Standardisation-function Kernel Stein Discrepancy: A Unifying View on Kernel Stein Discrepancy Tests for Goodness-of-fit

Wenkai Xu
Department of Statistics,
Oxford University

Abstract

Non-parametric goodness-of-fit testing procedures based on kernel Stein discrepancies (KSD) are promising approaches to validate general unnormalised distributions in various scenarios. Existing works focused on studying kernel choices to boost test performances. However, the choices of (non-unique) Stein operators also have considerable effect on the test performances. Inspired by the standardisation technique that was originally developed to better derive approximation properties for normal distributions, we present a unifying framework, called standardisation-function kernel Stein discrepancy (Sf-KSD), to study different Stein operators in KSD-based tests for goodness-of-fit. We derive explicitly how the proposed framework relates to existing KSD-based tests and show that Sf-KSD can be used as a guide to develop novel kernel-based non-parametric tests on complex data scenarios, e.g. truncated distributions or compositional data. Experimental results demonstrate that the proposed tests control type-I error well and achieve higher test power than existing approaches.

1 INTRODUCTION

Stein's method [Barbour and Chen, 2005] provides an elegant probabilistic tool for characterising and comparing distributions. Relevant techniques have been used to tackle various problems in statistical inference, random graph theory, and computational biology. Modern machine learning tasks may extensively involve modelling and learning with intractable densities, where the normalisation constant (or partition function) is unable to be obtained in closed form, e.g. density estimation [Hyvärinen, 2005], model criticism [Kim et al., 2016; Lloyd and Ghahramani, 2015], or generative modelling [Goodfellow et al., 2014]. *Stein operators* may only require to access distributions through the differential (or difference) of log density functions[1](or mass functions), which avoids the knowledge of normalisation constant. These Stein operators are particularly useful to study unnormalised models and recently caught attentions from the machine learning community in many aspects [Anastasiou et al., 2021] such as density estimation [Barp et al., 2019], sampling techniques [Chen et al., 2018, 2019; Gorham et al., 2019; Oates et al., 2019; Shi et al., 2021], numerical methods [Barp et al., 2018], approximate inference [Huggins and Mackey, 2018; Liu and Wang, 2016], and Bayesian inference [Fisher et al., 2021; Liu and Zhu, 2018].

The goodness-of-fit testing procedure aims to check the null hypothesis $\mathrm{H}_0 : q = p$, where q is the *known* target distribution and p is the *unknown* data distribution accessible only through a set of samples[2], $x_1, \ldots, x_n \sim p$. The non-parametric testing refers to the scenario where the assumptions made on the distributions p and q are minimal, i.e. the distributions in non-parametric testing are not assumed to be in any parametric family. By contrast, parametric tests (e.g. student t-test or normality test) assume pre-defined parametric family to be tested against and usually deal with summary statistics such as means or standard deviations, which can be restrictive in terms of comparing full distributions. Kernel-based methods have been applied to compare distributions via rich-enough reproducing kernel Hilbert spaces (RKHS) [Berlinet and Thomas, 2004] and achieved state-of-the-art results for

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

[1] Derivative of log density is also known as score-function.

[2] As only one set of samples are observed, the goodness-of-fit testing sometimes is also referred to as the *one-sample* test or one-sample problem. This is opposed to the two-sample problem where the distribution q is also unknown and appears in the sample form.

non-parametric two-sample test [Gretton et al., 2012a] or independence test [Gretton et al., 2008]. With well-defined *Stein operators*, kernel Stein discrepancy (KSD) [Gorham and Mackey, 2017] has been developed for non-parametric goodness-of-fit testing procedures for *unnormalised models* and demonstrated superior test performances in various scenarios including Euclidean data $\mathbb{R}^d$ [Chwialkowski et al., 2016; Liu et al., 2016], discrete data [Yang et al., 2018], point processes [Yang et al., 2019], latent variable models [Kanagawa et al., 2019], conditional densities [Jitkrittum et al., 2020], censored-data [Fernandez et al., 2020], directional data [Xu and Matsuda, 2020], and network data [Xu and Reinert, 2021]. It is worth to mention that previous KSD-based goodness-of-fit tests require developing case-specific Stein operators to address their corresponding data scenarios. The Stein operators can have diverse forms and may seem to be unrelated.

To improve the test performances, existing works have focused on data-adaptive methods for kernel learning [Gretton et al., 2012b; Sutherland et al., 2017; Liu et al., 2020] and kernel selection [Kübler et al., 2020; Lim et al., 2020]. In addition, using the techniques related to kernel mean embedding [Muandet et al., 2017], KSD-based tests also enable the extraction of distributional features to perform computationally efficient tests and model criticisms [Jitkrittum, 2017; Xu and Matsuda, 2021]. Nonetheless, the choice of Stein operators can also have a considerable effect for test performances but this aspect of the research has so far been ignored mostly. Previous works have only demonstrated the non-uniqueness of valid Stein operators, e.g. Yang et al. [2018] mentioned in their Section 3.2 that for random graph models the Stein operator can be built from indicator functions or normalised Laplacians. Moreover, Fernandez et al. [2020] also derived three distinct Stein operators for censored data based on orthogonal properties from survival analysis where their test performances are only compared empirically. The main contribution of this paper is threefold.

Unifying KSD-based Tests We first propose a simple-enough framework that provides a unifying view for various KSD-based tests. We explicitly show how the seemingly-unrelated Stein operators for different testing scenarios are connected, via *auxiliary functions* that will be defined and explained later in Section 3.

Comparing KSD-based Tests With the unifying framework, we then interpret our Stein operators via *auxiliary functions* and compare KSD tests using optimality conditions to be introduced in Section 4.

Guiding New KSD-based Tests Moreover, we provide a systematic approach to develop KSD-based tests on intended testing scenarios with application on domain constraint densities in Section 5. We derive the bounded-domain KSD (bd-KSD) and its corresponding statistical properties. We experimentally demonstrate the test performances with different operator choices on truncated distributions and compositional data, i.e. data defined on a simplex/hypersimplex.

We begin our discussion by introducing relevant existing Stein operators and KSD-based tests in Section 2 and conclude with future directions in Section 6.

2 PRELIMINARIES

We first review a set of existing Stein operators developed in various data scenarios, followed by a brief reminder on KSD-based goodness-of-fit testing procedures. We also introduce the notion of Itô's process for making connections between the generator approach and score-based Stein operators.

2.1 Stein Operators and Stein's Method

Given a distribution q, an operator $\mathcal{T}_q$ is called a Stein operator w.r.t. q if the following Stein's identity holds for *test function* f: $\mathbb{E}_q[\mathcal{T}_q f] = 0$. The class of such test function f is referred to as the Stein class.

Euclidean Stein operator on $\mathbb{R}^d$ The Stein operator for continuous densities in $\mathbb{R}^d$ with Cartesian coordinate has been previously introduced [Gorham and Mackey, 2015; Ley et al., 2017], which is also referred to as the Langevin-diffusion Stein operator [Barp et al., 2019][3]. Let $\mathcal{X} = \mathbb{R}^d$ and $f_i : \mathcal{X} \to \mathbb{R}$ for $i = 1, \dots, d$ be scalar-valued functions on $\mathcal{X}$. $\mathbf{f}(x) = (f_1(x), \dots, f_d(x))^\top \in \mathbb{R}^d$ defines a vector-valued function. Let q be a smooth probability density on $\mathcal{X}$ that vanishes at infinity. For a bounded smooth function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^d$, the Stein operator $\mathcal{T}_q$ is defined by

$$\mathcal{T}_q \mathbf{f}(x) = \mathbf{f}(x)^\top \nabla \log q(x) + \nabla \cdot \mathbf{f}(x). \tag{1}$$

Stein's identity holds for $\mathcal{T}_q$ in Eq. (1) due to the integration by parts on $\mathbb{R}^d$: $\mathbb{E}_q[\mathcal{T}_q \mathbf{f}] = \int_{\mathbb{R}^d} \mathcal{T}_q \mathbf{f}(x) dq(x) = \sum_i \int_{\mathbb{R}^d} \frac{\partial}{\partial x^i} (f_i(x) q(x))\, dx = 0$, where the last equality holds since $f_i(x)$ is bounded and $q(x)$ vanishes at infinity. Since the Stein operator $\mathcal{T}_q$ depends on the density only through the derivatives of $\log q$, it does not involve the normalisation constant of q: a useful property dealing with unnormalised models.

Censored-data Stein operator In practical data scenarios such as medical trials or e-commerce, we encounter data with censoring where the actual event

[3] Another important approach to develop Stein operator is via the (infinitesimal) generator [Barbour, 1988]. The connection can be established via Itô's process in Section 2.3.

time of interest (or *survival times*) is not accessible but, instead, a *bound* or *interval*, in which the event time is known to belong, is observed. Fernandez et al. [2020] proposed a set of Stein operators for right-censored data, where the lower bound of the event time is observed. The right-censored data is observed in the form of (t_i, δ_i) where for the survival time x_i and censoring time c_i, the observation time is $t_i = \min\{x_i, c_i\}$ and $\delta_i = \mathbb{1}_{\{x_i \le c_i\}}$ indicates if we are observing x_i or c_i. Denote μ_0 as the density of event time x; S_C as the survival function[4] of the censoring time C; the test function $\omega : \mathbb{R}_+ \to \mathbb{R}$ assumed to vanish at origin, i.e. $\omega(0) = 0$; Ω the set of functions $\mathbb{R}_+ \times \{0, 1\} \to \mathbb{R}$, and the operator $(\mathcal{T}_0\omega) \in \Omega$. The censored-data Stein operator is defined as

$$(\mathcal{T}_0\omega)(x, \delta) = \delta \frac{(\omega(x)S_C(x)\mu_0(x))'}{S_C(x)\mu_0(x)}. \tag{2}$$

Denote $\mathbb{E}_0$ as taking expectation w.r.t. the observation pair (x, δ) where $x \sim \mu_0$, which needs to distinguish from $\mathbb{E}_{\mu_0}$ that takes expectation over $x \sim \mu_0$. Note that the censoring distribution, e.g. S_C remains *unknown*. By Eq. (24) in Appendix A, we have Stein's identity $\mathbb{E}_0[(\mathcal{T}_0\omega)(x, \delta)] = 0$.

The key challenge for this Stein operator is that the survival function for censoring time S_C is *unknown* and *not included* in the null hypothesis. Hence, Fernandez et al. [2020] applied identities in survival analysis to derive a computationally feasible operator: the survival Stein operator. This operator have an unbiased estimation from the empirical observations. For the hazard function λ_0 associated with μ_0, the survival Stein operator $\mathcal{T}_0^{(s)}$ is defined as

$$(\mathcal{T}_0^{(s)}\omega)(x, \delta) = \delta \left(\omega'(x) + \frac{\lambda_0'(x)}{\lambda_0(x)}\omega(x)\right) - \lambda_0(x)\omega(x). \tag{3}$$

By the martingale identities, Fernandez et al. [2020] also proposed the martingale Stein operator,

$$(\mathcal{T}_0^{(m)}\omega)(x, \delta) = \delta \frac{\omega'(x)}{\lambda_0(x)} - \omega(x). \tag{4}$$

Details for known identities regarding survival analysis and martingales can be found in Appendix A.

