left(2\beta + D\right)\gamma/2 - (2\beta + D)\gamma = \|\omega\|^2\left(2\beta + D\right)/2 \|\omega\|^2 \to 0,
\]
so $\hat{\Psi}_{c',\beta}^{\text{IMQ}}(\hat{\Psi}_{c,\beta}^{\text{IMQ}})^{1/2} \in L^2$ as long as $c' = c/2 > c\lambda/2$ and $\lambda(2\beta + D) > -D$. The first condition holds by construction and second condition is always satisfied, since $2\beta + D \geq 0 > -D$.

The first part of Assumption D holds as well since $\hat{\Psi}_{c,\beta}(\omega)$ decreases exponentially as $\|\omega\|^2 \to \infty$ and $\hat{\Psi}_{c,\beta}(\omega) \sim 1$ as $\|\omega\|^2 \to 0$, so $\omega^2 \hat{\Psi}_{c,\beta}(\omega)$ is integrable.

Finally, to verify the second part of Assumption D we first note that $\hat{\Psi}_{c',\beta}(\omega)$ increases exponentially as $\|\omega\|^2 \to \infty$ and so $\hat{\Psi}_{c',\beta}(\omega) \sim 1$ as $\|\omega\|^2 \to 0$, so $\omega^2 \hat{\Psi}_{c',\beta}(\omega)$ is integrable.

Thus, $\hat{\Psi}_{c,\beta}(\omega)$ decreases exponentially as $\|\omega\|^2 \to \infty$ and $\hat{\Psi}_{c,\beta}(\omega) \sim 1$ as $\|\omega\|^2 \to 0$, so $\omega^2 \hat{\Psi}_{c,\beta}(\omega)$ is integrable.

Both these conditions hold by construction.

K Proofs of Proposition 4.1 and Theorem 4.3: Asymptotics of $R\Phi SD$

The proofs of Proposition 4.1 and Theorem 4.3 rely on the following asymptotic result.

**Theorem K.1.** Let $\xi_i : \mathbb{R}^D \times Z \to \mathbb{R}$, $i = 1, \ldots, I$ be a collection of functions; let $Z_{n,m} \overset{\text{indep}}{\sim} \nu_N$, where $\nu_N$ is a distribution on $Z$; and let $X_n \overset{i.d.}{\sim} \mu$, where $\mu$ is absolutely continuous with respect to Lebesgue measure. Define the random variables $\xi_{n,i,m} := \xi_i(X_n, Z_{n,m})$ and, for $r, s > 0$, the random variable
\[
F_{r,s,N} := \left(\sum_{i=1}^I \left(\sum_{m=1}^M \left|\sum_{n=1}^N \xi_{n,i,m} \right|^r \right)^{s/r}\right)^{2/s}.
\]

Assume that for all $N \geq 1, i \in [I]$, and $m \in [M]$, $\xi_{n,i,m}$ has a finite second moment that $\Sigma_{i,m} := \text{Cov}(\xi_{n,i,m}, \xi_{n,i,m'}) < \infty$ exists for all $i, i' \in [I]$ and $m, m' \in [M]$. Then the following statements hold.

1. If $\varrho_{n,i,m} := (\mu \times \nu_N)(\xi_i) = 0$ for all $i \in [N]$ then
\[
NF_{r,s,N} \overset{D}{\to} \left(\sum_{i=1}^I \left(\sum_{m=1}^M \left|\xi_{n,i,m} \right|^r \right)^{s/r}\right)^{2/s} \text{ as } N \to \infty, \quad (5)
\]
where $\zeta \sim A \gamma(0, \Sigma)$.

2. If $\varrho_{n,i,m} \neq 0$ for some $i$ and $m$, then
\[
NF_{r,s,N} \overset{a.s.}{\to} \infty \text{ as } N \to \infty.
\]

**Proof** Let $V_{n,i,m} = N^{-1/2} \sum_{n=1}^N \xi_{n,i,m}$. By assumption $\|\Sigma\| < \infty$. Hence, by the multivariate CLT,
\[
V_N - N^{1/2} \varrho N \overset{D}{\to} A \gamma(0, \Sigma).
\]

Observe that $NF_{r,s,N} = \left(\sum_{i=1}^I \left(\sum_{m=1}^M \left|V_{n,i,m} \right|^r \right)^{s/r}\right)^{2/s}$. Hence if $\varrho = 0$, (5) follows from the continuous mapping theorem.

