The Minimax Lower Bound of Kernel Stein Discrepancy Estimation

Jose Cribeiro-Ramallo
Karlsruhe Institute of Technology

Agnideep Aich
University of Louisiana at Lafayette

Florian Kalinke
Karlsruhe Institute of Technology

Ashit Baran Aich
Formerly of Presidency College

Zoltán Szabó
London School of Economics

Abstract

Kernel Stein discrepancies (KSDs) have emerged as a powerful tool for quantifying goodness-of-fit over the last decade, featuring numerous successful applications. To the best of our knowledge, all existing KSD estimators with known rate achieve $\sqrt{n}$-convergence. In this work, we present two complementary results (with different proof strategies), establishing that the minimax lower bound of KSD estimation is $n^{-1/2}$ and settling the optimality of these estimators. Our first result focuses on KSD estimation on $\mathbb{R}^d$ with the Langevin-Stein operator; our explicit constant for the Gaussian base kernel indicates that the difficulty of KSD estimation may increase exponentially with the dimensionality d. Our second result settles the minimax lower bound for KSD estimation on general domains.

1 INTRODUCTION

A fundamental problem in data science and statistics is quantifying the goodness-of-fit (GoF) between a known fixed target distribution and a sampling distribution (observed through samples only). A recent approach to tackle this challenging task employs the family of kernel Stein discrepancies (KSDs; Chwialkowski et al. 2016; Liu et al. 2016), which combine a so-called Stein operator (Stein, 1972; Chen, 2021; Anastasiou et al., 2023) with the flexibility and computational tractability of reproducing kernel Hilbert spaces (RKHSs; Aronszajn 1950) associated to kernels. These kernel functions have been designed on a wide variety of domains, rendering KSDs broadly applicable.

KSDs rely on kernel mean embeddings (Berlinet and Thomas-Agnan, 2004; Smola et al., 2007; Gretton et al., 2012), mapping probability measures to RKHSs without loss of information, under mild conditions. Considering the RKHS distance of two embedded probability distributions results in the maximum mean discrepancy (MMD), known to be equivalent (Sejdinovic et al., 2013b) to energy distance (Baringhaus and Franz, 2004; Székely and Rizzo, 2004, 2005) (also known as N-distance; Zinger et al. 1992; Klebanov 2005), and to be a specific instance of integral probability metrics (IPM; Zolotarev 1983; Müller 1997). The key property guaranteeing that MMD is a metric is that the underlying kernel function is characteristic (Fukumizu et al., 2007; Sriperumbudur et al., 2010b). When MMD is applied—with the product kernel—to the embeddings of a joint distribution and the product of its marginals, one obtains the Hilbert-Schmidt independence criterion (HSIC), originally designed for $M = 2$ components (Gretton et al., 2005a,b), and later extended to $M \geq 2$ components (Quadrianto et al., 2009; Sejdinovic et al., 2013a; Pfister et al., 2018). HSIC is a valid independence measure for $M = 2$ random variables if the kernel components are characteristic (Lyons, 2013); for $M > 2$, c_0-universality of the kernel components suffices (Szabó and Sriperumbudur, 2018). HSIC can also be interpreted as the RKHS norm of the covariance operator; it is also equivalent (Sejdinovic et al., 2013b) to distance covariance (Székely et al., 2007; Székely and Rizzo, 2009; Lyons, 2013). Related mean embedding-based approaches constructed to measure the interaction of random variables include the kernel Lancaster and Streitberg interactions (Sejdinovic et al., 2013a), which, alongside MMD, HSIC ($M = 2$), and maximum variance discrepancy (Makigusa, 2024), are specific cases of kernel cumulants (Bonnier et al., 2023; Liu et al., 2023).

Proceedings of the 29$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s). Corresponding author: jose.cribeiro@kit.edu.

Similarly, KSD uses the mean embeddings of the target and the sampling distribution, where the underlying kernel is chosen such that the mean embedding of the target distribution vanishes. On Euclidean spaces, one attractive property of the classical Langevin-Stein KSD is that the resulting GoF measure is agnostic of the normalization constant of the sampling distribution, which can be challenging to compute in applications. This independence has led to its widespread use and its extension to other domains. Applications include model validation (Gorham and Mackey, 2017; Futami et al., 2019; Hodgkinson et al., 2021; Wang et al., 2023), learning variational models (Liu and Wang, 2016, 2018; Chen et al., 2018, 2019; Korba et al., 2020, 2021), testing (Liu et al., 2016; Chwialkowski et al., 2016; Schrab et al., 2022; Baum et al., 2023; Hagrass et al., 2025), model comparison (Lim et al., 2019; Kanagawa et al., 2020), distribution compression (Li et al., 2024), and model explainability (Sarvmaili et al., 2025). KSD has also successfully been applied on discrete spaces (Yang et al., 2018), Riemannian manifolds (Xu and Matsuda, 2020, 2021; Barp et al., 2022), Hilbert spaces (Wynne et al., 2025), point-processes (Yang et al., 2019), and graph data (Xu and Reinert, 2021).

Despite their broad applicability, to the best our knowledge, convergence rates of KSD estimators have only been studied for V-statistic and Nyström-based estimators (Kalinke et al., 2025). In fact, under a sub-Gaussian assumption, both estimators achieve $\sqrt{n}$-convergence on general domains.[1] Whether faster rates for KSD estimation are achievable is an open problem and the main focus of this work.

Answering this question requires obtaining minimax lower bounds and contrasting them with the existing upper bounds. Related minimax lower bounds have been established for MMD (Tolstikhin et al., 2016), the mean embedding (Tolstikhin et al., 2017), covariance operators (Zhou et al., 2019), and HSIC (Kalinke and Szabó, 2024). While the proofs differ in all of the mentioned works, they (i) all assume the underlying kernel function to be bounded and (ii) rely on Le Cam's two point method (elaborated in Section 7) to establish the minimax lower bounds. Unfortunately, in the context of KSD, boundedness practically never holds, see, for example, (Kalinke et al., 2025, Example 1) and (Hagrass et al., 2025, Remark 2). Hence, existing results do not apply to the analysis of KSD estimation. In this work, we address this gap by making the following **contributions**.

(i) We establish the minimax lower bound $n^{-1/2}$ of KSD estimation on $\mathbb{R}^d$ with continuous bounded translation-invariant characteristic base kernels, with explicit constants for Gaussian base kernels.

(ii) Following a different proof strategy—by employing local perturbations—, we obtain the same lower bound for KSD estimation on general domains. Our imposed integrability conditions can be seen as a relaxation of the usual boundedness assumptions.

The paper is structured as follows. Notations are introduced in Section 2, followed by recalling the notion of KSD (Section 3). Section 4 is dedicated to existing KSD estimators with known convergence rates. After recalling the minimax estimation framework (Section 5), our minimax results on KSD estimation are presented (Section 6) alongside their proof sketches (Section 7). Detailed proofs are available in the appendix.

2 NOTATIONS

In this section, we introduce our notations: $\mathbb{N}_0$, $\mathbb{N}_{>0}$, $\mathbb{R}$, $[n]$, $\{\{\cdot\}\}$, $(\cdot)^\top$, $\langle\cdot,\cdot\rangle_2$, $\|\cdot\|_2$, $\|\cdot\|_\infty$, $\mathbf{1}_d$, $\mathbf{x}<\mathbf{y}$, $\mathbf{A}^-$, ∇, supp, $\bar{S}$, $\overline{z}$, $\overline{\mathbf{z}}$, $\operatorname{Re}(\cdot)$, $\operatorname{Im}(\cdot)$, $\mathbf{A}^*$, $\langle\mathbf{x},\mathbf{y}\rangle_{\mathbb{C}^d}$, $\|\mathbf{x}\|_{\mathbb{C}^d}$, $|\mathbf{x}|$, $\{\mathbf{e}_j\}_{j=1}^d$, $\frac{\partial f}{\partial x_j}$, $\mathbf{x}^{\boldsymbol{\alpha}}$, $D^{\boldsymbol{\alpha}}f$, $B(\mathcal{H})$, Span, $\mathcal{C}^s(\mathbb{R}^d)$, $\mathcal{C}(\mathcal{X})$, $\mathcal{C}_b(\mathcal{X})$, $\mathcal{M}_1^+(\mathcal{X})$, $\mathcal{B}(\mathcal{X})$, λ_d, δ_x, $\mathbb{E}_P[X]$, P^n, $\ll$, $\frac{\mathrm{d}Q}{\mathrm{d}P}$, $\mathrm{KL}(Q\|P)$, $M_{\boldsymbol{\alpha}}^P$, $\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, ψ_P, $\mathcal{H}_k$, $\mathcal{H}_k^d$, $\mathcal{O}$, Ω, Θ, $\mathcal{O}_P$.

The set of natural numbers is written as $\mathbb{N}_0 = \{0,1,2,3,\dots\}$; the set of positive integers is denoted by $\mathbb{N}_{>0}$; $\mathbb{R}$ stands for reals. Let $[n] \coloneqq \{1,\dots,n\}$ with $n \in \mathbb{N}_{>0}$. We write $\{\{\cdot\}\}$ for a multiset. The transpose of a vector $\mathbf{v}\in\mathbb{R}^d$ is written as $\mathbf{v}^\top \in \mathbb{R}^{1\times d}$. The inner product of vectors $\mathbf{u}=(u_j)_{j=1}^d, \mathbf{v}=(v_j)_{j=1}^d \in \mathbb{R}^d$ is $\langle\mathbf{u},\mathbf{v}\rangle_2 = \sum_{j=1}^d u_j v_j$. The Euclidean norm of $\mathbf{x}\in\mathbb{R}^d$ is $\|\mathbf{x}\|_2 = \sqrt{\langle\mathbf{x},\mathbf{x}\rangle_2}$; its supremum norm is $\|\mathbf{x}\|_\infty = \max_{j\in[d]}|x_j|$. The d-dimensional vector of ones is denoted by $\mathbf{1}_d$. For vectors $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, $\mathbf{x}<\mathbf{y}$ means that $x_j<y_j$ for all $j\in[d]$. The (Moore-Penrose) pseudo-inverse of a matrix $\mathbf{A}\in\mathbb{R}^{d_1\times d_2}$ is $\mathbf{A}^-\in\mathbb{R}^{d_2\times d_1}$. For a differentiable function $f:\mathbb{R}^d\to\mathbb{R}$, let $\nabla_{\mathbf{x}}f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_j}(\mathbf{x})\right)_{j=1}^d \in \mathbb{R}^d$ ($\mathbf{x}\in\mathbb{R}^d$). The support of a function $\varphi:\mathbb{R}^d\to\mathbb{R}$ is $\operatorname{supp}(\varphi) = \overline{\{\mathbf{x}\in\mathbb{R}^d : \varphi(\mathbf{x})\neq 0\}}$, where $\bar{S}$ stands for the closure of the set S. The conjugate of a complex number $z=a+ib\in\mathbb{C}$ is denoted by $\overline{z}=a-ib$ with $i=\sqrt{-1}$; its real part is $\operatorname{Re}(z)=a$, its complex part is $\operatorname{Im}(z)=b$. On a vector $\mathbf{z}=(z_j)_{j=1}^d\in\mathbb{C}^d$, conjugation, real part and complex part act coordinate-wise: $\overline{\mathbf{z}} = (\overline{z_j})_{j=1}^d$, $\operatorname{Re}(\mathbf{z}) = (\operatorname{Re}(z_j))_{j=1}^d$, $\operatorname{Im}(\mathbf{z}) = (\operatorname{Im}(z_j))_{j=1}^d$. The adjoint of a matrix $\mathbf{A}\in\mathbb{C}^{d_1\times d_2}$ is written as $\mathbf{A}^*\in\mathbb{C}^{d_2\times d_1}$. The inner product of vectors $\mathbf{x},\mathbf{y}\in\mathbb{C}^d$ is $\langle\mathbf{x},\mathbf{y}\rangle_{\mathbb{C}^d} = \mathbf{y}^*\mathbf{x}$; $\|\mathbf{x}\|_{\mathbb{C}^d} \coloneqq \sqrt{\langle\mathbf{x},\mathbf{x}\rangle_{\mathbb{C}^d}}$ for $\mathbf{x}\in\mathbb{C}^d$.

[1] As noted in the cited work, the $\sqrt{n}$-rate, while presented on $\mathbb{R}^d$, also holds on general domains.

Let $\mathbf{x} = (x_i)_{i=1}^d \in \mathbb{R}^d$; we write $|\mathbf{x}| := \sum_{j\in[d]} |x_j|$. Let $\{\mathbf{e}_j\}_{j=1}^d \subset \mathbb{R}^d$ be the canonical basis of $\mathbb{R}^d$. For $f : \mathbb{R}^d \to \mathbb{C}$, we define $\frac{\partial f}{\partial x_j}(\mathbf{x}) = \lim\limits_{h\to 0} \frac{f(\mathbf{x}+h\mathbf{e}_j)-f(\mathbf{x})}{h} = \lim\limits_{h\to 0} \frac{\mathrm{Re}(f(\mathbf{x}+h\mathbf{e}_j))-\mathrm{Re}(f(\mathbf{x}))}{h} + i \lim\limits_{h\to 0} \frac{\mathrm{Im}(f(\mathbf{x}+h\mathbf{e}_j))-\mathrm{Im}(f(\mathbf{x}))}{h}$ as the partial derivative of f on x_j, and $\nabla_{\mathbf{x}} f(\mathbf{x}) = \left(\frac{\partial f}{\partial x_j}(\mathbf{x})\right)_{j=1}^d \in \mathbb{C}^d$ as the gradient of f ($\mathbf{x} \in \mathbb{R}^d$). Let $\boldsymbol{\alpha} = (\alpha_j)_{j=1}^d \in \mathbb{N}_0^d$ and $\mathbf{x} \in \mathbb{R}^d$. We write $\mathbf{x}^{\boldsymbol{\alpha}} := \prod_{j=1}^d x_j^{\alpha_j}$ and $D^{\boldsymbol{\alpha}} f := \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial \mathbf{x}^{\boldsymbol{\alpha}}} = \frac{\partial^{|\boldsymbol{\alpha}|} f}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}$. Let $\mathcal{H}$ be a Hilbert space; $B(\mathcal{H}) = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \le 1\}$ denotes its unit ball centered at the origin. For a set S in a vector space, $\mathrm{Span}(S)$ stands for its linear hull. For $s \in \mathbb{N}_0$, the space of s-times continuously differentiable real-valued functions on $\mathbb{R}^d$ is denoted by $\mathcal{C}^s(\mathbb{R}^d)$. Let $\mathcal{X}$ be a topological space. The set of real-valued continuous functions on $\mathcal{X}$ is denoted by $\mathcal{C}(\mathcal{X})$. The subspace of $\mathcal{C}(\mathcal{X})$ consisting of bounded functions is denoted by $\mathcal{C}_b(\mathcal{X})$. The set of Borel probability measures on $\mathcal{X}$ is denoted by $\mathcal{M}_1^+(\mathcal{X})$, with $\mathcal{B}(\mathcal{X})$ standing for the Borel sigma algebra on $\mathcal{X}$. Let λ_d denote the Lebesgue measure on $\mathbb{R}^d$. The Dirac measure centered at $x \in \mathcal{X}$ is denoted by δ_x. The expectation of a random variable $X \sim P \in \mathcal{M}_1^+(\mathcal{X})$ is $\mathbb{E}_P[X] = \int_{\mathcal{X}} x \mathrm{d}P(x)$. The n-fold product measure of P is denoted by $P^n = \otimes_{j=1}^n P$. Let $Q, P \in \mathcal{M}_1^+(\mathcal{X})$, and let Q absolutely continuous w.r.t. P ($Q \ll P$, with the corresponding Radon-Nikodym derivative denoted by $\frac{\mathrm{d}Q}{\mathrm{d}P}$), their Kullback–Leibler divergence is defined as $\mathrm{KL}(Q\|P) = \int_{\mathcal{X}} \ln\left(\frac{\mathrm{d}Q(x)}{\mathrm{d}P(x)}\right) \mathrm{d}Q(x)$. Given a probability measure $P \in \mathcal{M}_1^+(\mathbb{R}^d)$, we denote its moment of order $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ as $M_{\boldsymbol{\alpha}}^P = \int_{\mathbb{R}^d} \mathbf{x}^{\boldsymbol{\alpha}} \mathrm{d}P(\mathbf{x})$. Normal random variables with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are denoted by $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The function $\psi_P(\boldsymbol{\omega}) = \mathbb{E}_P[e^{i\langle X, \boldsymbol{\omega}\rangle_2}]$ is known as the characteristic function of P. A Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ is a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$ associated to a kernel $k \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ if $k(\cdot, x) \in \mathcal{H}_k$ for all $x \in \mathcal{X}$ and the reproducing property $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}_k}$ holds for all $f \in \mathcal{H}_k$ and all $x \in \mathcal{X}$.[2] Let $\mathcal{H}_k^d = \mathcal{H}_k \times \cdots \times \mathcal{H}_k$ be the product RKHS with inner product $\langle \mathbf{f}, \mathbf{g}\rangle_{\mathcal{H}_k^d} = \sum_{j=1}^d \langle f_j, g_j\rangle_{\mathcal{H}_k}$ for $\mathbf{f} = (f_j)_{j=1}^d, \mathbf{g} = (g_j)_{j=1}^d \in \mathcal{H}_k^d$. For positive sequences $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$, (i) $a_n = \mathcal{O}(b_n)$ if there exist $C > 0$ and $n_0 \in \mathbb{N}_{>0}$ such that $a_n \le C b_n$ for all $n \ge n_0$, (ii) $a_n = \Omega(b_n)$ if $b_n = \mathcal{O}(a_n)$, (iii) $a_n = \Theta(b_n)$ if $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$. For a sequence of independent identically distributed (i.i.d.) real-valued random variables $(X_n)_{n=1}^\infty$, $X_n \sim P$ and a sequence of positive reals $(a_n)_{n=1}^\infty$ ($a_n > 0$ for all n), $X_n = \mathcal{O}_P(a_n)$ means that $\left(\frac{X_n}{a_n}\right)_{n=1}^\infty$ is bounded in probability.

3 KERNEL STEIN DISCREPANCY

We now introduce our quantity of interest, the kernel Stein discrepancy (KSD). To simplify exposition, we split the presentation into the Langevin-Stein KSD (with domain $\mathcal{X} = \mathbb{R}^d$; Section 3.1) and into the more abstract case of KSD on general domains $\mathcal{X}$ (Section 3.2); our results presented in Section 6 are structured similarly.

3.1 Langevin-Stein KSD on $\mathbb{R}^d$

Recall that we aim to compare a known and fixed distribution P_0 to an unknown distribution P, of which one obtains samples. Throughout this section, we assume that $P_0 \in \mathcal{M}_1^+(\mathbb{R}^d)$ and $P \in \mathcal{M}_1^+(\mathbb{R}^d)$. Also, assume that P_0 and P are absolutely continuous w.r.t. the Lebesgue measure with pdfs p_0 and p, respectively. One can tackle this problem by constructing a goodness-of-fit measure, such as Langevin-Stein KSD (Chwialkowski et al., 2016; Liu et al., 2016), which we detail below.