Latent-variable Stein operator Latent variable models are powerful tools in generative modelling and statistical inference. However, such models generally do not have closed form density expressions due to the integral operation on latent spaces. Latent-variable Stein operator [Kanagawa et al., 2019] is constructed via sampling the latent variables and the corresponding conditional densities . Let $q(x) \propto \int q(x|z)\pi(z)dz$ be the target distribution which is not accessible in closed form, even its unnormalised version. Sample $z_1, \ldots, z_m \sim q(z|x)$. The latent variable Stein operator is defined as

$$\mathcal{T}_{q,z}\mathbf{f}(x) = \frac{1}{m}\sum_{j=1}^{m} \mathcal{T}_{q(x|z_j)}\mathbf{f}(x) \tag{5}$$

A closely related construction is the *Stochastic Stein Operator* [Gorham et al., 2020], which has been developed for computationally efficient posterior sampling. Additional details and comparisons with latent variable Stein operator are included in Appendix D.2.

Second-order Stein operator To respect the geometry for distributions defined on Riemannian manifolds, Stein operators involving second-order differential operators have been studied [Barp et al., 2018; Le et al., 2020]. For a smooth Riemannian manifold $\mathcal{M}$ and scalar-valued function $\tilde{f} : \mathcal{M} \to \mathbb{R}$, the second-order Stein operator is

$$\mathcal{T}_q^{(2)}\tilde{f}(x) = \nabla\tilde{f}(x)^\top \nabla \log q + \Delta\tilde{f}(x), \quad x \in \mathcal{M}, \tag{6}$$

where $\Delta\tilde{f} = \nabla \cdot \nabla\tilde{f}$ denotes the corresponding Laplace-Beltrami operator[5]. This second-order operator can be seen as a natural consequence of diffusion process in $\mathcal{M}$ [Le et al., 2020]. Euclidean manifold is a special case of $\mathcal{M}$, on which the diffusion process would induce a similar second-order Stein operator: an example shown in Section 2.3. In the Euclidean case, another connection with $\mathcal{T}_q$ in Eq. (1) is via choosing test function in particular form: $\mathbf{f} = \nabla\tilde{f}$.

Coordinate-dependent Stein operator Consider coordinate system $(\theta^1, \ldots, \theta^d)$ that is almost everywhere in $\mathcal{M}$. For a density q on $\mathcal{M}$, the Stein operator with the chosen coordinate is defined as

$$\mathcal{T}_q^{(1)}\mathbf{f} = \sum_{i=1}^{d} \left(\frac{\partial f^i}{\partial\theta^i} + f^i \frac{\partial}{\partial\theta^i}\log(qJ)\right), \tag{7}$$

where $J = (\det G)^{1/2}$ is the volume element. $\mathcal{T}_q^{(1)}$ can be shown to satisfy Stein's identity using differential forms and the corresponding Stoke's theorem [Xu and Matsuda, 2020].

[4] Survival function is defined as $S(x) = 1 - F(x)$ where $F(x) = \int_0^x \mu(s)ds$ denotes the c.d.f. of the event time.

[5] Specifically, consider ∂x^i as the basis vector on Tangent space of x, $\mathrm{T}_x\mathcal{M}$. $\nabla\tilde{f} = \sum_{i,j}[G^{-1}]_{i,j}\frac{\partial\tilde{f}}{\partial x^j}\partial x^i$, denotes the (Riemannian) gradient operator, where $G \in \mathbb{R}^{d\times d}$ denotes the metric tensor matrix; the divergence operator is $\nabla \cdot \mathbf{s} = \sum_i \frac{\partial s_i}{\partial x^i} + s_i \frac{\partial}{\partial x^i}\log\sqrt{\det(G)}$ for $\mathbf{s} = s_1\partial x^1, \ldots, s_d\partial x^d$. In the Euclidean case, $G \equiv I$, the identity matrix, which is independent of x.

Stein's method and standardisation techniques Stein's method [Stein et al., 1972] refers to the characterisation related to Stein operators for analysing distributions properties or distribution approximation [Barbour and Chen, 2005; Barbour, 2005; Barbour et al., 2018] via solving the following Stein equation:

$$\mathcal{T}_q f_h(x) = h(x) - \mathbb{E}_q[h(x)] \tag{8}$$

where the subscript h of f explicitly refers to that function f is associated with h under regularity conditions. $\mathcal{T}_q$ denotes the corresponding Stein operator. Using centered Gaussian distribution in $\mathbb{R}^d$ as an example, $p(x) \propto \exp\{-\frac{1}{2}x^\top \Sigma^{-1} x\}$ where Σ denotes the covariance matrix. The Stein operator derived from Eq. (1) has the form

$$\mathcal{T}_q \mathbf{f}(x) = -(\Sigma^{-1}x)^\top \mathbf{f}(x) + \nabla \cdot \mathbf{f}(x). \tag{9}$$

However, during analysis and computation, the inverse covariance (or the precision) matrix can cause potential issues. The standaradisation techniques have been studied [Ley et al., 2017; Mijoule et al., 2021]. For Gaussian distributions with constant convariance matrix, pre-multiply Σ on the Stein operator would not destroy Stein's identity and the Stein operator reads $-x^\top \mathbf{f}(x) + \nabla \cdot \Sigma \mathbf{f}(x)$ that can be analysed nicely. The score function that interacts with $\mathbf{f}$ in $\mathcal{T}_q$ now becomes the score of **standard** Gaussian. Beyond Gaussian case, general standardisation Stein operators has been recently studied [Ernst et al., 2020, Definition 2.6].

2.2 KSD Tests for Goodness-of-fit

With any well-defined Stein operator, we can choose an appropriate RKHS w.r.t. the data scenario to construct its corresponding KSD. Let p, q be distributions satisfying regularity conditions for the relevant testing scenarios and the test function class to be the unit ball RKHS, $B_1(\mathcal{H})$, KSD between distributions p and q is defined as

$$\mathrm{KSD}(p\|q; \mathcal{H}) = \sup_{\mathbf{f} \in B_1(\mathcal{H})} \mathbb{E}_p[\mathcal{T}_q \mathbf{f}]. \tag{10}$$

It is known from Stein's identity that for any test functions in the Stein class, $p = q$ implies $\mathrm{KSD}(p\|q) = 0$. In the testing procedure, a desirable property of the discrepancy measure is that $\mathrm{KSD}(q\|p) = 0$ if and only if $p = q$. As such, we require our RKHS to be sufficiently large to capture any possible discrepancies between p and q, which requires mild regularity conditions [Chwialkowski et al., 2016, Theorem 2.2] for KSD to be a proper discrepancy measure. Algebraic manipulations produce the following quadratic form:

$$\mathrm{KSD}^2(p\|q) = \mathbb{E}_{x,\tilde{x}\sim p}[h_q(x, \tilde{x})], \tag{11}$$

where $h_q(x, \tilde{x}) = \langle \mathcal{T}_q k(x, \cdot), \mathcal{T}_q k(\tilde{x}, \cdot) \rangle_{\mathcal{H}}$ does not involve p; $k(x, \cdot)$ denotes the kernel associated with RKHS $\mathcal{H}$.

Testing procedure Now, suppose we have relevant samples $x_1, \ldots, x_n$ from the *unknown* distribution p. To test the null hypothesis $\mathrm{H}_0 : p = q$ against the (broad class of) alternative hypothesis $\mathrm{H}_1 : p \neq q$, KSD can be empirically estimated via Eq. (11) using U-statistics or V-statistics [Van der Vaart, 2000]; given the significance level of the test, the critical value can be determined by wild-bootstrap procedures [Chwialkowski et al., 2014] or spectral estimation [Jitkrittum et al., 2017] based on the Stein kernel matrix $H_{rs} = h_q(x_r, x_s)$; the rejection decision is then made by comparing empirical test statistics with the critical value. In this way, a systematic non-parametric goodness-of-fit testing procedure is obtained, which is applicable to unnormalised models.

2.3 Itô's Process and Infinitesimal Generator

Itô's process is fundamentally studied in stochastic differential equation (SDE) and can be described via continuous-time Markov process of the following form,

$$dX_t = b(X_t)dt + \sqrt{2}\sigma(X_t)d\mathbf{B}_t, \tag{12}$$

where $b(X_t)$, $\sigma(X_t)$ denote the drift and diffusion functions of the process respectively; $\mathbf{B}_t$ denotes the standard Brownian motion. Under regularity conditions, the process reaches stationary distribution $\pi(x)$ for $b(x) = -(\log \pi(x))'\sigma(x)^2 + 2\sigma(x)\sigma(x)'$, where $'$ denotes the derivative [Stramer and Tweedie, 1999]. For $\sigma \equiv 1$, the process is referred to as Langevin-diffusion. Vector-valued diffusion can be also established [Gorham et al., 2019, Definition 1] [Mijoule et al., 2021, Example 3.20].

If $(X_t)_{t\geq 0}$ is Feller process with invariant measure π, the (infinitesimal) generator for the process [Anastasiou et al., 2021, Section 2.2.1] is defined (pointwise) as

$$(\mathcal{A}f)(x) = \lim_{t \downarrow 0} \mathbb{E}[f(X_t)|X_0 = x] - f(x). \tag{13}$$

It is not hard to see Stein's identity hold $\mathbb{E}_\pi[(\mathcal{A}f)(x)] = 0$ where the generator is a valid Stein operator. Moreover, if the test function f is twice differentiable, the generator for the process in Eq. (12) admits the following form

$$(\mathcal{A}f)(x) = b(x)f'(x) + \sigma(x)^2 f''(x). \tag{14}$$

For Langevin-diffusion where $\sigma(x) \equiv 1$, the generator for the process with stationary distribution q becomes $(\mathcal{A}_q f)(x) = \log q(x)' f'(x) + f''(x)$, which retrieves the second order Stein operator $\mathcal{T}_q^{(2)}$ in Eq. (6) for $\mathcal{X} = \mathbb{R}$.

3 STANDARDISATION-FUNCTION KERNEL STEIN DISCREPANCY

3.1 Standardisation-function Stein Operator

We introduce a simple-enough framework that is capable to unify the Stein operators introduced in Section

2.1. The idea is inspired from the standardisation technique for Eq. (9). We note that standardisation for analysing the Gaussian case in Eq. (9) is done by pre-multiplying a *constant* covariance function, which only have scaling effect on the Stein operator[6]. Here, we consider the Stein operator standardised by matrix-valued *auxiliary function* $G(x) : \mathcal{X} \to \mathbb{R}^{d\times d}$.