Assume $\varrho_{n,i,j} \neq 0$ for some $i$ and $j$ and all $N \geq 0$. By the strong law of large numbers, $N^{-1/2}V_N \overset{a.s.}{\to} 0\infty$. Together with the continuous mapping theorem conclude that $F_{r,s,N} \overset{a.s.}{\to} c$ for
some \( c > 0 \). Hence \( NF_{r,s,N} \xrightarrow{a.s.} \infty. \)

When \( r = s = 2 \), the RΦSD is a degenerate \( V \)-statistic, and we recover its well-known distribution [24, Sec. 6.4, Thm. B] as a corollary. A similar result was used in Jitkrittum et al. [16] to construct the asymptotic null for the FSSD, which is degenerate \( U \)-statistic.

**Corollary K.2.** Under the hypotheses of Theorem K.1(1),

\[
NF_{2,2,N} \xrightarrow{D} \sum_{i=1}^{I} \sum_{m=1}^{M} \lambda_{im} \omega_{m}^{2} \text{ as } N \to \infty,
\]

where \( \lambda = \text{eig}(\Sigma) \) and \( \omega_{ij} \sim \mathcal{N}(0,1) \).

To apply these results to RΦSDs, take \( s = 2 \) and apply Theorem K.1 with \( I = D, \xi_{N,dm} = \xi_{r,N,dm} \). Under \( H_0 : \mu = P, P(\xi_{r,N,dm}) = 0 \) for all \( d \in [D] \) and \( m \in [M] \), so part 1 of Theorem K.1 holds. On the other hand, when \( \mu \neq P \), there exists some \( m \) and \( d \) for which \( \mu(\xi_{r,dm}) \neq 0 \). Thus, under \( H_1 : \mu \neq P \) part 2 of Theorem K.1 holds.

The proof of Theorem 4.3 is essentially identical to that of Jitkrittum et al. [16, Theorem 3].

## L Hyperbolic secant properties

Recall that the hyperbolic secant function is given by \( \text{sech}(a) = \frac{2}{e^{a} + e^{-a}}. \) For \( x \in \mathbb{R}^{d} \), define the hyperbolic secant kernel

\[
\Psi_{a}^{\text{sech}}(x) := \text{sech} \left( \sqrt{\frac{\pi}{2} ax} \right) := \prod_{i=1}^{d} \text{sech} \left( \sqrt{\frac{\pi}{2} ax_{i}} \right).
\]

It is a standard result that

\[
\hat{\Psi}_{a}^{\text{sech}}(\omega) = a^{-D} \Psi_{1/a}^{\text{sech}}(\omega).
\]

We can relate \( \Psi_{a}^{\text{sech}}(x) \) to \( \Psi_{a^\zeta}^{\text{sech}}(x) \), but to do so we will need the following standard result:

**Lemma L.1.** For \( a, b \geq 0 \) and \( \xi \in (0, 1] \),

\[
\frac{a^{\xi} + b^{\xi}}{2^{1-\xi}} \leq (a + b)^{\xi} \leq a^{\xi} + b^{\xi}.
\]

**Proof** The lower bound follows from an application of Jensen’s inequality and the upper bound follows from the concavity of \( a \mapsto a^{\xi}. \)

**Proposition L.2.** For \( \xi \in (0, 1] \),

\[
2^{-d(1-\xi)} \hat{\Psi}_{a/\xi}^{\text{sech}}(x)^{\xi} \leq \hat{\Psi}_{a}^{\text{sech}}(x)^{\xi} \leq 2^{d(1-\xi)} \hat{\Psi}_{a/\xi}^{\text{sech}}(x)^{\xi}.
\]

Thus, \( \Psi_{a/\xi}^{\text{sech}} \) is equivalent to \((\Psi_{a}^{\text{sech}})^{\xi}\).