KSD is a specific IPM; indeed, considering $\mathcal{F} = \left\{\mathcal{A}_{p_0}\mathbf{f} : \mathbf{f} \in B\big(\mathcal{H}_k^d\big)\right\}$,

$$\begin{aligned} \mathrm{KSD}(P_0, P) &= \sup_{f\in\mathcal{F}} \left|\mathbb{E}_{P_0}[f(X)] - \mathbb{E}_P[f(X)]\right| \\ &= \sup_{\mathbf{f}\in B(\mathcal{H}_k^d)} \left|\mathbb{E}_{P_0}[(\mathcal{A}_{p_0}\mathbf{f})(X)] - \mathbb{E}_P[(\mathcal{A}_{p_0}\mathbf{f})(X)]\right|, \quad (1) \end{aligned}$$

where the operator $\mathcal{A}_{p_0}$ is constructed to guarantee the mean-zero property (Gorham and Mackey 2015; Chwialkowski et al. 2016; Liu et al. 2016)

$$\mathbb{E}_{P_0}[(\mathcal{A}_{p_0}\mathbf{f})(X)] = 0 \text{ for all } \mathbf{f} \in B\big(\mathcal{H}_k^d\big)\,; \qquad (2)$$

this property, using the symmetry of $B\big(\mathcal{H}_k^d\big)$ [in other words, $\mathbf{f} \in B\big(\mathcal{H}_k^d\big) \implies -\mathbf{f} \in B\big(\mathcal{H}_k^d\big)$], simplifies (1) to

$$\mathrm{KSD}(P_0, P) = \sup_{\mathbf{f}\in B(\mathcal{H}_k^d)} \mathbb{E}_P[(\mathcal{A}_{p_0}\mathbf{f})(X)]. \qquad (3)$$

One well-known operator satisfying (2) is the so-called Langevin-Stein operator (Gorham and Mackey, 2015; Chwialkowski et al., 2016; Liu et al., 2016; Oates et al., 2017; Gorham and Mackey, 2017), defined for $\mathbf{f} = (f_j)_{j=1}^d \in \mathcal{H}_k^d$ as

$$(\mathcal{A}_{p_0}\mathbf{f})(\mathbf{x}) = \langle \nabla_{\mathbf{x}} \ln\big(p_0(\mathbf{x})\big), \mathbf{f}(\mathbf{x})\rangle_2 + \sum_{j=1}^d \frac{\partial f_j(\mathbf{x})}{\partial x_j}. \quad (4)$$

Notice that the computation of $\mathcal{A}_{p_0}$ relies on $\nabla_{\mathbf{x}} \ln\big(p_0(\mathbf{x})\big)$, hence one assumes that $p_0(\mathbf{x}) > 0$ for all

[2] $k(\cdot, x)$ denotes the function $x' \in \mathcal{X} \mapsto k(x', x) \in \mathbb{R}$ while keeping $x \in \mathcal{X}$ fixed.

$\mathbf{x} \in \mathbb{R}^d$ (written shortly as $p_0 > 0$)—this dependence means that it is sufficient to know p_0 up to a constant multiplier—and that p_0 is differentiable. For (2) to hold, one requires that $\lim_{\|\mathbf{x}\|_2 \to \infty} h(\mathbf{x})p_0(\mathbf{x}) = 0$ for all $h \in \mathcal{H}_k$ (Liu et al., 2016, Lemma 2.2); for this condition it is sufficient if p_0 is bounded and $\lim_{\|\mathbf{x}\|_2 \to \infty} h(\mathbf{x}) = 0$ for all $h \in \mathcal{H}_k$.

One can show (Chwialkowski et al., 2016, Theorem 2.2) that KSD is a valid goodness-of-fit measure in the sense of

$$\operatorname{KSD}(P_0, P) = 0 \iff P_0 = P \tag{5}$$

under mild conditions, particularly if the base kernel k is c_0-universal (Carmeli et al., 2010; Sriperumbudur et al., 2010a). In the KSD construction (and throughout the paper when considering the Langevin-Stein KSD), we assume that the base kernel k is twice continuously differentiable [$k \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d)$]. Indeed, regarding (4), by the reproducing property for kernel derivatives (Zhou 2008, Theorem 1; Aubin-Frankowski and Szabó 2022, Lemma 1), one can write $(\mathcal{A}_{p_0}\mathbf{f})(\mathbf{x})$ as an inner product

$$(\mathcal{A}_{p_0}\mathbf{f})(\mathbf{x}) = \langle \mathbf{f}, \xi_{p_0}(\mathbf{x}) \rangle_{\mathcal{H}_k^d}, \tag{6}$$

$$\mathcal{H}_k^d \ni \xi_{p_0}(\mathbf{x}) \coloneqq \nabla_{\mathbf{x}} \left[\ln\bigl(p_0(\mathbf{x})\bigr)\right] k(\cdot, \mathbf{x}) + \nabla_{\mathbf{x}} k(\cdot, \mathbf{x}), \tag{7}$$

for all $\mathbf{f} \in \mathcal{H}_k^d$ and $\mathbf{x} \in \mathbb{R}^d$, which gives rise to the alternative form of KSD:

$$\operatorname{KSD}(P_0, P) \overset{\text{(a)}}{=} \sup_{\mathbf{f} \in B(\mathcal{H}_k^d)} \mathbb{E}_P\left[\langle \mathbf{f}, \xi_{p_0}(X) \rangle_{\mathcal{H}_k^d}\right]$$

$$\overset{\text{(b)}}{=} \sup_{\mathbf{f} \in B(\mathcal{H}_k^d)} \langle \mathbf{f}, \mathbb{E}_P[\xi_{p_0}(X)] \rangle_{\mathcal{H}_k^d} \overset{\text{(c)}}{=} \|\mathbb{E}_P[\xi_{p_0}(X)]\|_{\mathcal{H}_k^d}, \tag{8}$$

where (a) is implied by (3) and (6), (b) comes from swapping the inner product and the expectation (Steinwart and Christmann, 2008, (A.32)), and (c) follows from the Cauchy–Bunyakovsky–Schwarz (CBS) inequality.

The Stein kernel $K_0 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined based on ξ_{p_0} as $K_0(\mathbf{x}, \mathbf{y}) = \langle \xi_{p_0}(\mathbf{x}), \xi_{p_0}(\mathbf{y}) \rangle_{\mathcal{H}_k^d}$, ($\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$) which, by (7) and the reproducing property, takes the form

$$\begin{aligned} K_0(\mathbf{x}, \mathbf{y}) = {} & \bigl\langle \nabla_{\mathbf{x}} \ln\bigl(p_0(\mathbf{x})\bigr), \nabla_{\mathbf{y}} \ln\bigl(p_0(\mathbf{y})\bigr) \bigr\rangle_2 k(\mathbf{x}, \mathbf{y}) \\ & + \bigl\langle \nabla_{\mathbf{y}} \ln\bigl(p_0(\mathbf{y})\bigr), \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{y}) \bigr\rangle_2 \\ & + \bigl\langle \nabla_{\mathbf{x}} \ln\bigl(p_0(\mathbf{x})\bigr), \nabla_{\mathbf{y}} k(\mathbf{x}, \mathbf{y}) \bigr\rangle_2 \\ & + \sum_{j=1}^d \frac{\partial^2 k(\mathbf{x}, \mathbf{y})}{\partial x_j \partial y_j}. \end{aligned} \tag{9}$$

We assume that $p_0 \in \mathcal{C}^1(\mathbb{R}^d)$, which, together with the assumed property that $k \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d)$, implies the continuity of K_0 and, in turn, the separability of $\mathcal{H}_{K_0}$ (Steinwart and Christmann, 2008, Lemma 4.33). The following assumption summarizes our requirements for the Langevin-Stein KSD (i.e., the domain $\mathcal{X} = \mathbb{R}^d$).

Assumption 1 (Langevin-Stein KSD)**.** *Let $P_0 \in \mathcal{M}_1^+\bigl(\mathbb{R}^d\bigr)$ and $k \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d)$. Assume that (i) P_0 is absolutely continuous w.r.t. the Lebesgue measure with corresponding density p_0, (ii) p_0 is continuously differentiable: $p_0 \in \mathcal{C}^1(\mathbb{R}^d)$, (iii) p_0 is positive: $p_0 > 0$, and (iv) $\lim_{\|\mathbf{x}\|_2 \to \infty} h(\mathbf{x})p_0(\mathbf{x}) = 0$ for all $h \in \mathcal{H}_k$.*

3.2 General KSD

The construction in the preceding section can be extended to a topological space $(\mathcal{X}, \tau_{\mathcal{X}})$ by considering $P_0, P \in \mathcal{M}_1^+(\mathcal{X})$, $\mathcal{H}$ a Hilbert space of functions on $\mathcal{X}$, and $\Psi_{P_0} : \mathcal{X} \to \mathcal{H}$ such that the mean-zero property

$$\mathbb{E}_{P_0}[\Psi_{P_0}(X)] = 0 \tag{10}$$

holds.[3] One can then define the Stein operator T_{P_0} on $\mathcal{H}$ as

$$\left(T_{P_0} f\right)(x) = \langle \Psi_{P_0}(x), f \rangle_{\mathcal{H}}, \quad (f \in \mathcal{H}, x \in \mathcal{X}); \tag{11}$$

the operator inherits the mean-zero property (10)

$$\mathbb{E}_{P_0}[\left(T_{P_0} f\right)(X)] = \langle \mathbb{E}_{P_0}[\Psi_{P_0}(X)], f \rangle_{\mathcal{H}} = 0, \tag{12}$$

seen by interchanging the inner product with the expectation and using that $\mathbb{E}_{P_0}[\Psi_{P_0}(X)] = 0$. The KSD of P_0 (assumed to be fixed and known) and the sampling measure P is then defined as the IPM

$$\begin{aligned} & \operatorname{KSD}(P_0, P) \coloneqq \\ & \quad \sup_{f \in B(\mathcal{H})} \Bigl| \underbrace{\mathbb{E}_{P_0}[\left(T_{P_0} f\right)(X)]}_{=0} - \mathbb{E}_P[\left(T_{P_0} f\right)(X)] \Bigr| \\ & \overset{\text{(a)}}{=} \sup_{f \in B(\mathcal{H})} \mathbb{E}_P[\left(T_{P_0} f\right)(X)] && (13) \\ & \overset{(11)}{=} \sup_{f \in B(\mathcal{H})} \mathbb{E}_P \left\langle \Psi_{P_0}(X), f \right\rangle_{\mathcal{H}} && (14) \\ & \overset{\text{(b)}}{=} \|\mathbb{E}_P[\Psi_{P_0}(X)]\|_{\mathcal{H}} && (15) \\ & \overset{(c),(d),(17)}{=} \sqrt{\mathbb{E}_{P \otimes P}[K_0(X, X')]} \\ & \overset{(c),(d),(e)}{=} \left\| \int_{\mathcal{X}} K_0(\cdot, x) \mathrm{d}P(x) \right\|_{\mathcal{H}_{K_0}} && (16) \end{aligned}$$

(a) follows from the homogeneity of T_{P_0} and the expectation, and using the symmetry of $B(\mathcal{H})$. (b) follows as in (8). We use that the norm in a Hilbert space is induced by its inner product in (c), the expectation

[3] The existence of the l.h.s. requires that $\mathbb{E}_{P_0} \|\Psi_{P_0}(X)\|_{\mathcal{H}} < \infty$ (Diestel and Uhl, 1977, Theorem 2).

and the inner product are swapped in (d) and the reproducing property (18) implies (e); we also used the definition

$$K_0(x,x') \coloneqq \langle \Psi_{P_0}(x), \Psi_{P_0}(x') \rangle_{\mathcal{H}} \quad (x, x' \in \mathcal{X}). \tag{17}$$

As K_0 is a kernel, there exists an associated RKHS $\mathcal{H}_{K_0}$ for which K_0 is the (reproducing) kernel. Hence, for any $x, x' \in \mathcal{X}$ it holds that

$$K_0(x,x') = \langle K_0(\cdot,x), K_0(\cdot,x') \rangle_{\mathcal{H}_{K_0}}. \tag{18}$$

We note that $\Psi_{P_0}(x) \in \mathcal{H}$ and $K_0(\cdot,x) \in \mathcal{H}_{K_0}$ $(x \in \mathcal{X})$ but both yield the same Stein kernel K_0 [by (17) and (18)].

We collect our requirements for the general KSD in the following assumption.

Assumption 2 (General KSD)**.** *Assume that $(\mathcal{X}, \tau_{\mathcal{X}})$ is a topological space. Let $P_0 \in \mathcal{M}_1^+(\mathcal{X})$ and $\Psi_{P_0} : \mathcal{X} \to \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space. Let $K_0(x,y) = \langle \Psi_{P_0}(x), \Psi_{P_0}(y) \rangle_{\mathcal{H}}$ for $x, y \in \mathcal{X}$. Suppose that (i) Ψ_{P_0} is measurable, (ii) $\mathbb{E}_{P_0}[\Psi_{P_0}(X)] = 0$, and (iii) $\mathcal{H}_{K_0}$ is separable.*

We note that the measurability of $x \mapsto \Psi_{P_0}(x)$ for all $x \in \mathcal{X}$ is sufficient to guarantee the measurability of K_0 and $K_0(\cdot,x)$ $(x \in \mathcal{X})$ by the assumed separability of $\mathcal{H}_{K_0}$ (Steinwart and Christmann, 2008, Lemma 4.25). Further, $\mathbb{E}_{P_0}[\Psi_{P_0}(X)] = 0$ implies that $\mathbb{E}_{P_0}[K_0(\cdot,X)] = 0$ by the equality of (15) and (16).

Taking $\mathcal{X} = \mathbb{R}^d$, $\mathcal{H} = \mathcal{H}_k^d$, and $\Psi_{P_0}(\mathbf{x}) = \xi_{p_0}(\mathbf{x}) = \nabla_{\mathbf{x}}\big[\ln\big(p_0(\mathbf{x})\big)\big]k(\cdot,\mathbf{x}) + \nabla_{\mathbf{x}} k(\cdot,\mathbf{x}) \in \mathcal{H}_k^d$, where $\mathcal{H}_k$ is an RKHS with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, recovers the Langevin-Stein KSD on $\mathbb{R}^d$, derived independently in Section 3.1.

Besides Langevin-Stein KSD, the general construction detailed in this section encompasses, for example, KSD on Riemannian manifolds and KSD on Hilbert spaces (Hagrass et al., 2025, Example 2 and Example 3).

Remark 1. *KSDs can also be defined using the Petti's integral (Barp et al., 2024), which, for simplicity, we do not consider in this paper.*

4 KSD ESTIMATORS

In this section, we recall two existing KSD estimators with established convergence rates alongside their computational complexity. Let $X_{1:n} = (X_1, \ldots, X_n)$ be an i.i.d. sample from P (shortly, $X_{1:n} \sim P^n$) from which $\mathrm{KSD}(P_0, P)$ is estimated.

The squared KSD can be written in the form

$$\begin{aligned} \mathrm{KSD}^2(P_0,P) &\overset{(15)}{=} \|\mathbb{E}_P[\Psi_{P_0}(X)]\|_{\mathcal{H}}^2 \\ &\overset{(a),(b),(c),(b),(a)}{=} \|\mathbb{E}_P[K_0(\cdot,X)]\|_{\mathcal{H}_{K_0}}^2 \\ &\overset{(a),(b),(d)}{=} \mathbb{E}_{P\otimes P}[K_0(X,X')]. \end{aligned} \tag{19}$$

By making use of the fact that in a Hilbert space the norm is induced by the inner product in (a), swapping the expectation and the inner product in (b), using that $K_0(x,y) = \langle \Psi_{P_0}(x), \Psi_{P_0}(y) \rangle_{\mathcal{H}_k^d} = \langle K_0(\cdot,x), K_0(\cdot,y) \rangle_{\mathcal{H}_{K_0}}$ for all $x, y \in \mathcal{X}$ in (c), and leveraging the reproducing property in (d). We refer to $x \in \mathcal{X} \mapsto K_0(\cdot,x) \in \mathcal{H}_{K_0}$ as the Stein feature map.

V-statistic estimator. Replacing P in (19) with the empirical measure $\hat{P}_n = \frac{1}{n}\sum_{j=1}^n \delta_{X_j}$ yields the V-statistic-based KSD estimator (Chwialkowski et al., 2016) $\widehat{\mathrm{KSD}}_V^2(P_0,P) \coloneqq \mathrm{KSD}^2\big(P_0,\hat{P}_n\big) = \frac{1}{n^2}\sum_{i,j=1}^n K_0(X_i,X_j)$. This estimator has runtime complexity $\mathcal{O}\big(n^2\big)$ and under a sub-Gaussian assumption on the Stein feature map, one can show (Kalinke et al., 2025) that it has a convergence rate

$$\left|\widehat{\mathrm{KSD}}_V(P_0,P) - \mathrm{KSD}(P_0,P)\right| = \mathcal{O}_{P^n}\left(n^{-1/2}\right).$$

Nyström-KSD estimator. Recently, the Nyström technique has been adapted to design an accelerated KSD estimator (Kalinke et al., 2025). The idea of the approach is to consider a subsample (the sampling is carried out with replacement) $\{\{\tilde{X}_1, \ldots, \tilde{X}_m\}\}$ of the original sample $X_{1:n}$, giving rise to the subspace

$$\mathcal{H}_{K_0,m} = \mathrm{Span}\left(K_0\left(\cdot, \tilde{X}_j\right) \,:\, j \in [m]\right) \subset \mathcal{H}_{K_0}.$$

This subspace is then used to approximate $\mathbb{E}_{\hat{P}_n}[K_0(\cdot,X)]$ by taking the minimum norm solution of the optimization problem

$$\min_{\boldsymbol{\alpha}=(\alpha_j)_{j=1}^m \in \mathbb{R}^m} \left\| \mathbb{E}_{\hat{P}_n}[K_0(\cdot,X)] - \sum_{j=1}^m \alpha_j K_0\left(\cdot, \tilde{X}_j\right) \right\|_{\mathcal{H}_{K_0}},$$

attained by $\hat{\boldsymbol{\alpha}}$, resulting in the squared KSD estimator

$$\widehat{\mathrm{KSD}}_N^2(P_0,P) = \left\| \sum_{j=1}^m \hat{\alpha}_j K_0\left(\cdot, \tilde{X}_j\right) \right\|_{\mathcal{H}_{K_0}}^2.$$

The estimator can be computed as

$$\widehat{\mathrm{KSD}}_N^2(P_0,P) = \boldsymbol{\beta}^\top \mathbf{K}_{m,m}^- \boldsymbol{\beta}, \quad \boldsymbol{\beta} = \frac{1}{n}\mathbf{K}_{m,n}\mathbf{1}_n \in \mathbb{R}^m,$$

with the Gram matrices

$$\mathbf{K}_{m,m} = \left[K_0\left(\tilde{X}_a, \tilde{X}_b\right)\right]_{a,b=1}^{m} \in \mathbb{R}^{m\times m},$$
$$\mathbf{K}_{m,n} = \left[K_0\left(\tilde{X}_a, X_b\right)\right]_{a,b=1}^{m,n} \in \mathbb{R}^{m\times n}.$$

The runtime complexity of this estimator is $\mathcal{O}\left(mn + m^3\right)$. Under a sub-Gaussian assumption on the Stein feature map and given an appropriate spectral decay of its centered covariance operator and associated lower bound on m, the estimator achieves a convergence rate

$$\left|\widehat{\mathrm{KSD}}_N(P_0, P) - \mathrm{KSD}(P_0, P)\right| = \mathcal{O}_{P^n \otimes \Lambda^m}\left(n^{-1/2}\right),$$

with Λ^m encoding the Nyström sampling.