Let $\mathbf{f}$ be an appropriately test function as defined in Section 2.1. We consider the following Stein operator for density q,

$$\mathcal{T}_{q,G}\mathbf{f} = \mathcal{T}_q(G\mathbf{f}) = (G\mathbf{f})^\top \nabla \log q + \nabla \cdot (G\mathbf{f}), \qquad (15)$$

where $(G\mathbf{f}) \in \mathbb{R}^d$ stands for the matrix multiplication of $G(x)$ with $\mathbf{f}(x)$. We note this operator is analogous to the diffusion Stein operator [Barp et al., 2019; Mijoule et al., 2021] studied in the estimation context. For our unifying purpose, it is enough to reduce $G(x)$ as a diagonal form with diagonal entries $\mathbf{g} = (g_1, \ldots, g_d) : \mathbb{R}^d \to \mathbb{R}^d$ with corresponding $\mathcal{T}_{q,\mathbf{g}}$ being

$$(\mathcal{T}_{q,\mathbf{g}}\mathbf{f}) = \sum_{i=1}^{d} g_i(f_i \frac{\partial}{\partial x^i} \log q + \frac{\partial}{\partial x^i} f_i) + f_i \frac{\partial}{\partial x^i} g_i. \quad (16)$$

Stein's identity holds $\mathbb{E}_q[(\mathcal{T}_{q,\mathbf{g}}\mathbf{f})(x)] = 0$ holds for all bounded function $\mathbf{g}$, due to similar integration by parts argument used for Eq. (1). With Stein operator $\mathcal{T}_{q,\mathbf{g}}$, we define the standardisation-function kernel Stein discrepancy,

$$\text{Sf-KSD}_{\mathbf{g}}(p\|q;\mathcal{H}) = \sup_{\|\mathbf{f}\|_{\mathcal{H}}\leq 1} \mathbb{E}_p[\mathcal{T}_{q,\mathbf{g}}\mathbf{f}]. \qquad (17)$$

$\text{Sf-KSD}^2_{\mathbf{g}}$ admits the quadratic form similar to Eq. (11).

Proposition 1. *Let $\mathcal{H}$ be RKHS associated with kernel K. For fixed choice of bounded function $\mathbf{g}$,*

$$\text{Sf-KSD}^2_{\mathbf{g}}(p\|q;\mathcal{H}) = \mathbb{E}_{x,\tilde{x}\sim p}[h_{q,\mathbf{g}}(x,\tilde{x})], \qquad (18)$$

where $h_{q,\mathbf{g}}(x,\tilde{x}) = \langle \mathcal{T}_{q,\mathbf{g}}K(x,\cdot), \mathcal{T}_{q,\mathbf{g}}K(\tilde{x},\cdot)\rangle_{\mathcal{H}}$.

For different choice of auxiliary functions $\mathbf{g}$, Sf-KSD exhibits distinct diffusion pattern induced by $\mathbf{g}$, i.e. $\mathbf{g}$ corresponds to the diagonal of second moment for diffusion $\sigma(x)\sigma(x)^\top$. We note that, by choosing $g_i(x) \equiv 1, \forall i \in [d]$, Sf-KSD recovers the KSD with Stein operator in $\mathbb{R}^d$ in Eq. (1), induced by Langevin diffusion in form of Eq. (12). Related ideas have been discussed using interpretations on Fisher information metric [Mijoule et al., 2018, Section 7.2]; as well as the Bregman divergence induced mapping in mirrored descent Stein sampler [Shi et al., 2021]. Analogous ideas also appeared in Stein variational gradient descent (SVGD) [Liu and Wang, 2016] for discrete variables [Han et al., 2020]. We now proceed to show connections for existing KSDs proposed for various goodness-of-fit testing scenarios using Sf-KSD.

[6] In the form of Eq. (12), constant diffusion $\sigma(x) \equiv c$ applies.

3.2 Unifying Existing Stein Operators

The specific choice of $\mathbf{g}$ and its interplay with $\mathbf{f}$ can be helpful to understand the conditions in various testing scenarios. With appropriate choice of the auxiliary function $\mathbf{g}$, Sf-KSD is capable to provide a unifying view for the KSDs introduced in Section 2.1. To specify the equivalence notion, we denote $\triangleq$ as identical formulations beyond the equality in evaluations. All proofs and derivation details are included in the Appendix B.

Theorem 1 (Censored-data Stein operator)**.** *Let dimension of the data $d = 1$ and w.l.o.g., the test function ω vanishes at 0. For $g : \mathbb{R}_+ \to \mathbb{R}$, choosing $g(x) = S_C(x)$,* Sf-KSD *with $\mathcal{T}_{\mu_0,g}$ recovers the censored-data* KSD *with Stein operator defined in Eq. (2),*

$$\mathbb{E}_{\mu_0}[\mathcal{T}_{\mu_0,S_C}\omega] \triangleq \mathbb{E}_0[(\mathcal{T}_0\omega)(x,\delta)] = 0. \qquad (19)$$

It is not difficult to see that the result holds from directly applying identity in Eq. (24) (explained in Appendix A). However, it is worth noting that during the testing procedure, S_C is *unknown* so we do not have direct access to g here. Moreover, the expectation on l.h.s. of Eq. (19) is w.r.t. the density of survival time μ_0 for Sf-KSD, where the expectation on the r.h.s. is $\mathbb{E}_0$, w.r.t. the paired observation incorporating censoring information. Theorem 1 serves the purpose of explicitly demonstrating how a particular choice of auxiliary function can bridge the gap between censored-data Stein operator with the Stein operator on distributions without the presence censoring information. It will also be interesting to understand how the auxiliary function may explain the Stein operators in Eq. (3) and Eq. (4) when $\mathbb{E}_0$ applies to Sf-KSD when the expectation is taken over the paired variable (x,δ).

Theorem 2 (Martingale Stein operator)**.** *Assume the same setting as in Theorem 1. Further assume that the positive definite test function $\omega : \mathbb{R}_+ \to \mathbb{R}$ is integrable such that $\int_0^x \omega(s)ds < \infty, \forall x$; $\mu_0(x) > 0$ for the survival times so the inverse of its corresponding hazard function $\lambda_0(x) = \frac{\mu_0(x)}{S_0(x)}$ is then well-defined on $\mathbb{R}_+$. For $g : \mathbb{R}_+ \to \mathbb{R}$, choosing $g(x) = \delta\lambda_0(x)^{-1} + (1-\delta)\frac{\int_0^x \mu_0(s)\omega(s)ds}{\mu_0(x)\omega(x)}$,* Sf-KSD *with $\mathcal{T}_{\mu_0,g}$ recovers the martingale* KSD *with Stein operator defined in Eq. (4),*

$$\mathbb{E}_0[\mathcal{T}_{\mu_0,g}\omega] \triangleq \mathbb{E}_0[(\mathcal{T}_0^{(m)}\omega)(x,\delta)] = 0. \qquad (20)$$

The δ-dependent decomposition of g in Theorem 2 reveals the relationship between how censoring is incorporated in the martingale Stein operator, i.e. through the hazard function for uncensored data while through an interaction between the density μ_0 and the test function ω in the censored part. Similarly, choosing δ-dependent auxiliary function g can recover survival Stein operator in Eq. (3).

Corollary 1 (Survival Stein operator). *Assume the conditions in Theorem 2 hold. Consider* $\zeta(x) = \frac{\int_0^x \mu_0(s)\omega(s)\lambda_0(s)ds}{\mu_0(x)\omega(x)}$. *For* $g : \mathbb{R}_+ \to \mathbb{R}$, *choosing* $g(x) = \delta + (1-\delta)\zeta(x)$, Sf-KSD *with* $\mathcal{T}_{\mu_0,g}$ *recovers the survival* KSD *with Stein operator defined in Eq. (3),* $\mathbb{E}_0[\mathcal{T}_{\mu_0,g}\omega] \triangleq \mathbb{E}_0[(\mathcal{T}_0^{(s)}\omega)(x,\delta)] = 0$.

Comparisons between $\mathcal{T}_0^{(m)}$ and $\mathcal{T}_0^{(s)}$ From Theorem 2 and Corollary 1 above, we are able to explicitly see the following.

1) *in the uncensored part:* the diffusion for martingale Stein operator $\mathcal{T}_0^{(m)}$ is through the inverse of hazard function while the survival Stein operator $\mathcal{T}_0^{(s)}$ has constant auxiliary function, replicating constant diffusion in Eq. (12).

2) *in the censored part:* both Stein operators rely on the integral form where density μ_0 and test function ω interacts, while $\mathcal{T}_0^{(s)}$ involves the hazard function λ_0 within the integral that can be harder to estimate empirically.

Our results show that for $\mathcal{T}_0^{(s)}$, the censoring information is only extracted from the censored part of data ($\delta = 0$) while the uncensored part of data ($\delta = 1$) are treated exactly the same as Eq. (1). However for $\mathcal{T}_0^{(m)}$, the censoring information is re-weighted (or re-scaled) via both censored part and uncensored part of data, which results in more accurate empirical estimation compared to that of $\mathcal{T}_0^{(s)}$. The theoretical interpretations corroborate the empirical findings reported in Fernandez et al. [2020]. Additional comparisons and interpretation based on optimality conditions for KSD test in Section 4 are detailed in Appendix F.

Theorem 3 (Latent-variable Stein operator). *Assume* $q(x|z)$ *vanishes at infinity* $\forall z$. *Given sample* $z = \{z_j\}_{j\in[m]} \sim q(z|x)$, *the* z*-dependent function* $\mathbf{g}^z : \mathbb{R}^d \to \mathbb{R}^d$ *is chosen as* $g_i^z(x) = \sum_j \delta_{z_j}(x,z)$[7]. Sf-KSD *recovers the latent-variable Stein operator in Eq. (5),* $\mathbb{E}_q[\mathcal{T}_{q,\mathbf{g}^z}\mathbf{f}] \triangleq \sum_j \mathbb{E}_{q(x|z_j)}[\mathcal{T}_{q(x|z_j)}\mathbf{f}] = 0$.

By choosing the auxiliary function as the finite sum of delta measures on the latent variable locations, the auxiliary function is effectively performing the sampling procedure to construct the random kernel for the latent-variable Stein operator in Kanagawa et al. [2019], which surpasses the intractability arising from integral operation over the latent variables.

Theorem 4 (Second-order Stein operator). *Let* $g(x) = \frac{\partial}{\partial x}\log f(x)$, Sf-KSD *recovers the second-order Stein operator defined in Eq. (6),* $\mathcal{T}_{q,\log f'} f \triangleq \mathcal{T}_q^{(2)} f$.

[7] For test function $F(x,z)$, the bivariate delta function (of the second argument) satisfies $\int \delta_{z_j}(x,z)F(x,z)dz = F(x,z_j)$.

Using the fact that $g \cdot f = (\log f)' f = f'$, having $g = (\log f)'$ produces the extra order on the differential operator. For the more general formulation based on the linear operator are discussed in Appendix E. By choosing the linear operator itself to be the differential operator will automatically recover such second-order Stein operator. The multivariate version and the Riemannian manifold version are also applicable. Details are included in the Appendix E.

Theorem 5 (Coordinate-dependent Stein operator). *Let* $g_i(x) = \log J, \forall i \in [d]$, Sf-KSD *recovers the Stein operator defined in Eq. (7),*

$$\sum_i \int_{\mathcal{X}} \mathcal{T}_q(f_i \log J)dqJ = \sum_i \int_{\mathcal{X}} \frac{\partial}{\partial\theta^i}(f_i qJ) = 0.$$

Proof. The result follows from separating the last term in Eq. (7): $\log(qJ) = \log q + \log J$. □

With the particular choice of coordinate system, choosing the auxiliary function g to be the log of Jacobian can be interpreted as changing the diffusion pattern to incorporate coercive expectation w.r.t. taking expectation over the density. This can be explicitly shown using Stoke's theorem with differential forms [Xu and Matsuda, 2020]. In addition, the idea of using auxiliary function to incorporate domain properties or constraints can be very useful for problems where the data has a complicated and irregular domain. We provide case study on domain constraint distributions in Section 5.