**Proof** Apply Lemma L.1 and (6).

**Proposition L.3.** For all \( x, y \in \mathbb{R}^{d} \) and \( a > 0 \),

\[
\Psi_{a}^{\text{sech}}(x-z) \leq e^{\sqrt{a\|x\|_{1}}} \Psi_{a}^{\text{sech}}(z).
\]

**Proof** Take \( d = 1 \) since the general case follows immediately. Without loss of generality assume that \( x \geq 0 \) and let \( a' = \sqrt{\frac{a}{2}}. \) Then

\[
\frac{\Psi_{a}^{\text{sech}}(x-z)}{\Psi_{a}^{\text{sech}}(z)} = \frac{e^{a'z} + e^{-a'z}}{e^{a'(x-z)} + e^{-a'(x-z)}} = \frac{e^{a'z} + e^{-a'z}}{e^{-a'z} + e^{2a'z} e^{a'z}} \leq e^{a'z}.
\]
M Concentration inequalities

Theorem M.1 (Chung and Lu [4, Theorem 2.9]). Let $X_1, \ldots, X_m$ be independent random variables satisfying $X_i > -A$ for all $i = 1, \ldots, m$. Let $X := \sum_{i=1}^{m} X_i$ and $\overline{X}^2 := \sum_{i=1}^{m} \mathbb{E}[X_i^2]$. Then for all $t > 0$,

$$
\mathbb{P}(X \leq \mathbb{E}[X] - t) \leq e^{-\frac{1}{2}t^2/(\overline{X}^2 + At/3)}.
$$

Let $\hat{X} := \frac{1}{m} \sum_{i=1}^{m} X_i$.

Corollary M.2. Let $X_1, \ldots, X_m$ be i.i.d. nonnegative random variables with mean $\overline{X} := \mathbb{E}[X_1]$. Assume there exist $c > 0$ and $\gamma \in [0, 2]$ such that $\mathbb{E}[X_1^2] \leq c \overline{X}^{2-\gamma}$. If, for $\delta \in (0, 1)$ and $\varepsilon \in (0, 1)$,

$$
m \geq \frac{2c \log(1/\delta)}{\varepsilon^2} \overline{X}^{-\gamma},
$$

then with probability at least $1 - \delta$, $\hat{X} \geq (1 - \varepsilon) \overline{X}$.

Proof Applying Theorem M.1 with $t = m \varepsilon \overline{X}$ and $A = 0$ yields

$$
\mathbb{P}(\hat{X} \leq (1 - \varepsilon) \overline{X}) \leq e^{-\frac{1}{2} \varepsilon^2 m \overline{X}^2 / (c \mathbb{E}[X_1^2])} \leq e^{-\frac{1}{2} \varepsilon^2 m \overline{X}^2}.
$$

Upper bounding the right hand side by $\delta$ and solving for $m$ yields the result. \qed

Corollary M.3. Let $X_1, \ldots, X_m$ be i.i.d. nonnegative random variables with mean $\overline{X} := \mathbb{E}[X_1]$. Assume there exists $c > 0$ and $\gamma \in [0, 2]$ such that $\mathbb{E}[X_1^2] \leq c \overline{X}^{2-\gamma}$. Let $\epsilon' = |X^* - \overline{X}|$ and assume $\epsilon' \leq \eta X^*$ for some $\eta \in (0, 1)$. If, for $\delta \in (0, 1)$,

$$
m \geq \frac{2c \log(1/\delta)}{\varepsilon^2} \overline{X}^{-\gamma},
$$

then with probability at least $1 - \delta$, $\hat{X} \geq (1 - \varepsilon) X^*$. In particular, if $\epsilon' \leq \frac{X^*}{\sqrt{n}}$ and $X^* \geq \frac{\varepsilon^2}{\sigma^2 n}$, then with probability at least $1 - \delta$, $\hat{X} \geq (1 - \varepsilon) X^*$ as long as

$$
m \geq \frac{2c(1 - \eta)^2 \eta^2 \gamma}{\varepsilon^2 \sigma^2 \log(1/\delta)} n^\gamma.
$$

Proof Apply Corollary M.2 with $\frac{X^*}{\overline{X}}$ in place of $\varepsilon$. \qed

Example M.1. If we take $\gamma = 1/4$ and $\eta = \varepsilon = 1/2$, then $X^* \geq \frac{4\sigma^2}{n}$ and $m \geq \frac{\sqrt{2c \log(1/\delta)}}{\sigma^{1/4} n^{1/4}}$ guarantees that $\hat{X} \geq \frac{1}{2} X^*$ with probability at least $1 - \delta$. 

21