The main result of this paper is that no KSD estimator can achieve faster convergence rate than $n^{-1/2}$, specifically showing that the V-statistic and the Nyström-KSD estimators are rate-optimal.[4]

5 MINIMAX ESTIMATION

Before presenting our results, let us recall the framework of **minimax estimation** in our context. Our goal is to estimate $\mathrm{KSD}(P_0, P)$ based on samples $X_{1:n} \sim P^n$, given a target P_0. An estimator, denoted by $\hat{F}_n = \hat{F}_n(X_{1:n})$, is any (measurable) real-valued function of the observed data $X_{1:n}$ that approximates $\mathrm{KSD}(P_0, P)$. The performance of an estimator $\hat{F}_n$ (referred to as risk) is defined as the expected absolute difference between the estimate and the true value:

$$r_n\left(\hat{F}_n, P_0, P\right) = \mathbb{E}_{P^n}\left|\hat{F}_n(X_{1:n}) - \mathrm{KSD}(P_0, P)\right|.$$

However, a good estimator should perform well not just for a single P_0 and P, but uniformly well over a range of plausible distributions. This leads to a worst-case analysis, where one considers the maximum risk of an estimator over a large class of (P_0, P)-pairs. Indeed, we let $\mathcal{T}$ be the set of probability measures such that any $P_0 \in \mathcal{T}$ satisfies Assumption 1 for a fixed base kernel k in the case of Langevin-Stein KSD (resp. satisfies Assumption 2 in the general case), guaranteeing that KSD is well-defined. To each P_0, we associate the sampling probability measures $\mathcal{S}_{P_0}$ for which $\mathrm{KSD}(P_0, P)$ is finite for any $P \in \mathcal{S}_{P_0}$:

$$\begin{aligned}\mathcal{S}_{P_0} &\coloneqq \left\{P \in \mathcal{M}_1^+(\mathcal{X}) \,:\, \mathrm{KSD}(P_0, P) < \infty\right\}\\ &\overset{(\dagger)}{=} \left\{P \in \mathcal{M}_1^+(\mathcal{X}) : \mathbb{E}_P\sqrt{K_0(X, X)} < \infty\right\}. \quad (20)\end{aligned}$$

$(\dagger)$ holds as by (19) and the properties of the Bochner integral, one has that

$$\begin{aligned}\mathrm{KSD}\left(P_0, P\right) = \left\|\mathbb{E}_P[K_0(\cdot, X)]\right\|_{\mathcal{H}_{K_0}} < \infty \iff\\ \infty > \mathbb{E}_P\left\|K_0(\cdot, X)\right\|_{\mathcal{H}_{K_0}}\\ \overset{(a)}{=} \mathbb{E}_P\sqrt{\langle K_0(\cdot, X), K_0(\cdot, X)\rangle_{\mathcal{H}_{K_0}}}\\ \overset{(b)}{=} \mathbb{E}_P\sqrt{K_0(X, X)}, \quad (21)\end{aligned}$$

where (a) follows from the fact that in a Hilbert space the norm is induced by the inner product, and (b) is implied by the reproducing property.

The maximum risk of an estimator $\hat{F}_n$ is its worst-case performance over the (P_0, P)-pairs so constructed:

$$\begin{aligned}R_n\left(\hat{F}_n\right) &= \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} r_n\left(\hat{F}_n, P_0, P\right)\\ &= \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} \mathbb{E}_{P^n}\left|\hat{F}_n\left(X_{1:n}\right) - \mathrm{KSD}(P_0, P)\right|. \quad (22)\end{aligned}$$

Note that we require two supremums in (22) due to the valid P-s depending on the choice of P_0.

Finally, the minimax risk R_n^* is the smallest possible maximum risk achievable by *any* estimator. The term “minimax” reflects this two-step logic: one first takes the *max*imum risk for a given estimator and then finds the estimator that *min*imizes this maximum risk. Formally, it is the infimum of the maximum risk over all possible estimators $\hat{F}_n$:

$$\begin{aligned}R_n^* &= \inf_{\hat{F}_n} R_n\left(\hat{F}_n\right)\\ &= \inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} \mathbb{E}_{P^n}\left|\hat{F}_n\left(X_{1:n}\right) - \mathrm{KSD}(P_0, P)\right|.\end{aligned}$$

The quantity R_n^* represents the intrinsic statistical difficulty of the estimation problem and our goal is to establish a lower bound on R_n^*. To achieve this goal, we apply Markov's inequality, obtaining, for any $s_n > 0$,

$$\begin{aligned}&s_n^{-1} R_n^* \geq\\ &\inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n\Big(\underbrace{\left|\hat{F}_n(X_{1:n}) - \mathrm{KSD}(P_0, P)\right|}_{=:\hat{\Delta}_n} \geq s_n\Big), \quad (23)\end{aligned}$$

and control the r.h.s. using Le Cam's two-point method (outlined in Theorem 3). In the next section, we establish a positive lower bound on (23) with $s_n = \Theta\left(n^{-1/2}\right)$, implying lower bounds for the minimax risk of KSD estimation. Further, recalling from Section 4 that known KSD estimation rates are $\mathcal{O}(s_n)$ with $s_n = n^{-1/2}$, our results settle the statistical optimality of these estimators.

[4] One can also obtain the same $n^{-1/2}$ convergence rate (up to logarithmic factors) in the context of distribution compression with KSD (Li et al., 2024).

6 RESULTS

Next, we present our lower bounds on the minimax estimation of KSD, both for the Langevin-Stein KSD on $\mathbb{R}^d$ (Section 6.1) and for general domains (Section 6.2).

6.1 Langevin-Stein KSD

In this section, we consider $\mathcal{X} = \mathbb{R}^d$ with the usual topology and K_0 as in (9). Before stating our result, we make the following assumption, which, with the continuity of k, implies that k has a Bochner representation (detailed in Theorem C.1).

Assumption 3 (Langevin-Stein KSD; additional kernel assumptions)**.** *Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel. Assume that k is translation-invariant ($\exists$ positive definite κ such that $k(\mathbf{x}, \mathbf{y}) = \kappa(\mathbf{x} - \mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$).*[5]

Remark 2. *Kernels satisfying Assumptions 1 and 3 are, for example, the IMQ kernel and the whole class of twice-differentiable Matérn kernels, in particular, Gaussian kernels.*

Our result on the minimax lower bound of Langevin-Stein KSD reads as follows.

Theorem 1 (minimax lower bound of Langevin-Stein KSD)**.** *Suppose that Assumptions 1 and 3 hold, and that k is characteristic. Let $\hat{F}_n$ be any estimator of* $\mathrm{KSD}(P_0, P)$ *using $n \in \mathbb{N}_{>0}$ samples from $P \in \mathcal{S}_{P_0}$ ($P_0 \in \mathcal{T}$), where $\mathcal{S}_{P_0}$ is defined in* (20) *with $\mathcal{X} = \mathbb{R}^d$. Then, there exists a universal constant $c > 0$ such that*

$$\inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n\left(\hat{\Delta}_n \geq \frac{c}{\sqrt{n}}\right) > 0, \qquad (24)$$

with $\hat{\Delta}_n$ as defined in (23)*. In particular, by* (23)*, $n^{1/2} c^{-1} R_n^* > 0$.*

Remark 3.

(i) Note that the characteristic property of k is sufficient; we do not require c_0-universality [as discussed below (5)*] for Theorem 1 to hold.*

(ii) This result shows that the minimax lower bound of KSD estimation on $\mathbb{R}^d$ is $s_n = \Theta\big(n^{-1/2}\big)$, and specifically establishes the rate optimality of the V-statistic and Nyström-based KSD estimators given their matching rate of convergence recalled in Section 4.

For a Gaussian base kernel k, our following corollary makes the constant $c > 0$ explicit.

Corollary 1. *In the setting of Theorem 1, suppose that $k(\mathbf{x}, \mathbf{y}) = e^{-\gamma \|\mathbf{x} - \mathbf{y}\|_2^2}$ for some $\gamma > 0$ ($\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$). Then,* (24) *holds with $c = (4\gamma + 1)^{-d/4}/2$.*

Note that the constant c presented in the corollary increases exponentially with the dimension d, highlighting that the difficulty of KSD estimation can increase exponentially with d.

6.2 General KSD

In this section, $(\mathcal{X}, \tau_{\mathcal{X}})$ is a topological space and we impose the following additional assumption, ensuring that (i) KSD is valid for at least one P_0 and that (ii) $(\mathcal{X}, \tau_{\mathcal{X}})$ is sufficiently equipped with continuous bounded functions (used throughout the proof; see Lemma B.4).

Assumption 4 (General KSD; weak validity)**.** *Assume that, for at least one $P_0 \in \mathcal{T}$, KSD is valid in the sense of* (5) *for all $P \in \mathcal{S}_{P_0}$. In other words, for all $P \in \mathcal{S}_{P_0}$, $P \neq P_0$ iff.* $\mathrm{KSD}(P_0, P) > 0$*. Further assume that there exists a $\varphi_0 \in \mathcal{C}_b(\mathcal{X})$ such that there exists no $c \in \mathbb{R}$ such that $\varphi_0 = c$ holds P_0-almost surely.*

Our minimax lower bound result for general KSD is as follows.

Theorem 2 (minimax lower bound of general KSD)**.** *Let Assumptions 2 and 4 hold. Then, there exists a constant $B > 0$ such that*

$$\liminf_{n \to \infty} \inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n\left(\hat{\Delta}_n \geq \frac{B}{\sqrt{n}}\right) > 0,$$

with $\hat{\Delta}_n$ as defined in (23)*. In particular, by* (23)*, $\liminf_{n \to \infty} n^{1/2} B^{-1} R_n^* > 0$.*

Remark 4.

(i) This result shows that the minimax lower bound of KSD estimation on a general topological space $(\mathcal{X}, \tau_{\mathcal{X}})$ is $n^{-1/2}$, given Assumptions 2 and 4; in other words, no KSD estimator can achieve a faster rate in the minimax sense.

(ii) We also note that the bound in Theorem 1 is achieved for any $n \in \mathbb{N}_{>0}$, while Theorem 2 provides an asymptotic bound for the risk.

We proceed by sketching the main ideas of the proofs of our main results (Theorem 1 and Theorem 2), with the full proofs deferred to the appendices.

7 PROOF SKETCHES

Both of our results use Le Cam's two-point method. The core idea of this technique is to reduce the problem

[5] Note that translation-invariance implies the boundedness of the kernel ($\sup_{\mathbf{x} \in \mathbb{R}^d} \sqrt{k(\mathbf{x}, \mathbf{x})} < \infty$).

of finding a lower bound over a large class of distributions $P_0 \in \mathcal{T}$ and $P \in \mathcal{S}_{P_0}$ to the problem of finding a carefully crafted adversarial sequence of distributions; the key technical challenge and one contribution of our work is the construction of this adversarial sequence. Le Cam's two-point approach, following directly from Tsybakov (2009, (2.9) and Theorem 2.2), is as follows.

Theorem 3 (Theorem 2.2; Tsybakov 2009)**.** *Let $\mathcal{Y}$ be a measurable space, (Θ, d) a semi-metric space, and $\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\}$ a class of probability measures on $\mathcal{Y}$ indexed by Θ. We observe data $D \sim P_\theta \in \mathcal{P}_\Theta$ with some unknown parameter θ. The goal is to estimate θ. Let $\hat{\theta} = \hat{\theta}(D)$ be an estimator of θ based on D. Assume that there exist $\theta_0, \theta_1 \in \Theta$ such that $d(\theta_0, \theta_1) \geq 2s > 0$ and $\mathrm{KL}(P_{\theta_1} || P_{\theta_0}) \leq \alpha < \infty$ for $\alpha > 0$. Then*

$$\inf_{\hat{\theta}} \sup_{\theta \in \Theta} P_\theta\big(d(\hat{\theta}, \theta) \geq s\big) \geq f(\alpha),$$

with $f(\alpha) := \max\big\{ \exp(-\alpha)/4, (1 - \sqrt{\alpha/2})\big\} > 0$.

We now elaborate the main ideas behind our results.

7.1 Proof Sketch for Theorem 1

After recalling from (23) that

$$R_n^* \geq \inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n\big(\hat{\Delta}_n \geq C\big)$$

by Markov's inequality, and noticing that all Gaussian distributions are in $\mathcal{T}$ for a bounded k, we first obtain the bound $R_n^* \geq \inf_{\hat{F}_n} \sup_{P \in \mathcal{S}_{P_0}} P^n\big(\hat{\Delta}_n \geq C\big)$, with $P_0 = \mathcal{N}\big(\mathbf{0}_d, \mathbf{I}_d\big)$ now fixed. Using this probabilistic form, we proceed by applying Le Cam's method. In our case, this boils down to designing an adversarial distribution pair $(\mathbb{P}, \mathbb{Q})$ such that

$$\big|\,\mathrm{KSD}(P_0, \mathbb{P}) - \mathrm{KSD}(P_0, \mathbb{Q})\big| \geq \frac{2c}{\sqrt{n}} \qquad (25)$$

while $\mathrm{KL}\left(\mathbb{Q}^n \| \mathbb{P}^n\right) \leq \alpha$ with $0 < c, \alpha < \infty$.

To achieve this goal, we let $\mathbb{P} = \mathcal{N}\big(n^{-1/2}\mathbf{e}_j, \mathbf{I}_d\big)$ and $\mathbb{Q} = \mathcal{N}\big(\mathbf{0}_d, \mathbf{I}_d\big) = P_0$. Notice that, in this case, $\mathrm{KSD}(P_0, \mathbb{Q}) = \mathrm{KSD}(P_0, P_0) = 0$, and thus (25) reduces to $\mathrm{KSD}(P_0, \mathbb{P}) \geq 2c/\sqrt{n}$.

Controlling the distance. To control the distance, we rely on two auxiliary lemmas. Our first lemma shows that in case of a standard normal target P_0, $\mathrm{KSD}(P_0, P)$ can be expressed in terms of the characteristic function of the sampling distribution P, if P satisfies weak moment conditions.

Lemma 1 (KSD in terms of characteristic functions)**.** *Suppose that Assumption 3 holds. Further assume that $k \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d)$. Let k have Bochner representation $k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y}, \boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega})$. Let $P_0 = \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d) \in \mathcal{M}_1^+(\mathbb{R}^d)$ and suppose $P \in \mathcal{M}_1^+(\mathbb{R}^d)$ is such that $M_{\boldsymbol{\alpha}}^P < \infty$ for all $|\boldsymbol{\alpha}| \leq 2$ ($\boldsymbol{\alpha} \in \mathbb{N}_0^d$). Then it holds that*

$$\mathrm{KSD}^2(P_0, P) = \int_{\mathbb{R}^d} \|\nabla_{\boldsymbol{\omega}} \psi_P(\boldsymbol{\omega}) + \boldsymbol{\omega}\psi_P(\boldsymbol{\omega})\|_{\mathbb{C}^d}^2 \,\mathrm{d}\Lambda(\boldsymbol{\omega}).$$

Remark 5. *Wynne et al. (2025, Example 4.2) contains a similar expression of the KSD. However, their formula relies on the characteristic functions of the target (ψ_{P_0}) and the sampling distribution (ψ_P).*

If P is a multivariate Gaussian, Lemma 1 simplifies as shown in the following corollary.

Lemma 2 (Lemma 1 with $P = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$)**.** *In the setting of Lemma 1, let $P = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then KSD is*

$$\mathrm{KSD}^2(P_0, P) = \int_{\mathbb{R}^d} \left(\|\boldsymbol{\mu}\|_2^2 + \|\boldsymbol{\omega} - \boldsymbol{\Sigma}\boldsymbol{\omega}\|_2^2\right) \|\psi_P(\boldsymbol{\omega})\|_{\mathbb{C}}^2 \,\mathrm{d}\Lambda(\boldsymbol{\omega}).$$

Therefore, by using that $\mathrm{KSD}(P_0, \mathbb{P}) = 0$ and invoking Lemma 2, we obtain

$$\mathrm{KSD}(P_0, \mathbb{P}) = n^{-1} \int_{\mathbb{R}^d} \|\psi_{\mathbb{P}}(\boldsymbol{\omega})\|_{\mathbb{C}}^2 \,\mathrm{d}\Lambda(\boldsymbol{\omega}),$$

which, after establishing the positivity of the integral (by the characteristic property of k) and taking the positive square root on both sides, implies (25).

Controlling the KL divergence. Utilizing the known expressions for the KL divergence of product measures and the KL divergence of Gaussians (recalled in Lemma C.1 and Lemma C.2, respectively) yields $\mathrm{KL}(\mathbb{Q}^n \| \mathbb{P}^n) \leq 1/2 =: \alpha$ for all $n \in \mathbb{N}_{>0}$.

We conclude by invoking Theorem 3 using both controlled quantities.

7.2 Proof Sketch for Theorem 2

Let P_0 be as in Assumption 4. The proof starts by observing that $\{P_0\} \subset \mathcal{T}$ implies

$$\begin{aligned} R_n^* &= \inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} C^{-1}\mathbb{E}_{P^n}\big[\hat{\Delta}_n\big] \\ &\geq \inf_{\hat{F}_n} \sup_{P \in \mathcal{S}_{P_0}} P^n\big(\hat{\Delta}_n \geq C\big). \end{aligned}$$

Then, we obtain a lower bound by applying Le Cam's method with the adversarial distribution pair $\mathbb{P} = P_0$ and $\mathbb{Q} = P_n$. P_n is defined as a perturbation of P_0:[6]

$$P_n(A) = \int_A 1 + \epsilon_n \varphi(x) \mathrm{d}P_0(x), \quad \forall A \in \mathcal{B}(\mathcal{X}), (26)$$

[6] Similar perturbations are known in the testing literature (Anderson et al., 1994).

where $\varphi \in \mathcal{C}_b(\mathcal{X})$, $\mathbb{E}_{P_0}[\varphi(X)] = 0$, φ is not identically zero P_0-almost surely, and $\epsilon_n = cn^{-1/2}$ with $c > 0$.[7]

We start by showing that P_n belongs to $\mathcal{S}_{P_0}$. Then, to apply Le Cam's method, we establish that $|\operatorname{KSD}(P_0, \mathbb{P}) - \operatorname{KSD}(P_0, \mathbb{Q})| \geq 2cn^{-1/2}$ and that $\operatorname{KL}(\mathbb{P}^n \| \mathbb{Q}^n) \leq \alpha$, with $0 < \alpha < \infty$.

P_n is a probability measure. $P_n(\mathcal{X}) = 1$ holds by the definition of P_n. To show that P_n is non-negative, it suffices to note that

$$1 + \epsilon_n \varphi(x) \geq 1 + \epsilon_n L \text{ for all } x \in \mathcal{X}, \tag{27}$$

where $L = \inf_{x \in \mathcal{X}} \varphi(x) \in (-\infty, 0]$. The r.h.s. of (27) is non-negative for n large enough; hence, P_n is non-negative for n large enough.

$\operatorname{KSD}(P_0, P_n) < \infty$. Rewriting $\mathbb{E}_{P_n}\sqrt{K_0(X,X)} = \int_{\mathcal{X}} \sqrt{K_0(x,x)}\mathrm{d}P_0(x) + \epsilon_n \int_{\mathcal{X}} \sqrt{K_0(x,x)}\varphi(x)\mathrm{d}P_0(x)$, the first integral is finite by (21); the finiteness of the second term follows by using that φ is bounded, (21), and $\epsilon_n < \infty$. As $P_n \in \mathcal{M}_1^+(\mathcal{X})$ and $\operatorname{KSD}(P_0, P_n) < \infty$, we have shown that $P_n \in \mathcal{S}_{P_0}$.

Controlling the distance. Recall from (15) that for all $P \in \mathcal{S}_{P_0}$, $\operatorname{KSD}(P_0, P) = \|\mathbb{E}_P \Psi_{P_0}(X)\|_{\mathcal{H}}$. Our specific choice of P_n [(26)] allows to write $\operatorname{KSD}(P_0, P_n) = \epsilon_n \|\mathbb{E}_{P_0} \varphi(X) \Psi_{P_0}(X)\|_{\mathcal{H}} =: \epsilon_n C_\varphi > 0$, where the positivity follows from (i) the assumed validity of KSD in Assumption 4 and (ii) $\epsilon_n > 0$.

Controlling the KL divergence. The definition of P_n implies that

$$\operatorname{KL}(P_n \| P_0) = \mathbb{E}_{P_0}\left[\left(1 + \epsilon_n \varphi(X)\right) \ln\left(1 + \epsilon_n \varphi(X)\right)\right].$$

Then, by the fact that $\ln(1 + x) \leq x$, when $x > -1$, we obtain the bound on the KL divergence

$$\operatorname{KL}(P_n \| P_0) \leq \epsilon_n \underbrace{\mathbb{E}_{P_0}[\varphi(X)]}_{=0} + \varepsilon_n^2 \underbrace{\mathbb{E}_{P_0}[\varphi^2(X)]}_{=:M} = cn^{-1}M.$$

Therefore, by the formula of the KL divergence of product measures (Lemma C.1), we get the bound $\operatorname{KL}(\mathbb{Q}^n, \mathbb{P}^n) = \operatorname{KL}(P_n^n \| P_0^n) = n \operatorname{KL}(P_n \| P_0) \leq cM < \infty$.