4 OPTIMALITY CONDITIONS

We have analysed how the choice of fixed auxiliary function can provide a unifying view on KSD-tests in different testing scenarios. However, Stein operators for the same test scenario is non-unique and under what conditions the choice of Stein operators is the optimal? As we are thinking of optimality conditions for the auxiliary functions, we tackle this point via a *calculus of variation* approach [Gelfand et al., 2000].

Consider a cost function $L(y,\dot{y})$ (usually in integral form) that depends on a path function $y(x)$ and its derivative $\dot{y}(x) = \frac{d}{dx}y(x)$. The calculus variation approach perturbs the path function y by a minimal amount and ask when is y a stationary path w.r.t. $L(y,\dot{y})$. Then condition is then characterised via the following Euler-Lagrangian equation,

$$\frac{\partial}{\partial y}L(y(x),\dot{y}(x)) - \frac{d}{dx}\frac{\partial}{\partial\dot{y}}L(y(x),\dot{y}(x)) = 0. \quad (21)$$

Note that the differential w.r.t y is the path perturbation instead of moving variable x. Denote $\dot{f}_i = \frac{\partial f_i}{\partial x^i}$

and $\dot{\mathbf{f}} = (\dot{f}_1, \ldots, \dot{f}_d)$ (the same notion applies to $\mathbf{g}$). The cost Sf-KSD has the path integral form $L(\mathbf{f}, \dot{\mathbf{f}}, \mathbf{g}, \dot{\mathbf{g}}; p, q) = \mathbb{E}_p[\mathcal{T}_{q,\mathbf{g}}\mathbf{f}]$ based on Eq. (16).

Theorem 6 (Optimality conditions)**.** Sf-KSD *in Eq.* (17) *admits a stationary point w.r.t.* $\mathbf{f}$ *when the following holds,*

$$\mathbb{E}_p[g_i(x)\frac{\partial}{\partial x^i}\log q(x)] = 0, \quad \forall i \in [d]. \tag{22}$$

The proof follows from applying the Euler-Langrangian equation of Eq. (21) for each f_i, $i \in [d]$. As we choose f_i to be in the same function space and only interact with g_i due to Eq. (9) for all i, the symmetry on optimality conditions on g_i is expected.

Variance constraints and connections to the infinitesimal generator The optimality condition in Eq. (22) holds with flexibility in scaling. Following the fashion where KSD in Eq. (10) uses unit ball RKHS constraint, we would like to regularise the diffusion variance w.r.t. data distribution p. Making connection with diffusion process, its generator in Eq. (14) variance implies the diffusion function $\sigma(x)^2$ corresponds to $g(x)$. As such, we would like to regularise $\mathbb{V}ar_p[dX_t] \leq 1$. As Sf-KSD is monotonic w.r.t. scaling, our variance regularisation becomes equality constraint $\mathbb{E}_p[g(x)] = 1$.

It is not hard to check that the density ratio function, or the importance weight, $g(x) = \frac{q(x)}{p(x)}$ satisfies both the optimality conditions in Eq. (22) and the variance constraint. However, in goodness-of-ft testing, we do not know data distribution p (only accessible through sample form), we are unable to derive g directly. Density ratio and its estimation have been extensively studied in the literature [Hido et al., 2011; Kanamori et al., 2009; Sugiyama et al., 2012] and its link with kernel methods was pioneered in Kanamori et al. [2012]. Regarding density ratio argument, it is also interesting to see that $g(x) \equiv 1$ satisfies the optimality condition when $p = q$, which corresponds to the null hypothesis.

5 APPLICATION: TESTING WITH DOMAIN CONSTRAINTS

We show that different choices of auxiliary functions are able to induce appropriate Stein operators for various scenarios. Sf-KSD can then be useful to develop a systematic approach for new kernel Stein tests when appropriately auxiliary functions are used. In this section, we apply the Sf-KSD framework for testing data with general domain constraint.

Let q be a probability distribution defined on a compact domain[8] V with boundary ∂V. Denote the unnormalised density $\tilde{q}(x) \propto q(x), x \in V$. Common examples include the truncated Gaussian distribution on the interval $[a, b]$ or compositional data that defined on a simplex/hyper-simplex. Complex boundaries such as polygon [Liu and Kanamori, 2019; Yu et al., 2020] or non-negative constraint for graphical models [Yu et al., 2018] have been studied. Such settings are common when the observed data is only a subset of a larger domain or consists of structural constraint such as composition. For instance, if a local government would like to study the spread of the disease during pandemic if information about infections is not accessible from other countries, one may need to validate model assumptions with the domain truncated by the designated border.

5.1 Stein Operators on Compact Domains

To create the KSD-type test for data on domain V, we first consider Stein operators for densities on V. We develop such a Stein operator guided by the Sf-Stein operator in Eq. (16). Unlike densities on unbounded domains that are commonly assumed to vanish at infinity, densities on compact domains may not usually vanish at the boundary. Hence, direct application of a Stein operator on $\mathbb{R}^d$ may require the knowledge of *normalised* density at the boundary, which defeats the purpose of KSD testing for unnormalised models. To address this issue, we consider a bounded smooth function $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ such that $g_i(\partial V) = 0, \forall i \in [d]$ and for unnormalised $\tilde{q}$ on V, the bounded-domain Stein operator is defined as $\widetilde{\mathcal{T}}_{q,\mathbf{g}}\mathbf{f}(x) = \left(\frac{1}{q}\sum_i \frac{\partial}{\partial x^i}(qg_i f_i)\right)(x)$. With the aid from auxiliary function $\mathbf{g}$, it is not hard to check Stein's identity holds w.r.t. q. The recent advances in sampling with domain constraint [Shi et al., 2021] also utilise the Stein operator with mirror descent and perform SVGD [Liu and Wang, 2016] in ψ-transformed space. Based on the generator on similar Itô's process [Shi et al., 2021, Theorem 10], their Mirrored Stein operator corresponds to $\mathcal{T}_{q,G}$ in Eq. (15) with $G = \nabla^2\psi^{-1}$. In simplex case, they choose ψ to be negative-entropy to satisfy the boundary conditions.

5.2 Bounded-domain Kernel Stein Discrepancy

With the Stein operator $\widetilde{\mathcal{T}}_{\tilde{q},\mathbf{g}}$, we proceed to define the bounded-domain Kernel Stein Discrepancy (bd-KSD) for goodness-of-fit testing, similar to the Section 2.2. $\text{bd-KSD}_{\mathbf{g}}(\tilde{q}\|\tilde{p}) = \sup_{\mathbf{f}\in B_1(\mathcal{H})} \mathbb{E}_{\tilde{p}}[\widetilde{\mathcal{T}}_{q,\mathbf{g}}\mathbf{f}(x)]$. Standard reproducing property gives the quadratic form $\text{bd-KSD}_{\mathbf{g}}(\tilde{q}\|\tilde{p})^2 = \mathbb{E}_{x,x'\sim\tilde{q}}[h_{q,\mathbf{g}}(x, x')]$, where $h_{q,\mathbf{g}}(x, x') = \left\langle \widetilde{\mathcal{T}}_{q,\mathbf{g}}k(x, \cdot), \widetilde{\mathcal{T}}_{q,\mathbf{g}}k(x', \cdot)\right\rangle_{\mathcal{H}}$. Let $L(x) = (L_1(x), \ldots, L_d(x))^\top \in \mathbb{R}^d$ with $L_i(x) = \frac{\partial}{\partial x^i}\log\frac{q(x)}{p(x)}$, we

[8] It is common that V is embedded in some non-compact domain Ω, e.g. truncated distribution from $\mathbb{R}^d$.

show that under mild regularity conditions, bd-KSD is a proper discrepancy measure on V.

Goodness-of-fit testing with bd-KSD Similar procedure as introduced in Section 2.2 applies to test the null hypothesis $\mathrm{H}_0 : \tilde{p} = \tilde{q}$ against the alternative $\mathrm{H}_1 : \tilde{p} \neq \tilde{q}$. Observed samples $x'_1, \ldots, x'_n \sim \tilde{p}$ on V, the empirical U-statistic [Lee, 1990] can be computed, $\text{bd-KSD}_{\mathbf{g}}(\tilde{q}\|\tilde{p})^2_u = \frac{1}{n(n-1)} \sum_{i \neq j} [h_{q,\mathbf{g}}(x'_i, x'_j)]$. The asymptotic distribution is obtained via U-statistics theory [Van der Vaart, 2000] as follows. We denote the convergence in distribution by $\xrightarrow{d}$.

Theorem 7. *Consider U-statistic* $\text{bd-KSD}_{\mathbf{g}}(\tilde{q}\|\tilde{p})^2_u$. *1) Under* $H_0 : \tilde{p} = \tilde{q}$, $n \cdot \text{bd-KSD}_{\mathbf{g}}(\tilde{q}\|\tilde{p})^2_u \xrightarrow{d} \sum_{j=1}^{\infty} w_j (Z_j^2 - 1)$, *where* Z_j *are i.i.d. standard Gaussian random variables and* w_j *are the eigenvalues of the Stein kernel* $h_{q,\mathbf{g}}(x, x')$ *under* $\tilde{p}(x')$: $\int_V h_{q,\mathbf{g}}(x, x') \phi_j(x') \tilde{q}(x') \mathrm{d}x' = w_j \phi_j(x)$, *where* $\phi_j(x) \neq 0$ *is the non-trivial eigenfunction for Stein kernel operator* $h_{q,\mathbf{g}}$. *2) Under* $H_1 : \tilde{p} \neq \tilde{q}$,

$$\sqrt{n} \cdot \big(\text{bd-KSD}_{\mathbf{g}}(\tilde{q}\|\tilde{p})^2_u - \text{bd-KSD}_{\mathbf{g}}(\tilde{q}\|\tilde{p})^2\big) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{H_1}),$$

where $\sigma^2_{H_1} = \mathrm{Var}_{x \sim p}[\mathbb{E}_{\tilde{x} \sim p}[h_{q,\mathbf{g}}(x, x')]] > 0$ *produces the non-degenerate U-statistics.*

The goodness-of-fit testing then follows the standard procedures in Section 2.2 by applying bd-KSD.

5.3 Case Studies on Compact Domains

We first consider **truncated distributions**. In Fig. 1 (left), an example of two-components Gaussian mixture truncated in a unit ball is plotted in $B_1(\mathbb{R}^2)$. It is obvious that the density $q(x)$ does not necessarily vanish at the truncation boundary. Truncated distributions, including truncated Gaussian distributions [Horrace, 2005, 2015], truncated Pareto distributions [Aban et al., 2006], or truncated power-law distributions [Deluca and Corral, 2013] have been studied. In particular, left-truncated distributions are of special interest in survival analysis [Klein and Moeschberger, 2006]. To the best of our knowledge, goodness-of-fit testing procedures for general truncated distributions has not yet been established.

We also consider **compositional data** where distributions are defined on a simplex, $\mathrm{S}^{d-1} = \{x^i \in [0,1], \sum_{i=1}^d x^i = 1\}$, which is a compact domain. A common example for compositional distribution is the Dirichlet distribution, with *unnormalised* density of the form $\tilde{q}(x) \propto \Pi_{i=1}^d (x^i)^{\alpha_i - 1}$, $\forall x \in \mathrm{S}^{d-1}$, where $\alpha_i > 0$ are the *concentration parameters*. An example Dirichlet distribution on S^2 is illustrated in Fig. 1 (right). It is also obvious that $q(x)$ does not necessarily vanish at the boundary[9]. Recently, score matching procedures have been proposed to estimate unnormalised models for compositional data [Scealy and Wood, 2021]. To the best of our knowledge, goodness-of-fit testing procedures for general *unnormalised* compositional distributions has not been well studied.

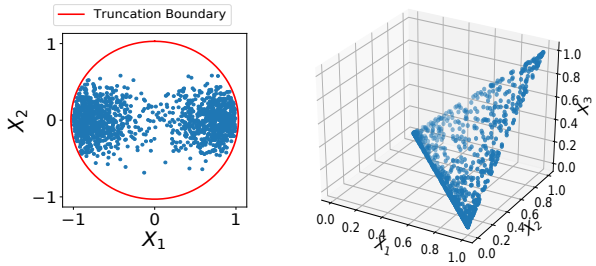


Figure 1: Sample distributions with compact domains. Left: two-component Gaussian mixtures truncated within a unit circle; Right: Dirichlet distribution on a simplex with $\boldsymbol{\alpha} = (0.2, 0.5, 0.5)$.

Comparing tests for goodness-of-fit An important aspect of applying bd-KSD is to choose the appropriate auxiliary functions in each data scenario, which we now specify. Simulation results are shown in Table. 1. We compare the KSD-based approach to the maximum mean discrepancy (MMD) [Gretton et al., 2012a] based approach where samples are generated from the null model and a two-sample test is then performed. Such strategy to test goodness-of-fit has been considered previously [Jitkrittum et al., 2017; Xu and Matsuda, 2020]. Another data-adaptive test based on kernel-embedding, named as Moderated MMD (M3D) [Balasubramanian et al., 2021], is also compared. Recently, Shi et al. [2021] has developed mirrored Stein discrepancy (MSD) for sampling purposes with applications for compositional data. We adapt and modify their formulation to perform goodness-of-fit testing procedure as comparison for our compositional data example, where the neg-entropy function is chosen as the mirrored map.

5.3.1 Simulation results on synthetic data

Truncated Distributions in Unit Ball $B_1(\mathbb{R}^d)$

Requiring to vanish at the boundary and take into account the rotational invariance of the unit ball, the auxiliary functions can be chosen as $g_i^{(p)}(x) = 1 - \|x\|^p$, which relates to the Euclidean distance from the boundary raising to chosen power p. Similar form of auxiliary function, with $p = 1$, was discussed for density estimation on truncated domains [Liu and Kanamori, 2019]. For larger p, more weights are concentrated to

[9] Specifically, $q(x) = 0$ at boundary $D_i = \{x | x^i = 0\}$ for $\alpha_i > 1$; while $q(x) > 0$ on D_i for $\alpha_i \in (0, 1]$.

Truncated Data on $B_1(\mathbb{R}^3)$

	$\nu = 0.1$	$\nu = 0.3$	$\nu = 1.0$
bd-KSD($g^{(1)}$)	0.235	0.760	**1.000**
bd-KSD($g^{(2)}$)	**0.305**	**0.910**	**1.000**
MMD	0.045	0.210	**1.000**
M3D	0.120	0.310	**1.000**

Compositional Data on S^2

	$\nu = 0.1$	$\nu = 0.3$	$\nu = 1.0$
bd-KSD($g^{(1)}$)	**0.430**	**0.715**	**0.905**
bd-KSD($g^{(2)}$)	0.325	0.565	0.855
MMD	0.090	0.425	0.730
M3D	0.135	0.505	0.805
MSD	0.230	0.610	0.895

Table 1: Test power for simulation results. perturbed level ν for the alternatives; sample size $n = 200$, test size $\alpha = 0.01$; repeat 200 trials. Bold number indicates the best power.

the center of the ball. We present the case where the null is a two-component Gaussian mixture with identity variances and the alternative is the two-component Gaussian mixture with correlation coefficient perturbed by ν. With such alternative distributions, $g^{(2)}$ is closer to the density ratio for optimality condition in Eq. (22), making bd-KSD($g^{(2)}$) a more powerful test as shown in Table. 1, outperform MMD-based tests. M3D improves from MMD tests with better data-adaptive scheme, while is still outperformed by bd-KSD based tests.

Compositional Distributions To vanish at the boundary $D_i = \{x|x^i = 0\}$, a natural choice of the auxiliary function is the geometric mean function $g_i^{(1)}(x) = (\Pi_{i=1}^d x^i)^{1/d}$. Moreover, minimum distance-to-boundary function $g_i^{(2)}(x) = \min\{\|x - z\| | z \in D_i\}$ satisfies the boundary conditions. We present Dirichlet distribution with concentration parameters $\alpha_i = 0.5$ as the null and perturb $\alpha_1 = 0.5 + \nu$ as the alternative. The geometric mean function $g^{(1)}$ is closer to the density ratio function, making it more sensitive detect the difference on the boundary compared to the minimum distance-to-boundary function $g^{(2)}$, bd-KSD($g^{(1)}$) produces higher power as shown in Table. 1. Moreover, results in Table. 1 also show that bd-KSD based tests outperforms the MMD based tests. Similar trends apply, M3D improves from MMD but is not more powerful than bd-KSD. MSD, being a specific form of Sf-KSD, performs competitively as compared to bd-KSD when the perturbation of the alternative becomes larger. We include additional details, simulation results and insights on $\mathbf{g}$ choices in Appendix C.

5.3.2 Real data experiments

We apply bd-KSD tests on two real datasets to assess the model goodness-of-fit. Models are fitted by density estimation techniques for these two data scenarios.

1. *Chicago Crime Dataset*[10]

The dataset contains all crime locations within the city of Chicago, where we use "robbery" data in 2020. We consider TruncSM [Liu and Kanamori, 2019], a score-matching based estimation objective, to fit a Gaussian mixture model[11]. We set auxiliary function as the Euclidean distance to the nearest boundary point (analogous to $g^{(2)}$). For a 2-component Gaussian mixture, bd-KSD gives p-values 0.002 which is clearly an inadequate fit; for a 20-component Gaussian mixture, p-value is 0.162 which indicates a good fit of the TruncSM estimated model.

2. *Three-composition AFM of 23 Aphyric Skye Lavas Dataset* [Aitchison and Lauder, 1985]

The variables A, F and M represent the relative proportions of $Na_2 + K_2O$, Fe_2O_3 and MgO, respectively. We fit the Gaussian kernel density estimation proposed by Chacón et al. [2011], using half of the data and test on the other half. We choose the auxiliary function $g^{(2)}$: the min distance to the closest boundary. The bd-KSD gives p-value 0.004 which rejects the null hypothesis, indicating the fit is not good enough.

6 CONCLUSION AND FUTURE DIRECTIONS

The present work studies a unifying framework Sf-KSD, to interpret and compare existing KSD-based goodness-of-fit tests; as well as to design new tests. Optimality conditions for developing Sf-KSD are studied, making connections between the score-based Stein operator and the generator approach. When performing goodness-of-fit tests, it is worth to note where the procedures may be mis-applied or wrongly interpreted in the wider scientific community where these must be guarded against. For instance, failure of p-value corrections from correlated samples can result in false positives, which is particularly risky in studies associated with healthcare. With density ratio satisfying the optimality conditions, it is an interesting unexplored area that how density ratio estimation or importance reweighting can be helpful in learning KSD in the context of goodness-of-fit testing, which we leave as a future direction.

[10] Data source https://data.cityofchicago.org.

[11] The model is fitted with half of the data and tested on the other half, which is referred to as "data-splitting" techniques [Jitkrittum et al., 2017; Kübler et al., 2020]

Acknowledgements

W.X. acknowledges Gesine Reinert for the enlightenment on Stein's method and introduction to the standardisation techniques that inspired this work and many helpful discussions. W.X. would like to thank Arthur Gretton and Nan Lu for helpful discussions and constructive comments. W.X would also like to thank anonymous reviewers for helpful comments. W.X. is supported by the EPSRC grant EP/T018445/1.

References

Aalen, O., Borgan, O., and Gjessing, H. (2008). *Survival and event history analysis: a process point of view*. Springer Science & Business Media.

Aban, I. B., Meerschaert, M. M., and Panorska, A. K. (2006). Parameter estimation for the truncated pareto distribution. *Journal of the American Statistical Association*, 101(473):270–277.

Aitchison, J. and Lauder, I. (1985). Kernel density estimation for compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 34(2):129–137.

Anastasiou, A., Barp, A., Briol, F.-X., Ebner, B., Gaunt, R. E., Ghaderinezhad, F., Gorham, J., Gretton, A., Ley, C., Liu, Q., et al. (2021). Stein's method meets statistics: A review of some recent developments. *arXiv preprint arXiv:2105.03481*.

Balasubramanian, K., Li, T., and Yuan, M. (2021). On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of machine learning research*, 22(1).

Barbour, A. (2005). Multivariate poisson-binomial approximation using stein's. *Stein's Method and Applications*, 5:131.

Barbour, A., Luczak, M. J., Xia, A., et al. (2018). Multivariate approximation in total variation, ii: Discrete normal approximation. *The Annals of Probability*, 46(3):1405–1440.

Barbour, A. D. (1988). Stein's method and poisson process convergence. *Journal of Applied Probability*, 25(A):175–184.

Barbour, A. D. and Chen, L. H. Y. (2005). *An introduction to Stein's method*, volume 4. World Scientific.

Barp, A., Briol, F.-X., Duncan, A., Girolami, M., and Mackey, L. (2019). Minimum stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, pages 12964–12976.

Barp, A., Oates, C., Porcu, E., and Girolami, M. (2018). A riemannian-stein kernel method. *arXiv preprint arXiv:1810.04946*.

Berlinet, A. and Thomas, C. (2004). *Reproducing kernel Hilbert spaces in Probability and Statistics*. Kluwer Academic Publishers.

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

Chacón, J. E., Duong, T., and Wand, M. (2011). Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, pages 807–840.

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. (2019). Stein point markov chain monte carlo. In *International Conference on Machine Learning*, pages 1011–1021. PMLR.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. (2018). Stein points. In *International Conference on Machine Learning*, pages 844–853. PMLR.

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

Chwialkowski, K. P., Sejdinovic, D., and Gretton, A. (2014). A wild bootstrap for degenerate kernel tests. In *Advances in neural information processing systems*, pages 3608–3616.

Deluca, A. and Corral, Á. (2013). Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. *Acta Geophysica*, 61(6):1351–1394.

Ernst, M., Reinert, G., and Swan, Y. (2020). First-order covariance inequalities via stein's method. *Bernoulli*, 26(3):2051–2081.

Fernandez, T. and Gretton, A. (2019). A maximum-mean-discrepancy goodness-of-fit test for censored data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2966–2975.

Fernandez, T., Rivera, N., Xu, W., and Gretton, A. (2020). Kernelized stein discrepancy tests of goodness-of-fit for time-to-event data. In *International Conference on Machine Learning*, pages 3112–3122. PMLR.