The proof concludes by invoking Theorem 3 with both controlled quantities.

Acknowledgments

The authors thank Lester Mackey for his constructive comments. FK is supported by the pilot program Core-Informatics of the Helmholtz Association (HGF).

[7] The existence of such φ is guaranteed by Lemma B.4.

References

Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E. Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, Lester Mackey, Chris J. Oates, Gesine Reinert, and Yvik Swan. Stein's method meets computational statistics: a review of some recent developments. *Statistical Science*, 38(1):120–139, 2023.

Niall H. Anderson, Peter Hall, and D. M. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 1994.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

Pierre-Cyril Aubin-Frankowski and Zoltán Szabó. Handling hard affine SDP shape constraints in RKHSs. *Journal of Machine Learning Research*, 23 (297):1–54, 2022.

Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88:190–206, 2004.

Alessandro Barp, Chris. J. Oates, Emilio Porcu, and Mark Girolami. A Riemann–Stein kernel method. *Bernoulli*, 28(4):2181 – 2208, 2022.

Alessandro Barp, Carl-Johann Simon-Gabriel, Mark Girolami, and Lester Mackey. Targeted separation and convergence with kernel discrepancies. *Journal of Machine Learning Research*, 25(378):1–50, 2024.

Jerome Baum, Heishiro Kanagawa, and Arthur Gretton. A kernel Stein test of goodness of fit for sequential models. In *International Conference on Machine Learning (ICML)*, pages 1936–1953, 2023.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer, 2004.

Patric Bonnier, Harald Oberhauser, and Zoltán Szabó. Kernelized cumulants: Beyond kernel mean embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11049–11074, 2023.

Claudio Carmeli, Ernesto De Vito, Alessandro Toigo, and Veronica Umanitá. Vector valued reproducing kernel Hilbert spaces and universality. *Analysis and Applications*, 8(1):19–61, 2010.

Louis H. Y. Chen. Stein's method of normal approximation: Some recollections and reflections. *The Annals of Statistics*, 49(4):1850–1863, 2021.

Wilson Ye Chen, Lester Mackey, Jackson Gorham, Francois-Xavier Briol, and Chris J. Oates. Stein points. In *International Conference on Machine Learning (ICML)*, pages 844–853, 2018.

Wilson Ye Chen, Alessandro Barp, Francois-Xavier Briol, Jackson Gorham, Mark Girolami, Lester Mackey, and Chris J. Oates. Stein point Markov chain Monte Carlo. In *International Conference on Machine Learning (ICML)*, pages 1011–1021, 2019.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning (ICML)*, pages 2606–2615, 2016.

Joseph Diestel and John J. Uhl, Jr. *Vector Measures.* American Mathematical Society, 1977.

John Duchi. Derivations for linear algebra and optimization. Technical report, Stanford University, 2007. (`https://ai.stanford.edu/~jduchi/projects/general_notes.pdf`).

Gerald B. Folland. *Real Analysis – Modern Techniques and Their Applications.* John Wiley & Sons, second edition, 1999.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 489–496, 2007.

Futoshi Futami, Zhenghang Cui, Issei Sato, and Masashi Sugiyama. Bayesian posterior approximation via greedy particle optimization. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 3606–3613, 2019.

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 226–234, 2015.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning (ICML)*, pages 1292–1301, 2017.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory (ALT)*, pages 63–78, 2005a.

Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(70):2075–2129, 2005b.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Omar Hagrass, Bharath Sriperumbudur, and Krishnakumar Balasubramanian. Minimax optimal goodness-of-fit testing with kernel Stein discrepancy. *Bernoulli*, 2025. (accepted; preprint: `https://arxiv.org/abs/2404.08278`).

Liam Hodgkinson, Robert Salomone, and Fred Roosta. The reproducing Stein kernel approach for post-hoc corrected sampling. Technical report, 2021. (`https://arxiv.org/abs/2001.09266`).

Florian Kalinke and Zoltán Szabó. The minimax rate of HSIC estimation for translation-invariant kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 108468–108489, 2024.

Florian Kalinke, Zoltán Szabó, and Bharath K. Sriperumbudur. Nyström kernel Stein discrepancy. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 388–396, 2025.

Heishiro Kanagawa, Wittawat Jitkrittum, Lester Mackey, Kenji Fukumizu, and Arthur Gretton. A kernel Stein test for comparing latent variable models. Technical report, 2020. (`https://arxiv.org/abs/1907.00586`).

Lev Klebanov. *N-Distances and Their Applications.* Charles University, Prague, 2005.

Anna Korba, Adil Salim, Michael Arbel, Giulia Luise, and Arthur Gretton. A non-asymptotic analysis for Stein variational gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4672–4682, 2020.

Anna Korba, Pierre-Cyril Aubin-Frankowski, Szymon Majewski, and Pierre Ablin. Kernel Stein discrepancy descent. In *International Conference on Machine Learning (ICML)*, pages 5719–5730, 2021.

Lingxiao Li, Raaz Dwivedi, and Lester Mackey. Debiased distribution compression. In *International Conference on Machine Learning (ICML)*, pages 27675–27731, 2024.

Jen Ning Lim, Makoto Yamada, Bernhard Schölkopf, and Wittawat Jitkrittum. Kernel Stein tests for multiple model comparison. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2243–2253, 2019.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2378–2386, 2016.

Qiang Liu and Dilin Wang. Stein variational gradient descent as moment matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8854–8863, 2018.

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests.

In *International Conference on Machine Learning (ICML)*, pages 276–284, 2016.

Zhaolu Liu, Robert L. Peach, Pedro A.M. Mediano, and Mauricio Barahona. Interaction measures, partition lattices and kernel tests for high-order interactions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 36991–37012, 2023.

Russell Lyons. Distance covariance in metric spaces. *The Annals of Probability*, 41:3284–3305, 2013.

Natsumi Makigusa. Two-sample test based on maximum variance discrepancy. *Communications in Statistics. Theory and Methods*, 53(15):5421–5438, 2024.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29:429–443, 1997.

Chris J. Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):5–31, 2018.

Novi Quadrianto, Le Song, and Alex Smola. Kernelized sorting. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1289–1296, 2009.

Mahtab Sarvmaili, Hassan Sajjad, and Ga Wu. Data-centric prediction explanation via kernelized Stein discrepancy. In *International Conference on Learning Representations (ICLR)*, 2025.

Zoltán Sasvári. *Multivariate Characteristic and Correlation Functions.* Walter de Gruyter & Co., 2013.

Antonin Schrab, Benjamin Guedj, and Arthur Gretton. KSD aggregated goodness-of-fit test. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 32624–32638, 2022.

Dino Sejdinovic, Arthur Gretton, and Wicher Bergsma. A kernel test for three-variable interactions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1124–1132, 2013a.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals of Statistics*, 41:2263–2291, 2013b.

Alexander Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *Algorithmic Learning Theory (ALT)*, pages 13–31, 2007.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 781–788, 2010a.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010b.

Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Berkeley Symposium on Mathematical Statistics and Probability*, pages 583–602, 1972.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines.* Springer, 2008.

Zoltán Szabó and Bharath K. Sriperumbudur. Characteristic and universal tensor product kernels. *Journal of Machine Learning Research*, 18(233):1–29, 2018.

Gábor J. Székely and Maria L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5: 1249–1272, 2004.

Gábor J. Székely and Maria L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93:58–80, 2005.

Gábor J. Székely and Maria L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3:1236–1265, 2009.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35:2769–2794, 2007.

Ilya Tolstikhin, Bharath K. Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximal mean discrepancy with radial kernels. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1930–1938, 2016.

Ilya Tolstikhin, Bharath K. Sriperumbudur, and Krikamol Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(86):1–47, 2017.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation.* Springer, 2009.

Congye Wang, Wilson Ye Chen, Heishiro Kanagawa, and Chris J. Oates. Stein Π-importance sampling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 71948–71994, 2023.

Holger Wendland. *Scattered Data Approximation.* Cambridge University Press, 2005.

George Wynne, Mikołaj J. Kasprzak, and Andrew B. Duncan. A Fourier representation of kernel Stein discrepancy with application to goodness-of-fit tests for measures on infinite dimensional Hilbert spaces. *Bernoulli*, 31(2):868–893, 2025.

Wenkai Xu and Takeru Matsuda. A Stein goodness-of-fit test for directional distributions. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 320–330, 2020.

Wenkai Xu and Takeru Matsuda. Interpretable Stein goodness-of-fit tests on Riemannian manifold. In *International Conference on Machine Learning (ICML)*, pages 11502–11513, 2021.

Wenkai Xu and Gesine Reinert. A Stein goodness-of-test for exponential random graph models. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 415–423, 2021.

Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *International Conference on Machine Learning (ICML)*, pages 5561–5570, 2018.

Jiasen Yang, Vinayak A. Rao, and Jennifer Neville. A Stein-Papangelou goodness-of-fit test for point processes. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 226–235, 2019.

Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1-2): 456–463, 2008.

Yang Zhou, Di-Rong Chen, and Wei Huang. A class of optimal estimators for the covariance operator in reproducing kernel Hilbert spaces. *Journal of Multivariate Analysis*, 169:166–178, 2019.

Abram A. Zinger, Ashot V. Kakosyan, and Lev B. Klebanov. A characterization of distributions by mean values of statistics and certain probabilistic metrics. *Journal of Soviet Mathematics*, 59(4):914–920, 1992.

V. Zolotarev. Probability metrics. *Theory of Probability and its Applications*, 28:278–302, 1983.

The Minimax Lower Bound of Kernel Stein Discrepancy Estimation: Supplementary Materials

A PROOFS

This section is dedicated to the proofs of our statements in the main text. The proof of Lemma 1 is in Appendix A.1, that of Lemma 2 is in Appendix A.2, and that of Theorem 1 is in Appendix A.3. Corollary 1 is proved in Appendix A.4. We prove the general KSD lower bound (Theorem 2) in Appendix A.5.

A.1 Proof of Lemma 1

Recall that $k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda(\boldsymbol{\omega})$ by Theorem C.1. Let $\Lambda' = \Lambda/\Lambda(\mathbb{R}^d)$ and note that $\Lambda' \in \mathcal{M}_1^+(\mathbb{R}^d)$. We first show that

$$M_{\boldsymbol{\alpha}}^{\Lambda'} < \infty, \tag{A.1}$$

with $|\boldsymbol{\alpha}| \leq 2$ and $\boldsymbol{\alpha} \in \mathbb{N}_0^d$, which we will use multiple times throughout the remaining proof. Indeed, notice that $k(\mathbf{x},\mathbf{y}) = \kappa(\mathbf{x}-\mathbf{y}) = \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda(\boldsymbol{\omega}) = \Lambda(\mathbb{R}^d)\psi_{\Lambda'}(\mathbf{y}-\mathbf{x})$; hence,

$$k \in \mathcal{C}^2(\mathbb{R}^d \times \mathbb{R}^d) \implies \kappa \in \mathcal{C}^2(\mathbb{R}^d) \implies \psi_{\Lambda'} \in \mathcal{C}^2(\mathbb{R}^d), \tag{A.2}$$

where the first implication holds by the composition of functions. Given that $\psi_{\Lambda'} \in \mathcal{C}^2(\mathbb{R}^d)$, the application of Theorem C.3 now yields (A.1).

To obtain the expression presented in Lemma 1, we rewrite KSD as

$$\begin{aligned}
\mathrm{KSD}^2(P_0,P) &\overset{\text{(a)}}{=} \mathbb{E}_{P\otimes P}K_0(X,Y)\\
&\overset{\text{(b)}}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \nabla_{\mathbf{x}} \log p_0(\mathbf{x}), \nabla_{\mathbf{y}} \log p_0(\mathbf{y})\rangle_{\mathbb{C}^d}\, k(\mathbf{x},\mathbf{y}) + \langle \nabla_{\mathbf{x}} \log p_0(\mathbf{x}), \nabla_{\mathbf{y}} k(\mathbf{x},\mathbf{y})\rangle_{\mathbb{C}^d}\\
&\qquad + \langle \nabla_{\mathbf{y}} \log p_0(\mathbf{y}), \nabla_{\mathbf{x}} k(\mathbf{x},\mathbf{y})\rangle_{\mathbb{C}^d} + \sum_{j=1}^d \frac{\partial^2 k(\mathbf{x},\mathbf{y})}{\partial x_j \partial y_j}\mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y})\\
&\overset{\text{(c)}}{=} \underbrace{\int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \mathbf{x},\mathbf{y}\rangle_{\mathbb{C}^d}\, k(\mathbf{x},\mathbf{y})\mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y})}_{=:t_1} - \underbrace{\int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \mathbf{x}, \nabla_{\mathbf{y}} k(\mathbf{x},\mathbf{y})\rangle_{\mathbb{C}^d}\, \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y})}_{=:t_2}\\
&\qquad - \underbrace{\int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \mathbf{y}, \nabla_{\mathbf{x}} k(\mathbf{x},\mathbf{y})\rangle_{\mathbb{C}^d}\, \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y})}_{=:t_3} + \underbrace{\int_{\mathbb{R}^d\times\mathbb{R}^d} \sum_{j=1}^d \frac{\partial^2 k(\mathbf{x},\mathbf{y})}{\partial x_j \partial y_j}\mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y})}_{=:t_4}\\
&\overset{\text{(d)}}{=} \underbrace{\int_{\mathbb{R}^d} \langle \nabla_{\boldsymbol{\omega}} \psi_P(-\boldsymbol{\omega}), \nabla_{\boldsymbol{\omega}} \psi_P(-\boldsymbol{\omega})\rangle_{\mathbb{C}^d}\, \mathrm{d}\Lambda(\boldsymbol{\omega})}_{\overset{\text{(A.3)}}{=} t_1} + \underbrace{\int_{\mathbb{R}^d} \langle \nabla_{\boldsymbol{\omega}} \psi_P(\boldsymbol{\omega}), \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\, \psi_P(-\boldsymbol{\omega})\mathrm{d}\Lambda(\boldsymbol{\omega})}_{\overset{\text{(A.4)}}{=} -t_2}\\
&\qquad + \underbrace{\int_{\mathbb{R}^d} \langle \nabla_{\boldsymbol{\omega}} \psi_P(-\boldsymbol{\omega}), \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\, \psi_P(\boldsymbol{\omega})\mathrm{d}\Lambda(\boldsymbol{\omega})}_{\overset{\text{(A.5)}}{=} -t_3} + \underbrace{\int_{\mathbb{R}^d} \left\|\boldsymbol{\omega}\right\|_2^2 \psi_P(-\boldsymbol{\omega})\psi_P(\boldsymbol{\omega})\mathrm{d}\Lambda(\boldsymbol{\omega})}_{\overset{\text{(A.6)}}{=} t_4}\\
&\overset{\text{(e)}}{=} \int_{\mathbb{R}^d} \langle \nabla_{\boldsymbol{\omega}} \psi_P(-\boldsymbol{\omega}), \nabla_{\boldsymbol{\omega}} \psi_P(-\boldsymbol{\omega})\rangle_{\mathbb{C}^d} + \langle \nabla_{\boldsymbol{\omega}} \psi_P(\boldsymbol{\omega}), \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\, \psi_P(-\boldsymbol{\omega})\\
&\qquad + \langle \nabla_{\boldsymbol{\omega}} \psi_P(-\boldsymbol{\omega}), \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\, \psi_P(\boldsymbol{\omega}) + \|\boldsymbol{\omega}\|_2^2 \psi_P(\boldsymbol{\omega})\psi_P(-\boldsymbol{\omega})\mathrm{d}\Lambda(\boldsymbol{\omega}) \overset{\text{(f)}}{=} \int_{\mathbb{R}^d} \left\|\nabla_{\boldsymbol{\omega}}\psi_P(\boldsymbol{\omega}) + \boldsymbol{\omega}\psi_P(\boldsymbol{\omega})\right\|_{\mathbb{C}^d}^2 \mathrm{d}\Lambda(\boldsymbol{\omega}),
\end{aligned}$$

with the following details. In (a), we use the definition of $\mathrm{KSD}^2(P_0, P)$ and in (b) the definition of K_0 [(9)]. Note that P_0 has (Lebesgue) density $p_0(\mathbf{x}) \propto e^{-\|\mathbf{x}\|_2^2/2}$ by assumption; to obtain (c), we use that $\nabla_{\mathbf{x}} \log p_0(\mathbf{x}) = -\mathbf{x}$ (resp. $\nabla_{\mathbf{y}} \log p_0(\mathbf{y}) = -\mathbf{y}$) together with linearity of the inner product and the expectation. We tackle terms t_1–t_4 separately below [in particular, we verify that we can (i) apply Fubini's theorem and (ii) flip the order of integration and differentiation, respectively] and combine them afterwards, to obtain (d). (e) follows from the linearity of the integration. Properties of the norm on $\mathbb{C}^d$ yield (f), as we show in the following. Indeed, abbreviate $\mathbf{z} = \nabla_{\boldsymbol{\omega}}\psi_P(\boldsymbol{\omega}) \in \mathbb{C}^d$, $\boldsymbol{\omega} \in \mathbb{R}^d$, and $c = \psi_P(\boldsymbol{\omega}) \in \mathbb{C}$. Then it follows that

$$\begin{aligned}
&\|\nabla_{\boldsymbol{\omega}}\psi_P(\boldsymbol{\omega}) + \boldsymbol{\omega}\psi_P(\boldsymbol{\omega})\|_{\mathbb{C}^d}^2 = \|\mathbf{z} + c\boldsymbol{\omega}\|_{\mathbb{C}^d}^2 \\
&\quad \overset{(a)}{=} \|\mathbf{z}\|_{\mathbb{C}^d}^2 + \|c\boldsymbol{\omega}\|_{\mathbb{C}^d}^2 + \langle \mathbf{z}, c\boldsymbol{\omega}\rangle_{\mathbb{C}^d} + \langle c\boldsymbol{\omega}, \mathbf{z}\rangle_{\mathbb{C}^d} \\
&\quad \overset{(b)}{=} \|\mathbf{z}\|_{\mathbb{C}^d}^2 + \|c\boldsymbol{\omega}\|_{\mathbb{C}^d}^2 + \langle \mathbf{z}, \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\,\overline{c} + \langle \boldsymbol{\omega}, \mathbf{z}\rangle_{\mathbb{C}^d}\, c \\
&\quad \overset{(c)}{=} \|\mathbf{z}\|_{\mathbb{C}^d}^2 + \|\boldsymbol{\omega}\|_{\mathbb{C}^d}^2\, \|c\|_{\mathbb{C}}^2 + \langle \mathbf{z}, \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\,\overline{c} + \langle \boldsymbol{\omega}, \mathbf{z}\rangle_{\mathbb{C}^d}\, c \\
&\quad \overset{(d)}{=} \|\mathbf{z}\|_{\mathbb{C}^d}^2 + \|\boldsymbol{\omega}\|_{\mathbb{C}^d}^2\, \|c\|_{\mathbb{C}}^2 + \langle \mathbf{z}, \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\,\overline{c} + \langle \overline{\mathbf{z}}, \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\, c \\
&\quad \overset{(e)}{=} \|\nabla_{\boldsymbol{\omega}}\psi_P(\boldsymbol{\omega})\|_{\mathbb{C}^d}^2 + \|\boldsymbol{\omega}\|_2^2\, \|\psi_P(\boldsymbol{\omega})\|_{\mathbb{C}}^2 + \langle \nabla_{\boldsymbol{\omega}}\psi_P(\boldsymbol{\omega}), \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\,\psi_P(-\boldsymbol{\omega}) + \langle \nabla_{\boldsymbol{\omega}}\psi_P(-\boldsymbol{\omega}), \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\,\psi_P(\boldsymbol{\omega}).
\end{aligned}$$

In (a), we used that the norm in $\mathbb{C}^d$ is induced by the inner product, (b) follows from linearity in the 1st argument and the conjugate-linearity in the 2nd argument of the complex inner product, (c) is implied by the homogeneity of norms, (d) holds by $\langle \boldsymbol{\omega}, \mathbf{z}\rangle_{\mathbb{C}^d} = \mathbf{z}^*\boldsymbol{\omega} = \sum_{j\in[d]} \overline{z}_j\omega_j = \sum_{j\in[d]} \overline{\omega}_j\overline{z}_j = \boldsymbol{\omega}^*\overline{\mathbf{z}} = \langle \overline{\mathbf{z}}, \boldsymbol{\omega}\rangle_{\mathbb{C}^d}$ using that $\boldsymbol{\omega} \in \mathbb{R}^d$, and, in (e), we substituted the abbreviated quantities and used that $\psi_P(-\boldsymbol{\omega}) = \overline{\psi_P(\boldsymbol{\omega})}$.