Fisher, M., Nolan, T., Graham, M., Prangle, D., and Oates, C. (2021). Measure transport with kernel stein discrepancy. In *International Conference on Artificial Intelligence and Statistics*, pages 1054–1062. PMLR.

Gelfand, I. M., Silverman, R. A., et al. (2000). *Calculus of variations*. Courier Corporation.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Gorham, J., Duncan, A. B., Vollmer, S. J., and Mackey, L. (2019). Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928.

Gorham, J. and Mackey, L. (2015). Measuring sample quality with stein's method. In *Advances in Neural Information Processing Systems*, pages 226–234.

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

Gorham, J., Raj, A., and Mackey, L. (2020). Stochastic stein discrepancies. In *NeurIPS*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012a). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. (2008). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, pages 585–592.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213.

Han, J., Ding, F., Liu, X., Torresani, L., Peng, J., and Liu, Q. (2020). Stein variational inference for discrete distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 4563–4572. PMLR.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., and Kanamori, T. (2011). Statistical outlier detection using direct density ratio estimation. *Knowledge and information systems*, 26(2):309–336.

Horrace, W. C. (2005). Some results on the multivariate truncated normal distribution. *Journal of multivariate analysis*, 94(1):209–221.

Horrace, W. C. (2015). Moments of the truncated normal distribution. *Journal of Productivity Analysis*, 43(2):133–138.

Huggins, J. H. and Mackey, L. (2018). Random feature stein discrepancies. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 1903–1913.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709.

Jitkrittum, W. (2017). *Kernel-based distribution features for statistical tests and Bayesian inference*. PhD thesis, UCL (University College London).

Jitkrittum, W., Kanagawa, H., and Schölkopf, B. (2020). Testing goodness of fit of conditional density models with kernels. In *Conference on Uncertainty in Artificial Intelligence*, pages 221–230. PMLR.

Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K., and Gretton, A. (2017). A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems*, pages 262–271.

Kanagawa, H., Jitkrittum, W., Mackey, L., Fukumizu, K., and Gretton, A. (2019). A kernel stein test for comparing latent variable models. *arXiv preprint arXiv:1907.00586*.

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445.

Kanamori, T., Suzuki, T., and Sugiyama, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3):335–367.

Kim, B., Khanna, R., and Koyejo, O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2288–2296.

Klein, J. P. and Moeschberger, M. L. (2006). *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media.

Kübler, J., Jitkrittum, W., Schölkopf, B., and Muandet, K. (2020). Learning kernel tests without data splitting. *Advances in Neural Information Processing Systems*, 33.

Le, H., Lewis, A., Bharath, K., and Fallaize, C. (2020). A diffusion approach to stein's method on riemannian manifolds. *arXiv preprint arXiv:2003.11497*.

Lee, A. J. (1990). *U-Statistics: Theory and Practice*. CRC Press.

Ley, C., Reinert, G., Swan, Y., et al. (2017). Stein's method for comparison of univariate distributions. *Probability Surveys*, 14:1–52.

Lim, J. N., Yamada, M., Schölkopf, B., and Jitkrittum, W. (2020). Kernel stein tests for multiple model comparison. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, pages 2232–2242.

Liu, C. and Zhu, J. (2018). Riemannian stein variational gradient descent for bayesian inference. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. (2020). Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*, pages 6316–6326. PMLR.

Liu, Q., Lee, J., and Jordan, M. (2016). A kernelized stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, pages 276–284.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2378–2386.

Liu, S. and Kanamori, T. (2019). Estimating density models with complex truncation boundaries. *arXiv preprint arXiv:1910.03834*.

Lloyd, J. R. and Ghahramani, Z. (2015). Statistical model criticism using kernel two sample tests. In *NeurIPS*, pages 829–837.

Lyu, S. (2009). Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 359–366.

Mijoule, G., Reinert, G., and Swan, Y. (2018). Stein operators, kernels and discrepancies for multivariate continuous distributions. *arXiv preprint arXiv:1806.03478*.

Mijoule, G., Reinert, G., and Swan, Y. (2021). Stein's density method for multivariate continuous distributions. *arXiv preprint arXiv:2101.05079*.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1-2):1–141.

Oates, C. J., Cockayne, J., Briol, F.-X., and Girolami, M. (2019). Convergence rates for a class of estimators based on stein's method. *Bernoulli*, 25(2):1141–1159.

Scealy, J. L. and Wood, A. T. A. (2021). Score matching for compositional distributions. *Journal of the American Statistical Association*, 0(ja):1–32.

Shi, J., Liu, C., and Mackey, L. (2021). Sampling with mirrored stein operators. *arXiv preprint arXiv:2106.12506*.

Stein, C. et al. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California.

Stramer, O. and Tweedie, R. (1999). Langevin-type models i: Diffusions with given stationary distributions and their discretizations. *Methodology and Computing in Applied Probability*, 1(3):283–306.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

Sutherland, D., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A. J., and Gretton, A. (2017). Generative models and model criticism via optimized maximum mean discrepancy. In *ICLR (Poster)*.

Van der Vaart, A. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Xu, W. and Matsuda, T. (2020). A stein goodness-of-fit test for directional distributions. *The 23rd International Conference on Artificial Intelligence and Statistics*.

Xu, W. and Matsuda, T. (2021). Interpretable stein goodness-of-fit tests on riemannian manifolds. *International Conference on Machine Learning*.

Xu, W. and Reinert, G. (2021). A stein goodness-of-test for exponential random graph models. In *International Conference on Artificial Intelligence and Statistics*, pages 415–423. PMLR.

Yang, J., Liu, Q., Rao, V., and Neville, J. (2018). Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pages 5557–5566.

Yang, J., Rao, V., and Neville, J. (2019). A stein–papangelou goodness-of-fit test for point processes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 226–235.

Yu, S., Drton, M., and Shojaie, A. (2018). Graphical models for non-negative data using generalized score matching. In *International conference on artificial intelligence and statistics*, pages 1781–1790. PMLR.

Yu, S., Drton, M., and Shojaie, A. (2020). Generalized score matching for general domains. *arXiv preprint arXiv:2009.11428*.

Supplementary Material: Standardisation-function Kernel Stein Discrepancy: A Unifying View on Kernel Stein Discrepancy Tests for Goodness-of-fit

A Known Identities

Expectations in Survival Analysis

We know the following identities in survival analysis, which will be useful for discussions in the main text: for any measurable function ϕ,

$$\mathbb{E}_0[\Delta\phi(T)] = \int_0^\infty \phi(s)\mu_0(s)S_C(s)ds, \tag{23}$$

$$\mathbb{E}_0[(1-\Delta)\phi(T)] = \int_0^\infty \phi(s)\mu_C(s)S_0(s)ds. \tag{24}$$

where μ_C here denotes the p.d.f. of the censoring distribution and S_0 denotes the survival function w.r.t. μ_0.

Martingales in Survial Analysis

The following identity is useful to understand the martingale Stein operator in Fernandez et al. [2020]

$$\mathbb{E}_0\left[\Delta\phi(T) - \int_0^T \phi(t)\lambda_0(t)dt\right] = 0, \tag{25}$$

which holds under the null hypothesis, where λ_0 is the hazard function under the null μ_0. Let $N_i(x)$ and $Y_i(x)$ be the individual counting and risk processes, defined by by $N_i(x) = \delta_i \mathbb{1}_{\{T_i \le x\}}$ and $Y_i(x) = \mathbb{1}_{\{T_i \ge x\}}$, respectively. Then, the individual zero-mean martingale for the i-th individual corresponds to $M_i(x) = N_i(x) - \int_0^x Y_i(y)\lambda_0(y)dy$, where $\mathbb{E}_0[M_i(x)] = 0$ for all x.

Additionally, let $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that $\mathbb{E}_0\left|\int_0^x \phi(y)dM_i(y)\right| < \infty$ for all x, then $\int_0^x \phi(y)dM_i(y)$ is a zero-mean $(\mathcal{F}_x)$-martingale (see Chapter 2 of [Aalen et al., 2008]). Then, taking expectation, we have

$$\begin{aligned}\mathbb{E}_0\left[\int_0^\infty \phi(x)dM_i(x)\right] &= \mathbb{E}_0\left[\int_0^\infty \phi(x)(dN_i(x) - Y_i(x)\lambda_0(x)dx)\right] \\ &= \mathbb{E}_0\left[\Delta\phi(T) - \int_0^T \phi(x)\lambda_0(x)dx\right] = 0,\end{aligned}$$

as stated above. The martingale property is useful to derive the martingale Stein operator in Eq. (4). For more details, see Fernandez et al. [2020].

B Proofs and Derivations

Proof of Proposition 1

Proof. Standard reproducing properties and taking the supremum over unit ball RKHS apply,

$$\text{Sf-KSD}_{\mathbf{g}}(p\|q;\mathcal{H}) = \sup_{\|\mathbf{f}\|_{\mathcal{H}}\le 1} \mathbb{E}_p[\langle \mathcal{T}_{q,\mathbf{g}}K(x,\cdot), \mathbf{f}\rangle_{\mathcal{H}}] = \|\mathbb{E}_p[\mathcal{T}_{q,\mathbf{g}}K(x,\cdot)]\|_{\mathcal{H}}.$$

Specifically, assume $f_i(x) = \langle k(x,\cdot), f_i\rangle$, the setting in Chwialkowski et al. [2016]; Liu et al. [2016],

$$\mathcal{T}_{q,\mathbf{g}}K(x,\cdot) = \sum_{i=1}^{d} g_i(x)\left(\frac{\partial \log q(x)}{\partial x^i}k(x,\cdot) + \frac{\partial k(x,\cdot)}{\partial x^i}\right) + \frac{\partial g_i(x)}{\partial x^i}k(x,\cdot).$$

We can write $h_{q,\mathbf{g}}(x,\tilde{x})$ explicitly as

$$\begin{aligned} &h_{q,\mathbf{g}}(x,\tilde{x}) = \\ &\sum_{i=1}^{d}\left(\frac{\partial^2 k(x,\tilde{x})}{\partial x^i \partial \tilde{x}^i} + \frac{\partial \log q(x)}{\partial x^i}\frac{\partial k(x,\tilde{x})}{\partial \tilde{x}^i} + \frac{\partial \log q(\tilde{x})}{\partial \tilde{x}^i}\frac{\partial k(x,\tilde{x})}{\partial x^i} + \frac{\partial \log q(x)}{\partial x^i}\frac{\partial \log q(\tilde{x})}{\partial \tilde{x}^i}k(x,\tilde{x})\right)\times \\ &\quad g_i(x)g_i(\tilde{x}) + \frac{\partial g_i(x)}{\partial x^i}\frac{\partial g_i(\tilde{x})}{\partial \tilde{x}^i}k(x,\tilde{x}). \end{aligned}$$

which recovers the quadratic form, which only depends on density q but not p. □

Proof of Theorem 1

Proof. Note that the expectation on l.h.s. of Eq. (19) is integrating over the density of survival time μ_0 where the expectation on the r.h.s., having the multiplication of δ in $\mathcal{T}_0\omega$ in Eq. (2), is taken over the paired observation incorporating censoring information. Using the identity in Eq. (24), we have

$$\begin{aligned} \mathbb{E}_{\mu_0}[\mathcal{T}_{\mu_0,g}\omega] &= \int_{\mathbb{R}_+} \mathcal{T}_{\mu_0,g}\omega(s)\mu_0(s)ds \\ &= \int_{\mathbb{R}_+}\left(\omega'(s) + \omega(s)(\frac{g'(s)}{g(s)} + (\log\mu_0(s))')\right)g(s)\mu_0(s)ds \\ &= \int_{\mathbb{R}_+}\left(\omega'(s) + \frac{\omega(s)S_C'(s)}{S_C(s)} + \frac{\omega(s)\mu_0(x)'}{\mu_0(s)}\right)\mu_0(s)S_C(s)ds \\ &= \int_{\mathbb{R}_+}\frac{\omega'(s)S_C(s)\mu_0(s) + \omega(s)S_C'(s)\mu_0(s) + \omega(s)S_C(s)\mu_0(x)'}{S_C(s)\mu_0(s)}\mu_0(s)S_C(s)ds \\ &= \int_{\mathbb{R}_+}(\mathcal{T}_0\omega)(x,\delta)\mu_0(x)S_C(x)ds = \mathbb{E}_0[(\mathcal{T}_0\omega)(x,\delta)] = 0. \end{aligned}$$