Term t_1. We rewrite the first term as

$$\begin{aligned}
t_1 &\overset{(a)}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \mathbf{x}, \mathbf{y}\rangle_{\mathbb{C}^d} \underbrace{\int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y}, \boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega})}_{=k(\mathbf{x},\mathbf{y})} \mathrm{d}(P\otimes P)(\mathbf{x}, \mathbf{y}) \\
&\overset{(b)}{=} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d\times\mathbb{R}^d} \langle \mathbf{x}, \mathbf{y}\rangle_{\mathbb{C}^d}\, e^{-i\langle \mathbf{x}-\mathbf{y}, \boldsymbol{\omega}\rangle_2} \mathrm{d}(P\otimes P)(\mathbf{x}, \mathbf{y})\mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&\overset{(c)}{=} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d\times\mathbb{R}^d} \left\langle \mathbf{x}e^{-i\langle \mathbf{x}, \boldsymbol{\omega}\rangle_2}, \mathbf{y}e^{-i\langle \mathbf{y}, \boldsymbol{\omega}\rangle_2}\right\rangle_{\mathbb{C}^d} \mathrm{d}(P\otimes P)(\mathbf{x}, \mathbf{y})\mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&\overset{(d)}{=} \int_{\mathbb{R}^d} \left\langle \int_{\mathbb{R}^d} \mathbf{x}e^{-i\langle \mathbf{x}, \boldsymbol{\omega}\rangle_2}\mathrm{d}P(\mathbf{x}), \int_{\mathbb{R}^d} \mathbf{y}e^{-i\langle \mathbf{y}, \boldsymbol{\omega}\rangle_2}\mathrm{d}P(\mathbf{y})\right\rangle_{\mathbb{C}^d} \mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&\overset{(e)}{=} \int_{\mathbb{R}^d} \langle i\nabla_{\boldsymbol{\omega}}\psi_P(-\boldsymbol{\omega}), i\nabla_{\boldsymbol{\omega}}\psi_P(-\boldsymbol{\omega})\rangle_{\mathbb{C}^d}\,\mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&\overset{(f)}{=} \int_{\mathbb{R}^d} \langle \nabla_{\boldsymbol{\omega}}\psi_P(-\boldsymbol{\omega}), \nabla_{\boldsymbol{\omega}}\psi_P(-\boldsymbol{\omega})\rangle_{\mathbb{C}^d}\,\mathrm{d}\Lambda(\boldsymbol{\omega}), \qquad \text{(A.3)}
\end{aligned}$$

where Bochner's theorem (recalled in Theorem C.1) implies (a). In (b), we use the linearity of the integral and apply Fubini's theorem to change the order of integration, which we validate in Lemma B.1(i) after recalling (A.1). The properties of the exponential function ($e^{-i\langle \mathbf{x}-\mathbf{y}, \boldsymbol{\omega}\rangle_2} = e^{-i\langle \mathbf{x}, \boldsymbol{\omega}\rangle_2}e^{i\langle \mathbf{y}, \boldsymbol{\omega}\rangle_2}$), the conjugate-linearity of the complex inner product in the second argument with the fact that $\overline{e^{iz}} = e^{-iz}$ for $z \in \mathbb{R}$, and the linearity of the inner product in the first argument yield (c). The integrals were swapped with the inner product in (d). Invoking Lemma B.2 [validated in (A.2)] on both arguments of the inner product yields (e), the linearity and the conjugate-linearity of the complex inner product in the first and the second argument, respectively, and using that $i\bar{i} = -i^2 = 1$ give (f).

Term t_2. We obtain the alternative expression of the second term

$$
\begin{aligned}
t_2 &\overset{(a)}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \left\langle \mathbf{x}, \nabla_{\mathbf{y}} \underbrace{\int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega})}_{=k(\mathbf{x},\mathbf{y})} \right\rangle_{\mathbb{C}^d} \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y}) \\
&\overset{(b)}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \left\langle \mathbf{x}, i\int_{\mathbb{R}^d} \boldsymbol{\omega} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega}) \right\rangle_{\mathbb{C}^d} \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y}) \\
&\overset{(c)}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \int_{\mathbb{R}^d} \left\langle \mathbf{x}, i\boldsymbol{\omega} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \right\rangle_{\mathbb{C}^d} \mathrm{d}\Lambda(\boldsymbol{\omega}) \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y}) \\
&\overset{(d)}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \int_{\mathbb{R}^d} -i \left\langle \mathbf{x}, \boldsymbol{\omega}\right\rangle_{\mathbb{C}^d} e^{i\langle \mathbf{x},\boldsymbol{\omega}\rangle_2} e^{-i\langle \mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega}) \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y}) \\
&\overset{(e)}{=} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} -i \left\langle \mathbf{x} e^{i\langle \mathbf{x},\boldsymbol{\omega}\rangle_2}, \boldsymbol{\omega}\right\rangle_{\mathbb{C}^d} e^{-i\langle \mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}P(\mathbf{y}) \mathrm{d}P(\mathbf{x}) \mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&\overset{(f)}{=} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} -i \left\langle \mathbf{x} e^{i\langle \mathbf{x},\boldsymbol{\omega}\rangle_2}, \boldsymbol{\omega}\right\rangle_{\mathbb{C}^d} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}P(\mathbf{y}) \mathrm{d}P(\mathbf{x}) \mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&\overset{(g)}{=} \int_{\mathbb{R}^d} -i \Big\langle \underbrace{\int_{\mathbb{R}^d} \mathbf{x} e^{i\langle \mathbf{x},\boldsymbol{\omega}\rangle_2} \mathrm{d}P(\mathbf{x})}_{\overset{\text{Lemma B.2(i)}}{=} -i\nabla_{\boldsymbol{\omega}}\psi_P(\boldsymbol{\omega})}, \boldsymbol{\omega} \Big\rangle_{\mathbb{C}^d} \psi_P(-\boldsymbol{\omega}) \mathrm{d}\Lambda(\boldsymbol{\omega}) \overset{(h)}{=} \int_{\mathbb{R}^d} -\left\langle \nabla_{\boldsymbol{\omega}}\psi_P(\boldsymbol{\omega}), \boldsymbol{\omega}\right\rangle_{\mathbb{C}^d} \psi_P(-\boldsymbol{\omega}) \mathrm{d}\Lambda(\boldsymbol{\omega}),
\end{aligned}
\tag{A.4}
$$

where (a) follows by Bochner's theorem (recalled in Theorem C.1) and (b) is shown in Lemma B.3(ii). In (c), we swap the inner product with the integral. The conjugate-linearity of the complex inner product in its second argument, the facts that $\overline{e^{iz}} = e^{-iz}$ $(z \in \mathbb{R})$ and $e^{i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} = e^{i\langle \mathbf{x},\boldsymbol{\omega}\rangle_2} e^{-i\langle \mathbf{y},\boldsymbol{\omega}\rangle_2}$ are used in (d). For (e), it suffices to apply Fubini's theorem, validated in Lemma B.1(ii) by using (A.1), to use the product structure of $P\otimes P$, and then the linearity of the complex inner product in its first argument. The linearity of the integration gives (f). By the definition of characteristic function, the linearity of integration, and exchanging the inner product and the integral, we obtain (g). Lemma B.2(i) [validated in (A.2)], the linearity of the complex inner product in the first argument, and $i^2 = -1$ yield (h).

Term t_3. Similarly to t_2, we have

$$
\begin{aligned}
t_3 &\overset{(a)}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \left\langle \mathbf{y}, \nabla_{\mathbf{x}} \underbrace{\int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega})}_{=k(\mathbf{x},\mathbf{y})} \right\rangle_{\mathbb{C}^d} \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y}) \\
&\overset{(b)}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \left\langle \mathbf{y}, -i\int_{\mathbb{R}^d} \boldsymbol{\omega} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega}) \right\rangle_{\mathbb{C}^d} \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y}) \\
&\overset{(c)}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \int_{\mathbb{R}^d} \left\langle \mathbf{y}, -i\boldsymbol{\omega} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \right\rangle_{\mathbb{C}^d} \mathrm{d}\Lambda(\boldsymbol{\omega}) \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y}) \\
&\overset{(d)}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \int_{\mathbb{R}^d} i \left\langle \mathbf{y}, \boldsymbol{\omega}\right\rangle_{\mathbb{C}^d} e^{i\langle \mathbf{x},\boldsymbol{\omega}\rangle_2} e^{-i\langle \mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega}) \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y}) \\
&\overset{(e)}{=} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} i \left\langle \mathbf{y} e^{-i\langle \mathbf{y},\boldsymbol{\omega}\rangle_2}, \boldsymbol{\omega}\right\rangle_{\mathbb{C}^d} e^{i\langle \mathbf{x},\boldsymbol{\omega}\rangle_2} \mathrm{d}P(\mathbf{x}) \mathrm{d}P(\mathbf{y}) \mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&\overset{(f)}{=} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} i \left\langle \mathbf{y} e^{-i\langle \mathbf{y},\boldsymbol{\omega}\rangle_2}, \boldsymbol{\omega}\right\rangle_{\mathbb{C}^d} \int_{\mathbb{R}^d} e^{i\langle \mathbf{x},\boldsymbol{\omega}\rangle_2} \mathrm{d}P(\mathbf{x}) \mathrm{d}P(\mathbf{y}) \mathrm{d}\Lambda(\boldsymbol{\omega}) \\
&\overset{(g)}{=} \int_{\mathbb{R}^d} i \Big\langle \underbrace{\int_{\mathbb{R}^d} \mathbf{y} e^{-i\langle \mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}P(\mathbf{y})}_{\overset{\text{Lemma B.2(ii)}}{=} i\nabla_{\boldsymbol{\omega}}\psi_P(-\boldsymbol{\omega})}, \boldsymbol{\omega} \Big\rangle_{\mathbb{C}^d} \psi_P(\boldsymbol{\omega}) \mathrm{d}\Lambda(\boldsymbol{\omega}) \overset{(h)}{=} \int_{\mathbb{R}^d} -\left\langle \nabla_{\boldsymbol{\omega}}\psi_P(-\boldsymbol{\omega}), \boldsymbol{\omega}\right\rangle_{\mathbb{C}^d} \psi_P(\boldsymbol{\omega}) \mathrm{d}\Lambda(\boldsymbol{\omega}),
\end{aligned}
\tag{A.5}
$$

where (a) follows by Bochner's theorem (recalled in Theorem C.1), (b) is shown in Lemma B.3(i). The integration is swapped with the inner product in (c). (d) follows from the conjugate-linearity of the complex inner product in the second argument, $\overline{e^{iz}} = e^{-iz}$ $(z \in \mathbb{R})$ and $e^{i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} = e^{i\langle \mathbf{x},\boldsymbol{\omega}\rangle_2} e^{-i\langle \mathbf{y},\boldsymbol{\omega}\rangle_2}$. Fubini's theorem, validated in Lemma B.1(iii) by using (A.1), the product structure of $P\otimes P$, and the linearity of the complex inner product in its first argument yield (e). The linearity of the integration gives (f). By the definition of characteristic function, the linearity of integration, and exchanging the inner product and the integral, we obtain (g). Lemma B.2(ii) [validated in (A.2)], the linearity of the complex inner product in the first argument, and $i^2 = -1$ yield (h).

Term t_4. Last, we rewrite t_4 as

$$
\begin{aligned}
t_4 &= \int_{\mathbb{R}^d\times\mathbb{R}^d} \sum_{j=1}^d \frac{\partial^2 k(\mathbf{x},\mathbf{y})}{\partial x_j \partial y_j} \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y}) \overset{\text{(a)}}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \sum_{j=1}^d \int_{\mathbb{R}^d} \boldsymbol{\omega}^{2\mathbf{e}_j} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega}) \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y}) \\
&\overset{\text{(b)}}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d} \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_2^2 \, e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega}) \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y}) \\
&\overset{\text{(c)}}{=} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\boldsymbol{\omega}\|_2^2 \, e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y}) \mathrm{d}\Lambda(\boldsymbol{\omega}) \overset{\text{(d)}}{=} \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_2^2 \, \psi_P(-\boldsymbol{\omega}) \psi_P(\boldsymbol{\omega}) \mathrm{d}\Lambda(\boldsymbol{\omega}).
\end{aligned}
\tag{A.6}
$$

Lemma B.3(iii) gives (a), while linearity of the integral and observing that $\sum_{j=1}^d \boldsymbol{\omega}^{2e_j} = \sum_{j=1}^d \omega_j^2 = \|\boldsymbol{\omega}\|_2^2$ yields (b). (c) follows by applying Fubini's theorem, verified in Lemma B.1(iv) by using (A.1). The product structure of $P\otimes P$, the property $e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} = e^{-i\langle \mathbf{x},\boldsymbol{\omega}\rangle_2} e^{i\langle \mathbf{y},\boldsymbol{\omega}\rangle_2}$, the linearity of the integration, and the definition of the characteristic function imply (d).

A.2 Proof of Lemma 2

As $P = \mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, it holds that $M_{\boldsymbol{\alpha}}^P < \infty$ for all $\boldsymbol{\alpha}\in\mathbb{N}_0^d$. Hence, by Lemma 1, we obtain that

$$
\mathrm{KSD}^2(P_0,P) = \int_{\mathbb{R}^d} \|\nabla_{\boldsymbol{\omega}} \psi_P(\boldsymbol{\omega}) + \boldsymbol{\omega}\psi_P(\boldsymbol{\omega})\|_{\mathbb{C}^d}^2 \, \mathrm{d}\Lambda(\boldsymbol{\omega}).
$$

Recall that the characteristic function of a multivariate normal is $\psi_P(\boldsymbol{\omega}) = e^{i\langle \boldsymbol{\mu},\boldsymbol{\omega}\rangle_2 - \frac{1}{2}\langle \boldsymbol{\omega},\boldsymbol{\Sigma}\boldsymbol{\omega}\rangle_2}$. Thus, $\nabla_{\boldsymbol{\omega}}\psi_P(\boldsymbol{\omega}) = (i\boldsymbol{\mu} - \boldsymbol{\Sigma}\boldsymbol{\omega}) e^{i\langle \boldsymbol{\mu},\boldsymbol{\omega}\rangle_2 - \frac{1}{2}\langle \boldsymbol{\omega},\boldsymbol{\Sigma}\boldsymbol{\omega}\rangle_2} = \mathbf{z}\,\psi_P(\boldsymbol{\omega})$, with $\mathbf{z} \coloneqq i\boldsymbol{\mu} - \boldsymbol{\Sigma}\boldsymbol{\omega}$. To obtain the stated expression, we rewrite the integrand as

$$
\begin{aligned}
\|\nabla_{\boldsymbol{\omega}}\psi_P(\boldsymbol{\omega}) + \boldsymbol{\omega}\psi_P(\boldsymbol{\omega})\|_{\mathbb{C}^d}^2 &= \|\mathbf{z}\psi_P(\boldsymbol{\omega}) + \boldsymbol{\omega}\psi_P(\boldsymbol{\omega})\|_{\mathbb{C}^d}^2 \overset{\text{(a)}}{=} \|\mathbf{z}+\boldsymbol{\omega}\|_{\mathbb{C}^d}^2 \, \|\psi_P(\boldsymbol{\omega})\|_{\mathbb{C}}^2 \\
&\overset{\text{(b)}}{=} \|i\boldsymbol{\mu} - \boldsymbol{\Sigma}\boldsymbol{\omega} + \boldsymbol{\omega}\|_{\mathbb{C}^d}^2 \, \|\psi_P(\boldsymbol{\omega})\|_{\mathbb{C}}^2 \overset{\text{(c)}}{=} \left(\|\boldsymbol{\mu}\|_2^2 + \|\boldsymbol{\omega} - \boldsymbol{\Sigma}\boldsymbol{\omega}\|_2^2\right) \|\psi_P(\boldsymbol{\omega})\|_{\mathbb{C}}^2 \,,
\end{aligned}
$$

In (a) we used the homogeneity of norms, (b) follows by the definition of $\mathbf{z}$, and using that $\|\mathbf{z}\|_{\mathbb{C}^d}^2 = \|\mathrm{Re}(\mathbf{z})\|_2^2 + \|\mathrm{Im}(\mathbf{z})\|_2^2$ yields (c).

A.3 Proof of Theorem 1

Fix $j\in[d]$, $n\in\mathbb{N}_{>0}$, and denote by $\mathcal{G} = \left\{\mathcal{N}(\rho\,\mathbf{e}_j, \mathbf{I}_d) : \ \rho \geq 0\right\} \subset \mathcal{M}_1^+(\mathbb{R}^d)$ a subset of the Gaussian measures on $\mathbb{R}^d$. As this family is parameterized by $\rho\geq 0$, we write $G_\rho \in \mathcal{G}$.[8] We proceed by lower bounding the l.h.s. of (24) and then applying Theorem 3. In particular, for any $C > 0$, we have

$$
\begin{aligned}
\inf_{\hat{F}_n} \sup_{P_0\in\mathcal{T}} \sup_{P\in\mathcal{S}_{P_0}} P^n\Big(\underbrace{\left|\mathrm{KSD}(P_0,P) - \hat{F}_n\right|}_{=\hat{\Delta}_n} > C\Big) &\overset{\text{(a)}}{\geq} \inf_{\hat{F}_n} \sup_{P_0\in\{G_0\}} \sup_{P\in\mathcal{S}_{P_0}} P^n\left(\left|\mathrm{KSD}(P_0,P) - \hat{F}_n\right| > C\right) \\
&\overset{\text{(b)}}{=} \inf_{\hat{F}_n} \sup_{P\in\mathcal{S}_{G_0}} P^n\left(\left|\mathrm{KSD}(G_0,P) - \hat{F}_n\right| > C\right) \overset{\text{(c)}}{\geq} \inf_{\hat{F}_n} \sup_{G\in\mathcal{G}} G^n\left(\left|\mathrm{KSD}(G_0,G) - \hat{F}_n\right| > C\right),
\end{aligned}
\tag{A.7}
$$

where we obtain (a) as $\mathcal{T} \supseteq \{G_0\}$ and (b) by noting that the supremum of a singleton is attained at its element. To prove the inclusion $\mathcal{S}_{G_0} \supseteq \mathcal{G}$ used in (c), we observe that for any $G\in\mathcal{G}$, we have

$$
\mathbb{E}_G\sqrt{K_0(X,X)} \overset{\text{(a)}}{=} \mathbb{E}_G \left\|K_0(\cdot,X)\right\|_{\mathcal{H}_{K_0}} \overset{\text{(b)}}{\leq} \left(\mathbb{E}_G \left\|K_0(\cdot,X)\right\|_{\mathcal{H}_{K_0}}^2\right)^{1/2} \overset{\text{(a)}}{=} \left(\mathbb{E}_G K_0(X,X)\right)^{1/2}.
$$

[8] Since k is bounded, all $f\in\mathcal{H}_k$ are bounded. Then, we have that $\lim_{\|\mathbf{x}\|_2\to\infty} g_0(\mathbf{x})f(\mathbf{x}) = 0$, with g_0 the density of G_0 w.r.t. the Lebesgue measure, implying that $G_0\in\mathcal{T}$.

(a) holds by the fact that in a Hilbert space the norm is induced by the inner product and by using the reproducing property, (b) is implied by Jensen's inequality. The final term satisfies the bound

$$\mathbb{E}_G K_0(X,X) \overset{(c)}{=} \int_{\mathbb{R}^d} \|\nabla_{\mathbf{x}} \log p_0(\mathbf{x})\|_2^2 \,\kappa(0)\mathrm{d}G(\mathbf{x}) \overset{(d)}{=} \int_{\mathbb{R}^d} \|\mathbf{x}\|_2^2 \,\kappa(0)\mathrm{d}G(\mathbf{x}) \overset{(e)}{<} \infty,$$

where the definition of K_0 implies (c), as $k(\mathbf{x},\mathbf{y}) = \kappa(\mathbf{x}-\mathbf{y}) = \kappa(0)$ is constant and thus its derivatives are zero. In (d), we recall (from the proof of Lemma 1 in Appendix A.1) that $\nabla_{\mathbf{x}} \log p_0(\mathbf{x}) = -\mathbf{x}$ as P_0 has (Lebesgue) density $p_0(\mathbf{x}) \propto e^{-\|\mathbf{x}\|_2^2/2}$ by assumption. Noticing that Gaussian G-s have finite second moments gives (e) and proves that $\mathbb{E}_G\sqrt{K_0(X,X)} < \infty$; hence, $G \in \mathcal{S}_{G_0}$, which was to be shown.

To bring ourselves into the setting of Theorem 3, we let $\mathcal{Y} = \left(\mathbb{R}^d\right)^n$, $\Theta = \{\theta_\rho = \mathrm{KSD}(G_0, G_\rho) \,:\, \rho \geq 0\}$, $d(x,y) = |x-y|$ $(x, y \in \mathbb{R})$, and $\mathcal{P}_\Theta = \left\{G_\rho^n \,:\, \rho \geq 0\right\} = \left\{G_\rho^n \,:\, G_\rho \in \mathcal{G}\right\} = \left\{G_\rho^n \,:\, \theta_\rho \in \Theta\right\}$ therein. Hence, the observed data $X_{1:n} \in \mathcal{Y}$ is distributed as $X_{1:n} \sim G_\rho^n \in \mathcal{P}_\Theta$ for some unknown $\theta_\rho \in \Theta$. Let $\hat{F}_n = \hat{F}_n(X_{1:n})$ be any estimator of $\mathrm{KSD}(G_0, G_\rho)$ based on the n samples $X_{1:n}$.

In this setting, we consider the adversarial pair $(\theta_\rho, \theta_0) = \left(\mathrm{KSD}(G_0,G_\rho), \mathrm{KSD}(G_0,G_0)\right) = \left(\mathrm{KSD}(G_0,G_\rho), 0\right)$ with our choice of $\rho = 1/\sqrt{n}$; it remains to lower bound $d(\theta_\rho, \theta_0)$ and to upper bound $\mathrm{KL}(G_\rho^n \| G_0^n)$.

(i) **Lower bound for $d(\theta_\rho, \theta_0)$.** We obtain for the squared distance that

$$\begin{aligned} d^2(\theta_\rho,\theta_0) &\overset{(a)}{=} \mathrm{KSD}^2(G_0,G_\rho) \overset{(b)}{=} \rho^2 \int_{\mathbb{R}^d} \left\|\psi_{G_\rho}(\boldsymbol{\omega})\right\|_{\mathbb{C}}^2 \mathrm{d}\Lambda(\boldsymbol{\omega}) \overset{(c)}{=} \rho^2 \int_{\mathbb{R}^d} e^{-\|\boldsymbol{\omega}\|_2^2} \mathrm{d}\Lambda(\boldsymbol{\omega}) \overset{(d)}{\geq} \rho^2 \int_A e^{-\|\boldsymbol{\omega}\|_2^2} \mathrm{d}\Lambda(\boldsymbol{\omega}) \\ &\overset{(e)}{\geq} \rho^2 \Lambda(A) \inf_{\boldsymbol{\omega}\in A} e^{-\|\boldsymbol{\omega}\|_2^2} \overset{(f)}{=} \rho^2\Lambda(A)e^{-\delta_0} \overset{(g)}{\geq} \rho^2 \underbrace{\Lambda(B)e^{-\delta_0}}_{=:4c^2} \overset{(h)}{=} \frac{4c^2}{n}, \end{aligned} \tag{A.8}$$

where our choice of (θ_ρ, θ_0) gives (a). (b) holds by Lemma 2, and (c) follows by recalling that $\psi_{\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}(\boldsymbol{\omega}) = e^{i\langle\boldsymbol{\mu},\boldsymbol{\omega}\rangle_2 - \frac{1}{2}\langle\boldsymbol{\omega},\boldsymbol{\Sigma}\boldsymbol{\omega}\rangle_2}$ implies that $\left\|\psi_{G_\rho}(\boldsymbol{\omega})\right\|_{\mathbb{C}}^2 = \psi_{G_\rho}(\boldsymbol{\omega})\overline{\psi_{G_\rho}(\boldsymbol{\omega})} = e^{-\|\boldsymbol{\omega}\|_2^2}$. We define the closed ball with fixed radius $0 < \delta_0 < \infty$, $A = \{\boldsymbol{\omega} \in \mathbb{R}^d \,:\, \|\boldsymbol{\omega}\|_2^2 \leq \delta_0\} \subset \mathbb{R}^d$, which is compact, and use the positivity of the exponential function with the monotonicity of the integral in (d). Considering the infimum of the integrand with the monotonicity of the integration, and the integration of constant functions gives (e). In (f), we use that a continuous function on a compact domain attains its infimum and the definition of A. Let $B \subset A$ be the interior of A; we then use the monotonicity of measures to obtain (g). Since k is characteristic, $\mathrm{supp}(\Lambda) = \mathbb{R}^d$ (Theorem C.4), implying that $\Lambda(B) > 0$ (as the interior B is open), ensuring that $c > 0$. (h) follows from our choice of $\rho = 1/\sqrt{n}$. Finally, taking the square root of (A.8), we have

$$d(\theta_\rho, \theta_0) \geq \frac{2c}{\sqrt{n}} =: 2s > 0. \tag{A.9}$$

(ii) **Upper bound for $\mathrm{KL}(G_\rho^n \| G_0^n)$.** We have the chain of equalities

$$\mathrm{KL}(G_\rho^n\|G_0^n) \overset{(a)}{=} \sum_{j=1}^n \mathrm{KL}(G_\rho\|G_0) \overset{(b)}{=} \frac{n}{2}\left(d + \rho^2\|\mathbf{e}_j\|_2^2 - d + \ln(1)\right) \overset{(c)}{=} \frac{1}{2},$$

where (a) holds by Lemma C.1 and (b) by Lemma C.2. In (c), we use our choice of ρ. Hence, letting $\alpha \coloneqq \frac{1}{2}$, we have

$$\mathrm{KL}(G_\rho^n\|G_0^n) \leq \alpha = \frac{1}{2}.$$

Then, by invoking Theorem 3, we obtain for (A.7) using $C = s = c/\sqrt{n}$, with s defined in (A.9), that

$$\inf_{\hat{F}_n}\sup_{G\in\mathcal{G}} G^n\left(\left|\mathrm{KSD}(G_0,G) - \hat{F}_n\right| > \frac{c}{\sqrt{n}}\right) \geq \max\left(\frac{e^{-1/2}}{4}, \frac{1-\sqrt{1/4}}{2}\right) = \frac{1}{4},$$

concluding the proof.

A.4 Proof of Corollary 1

By the proof of Theorem 1 (Appendix A.3), in particular (A.8) and (A.9), it is sufficient to make the dependence of s_n on our choice of $k(\mathbf{x},\mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|_2^2}$ ($\mathbf{x},\mathbf{y} \in \mathbb{R}^d$) explicit. We proceed in two steps:

(i) First, we obtain a closed-form expression for $\frac{\mathrm{d}\Lambda}{\mathrm{d}\lambda_d}$, with Λ corresponding to the spectral measure associated to the Gaussian kernel.

(ii) Second, we also obtain $d(\theta_\rho,\theta_0)$ in closed form, using the density obtained in (i), which will imply the stated result.

The details are as follows.

(i) **Closed-form of $\mathrm{d}\Lambda/\mathrm{d}\lambda_d$.** Recall that by Bochner's theorem (Theorem C.1),

$$k(\mathbf{x},\mathbf{y}) = \kappa(\mathbf{x}-\mathbf{y}) = \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} \cos\big(\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2\big)\mathrm{d}\Lambda(\boldsymbol{\omega}), \tag{A.10}$$

where the last equation is implied by Euler's formula ($e^{ix} = \cos(x) + i\sin(x)$ for $x \in \mathbb{R}$), the definition of the complex integral, and as k is real-valued. By Sriperumbudur et al. (2010b, (4) and Table 2) κ has Fourier transform $\mathcal{F}\kappa$ given by (with $\gamma = 1/(2\sigma^2)$ therein)

$$(\mathcal{F}\kappa)(\boldsymbol{\omega}) = \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} e^{-i\langle \mathbf{z},\boldsymbol{\omega}\rangle_2}\kappa(\mathbf{z})\mathrm{d}\mathbf{z} = \sigma^d e^{-\frac{\sigma^2\|\boldsymbol{\omega}\|_2^2}{2}}. \tag{A.11}$$

Using this expression, the Fourier inversion theorem now implies that

$$\begin{aligned} k(\mathbf{x},\mathbf{y}) = \kappa(\mathbf{x}-\mathbf{y}) = \mathcal{F}^{-1}(\mathcal{F}\kappa)(\mathbf{x}-\mathbf{y}) &= \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} e^{i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}(\mathcal{F}\kappa)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega} \\ &= \frac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d} \cos\big(\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2\big)(\mathcal{F}\kappa)(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}, \end{aligned} \tag{A.12}$$

where Euler's formula, κ and $\mathcal{F}\kappa$ being real-valued, and the definition of the complex integral imply the last expression.
As (A.10) and (A.12) are equal, we obtain that

$$\frac{\mathrm{d}\Lambda}{\mathrm{d}\lambda_d}(\boldsymbol{\omega}) = \frac{1}{(2\pi)^{d/2}}(\mathcal{F}\kappa)(\boldsymbol{\omega}) \overset{\text{(a)}}{=} \frac{\sigma^d}{(2\pi)^{d/2}}e^{-\frac{\sigma^2\|\boldsymbol{\omega}\|_2^2}{2}} \overset{\text{(b)}}{=} \frac{1}{(4\pi\gamma)^{d/2}}e^{-\frac{\|\boldsymbol{\omega}\|_2^2}{4\gamma}}, \tag{A.13}$$

by using the explicit form of $\mathcal{F}\kappa$ [(A.11)] in (a) and $\gamma = 1/(2\sigma^2)$ in (b).

(ii) **Closed-form of $d(\theta_\rho,\theta_0)$.** From (A.8)(c), we have

$$d^2(\theta_\rho,\theta_0) = \rho^2\int_{\mathbb{R}^d} e^{-\|\boldsymbol{\omega}\|_2^2}\mathrm{d}\Lambda(\boldsymbol{\omega}) \overset{\text{(a)}}{=} \rho^2 c_1\int_{\mathbb{R}^d} e^{-\|\boldsymbol{\omega}\|_2^2}e^{-\frac{\|\boldsymbol{\omega}\|_2^2}{4\gamma}}\mathrm{d}\boldsymbol{\omega} \overset{\text{(b)}}{=} \rho^2 c_1\int_{\mathbb{R}^d} e^{-c_2\|\boldsymbol{\omega}\|_2^2}\mathrm{d}\boldsymbol{\omega}, \tag{A.14}$$

with (a) following from (A.13) and letting $c_1 := 1/(4\pi\gamma)^{d/2}$, and in (b) setting $c_2 := 1 + \frac{1}{4\gamma}$. Recall that the Gaussian integral has closed-form solution $\int_{\mathbb{R}} e^{-ax^2}\mathrm{d}x = (\pi/a)^{1/2}$ for $a > 0$; hence

$$\int_{\mathbb{R}^d} e^{-c_2\|\boldsymbol{\omega}\|_2^2}\mathrm{d}\boldsymbol{\omega} = \prod_{j=1}^d \int_{\mathbb{R}} e^{-c_2\omega_j^2}\mathrm{d}\omega_j = \prod_{j=1}^d \left(\frac{\pi}{c_2}\right)^{1/2} = \left(\frac{\pi}{c_2}\right)^{d/2},$$

which, continuing from (A.14), gives

$$d^2(\theta_\rho,\theta_0) = \rho^2 c_1\left(\frac{\pi}{c_2}\right)^{d/2} \overset{\text{(a)}}{=} \rho^2\left(\frac{1}{4\pi\gamma}\frac{\pi}{1+\frac{1}{4\gamma}}\right)^{d/2} \overset{\text{(b)}}{=} \rho^2\left(\frac{1}{4\gamma+1}\right)^{d/2},$$

using our definitions of c_1 and c_2 in (a) and simplifying in (b).
Our choice of $\rho = 1/\sqrt{n}$ and taking the positive square root yields that

$$d(\theta_\rho,\theta_0) = \frac{1}{\sqrt{n}}\left(\frac{1}{4\gamma+1}\right)^{d/4} =: 2s.$$

Following the notation in (A.9), one gets that $c := (4\gamma+1)^{-d/4}/2$.

A.5 Proof of Theorem 2

Observe that, for a P_0' defined as in Assumption 4, we have

$$\inf_{\hat{F}_n} \sup_{P_0 \in \mathcal{T}} \sup_{P \in \mathcal{S}_{P_0}} P^n\Big(\underbrace{\left| \mathrm{KSD}(P_0, P) - \hat{F}_n \right|}_{=\hat{\Delta}_n} \geq C \Big) \overset{\text{(a)}}{\geq} \inf_{\hat{F}_n} \sup_{P_0 \in \{P_0'\}} \sup_{P \in \mathcal{S}_{P_0}} P^n\Big(\left| \mathrm{KSD}(P_0, P) - \hat{F}_n \right| \geq C \Big)$$
$$\overset{\text{(b)}}{=} \inf_{\hat{F}_n} \sup_{P \in \mathcal{S}_{P_0'}} P^n\Big(\left| \mathrm{KSD}(P_0', P) - \hat{F}_n \right| \geq C \Big), \tag{A.15}$$

where (a) comes by the fact that $\{P_0'\} \subset \mathcal{T}$ and (b) by noting that the supremum of a singleton is attained at its element. In the following, we relabel P_0' as P_0; in other words, we write $P_0 = P_0'$.

To bring ourselves into the setting of Theorem 3, for any fixed $n \in \mathbb{N}_{>0}$, set $\mathcal{Y} := (\mathbb{R}^d)^n$, $\Theta := \{\theta_P := \mathrm{KSD}(P_0, P) : P \in \mathcal{S}_{P_0}\}$, $\mathcal{P}_\Theta := \{P^n : P \in \mathcal{S}_{P_0}\} = \{P^n : \theta_P \in \Theta\}$, and $d(x, y) := |x - y|$ $(x, y \in \mathbb{R})$. Let us define $F : \mathcal{S}_{P_0} \to \mathbb{R}$ by $P \mapsto \mathrm{KSD}(P_0, P)$, and let $\hat{F}_n$ denote the corresponding estimator based on n samples. We construct $(P_{\theta_0(n)}, P_{\theta_1(n)})$ for fixed n, where $P_{\theta_0(n)} := P_{\theta_0}$ with $\theta_0 := \theta_{P_0}$, $P_{\theta_1(n)} := P_{\theta_n}$ with $\theta_n := \theta_{P_n}$ and P_n specified below in (A.16). With these notations at hand, $d(\theta_0(n), \theta_1(n)) = |\mathrm{KSD}(P_0, P_0) - \mathrm{KSD}(P_0, P_n)| = |0 - \mathrm{KSD}(P_0, P_n)| = \mathrm{KSD}(P_0, P_n)$.

Next, we present the **construction of the adversarial sequence** P_n. Let $\varphi \in \mathcal{C}_b(\mathcal{X})$ be as constructed in Lemma B.4, that is, (i) satisfying $\mathbb{E}_{P_0}[\varphi(X)] = 0$ and (ii) guaranteeing that there exists $A' \in \mathcal{B}(\mathcal{X})$ with positive P_0-measure such that $\varphi(x) \neq 0$ for all $x \in A'$. We construct P_n as a perturbation of P_0 taking the form

$$P_n(A) = \int_A 1 + \epsilon_n \varphi(x) \mathrm{d}P_0(x) \text{ for any } A \in \mathcal{B}(\mathcal{X}), \tag{A.16}$$

with $\epsilon_n = cn^{-1/2}$, where the precise value of $c > 0$ will be specified later; we also note that $P_n \neq P_0$ by Lemma B.5. (A.16) implies that $P_n \ll P_0$ and the corresponding Radon-Nikodym derivative takes the form

$$\frac{\mathrm{d}P_n}{\mathrm{d}P_0} = 1 + \epsilon_n \varphi. \tag{A.17}$$

We show that $P_n \in \mathcal{S}_{P_0}$ for sufficiently large n. Indeed:

1. $P_n \geq 0$ for $n \geq n_{0,1}$: Recalling from (A.16) that for $A \in \mathcal{B}(\mathcal{X})$

$$P_n(A) = \int_A 1 + \epsilon_n \varphi(x) \mathrm{d}P_0(x),$$

it suffices to show that $1 + \epsilon_n \varphi(x) \geq 0$ for all $x \in \mathcal{X}$ and n large enough. As $\varphi \in \mathcal{C}_b(\mathcal{X})$, φ is bounded and

$$L := \inf_{x \in \mathcal{X}} \varphi(x) > -\infty.$$

Further, by the construction of φ, $\mathbb{E}_{P_0}[\varphi(X)] = 0$; hence

$$0 = \mathbb{E}_{P_0}[\varphi(X)] = \int_{\mathcal{X}} \varphi(x) \mathrm{d}P_0(x) \overset{\text{(a)}}{\geq} \int_{\mathcal{X}} \inf_{x \in \mathcal{X}} \varphi(x) \mathrm{d}P_0(x) \overset{\text{(b)}}{=} LP_0(\mathcal{X}) \overset{\text{(c)}}{=} L;$$

in other words, $L \leq 0$. (a) holds by the monotonicity of the integration, (b) follows from the definition of L and the integration of constants, (c) comes from $P_0 \in \mathcal{M}_1^+(\mathcal{X})$.
For any $x \in \mathcal{X}$, it holds that $1 + \epsilon_n \varphi(x) \geq \inf_{x \in \mathcal{X}}[1 + \epsilon_n \varphi(x)] = 1 + \epsilon_n L$, and we are done once we establish that the last term is non-negative:

$$1 + \epsilon_n L \geq 0 \iff 1 - \epsilon_n |L| \geq 0 \iff 1 \geq \epsilon_n |L| \iff \frac{1}{\epsilon_n} \geq |L|,$$

where we used that the non-positivity of L means that $L = -|L|$. By using that $\epsilon_n = cn^{-1/2}$ with $c > 0$, we have that $1/\epsilon_n = n^{1/2}/c \to \infty$ as $n \to \infty$, guaranteeing $1/\epsilon_n \geq |L|$ for n large enough (say, $n \geq n_{0,1}$).

2. $P_n(\mathcal{X}) = 1$: One has

$$P_n(\mathcal{X}) \overset{\text{(a)}}{=} \int_{\mathcal{X}} 1 + \epsilon_n \varphi(x) \mathrm{d}P_0(x) \overset{\text{(b)}}{=} 1 + \epsilon_n \underbrace{\int_{\mathcal{X}} \varphi(x) \mathrm{d}P_0(x)}_{\overset{\text{(c)}}{=}0} = 1.$$

   (a) follows from the definition of P_n [(A.16)]; (b) is by the linearity of integration and using that $\int_{\mathcal{X}} 1 \mathrm{d}P_0(x) = P_0(\mathcal{X}) = 1$; (c) uses the mean-zero property of φ w.r.t. P_0.