We also note that $S_C(0) = 1, S_C(\infty) = 0$ by definition of survival functions. As such, $g(x)$ is bounded almost everywhere in $\mathbb{R}_+$ which satisfy the conditions for testing. □

Proof of Theorem 2

Proof. To show the equivalence relation in the sense of Eq. (20), we need to consider the presence of indicator variable δ. This is essentially different from the proof of Theorem 1. Recall the following identity between hazard function and density: $\log\mu_0(s) = \frac{\mu_0'(x)}{\mu_0(x)} = \frac{\lambda_0'(x)}{\lambda_0(x)} - \lambda_0(x)$ since

$$\frac{\lambda_0'(x)}{\lambda_0(x)} = \frac{\mu_0'(x)}{S_0(x)\lambda_0(x)} + \frac{\mu_0(x)^2}{S_0(x)^2\lambda_0(x)} = \frac{\mu_0'(x)}{\mu_0(x)} + \lambda_0(x). \tag{26}$$

Denote $\zeta(x) = \frac{\int_0^x \mu_0(s)\omega(s)ds}{\mu_0(x)\omega(x)}$, such that we can write $g = \delta\lambda_0^{-1} + (1-\delta)\zeta$. Decompose the Stein operator $\mathcal{T}_{\mu_0,g}$ w.r.t. δ, we have

$$\mathcal{T}_{\mu_0,g}\omega = \delta\mathcal{T}_{\mu_0,\lambda_0^{-1}}\omega + (1-\delta)\mathcal{T}_{\mu_0,\zeta}\omega \tag{27}$$

as $\mathcal{T}_{\mu_0}$ is also linear operator w.r.t g. We now decompose the above two components $\mathcal{T}_{\mu_0,\lambda_0^{-1}}$ and $\mathcal{T}_{\mu_0,\zeta}$ using the form of Eq. (16),

$$\begin{aligned}\mathcal{T}_{\mu_0,\lambda_0^{-1}}\omega &= \lambda_0^{-1}\left(\omega' + \omega\log\mu_0'\right) + \lambda_0^{-1\prime}\omega \\ &= \lambda_0^{-1}\left(\omega' + \omega(\frac{\lambda_0'(x)}{\lambda_0(x)} - \lambda_0(x)) + \lambda_0\lambda_0^{-1\prime}\omega\right) \\ &= \lambda_0^{-1}\left(\omega' - \omega\lambda_0\right) \\ &= \lambda_0^{-1}\omega' - \omega\end{aligned}$$

the second line equality follows from Eq. (26) while the third line follows from $\lambda_0^{-1\prime} = -\frac{\lambda_0{}'}{\lambda_0^2}$. The derivation is interesting that it reveals that the uncensored data in the martingale Stein operator is connected to the Langevin-diffusion via the inverse of hazard function, i.e. when $\delta \equiv 1$ (or absence of censoring), $\mathcal{T}_{\mu_0,\lambda_0^{-1}}\omega = \mathcal{T}_0^{(m)}\omega$.

On the other hand, we rewrite the martingale Stein operator in Eq. (4) as $(\mathcal{T}_0^{(m)}\omega)(x,\delta) = \delta\frac{\omega'(x)}{\lambda_0(x)} - \omega(x) = \delta(\frac{\omega'(x)}{\lambda_0(x)} - \omega(x)) - (1-\delta)\omega(x)$. For Sf-KSD to match this operator, we need to find ζ such that $\mathcal{T}_{\mu_0,\zeta}\omega = \omega$.

$$\mathcal{T}_{\mu_0,\zeta}\omega = \omega\left(\zeta' + \zeta\log\mu_0'\right) + \omega'\zeta = \omega\left(\zeta' + \zeta\log\mu_0' + \zeta(\log\omega)'\right) = -\omega. \tag{28}$$

As $\omega(x) > 0$ for $\mu_0(x) > 0$ for $x > 0$, Eq. (28) gives the following autonomous differential equation form

$$\zeta' = -\zeta(\log\mu_0' + \log\omega') - 1, \tag{29}$$

solving which yields

$$\begin{aligned}\zeta(x)e^{\log\mu_0(x)+\log\omega(x)} &= \int_0^x -e^{\log\mu_0(s)+\log\omega(s)}ds \\ \zeta(x)\mu_0(x)\omega(x) &= \int_0^x \mu_0(s)\omega(s)ds\end{aligned}$$

as $\omega(0) = 0$ by assumption. $\zeta(x) = \frac{\int_0^x \mu_0(s)\omega(s)ds}{\mu_0(x)\omega(x)}$ as proposed. Putting together, we have

$$\mathcal{T}_{\mu_0,g}\omega = \delta\mathcal{T}_{\mu_0,\lambda_0^{-1}}\omega + (1-\delta)\mathcal{T}_{\mu_0,\zeta}\omega = \delta(\frac{\omega'(x)}{\lambda_0(x)} - \omega(x)) - (1-\delta)\omega = \mathcal{T}_0^{(m)}\omega$$

and Stein's identity result follows by taking the expectation of the same form, $\mathbb{E}_0[\mathcal{T}_{\mu_0,g}\omega] = \mathbb{E}_0[\mathcal{T}_0^{(m)}\omega] = 0$. □

Proof of Corollary 1

Proof. The proof follows from decomposing the survival Stein operator $\mathcal{T}_0^{(s)}\omega$ in to the uncensored part and censored part, similar to Eq. (27),

$$\begin{aligned}(\mathcal{T}_0^{(s)}\omega)(x,\delta) &= \delta\omega'(x) + \delta\omega(x)\frac{\lambda_0'(x)}{\lambda_0(x)} - \omega(x)\lambda_0(x) \\ &= \delta\omega'(x) + \delta\omega(x)\left(\frac{\lambda_0'(x)}{\lambda_0(x)} - \lambda_0(x)\right) - (1-\delta)\omega(x)\lambda_0(x) \\ &= \delta\omega'(x) + \delta\omega(x)\frac{\mu_0'(x)}{\mu_0(x)} - (1-\delta)\omega(x)\lambda_0(x) \\ &= \delta\Big(\omega'(x) + \omega(x)\log\mu_0(x)'\Big) - (1-\delta)\omega(x)\lambda_0(x).\end{aligned}$$

where the term involving δ, the uncensored part, is just the Langevin-diffusion Stein operator in 1d. Similar to Eq. (28), we solve the following autonomous differential equation for the censored part,

$$\mathcal{T}_{\mu_0,\zeta}\omega = \omega\left(\zeta' + \zeta\log\mu_0' + \zeta(\log\omega)'\right) = -\omega\lambda_0. \tag{30}$$

which simplifies to

$$\zeta' = -\zeta(\log \mu_0' + \log \omega') - \lambda_0. \tag{31}$$

Solving the differential equation with boundary condition $\omega(0) = 0$, we get $\zeta(x) = \frac{\int_0^x \mu_0(s)\omega(s)\lambda_0(s)ds}{\mu_0(x)\omega(x)}$. As such, using $g = \delta + (1-\delta)\zeta$, the result follows

$$\mathcal{T}_{\mu_0,g}\omega = \delta\mathcal{T}_{\mu_0}\omega + (1-\delta)\mathcal{T}_{\mu_0,\zeta}\omega = (\mathcal{T}_0^{(s)}\omega)(x,\delta).$$

□

Remarks In the main text, we discussed the advantages and disadvantages of KSD-based test with $\mathcal{T}_0^{(m)}$ and $\mathcal{T}_0^{(s)}$, which corroborate the empirical findings in Fernandez et al. [2020]. Moreover, Fernandez et al. [2020] studied the testing procedure via c.d.f. transformation followed by testing the uniform null density, which they call **model-free implementation**. This procedure has been shown to achieve higher test power. Similar testing strategy via c.d.f. transformation has been studied in Fernandez and Gretton [2019] using MMD-based test. Notice that since F_0 is monotone and $u_i = F_0(t_i) = \min\{F_0(x_i), F_0(c_i)\}$, thus δ_i remains consistent. Under this transformation, the null hypothesis is equivalent to test whether $F_0(x_i)$ is distributed as a uniform random variable. In this setting, the observations for the test is based on $\{(u_i, \delta_i)\}_{i\in[n]}$, where the Stein operator used is independent of density of $x \sim f_0$. Instead, $\mu_0(u) \equiv 1$ and $\lambda_0 = \lambda_{\mathcal{U}} = \frac{1}{1-x}$ are used to construct the Stein operator and

$$(\mathcal{T}_0^{(m)}\omega)(u,\delta) = \delta\omega'(u)(1-u) - \omega(u)$$

for $u = F_0(x)$ (notice that $F_0(0) = 0$). From our result, we see that with the particular choice of the uniform null, there is no more interaction between test function and density in the censored part, e.g. $\zeta(x) = \frac{\int_0^x \omega(s)/(1-s)ds}{\omega(x)}$, resulting a better estimation accuracy from the samples, thus higher test power.

Similarly, Fernandez et al. [2020] exploited another monotone transformation via the cumulative hazard function from the null Λ_0, such that $\Lambda_0(X) \sim \text{Exp}(1)$. In this case, $\mu_0(x) = \exp(-x)$ which still require interaction between μ_0 and λ_0 in ζ, resulting in decrease in estimation accuracy. Our results explain the empirical finding in Fernandez et al. [2020] that the test power from the model-free implementation of the test using cumulative hazard transformation is not higher than using the c.d.f. transformation.