3. $\mathbb{E}_{P_n}\sqrt{K_0(X,X)} < \infty$: One gets

$$\begin{aligned}\mathbb{E}_{P_n}\sqrt{K_0(X,X)} &\overset{\text{(a)}}{=} \int_{\mathcal{X}} \sqrt{K_0(x,x)}[1 + \epsilon_n\varphi(x)]\mathrm{d}P_0(x) \\ &\overset{\text{(b)}}{=} \underbrace{\int_{\mathcal{X}} \sqrt{K_0(x,x)}\mathrm{d}P_0(x)}_{=:t_1} + \epsilon_n \underbrace{\int_{\mathcal{X}} \sqrt{K_0(x,x)}\varphi(x)\mathrm{d}P_0(x)}_{=:t_2}.\end{aligned}$$

   The first step (a) is by the definition of the expectation and by the properties of the Radon-Nikodym derivative [(A.17)]. In (b), we use the linearity of the integral. Term t_1 is finite by applying (21) with $P = P_0$. For t_2, let $\sup_{x\in\mathcal{X}} |\varphi(x)| =: M < \infty$, where the finiteness of M holds by $\varphi \in \mathcal{C}_b(\mathcal{X})$. We have

$$\left|\int_{\mathcal{X}} \sqrt{K_0(x,x)}\varphi(x)\mathrm{d}P_0(x)\right| \overset{\text{(a)}}{\leq} \int_{\mathcal{X}} \sqrt{K_0(x,x)}|\varphi(x)|\mathrm{d}P_0(x) \overset{\text{(b)}}{\leq} M \int_{\mathcal{X}} \sqrt{K_0(x,x)}\mathrm{d}P_0(x) \overset{\text{(c)}}{=} Mt_1 \overset{\text{(d)}}{<} \infty,$$

   by applying in (a) Jensen's inequality and using the non-negativity of $\sqrt{K_0(x,x)}$ $(x \in \mathcal{X})$, in (b) the definition of M with the monotonicity and linearity of the integration, in (c) the definition of t_1, in (d) the finiteness of M and t_1.

Having defined P_n, we continue with the **control of the KSD value** $\mathrm{KSD}\left(P_0, P_n\right)$:

$$\begin{aligned}\mathrm{KSD}(P_0, P_n) &\overset{(15)}{=} \left\|\mathbb{E}_{P_n}[\Psi_{P_0}(X)]\right\|_{\mathcal{H}} \overset{\text{(a)}}{=} \left\|\mathbb{E}_{P_0}\left[\Psi_{P_0}(X)\big(1 + \epsilon_n\varphi(X)\big)\right]\right\|_{\mathcal{H}} \\ &\overset{\text{(b)}}{=} \big\|\underbrace{\mathbb{E}_{P_0}[\Psi_{P_0}(X)]}_{=0 \Longleftarrow (10)} + \epsilon_n\mathbb{E}_{P_0}[\varphi(X)\Psi_{P_0}(X)]\big\|_{\mathcal{H}} \overset{\text{(c)}}{=} \epsilon_n \underbrace{\left\|\mathbb{E}_{P_0}[\varphi(X)\Psi_{P_0}(X)]\right\|_{\mathcal{H}}}_{=:C_\varphi} \overset{\text{(d)}}{>} 0,\end{aligned}$$

where in (a) we used the definition of P_n and the property of the Radon-Nikodym derivative, (b) holds by the linearity of the expectation, (c) is implied by the homogeneity of norms and the positivity of ϵ_n, and (d) follows from the fact that $\epsilon_n > 0$ and that by $P_n \neq P_0$ we have $\mathrm{KSD}(P_0, P_n) > 0$ by the validity of KSD imposed in Assumption 4. Hence,

$$\mathrm{KSD}\left(P_0, P_n\right) = \epsilon_n C_\varphi \overset{\text{(a)}}{=} \Theta\Big(n^{-1/2}\Big), \tag{A.18}$$

where (a) holds by $\epsilon_n = cn^{-1/2}$ $(c > 0)$ and $C_\varphi > 0$.

We proceed by **controlling the KL divergence** $\mathrm{KL}\left(P_n \| P_0\right)$:

$$\mathrm{KL}\left(P_n \| P_0\right) \overset{\text{(a)}}{=} \mathbb{E}_{P_n} \ln\left[\frac{\mathrm{d}P_n}{\mathrm{d}P_0}(X)\right] \overset{\text{(b)}}{=} \mathbb{E}_{P_0}\left[(1 + \epsilon_n\varphi(X))\ln(1 + \epsilon_n\varphi(X))\right], \tag{A.19}$$

where in (a) the definition of the KL divergence was applied, (b) is implied by the definition of P_n [(A.17)] and the properties of the Radon-Nikodym derivative.

To gain control over the integral in (A.19), we recall that, for any $x > -1$, one has that $\ln(1 + x) \leq x$. Let n be large enough (say $n \geq n_{0,2}$) such that for all $x \in \mathcal{X}$ one has $|\epsilon_n\varphi(x)| < 1$; this is possible as φ is bounded. Then, we can upper bound (A.19) as

$$\begin{aligned}\mathbb{E}_{P_0}\big[\underbrace{(1 + \epsilon_n\varphi(X))}_{>0}\underbrace{\ln(1 + \epsilon_n\varphi(X))}_{\leq\epsilon_n\varphi(X)}\big] &\overset{\text{(a)}}{\leq} \mathbb{E}_{P_0}\left[(1 + \epsilon_n\varphi(X))\epsilon_n\varphi(X)\right] \overset{\text{(b)}}{=} \epsilon_n \underbrace{\mathbb{E}_{P_0}\left[\varphi(X)\right]}_{=0} + \epsilon_n^2 \underbrace{\mathbb{E}_{P_0}\left[\varphi^2(X)\right]}_{=:M_2<\infty} \\ &= M_2\epsilon_n^2 \overset{\text{(c)}}{=} \mathcal{O}(1/n).\end{aligned} \tag{A.20}$$

In (a), we use the monotonicity and in (b) the linearity of integration. The function φ has zero-mean w.r.t. P_0 by construction; it is also bounded, guaranteeing the finiteness of M_2. Our choice of $\epsilon_n = cn^{-1/2}$ yields (c) and we **choose** c in the following.

Indeed, from (A.20) and the definition of ϵ_n, one gets that

$$n\,\mathrm{KL}(P_n\|P_0) \le nM_2\frac{c^2}{n} = M_2c^2;$$

hence, by choosing $c \coloneqq \sqrt{\ln(2)}/\sqrt{M_2} > 0$, we arrive at

$$n\,\mathrm{KL}(P_n\|P_0) \le \ln(2).$$

Thus, for sufficiently large n (say $n \ge n_{0,2}$), the requirement $n\,\mathrm{KL}(P_n\|P_0) \le \ln(2) =: \alpha$ in Theorem 3 is fulfilled, and $n \ge n_0 \coloneqq \max(n_{0,1}, n_{0,2})$ incorporates all our n is large enough constraints. With our choice of c, by the definition of ϵ_n, (A.18) translates to

$$\mathrm{KSD}(P_0, P_n) = n^{-1/2}cC_\varphi =: 2s_n,$$

defining $s_n \coloneqq \frac{n^{-1/2}cC_\varphi}{2}$; $s_n > 0$ since $cC_\varphi > 0$ by $c > 0$ and $C_\varphi > 0$. Hence, Theorem 3 together with (A.15) implies that for all $n \ge n_0$

$$\inf_{\hat{F}_n}\sup_{P_0\in\mathcal{T}}\sup_{P\in\mathcal{S}_{P_0}} P^n\big(\hat{\Delta}_n \ge s_n\big) \ge f(\alpha),$$

with f defined in Theorem 3.[9] This means that for all $n \ge n_0$

$$\inf_{\hat{F}_n}\sup_{P_0\in\mathcal{T}}\sup_{P\in\mathcal{S}_{P_0}} P^n\Big(\hat{\Delta}_n \ge n^{-1/2}\underbrace{\frac{cC_\varphi}{2}}_{=:B>0}\Big) \ge f(\alpha),$$

which concludes the proof.

B AUXILIARY RESULTS

In this section, we collect a few auxiliary results. Lemma B.1 validates our applications of Fubini's theorem in the proof of Theorem 1. Lemma B.2 relates the gradient of a distribution's characteristic function to its moments. Lemma B.3 is about the derivatives of a continuous bounded translation-invariant kernel in terms of its Bochner representation. Lemma B.4 shows the existence of a bounded smooth perturbation function.

Lemma B.1 (Lebesgue integrability of key functions)**.** *Let $P \in \mathcal{M}_1^+(\mathbb{R}^d)$, Λ a finite non-negative measure on $\big(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d)\big)$, and $\Lambda' = \frac{\Lambda}{\Lambda(\mathbb{R}^d)}$.[10] Assume that for all $|\boldsymbol{\alpha}| \le 2$ with $\boldsymbol{\alpha} \in \mathbb{N}_0^d$, $M_{\boldsymbol{\alpha}}^P < \infty$ and $M_{\boldsymbol{\alpha}}^{\Lambda'} < \infty$. Then,*

(i) $\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \left|\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbb{C}^d}\, e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\right| \mathrm{d}(\Lambda\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y}) < \infty,$

(ii) $\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \left|\langle \mathbf{x}, \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\, e^{i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\right| \mathrm{d}(\Lambda\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y}) < \infty,$

(iii) $\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \left|\langle \mathbf{y}, \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\, e^{i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\right| \mathrm{d}(\Lambda\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y}) < \infty,$

(iv) $\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \left|\langle \boldsymbol{\omega}, \boldsymbol{\omega}\rangle_{\mathbb{C}^d}\, e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\right| \mathrm{d}(\Lambda\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y}) < \infty.$

Proof. We prove the finiteness of each integral separately.

[9] $f(\ln(2)) = \max\left(\frac{1}{8}, \frac{1-\sqrt{\frac{\ln(2)}{2}}}{2}\right) = \frac{1-\sqrt{\frac{\ln(2)}{2}}}{2} \approx 0.29.$

[10] This normalization implies that $\Lambda' \in \mathcal{M}_1^+\big(\mathbb{R}^d\big)$.

Integral (i). One has

$$\begin{aligned}
&\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \left|\langle \mathbf{x},\mathbf{y}\rangle_{\mathbb{C}^d}\, e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\right| \mathrm{d}(\Lambda\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y}) \overset{(a)}{=} \Lambda(\mathbb{R}^d)\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} |\langle \mathbf{x},\mathbf{y}\rangle_{\mathbb{C}^d}|\,\mathrm{d}(\Lambda'\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y})\\
&\overset{(b)}{\le} \Lambda(\mathbb{R}^d)\left[\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} |\langle \mathbf{x},\mathbf{y}\rangle_{\mathbb{C}^d}|^2\,\mathrm{d}(\Lambda'\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y})\right]^{1/2}\\
&\overset{(c)}{=} \Lambda\big(\mathbb{R}^d\big)\left[\int_{\mathbb{R}^d\times\mathbb{R}^d} |\langle \mathbf{x},\mathbf{y}\rangle_{\mathbb{C}^d}|^2\,\mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y})\right]^{1/2}\\
&\overset{(d)}{\le} \Lambda\big(\mathbb{R}^d\big)\left[\int_{\mathbb{R}^d\times\mathbb{R}^d} \|\mathbf{x}\|_2^2\,\|\mathbf{y}\|_2^2\,\mathrm{d}(P\otimes P)(\mathbf{x},\mathbf{y})\right]^{1/2} \overset{(e)}{=} \Lambda\big(\mathbb{R}^d\big)\left(\int_{\mathbb{R}^d}\|\mathbf{x}\|_2^2\,\mathrm{d}P(\mathbf{x})\right)^{1/2}\left(\int_{\mathbb{R}^d}\|\mathbf{y}\|_2^2\,\mathrm{d}P(\mathbf{y})\right)^{1/2}\\
&\overset{(f)}{=} \Lambda\big(\mathbb{R}^d\big)\left(\sum_{j=1}^d M^P_{2\mathbf{e}_j}\right)^{1/2}\left(\sum_{j=1}^d M^P_{2\mathbf{e}_j}\right)^{1/2} \overset{(g)}{<} \infty,
\end{aligned}$$

where (a) follows by noting that $\left|e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\right| = 1$ and the definition of Λ'. The monotonicity of L_p norms w.r.t. p with probability measures yields (b). In (c), we use the product structure of $\Lambda'\otimes P\otimes P$, and that $\Lambda'\big(\mathbb{R}^d\big) = 1$ as $\Lambda'\in \mathcal{M}_1^+\left(\mathbb{R}^d\right)$. To obtain (d), we apply the CBS inequality and that $\|\mathbf{x}\|_{\mathbb{C}^d} = \|\mathbf{x}\|_2$ and $\|\mathbf{y}\|_{\mathbb{C}^d} = \|\mathbf{y}\|_2$ when $\mathbf{x},\mathbf{y}\in\mathbb{R}^d$, (e) is by independence. (f) comes from the definition of $\|\cdot\|_2$, the linearity of integration, and the definition of $M^P_{\boldsymbol{\alpha}}$. (g) follows by observing that Bochner's theorem guarantees the finiteness of $\Lambda\big(\mathbb{R}^d\big)$ and since $M^P_{\boldsymbol{\alpha}} < \infty$ for $|\boldsymbol{\alpha}|\le 2$ by assumption.

Integral (ii). Observe that

$$\begin{aligned}
&\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \left|\langle \mathbf{x},\boldsymbol{\omega}\rangle_{\mathbb{C}^d}\, e^{i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\right| \mathrm{d}(\Lambda\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y}) \overset{(a)}{=} \Lambda(\mathbb{R}^d)\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} |\langle \mathbf{x},\boldsymbol{\omega}\rangle_{\mathbb{C}^d}|\,\mathrm{d}(\Lambda'\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y})\\
&\overset{(b)}{\le} \Lambda(\mathbb{R}^d)\left[\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} |\langle \mathbf{x},\boldsymbol{\omega}\rangle_{\mathbb{C}^d}|^2\,\mathrm{d}(\Lambda'\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y})\right]^{1/2} \overset{(c)}{=} \Lambda(\mathbb{R}^d)\left[\int_{\mathbb{R}^d\times\mathbb{R}^d} |\langle \mathbf{x},\boldsymbol{\omega}\rangle_2|^2\,\mathrm{d}(\Lambda'\otimes P)(\boldsymbol{\omega},\mathbf{x})\right]^{1/2}\\
&\overset{(d)}{\le} \Lambda(\mathbb{R}^d)\left(\int_{\mathbb{R}^d}\|\mathbf{x}\|_2^2\,\mathrm{d}P(\mathbf{x})\right)^{1/2}\left(\int_{\mathbb{R}^d}\|\boldsymbol{\omega}\|_2^2\,\mathrm{d}\Lambda'(\boldsymbol{\omega})\right)^{1/2} \overset{(e)}{=} \Lambda(\mathbb{R}^d)\left(\sum_{j=1}^d M^P_{2\mathbf{e}_j}\right)^{1/2}\left(\sum_{j=1}^d M^{\Lambda'}_{2\mathbf{e}_j}\right)^{1/2} \overset{(f)}{<} \infty, \qquad \text{(B.21)}
\end{aligned}$$

where (a) comes by noting that $|e^{i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}| = 1$ and the definition of Λ', and (b) by applying the monotonicity of L_p norms as in part (i). Noticing that $P(\mathbb{R}^d) = 1$ and that $\langle \mathbf{x},\boldsymbol{\omega}\rangle_{\mathbb{C}^d} = \langle \mathbf{x},\boldsymbol{\omega}\rangle_2$ for real vectors yields (c). To get (d), we apply the CBS inequality and independence. (e) follows from the definition of $\|\cdot\|_2$, the linearity of the integral, and by the definition of $M^P_{\boldsymbol{\alpha}}$. To obtain (f), note that (i) $\Lambda(\mathbb{R}^d) < \infty$ by Bochner's theorem, and (ii) $M^{\Lambda'}_{2\mathbf{e}_j} < \infty$ and $M^P_{2\mathbf{e}_j} < \infty$ for all $j\in[d]$ by assumption.

Integral (iii). We have

$$\begin{aligned}
&\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \left|\langle \mathbf{y},\boldsymbol{\omega}\rangle_{\mathbb{C}^d}\, e^{i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\right| \mathrm{d}(\Lambda\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y})\\
&\quad\overset{(a)}{=} \Lambda(\mathbb{R}^d)\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} |\langle \mathbf{y},\boldsymbol{\omega}\rangle_{\mathbb{C}^d}|\,\mathrm{d}(\Lambda'\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y}) \overset{(b)}{<} \infty
\end{aligned}$$

where (a) comes from $|\,\langle \mathbf{y},\boldsymbol{\omega}\rangle_{\mathbb{C}^d}\, e^{i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}| = |\,\langle \mathbf{y},\boldsymbol{\omega}\rangle_{\mathbb{C}^d}\,|$ and the definition of Λ'. With a change of the variables $\mathbf{y}$ and $\mathbf{x}$, (B.21) yields (b).

Integral (iv). We obtain bounds for the last integral by noting that $\langle \boldsymbol{\omega},\boldsymbol{\omega}\rangle_{\mathbb{C}^d} = \|\boldsymbol{\omega}\|^2_{\mathbb{C}} = \|\boldsymbol{\omega}\|_2^2$ for $\boldsymbol{\omega}\in\mathbb{R}^d$ and considering

$$\int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \left|\|\boldsymbol{\omega}\|_2^2\, e^{-i\langle\mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\right| \mathrm{d}(\Lambda\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y}) \overset{\text{(a)}}{=} \int_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \|\boldsymbol{\omega}\|_2^2\, \mathrm{d}(\Lambda\otimes P\otimes P)(\boldsymbol{\omega},\mathbf{x},\mathbf{y})$$
$$\overset{\text{(b)}}{=} \int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_2^2\, \mathrm{d}\Lambda(\boldsymbol{\omega}) \overset{\text{(c)}}{=} \Lambda(\mathbb{R}^d)\int_{\mathbb{R}^d} \|\boldsymbol{\omega}\|_2^2\, \mathrm{d}\Lambda'(\boldsymbol{\omega}) \overset{\text{(d)}}{=} \Lambda(\mathbb{R}^d)\int_{\mathbb{R}^d}\sum_{j=1}^{d} \boldsymbol{\omega}^{2\mathbf{e}_j}\mathrm{d}\Lambda'(\boldsymbol{\omega}) \overset{\text{(e)}}{=} \Lambda(\mathbb{R}^d)\sum_{j=1}^{d} M_{2\mathbf{e}_j}^{\Lambda'} \overset{\text{(f)}}{<} \infty,$$

where (a) uses that $|e^{iz}| = 1$ for any $z \in \mathbb{R}$. (b) follows from the product structure of $\Lambda\otimes P\otimes P$ and the property $P(\mathbb{R}^d) = 1$. Our definition of $\Lambda = \Lambda(\mathbb{R}^d)\Lambda'$ gives (c) and we make the definition of $\|\cdot\|_2^2$ explicit in (d). We swap the integral with the sum by using the linearity of the integration in (e) and use the notation for moments. (f) is implied by the assumed finiteness of $M_{2\mathbf{e}_j}^{\Lambda'}$ for all $j \in [d]$. □

Lemma B.2 (Gradient of characteristic function)**.** *Let $Q \in \mathcal{M}_1^+(\mathbb{R}^d)$ with characteristic function ψ_Q. If $D^{\mathbf{e}_j}\psi_Q$ exists for all $j \in [d]$, then for all $\boldsymbol{\omega} \in \mathbb{R}^d$, one has*

(i) $\nabla_{\boldsymbol{\omega}}\psi_Q(\boldsymbol{\omega}) = i\int_{\mathbb{R}^d} \mathbf{x}e^{i\langle\mathbf{x},\boldsymbol{\omega}\rangle_2}\mathrm{d}Q(\mathbf{x})$*, and*

(ii) $\nabla_{\boldsymbol{\omega}}\psi_Q(-\boldsymbol{\omega}) = -i\int_{\mathbb{R}^d} \mathbf{x}e^{-i\langle\mathbf{x},\boldsymbol{\omega}\rangle_2}\mathrm{d}Q(\mathbf{x})$*.*

Proof. Observing that $D^{\mathbf{e}_j}\psi_Q(\boldsymbol{\omega}) = i\int_{\mathbb{R}^d} \mathbf{x}^{\mathbf{e}_j}e^{i\langle\mathbf{x},\boldsymbol{\omega}\rangle_2}\mathrm{d}Q(\mathbf{x})$ by Theorem C.2 and that the expectation of a vector is the vector of expectations yield (i). We obtain (ii) by writing

$$\nabla_{\boldsymbol{\omega}}\psi_Q(-\boldsymbol{\omega}) \overset{\text{(a)}}{=} \nabla_{\boldsymbol{\omega}}\overline{\psi_Q(\boldsymbol{\omega})} \overset{\text{(b)}}{=} \overline{\nabla_{\boldsymbol{\omega}}\psi_Q(\boldsymbol{\omega})} \overset{\text{(c)}}{=} -i\int_{\mathbb{R}^d} \mathbf{x}e^{-i\langle\mathbf{x},\boldsymbol{\omega}\rangle_2}\mathrm{d}Q(\mathbf{x}),$$

where (a) comes by the definition of the characteristic function, (b) follows from the fact that the derivative of the conjugate is the conjugate of the derivative, and (c) is implied by taking the conjugate of the result obtained in (i). □

Lemma B.3 (Derivatives of the kernel via its Bochner's representation)**.** *Let k be a kernel satisfying Assumption 3 and $k \in \mathcal{C}^2(\mathbb{R}^d\times\mathbb{R}^d)$ with Bochner representation $k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} e^{-i\langle\mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda(\boldsymbol{\omega})$. Then,*

(i) $\nabla_{\mathbf{x}}k(\mathbf{x},\mathbf{y}) = -i\int_{\mathbb{R}^d} \boldsymbol{\omega}e^{-i\langle\mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda(\boldsymbol{\omega})$*,*

(ii) $\nabla_{\mathbf{y}}k(\mathbf{x},\mathbf{y}) = i\int_{\mathbb{R}^d} \boldsymbol{\omega}e^{-i\langle\mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda(\boldsymbol{\omega})$*,*

(iii) $\frac{\partial}{\partial\mathbf{x}^{\mathbf{e}_j}\partial\mathbf{y}^{\mathbf{e}_j}}k(\mathbf{x},\mathbf{y}) = \int_{\mathbb{R}^d} \boldsymbol{\omega}^{2\mathbf{e}_j}e^{-i\langle\mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda(\boldsymbol{\omega})$*.*

Proof. Throughout the proof, let $\Lambda' = \frac{\Lambda}{\Lambda(\mathbb{R}^d)}$, where we note that $\Lambda' \in \mathcal{M}_1^+(\mathbb{R}^d)$. Furthermore, let $g(\mathbf{x},\mathbf{y}) = \mathbf{y}-\mathbf{x}$. We show each statement separately.

Part (i). Considering the Bochner representation of $k(\mathbf{x},\mathbf{y})$ allows us to write

$$\nabla_{\mathbf{x}}k(\mathbf{x},\mathbf{y}) = \nabla_{\mathbf{x}}\int_{\mathbb{R}^d} e^{-i\langle\mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda(\boldsymbol{\omega}) \overset{\text{(a)}}{=} \begin{pmatrix} \Lambda(\mathbb{R}^d)\frac{\partial}{\partial\mathbf{x}^{\mathbf{e}_1}}\int_{\mathbb{R}^d} e^{i\langle g(\mathbf{x},\mathbf{y}),\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda'(\boldsymbol{\omega}) \\ \vdots \\ \Lambda(\mathbb{R}^d)\frac{\partial}{\partial\mathbf{x}^{\mathbf{e}_d}}\int_{\mathbb{R}^d} e^{i\langle g(\mathbf{x},\mathbf{y})\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda'(\boldsymbol{\omega}) \end{pmatrix} \overset{\text{(b)}}{=} \begin{pmatrix} \Lambda(\mathbb{R}^d)\frac{\partial}{\partial\mathbf{x}^{\mathbf{e}_1}}\psi_{\Lambda'}(g(\mathbf{x},\mathbf{y})) \\ \vdots \\ \Lambda(\mathbb{R}^d)\frac{\partial}{\partial\mathbf{x}^{\mathbf{e}_d}}\psi_{\Lambda'}(g(\mathbf{x},\mathbf{y})) \end{pmatrix}$$
$$\overset{\text{(c)}}{=} \begin{pmatrix} \Lambda(\mathbb{R}^d)\frac{\partial g(\mathbf{x},\mathbf{y})}{\partial\mathbf{x}^{\mathbf{e}_1}}\, D^{\mathbf{e}_1}\psi_{\Lambda'}(\mathbf{t})\big|_{\mathbf{t}=\mathbf{y}-\mathbf{x}} \\ \vdots \\ \Lambda(\mathbb{R}^d)\frac{\partial g(\mathbf{x},\mathbf{y})}{\partial\mathbf{x}^{\mathbf{e}_d}}\, D^{\mathbf{e}_d}\psi_{\Lambda'}(\mathbf{t})\big|_{\mathbf{t}=\mathbf{y}-\mathbf{x}} \end{pmatrix} \overset{\text{(d)}}{=} \begin{pmatrix} \Lambda(\mathbb{R}^d)(-i^{|\mathbf{e}_1|})\int_{\mathbb{R}^d} \boldsymbol{\omega}^{\mathbf{e}_1}e^{i\langle\mathbf{y}-\mathbf{x},\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda'(\boldsymbol{\omega}) \\ \vdots \\ \Lambda(\mathbb{R}^d)(-i^{|\mathbf{e}_d|})\int_{\mathbb{R}^d} \boldsymbol{\omega}^{\mathbf{e}_d}e^{i\langle\mathbf{y}-\mathbf{x},\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda'(\boldsymbol{\omega}) \end{pmatrix}$$
$$\overset{\text{(e)}}{=} -i\int_{\mathbb{R}^d} \boldsymbol{\omega}e^{-i\langle\mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda(\boldsymbol{\omega}),$$

where (a) comes by the definitions of $\nabla_{\mathbf{x}}$, Λ', g, and the linearity of the inner product. (b) stems from the definition of the characteristic function and (c) follows from the chain rule. Theorem C.2 and the substitution $\mathbf{t} = \mathbf{y}-\mathbf{x}$ yield (d). Last, we recall that the expectation of a random vector equals the vector of the expectations of its components, which, together with the definition of Λ' and the linearity of the inner product, imply (e).

Part (ii). Observing that $\frac{\partial g(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}^{\mathbf{e}_j}} = 1$, we can write

$$\nabla_{\mathbf{y}} k(\mathbf{x},\mathbf{y}) = \nabla_{\mathbf{y}} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega}) \overset{\text{(a)}}{=} \begin{pmatrix} \Lambda(\mathbb{R}^d)\frac{\partial}{\partial \mathbf{y}^{\mathbf{e}_1}} \int_{\mathbb{R}^d} e^{i\langle g(\mathbf{x},\mathbf{y}),\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda'(\boldsymbol{\omega}) \\ \vdots \\ \Lambda(\mathbb{R}^d)\frac{\partial}{\partial \mathbf{y}^{\mathbf{e}_d}} \int_{\mathbb{R}^d} e^{i\langle g(\mathbf{x},\mathbf{y})\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda'(\boldsymbol{\omega}) \end{pmatrix} \overset{\text{(b)}}{=} \begin{pmatrix} \Lambda(\mathbb{R}^d)\frac{\partial}{\partial \mathbf{y}^{\mathbf{e}_1}} \psi_{\Lambda'}(g(\mathbf{x},\mathbf{y})) \\ \vdots \\ \Lambda(\mathbb{R}^d)\frac{\partial}{\partial \mathbf{y}^{\mathbf{e}_d}} \psi_{\Lambda'}(g(\mathbf{x},\mathbf{y})) \end{pmatrix}$$

$$\overset{\text{(c)}}{=} \begin{pmatrix} \Lambda(\mathbb{R}^d)\frac{\partial g(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}^{\mathbf{e}_1}} \ D^{\mathbf{e}_1}\psi_{\Lambda'}(\mathbf{t})\big|_{\mathbf{t}=\mathbf{y}-\mathbf{x}} \\ \vdots \\ \Lambda(\mathbb{R}^d)\frac{\partial g(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}^{\mathbf{e}_d}} \ D^{\mathbf{e}_d}\psi_{\Lambda'}(\mathbf{t})\big|_{\mathbf{t}=\mathbf{y}-\mathbf{x}} \end{pmatrix} \overset{\text{(d)}}{=} \begin{pmatrix} \Lambda(\mathbb{R}^d) i^{|\mathbf{e}_1|} \int_{\mathbb{R}^d} \boldsymbol{\omega}^{\mathbf{e}_1} e^{i\langle \mathbf{y}-\mathbf{x},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda'(\boldsymbol{\omega}) \\ \vdots \\ \Lambda(\mathbb{R}^d) i^{|\mathbf{e}_d|} \int_{\mathbb{R}^d} \boldsymbol{\omega}^{\mathbf{e}_d} e^{i\langle \mathbf{y}-\mathbf{x},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda'(\boldsymbol{\omega}) \end{pmatrix}$$

$$\overset{\text{(e)}}{=} i \int_{\mathbb{R}^d} \boldsymbol{\omega} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega}),$$

where (a), (b), (c), (d), and (e) were obtained as in part (i).

Part (iii). Consider the Bochner representation of $k(\mathbf{x},\mathbf{y})$. Then,

$$\begin{aligned} \frac{\partial}{\partial \mathbf{x}^{\mathbf{e}_j} \partial \mathbf{y}^{\mathbf{e}_j}} k(\mathbf{x},\mathbf{y}) &= \frac{\partial^2}{\partial \mathbf{x}^{e_j} \partial \mathbf{y}^{e_j}} \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega}) \overset{\text{(a)}}{=} \Lambda(\mathbb{R}^d) \frac{\partial^2}{\partial \mathbf{x}^{e_j} \partial \mathbf{y}^{e_j}} \int_{\mathbb{R}^d} e^{i\langle \mathbf{y}-\mathbf{x},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda'(\boldsymbol{\omega}) \\ &\overset{\text{(b)}}{=} \Lambda(\mathbb{R}^d) \frac{\partial^2 \psi_{\Lambda'}(g(\mathbf{x},\mathbf{y}))}{\partial \mathbf{x}^{e_j} \partial \mathbf{y}^{e_j}} \overset{\text{(c)}}{=} \Lambda(\mathbb{R}^d) \underbrace{\frac{\partial g(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}^{e_j}}}_{=-1} \underbrace{\frac{\partial g(\mathbf{x},\mathbf{y})}{\partial \mathbf{y}^{e_j}}}_{=1} D^{2\mathbf{e}_j} \psi_{\Lambda'}(\mathbf{t})\big|_{\mathbf{t}=\mathbf{y}-\mathbf{x}} \\ &\overset{\text{(d)}}{=} -i^2 \int_{\mathbb{R}^d} \boldsymbol{\omega}^{2\mathbf{e}_j} e^{-i\langle \mathbf{x}-\mathbf{y},\boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega}), \end{aligned}$$

where (a) comes by $\Lambda = \Lambda(\mathbb{R}^d)\Lambda'$, the linearity of the integral, the partial derivative, and the inner product. The definitions of g and characteristic function yield (b). The chain rule gives (c) and Theorem C.2 with $\boldsymbol{\alpha} = 2\mathbf{e}_j$ implies (d). Noting that $i^2 = -1$ leads to the claimed result. □

Lemma B.4 (Existence of perturbation function)**.** *Let $(\mathcal{X}, \tau_{\mathcal{X}})$ be a topological space, $P_0 \in \mathcal{M}_1^+(\mathcal{X})$, and $\varphi_0 \in \mathcal{C}_b(\mathcal{X})$ such that there exists no $c \in \mathbb{R}$ such that $\varphi_0 = c$ holds P_0-almost surely. Then there exists $\varphi \in \mathcal{C}_b(\mathcal{X})$ such that $\mathbb{E}_{P_0}[\varphi(X)] = 0$ and there exists a set $A \in \mathcal{B}(\mathcal{X})$ with positive P_0-measure such that $\varphi(x) \neq 0$ for all $x \in A$.*

Proof. Since $\varphi_0 \in \mathcal{C}_b(\mathcal{X})$, φ_0 is integrable w.r.t. P_0 and $\mu_0 := \mathbb{E}_{P_0}[\varphi_0(X)] < \infty$. By centering φ_0 as $\varphi := \varphi_0 - \mu_0$, one has $\mathbb{E}_{P_0}[\varphi(X)] = \mathbb{E}_{P_0}[\varphi_0(X)] - \mu_0 = \mu_0 - \mu_0 = 0$. Also, as φ_0 is not constant P_0-almost surely, the property of $\varphi = 0$ P_0-almost surely does not hold, implying the existence of the stated $A \in \mathcal{B}(\mathcal{X})$. To see that $\varphi \in \mathcal{C}_b(\mathcal{X})$, it suffices to note that (i) φ is the sum of continuous functions and thus continuous, and (ii) $\sup_{x\in\mathcal{X}} |\varphi(x)| \le \sup_{x\in\mathcal{X}} |\varphi_0(x)| + |\mu_0| < \infty$ by the triangle inequality, hence φ is also bounded. □

Lemma B.5 (Perturbed measures are distinct)**.** *Assume $P_0 \in \mathcal{M}_1^+(\mathcal{X})$, let $\varphi \in \mathcal{C}_b(\mathcal{X})$ be such that there exists $A' \in \mathcal{B}(\mathcal{X})$ with positive P_0-measure, such that $\varphi(x) \neq 0$ for all $x \in A'$, and define the measure P_n as $P_n(A) = \int_A 1 + \epsilon_n \varphi(x) \mathrm{d}P_0(x)$, with $\epsilon_n > 0$, for all $A \in \mathcal{B}(\mathcal{X})$. Then, $P_0 \neq P_n$.*

Proof. We argue by contradiction. Assume that $P_0 = P_n$. Then,

$$P_0 = P_n \overset{\text{(a)}}{\Longrightarrow} 1 = \frac{\mathrm{d}P_n}{\mathrm{d}P_0} = 1 + \epsilon_n \varphi \ P_0\text{-almost surely} \overset{\text{(b)}}{\Longrightarrow} \varphi = 0 \ P_0\text{-almost surely},$$

where (a) uses the definition of the Radon-Nikodym derivative and (b) follows as $\epsilon_n > 0$. This contradicts the assumption imposed on φ, concluding the proof. □

C EXTERNAL STATEMENTS

To ensure self-completeness, this section collects the external statements that we use. Theorem C.1 fully characterizes continuous bounded translation-invariant kernels. Theorem C.2 relates the differentiability of the characteristic function of a random variable to its moments; we include only the part relevant to our proofs for brevity. Under certain conditions, the converse also holds, detailed in Theorem C.3. Theorem C.4 gives a necessary and sufficient condition for a continuous bounded translation-invariant kernel to be characteristic. We recall Fubini's theorem in Theorem C.5. Lemma C.1 and Lemma C.2 collect properties of the KL divergence.

Theorem C.1 (Bochner; Theorem 6.6; Wendland 2005). *A continuous function $\kappa : \mathbb{R}^d \to \mathbb{R}$ is positive definite if and only if it is the Fourier transform of a finite nonnegative Borel measure Λ on $\mathbb{R}^d$, that is,*

$$\kappa(\mathbf{x}) = \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}, \boldsymbol{\omega}\rangle_2} \mathrm{d}\Lambda(\boldsymbol{\omega}) \quad \textit{for all } \mathbf{x} \in \mathbb{R}^d.$$

Theorem C.2 (Differentiability characteristic function; Theorem 1.2.1(i); Sasvári 2013). *Let $X \sim P \in \mathcal{M}_1^+\left(\mathbb{R}^d\right)$ and $\boldsymbol{\alpha} \in \mathbb{N}_0^d$ such that the moment $M_{\boldsymbol{\alpha}}^P$ of P exists. Then the partial derivative $D^{\boldsymbol{\alpha}}\psi_P$ exists and one has $D^{\boldsymbol{\alpha}}\psi_P(\mathbf{t}) = i^{|\boldsymbol{\alpha}|}\int_{\mathbb{R}^d} \mathbf{x}^{\boldsymbol{\alpha}} e^{i\langle \mathbf{t}, \mathbf{x}\rangle_2}\mathrm{d}P(\mathbf{x})$ ($\mathbf{t} \in \mathbb{R}^d$).*

Theorem C.3 (Existence of the moments of P; Theorem 1.2.9; Sasvári 2013). *Let $X \sim P \in \mathcal{M}_1^+\left(\mathbb{R}^d\right)$ and $\boldsymbol{\alpha} \in \mathbb{N}_0^d \setminus \{\mathbf{0}\}$ such that all partial derivatives $D^{\boldsymbol{\beta}} Re(\psi_P)(\mathbf{t})$, $\boldsymbol{\beta} < 2\boldsymbol{\alpha}$ exist in an open neighborhood of zero. If $D^{2\boldsymbol{\alpha}} Re(\psi_P)(\mathbf{t})$ exists at zero, then the moment $M_{2\boldsymbol{\alpha}}^P$ of P exists.*

Theorem C.4 (Theorem 9; Sriperumbudur et al. 2010b). *Suppose $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a continuous bounded translation-invariant kernel. Then k is characteristic if and only if $\mathrm{supp}(\Lambda) = \mathbb{R}^d$, with Λ defined according to Theorem C.1 as $k(\mathbf{x}, \mathbf{y}) = \int_{\mathbb{R}^d} e^{-i\langle \mathbf{x}-\mathbf{y}, \boldsymbol{\omega}\rangle_2}\mathrm{d}\Lambda(\boldsymbol{\omega})$ ($\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$).*

The following theorem allows to exchange the order of integration. We recall that σ-finiteness always holds for any Borel probability measure.

Theorem C.5 (Fubini-Tonelli; Theorem 2.37.b; Folland 1999). *Suppose that $(\mathcal{X}, \mathcal{M}, \mu)$ and $(\mathcal{Y}, \mathcal{N}, \nu)$ are σ-finite measure spaces. Let $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. If $\int_{\mathcal{X}\times\mathcal{Y}} |f(x,y)|\mathrm{d}(\mu\otimes\nu)(x,y) < \infty$, then*

$$\int_{\mathcal{X}\times\mathcal{Y}} f(x,y)\mathrm{d}(\mu\otimes\nu)(x,y) = \int_{\mathcal{X}}\left[\int_{\mathcal{Y}} f(x,y)\mathrm{d}\nu(y)\right]\mathrm{d}\mu(x) = \int_{\mathcal{Y}}\left[\int_{\mathcal{X}} f(x,y)\mathrm{d}\mu(x)\right]\mathrm{d}\nu(y).$$

Lemma C.1 (KL divergence of product measures; p. 85; Tsybakov 2009). *Let $P = \otimes_{j=1}^n P_j$ and $Q = \otimes_{j=1}^n Q_j$. Then*

$$\mathrm{KL}(P||Q) = \sum_{j=1}^n \mathrm{KL}(P_j||Q_j).$$

Lemma C.2 (KL divergence of Gaussians; p. 13; Duchi 2007). *The KL divergence of two normal distributions $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ on $\mathbb{R}^d$ is*

$$\mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)) = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) - d + \ln\left(\frac{|\boldsymbol{\Sigma}_0|}{|\boldsymbol{\Sigma}_1|}\right)}{2}.$$