Proof of Theorem 3

Proof. As g_i consists of finite sum of delta measures on locations z_j sampled from $q(z|x)$, it's derivative is 0 everywhere excluding a finite number of points which is a set of measure zero. Recall that the marginalisation of the density $q(x) = \int q(x,z)dz$,

$$\begin{aligned}
\mathbb{E}_q[\mathcal{T}_{q,\mathbf{g}}\mathbf{f}] &= \int \sum_i \mathcal{T}_q(g_i^z(x)f_i(x))q(x,z)dxdz \\
&\overset{(a)}{=} \int_{\mathcal{Z}} \sum_i \Big(\int_{\mathcal{X}} \left(\log q(x)'(g_i^z(x)f_i(x)) + (g_i^z(x)f_i(x))'\right) q(x|z)dx \Big)\pi(z)dz \\
&\overset{(b)}{=} \sum_i \int \sum_j \left(\log q(x|z_j)' f_i(x) + f_i(x)'\right) q(x|z_j)dx \\
&= \sum_j \mathbb{E}_{q(x|z_j)}[\mathcal{T}_{q(x|z_j)}\mathbf{f}] = \mathbb{E}_q[\mathcal{T}_{q,z}\mathbf{f}]
\end{aligned}$$

where equality (a) uses the marginalisation and equality (b) utilises the the fact that z_j are samples from $q(z|x) \propto \pi(z)q(x|z)$ and the bivariate delta function gives $\int \delta_{z_j}(x,z)q(x,z)dz = q(x,z_j) = q(x|z_j)\pi(z_j)$. □

Proof of Theorem 4

Proof. It is not difficult to see that

$$\mathcal{T}_q(fg) = \mathcal{T}_q(f(\log f)') = \mathcal{T}_q(f') = f'' + \log q' f' \tag{32}$$

which is the second-order operator in 1d.

For the multivariate case, choosing $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ such that $g_i(x) = \frac{\partial}{\partial x^i} f_i(x) \partial x^i$, where ∂x^i is the differential form defined in the main text. Similar derivation as Eq. (32) above, we have

$$\mathcal{T}_{q,\mathbf{g}} \mathbf{f} = \sum_i \mathcal{T}_{q,(\frac{\partial}{\partial x^i} \log_i f_i)} f_i = \sum_i \frac{\partial^2}{\partial {x^i}^2} f_i \partial x^i + \log q' \frac{\partial}{\partial x^i} f_i \partial x^i = \mathcal{T}_q \nabla \cdot \mathbf{f}.$$

□

Theorem 8 (Characterisation of bd-KSD)**.** *Let $\tilde{p}$, $\tilde{q}$ be smooth densities defined on V. Assume: 1) kernel k is compact universal* [Carmeli et al., 2010, Definition 2(ii)]*; 2) $\mathbb{E}_{x,x' \sim \tilde{q}}[h_{q,\mathbf{g}}(x,x')^2] < \infty$; 3) $\mathbb{E}_{\tilde{q}} \|L(x)\|^2 < \infty$; 4) $g_i(x) > 0$ whenever $q(x) > 0$. Then,* bd-KSD$_{\mathbf{g}}(\tilde{q} \| \tilde{p}) \geq 0$ *and* bd-KSD$_{\mathbf{g}}(\tilde{q} \| \tilde{p}) = 0$ *if and only if $\tilde{p} = \tilde{q}$.*

Proof of Theorem 8

Proof. The Theorem extends from [Chwialkowski et al., 2016, Theorem 2.2] with additional assumptions $g_i(x) > 0$ if $\tilde{q}(x) > 0$, together with appropriate compact universality condition for the kernel. For more general settings, having a proper notion of universal kernel would extend Theorem 8 to show that bd-KSD is a proper discrepancy in desired testing scenarios.

Denote $\mathbf{s}_{q,\mathbf{g}}(\cdot) = \mathbb{E}_{x \sim p}[\widetilde{\mathcal{T}}_{q,\mathbf{g}} k(x, \cdot)] \in \mathcal{H}$ and we can write the quadratic form bd-KSD as bd-KSD$_{\mathbf{g}}(\tilde{q} \| \tilde{p})^2 = \|\mathbf{s}_{q,\mathbf{g}}(\cdot)\|^2_{\mathcal{H}} \geq 0$. If $p = q$, then bd-KSD$_{\mathbf{g}}(\tilde{q} \| \tilde{p})^2 = 0$ from Stein's identity.

Conversely, if bd-KSD$_{\mathbf{g}}(\tilde{q} \| \tilde{p})^2 = 0$, then $\mathbf{s}_{q,\mathbf{g}}(x) = \mathbf{0}$, $\forall x$, s.t. $p(x) > 0$. Then, from $\log(q/p) = \log(\tilde{q}) - \log(\tilde{p})$, we obtain,

$$\mathbb{E}_{x' \sim \tilde{p}} \left[L_i(x') k(x', x) \right] = (\mathbf{s}_{q,\mathbf{g}})_i(x) - \mathbb{E}_{x' \sim \tilde{p}} \left[\widetilde{\mathcal{T}}_{q,\mathbf{g}} k(x', x) \right] = 0,$$

for every x with positive densities. As $g_i(x) > 0$ for $q(x) > 0$, and k is compact-universal at V, the injectivity result in [Carmeli et al., 2010, Theorem 4(b)] implies that $L_i = 0, \forall i \in [d]$. Therefore, $\log(\tilde{q}/\tilde{p})$ is constant on V. Since both $\tilde{p}$ and $\tilde{q}$ are both densities on V that integrate to one, we conclude $\tilde{p} = \tilde{q}$. □

C Additional Simulation Results

We investigate the test performances on various problems, comparing different choice of $\mathbf{g}$. In the following problems, we apply the Gaussian kernel with median distance as bandwidth [Gretton et al., 2012a].

1. Truncated Gaussian distribution in $B_1(\mathbb{R}^3)$

$$\tilde{q}_\nu(x) \propto \mathcal{N}(x|0, \Sigma_\nu), \forall x \in B_1(\mathbb{R}^3), \Sigma_\nu = \begin{pmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{33}$$

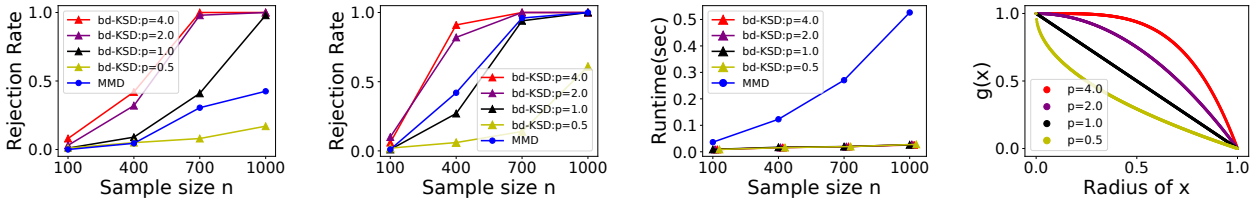
where $\nu > -1$. We test the null model $\tilde{q}_0$ against the alternative $\tilde{q}_1$ with perturbation of variance parameter $\nu = 1.0$. The test power of bd-KSD, with different choice of $g^{(p)}$, is shown in Fig. 2(a). All tests have increasing test powers as the simple size increases, which is what we expect. As the alternative is having variance difference in the first direction x^1, the test captures such difference better when more weights are put on near the origin (refer Fig. 2(d) for $g(x)$ values). Hence, the test power increases with the increase of parameter (p) for auxiliary function $g^{(p)}$. For $p > 1$, the bd-KSD tests outperforms the MMD test[12].

The rejection rate under the null is reported in Table 2. The bd-KSD tests, with appropriately choice of g incorporating the boundary conditions, achieve well-controlled Type-I errors. MMD-based tests also achieves the well-controlled Type-I errors. However, applying the KSD tests in Eq. (11) are unable to have controlled test level due to the violation of Stein's identity, which is what we expect.

[12] For all MMD-based tests, we draw n samples from the null samples when the sample size of observed data is n.

	n=100	n=400	n=700	n=1000
bd-KSD(p=4)	0.004	0.010	0.030	0.018
bd-KSD(p=2)	0.008	0.018	0.022	0.012
bd-KSD(p=1)	0.010	0.016	0.020	0.028
bd-KSD(p=0.5)	0.006	0.008	0.010	0.030
MMD	0.022	0.014	0.018	0.032
KSD($g(x) \equiv 1$)	1.00	1.00	1.00	1.00

Table 2: Rejection rate under the null; $\alpha = 0.01$; 500 trials.



(a) Gaussian distributions (b) Gaussian mixture (c) Computation runtime (d) Auxiliary function

Figure 2: Truncated distribution in unit ball $B_1(\mathbb{R}^3)$; test level $\alpha = 0.01$; test repeats 200 trials.

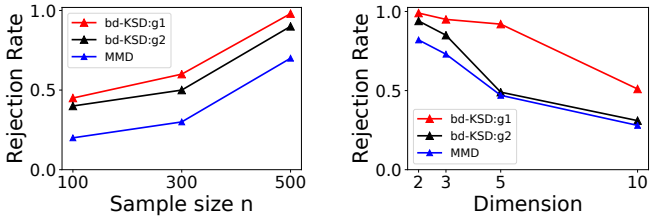


Figure 3: Dirichlet distributions on simplex S^{d-1}; test level $\alpha = 0.01$; test repeats 200 trials.

2. Truncated Gaussian mixture in $B_1(\mathbb{R}^2)$

$$\tilde{q}_\nu(x) \propto \frac{1}{2}\mathcal{N}(x|\mu_1, \Sigma_\nu) + \frac{1}{2}\mathcal{N}(x|\mu_2, \Sigma_\nu), \forall x \in B_1(\mathbb{R}^2),$$

where $\mu_1 = (-1, 0, 0)$, $\mu_2 = (1, 0, 0)$ and Σ_ν (as defined above) is shared between two components. We test the null model $\tilde{q}_0$ against the alternative $\tilde{q}_{-\frac{1}{2}}$. The test power with increasing sample size is shown in Fig. 2(b). Similar as the previous case, the bd-KSD of $g^{(p)}$ with larger parameter value p achieves better test power. MMD-based tests perform slightly better that bd-KSD with $g^{(1)}$. However, MMD suffers from slow computational time due to the sampling procedure in a bounded domain, as shown in Fig. 2(c).

3. Dirichlet distributions in S^{d-1}

$$\tilde{q}_\nu(x) \propto (x^1)^\nu \cdot \Pi_{i=1}^d (x^i)^{(-\frac{1}{2})}, \forall x \in \mathrm{S}^{d-1}. \tag{34}$$

We test the null model $\tilde{q}_0$ against the alternative $\tilde{q}_{-\frac{1}{3}}$ with perturbation of the concentration parameter in the first dimension (x^1). The test power of bd-KSD, with the choice of geometric mean function $g^{(1)}$ and the minimum distance-to-boundary function $g^{(2)}$, and the MMD-based test are shown in Fig. 3. From the result, we see that the bd-KSD tests have higher test power compared to MMD-based test power. The test power increases as sample size increases (Fig. 3 left) and decreases as the dimension of the problem increases (Fig. 3 right), which is what would we expect. We also see that geometric mean function $g^{(1)}$ induces a better Stein operator for bd-KSD testing on compositional data compared to the minimum distance-to-boundary function $g^{(2)}$.

D Additional KSD Details

D.1 Kernel Discrete Stein Discrepancy

In this section, we briefly review the kernel discrete Stein discrepancy (KDSD) introduced in [Yang et al., 2018]. First we need some definitions.

Definition 1. *[Definition 1 [Yang et al., 2018]](Cyclic permutation). For a set X of finite cardinality, a cyclic permutation $\neg : X \to X$ is a bijective function such that for some ordering $x^{[1]}, x^{[2]}, \ldots, x^{[|X|]}$ of the elements in X , $\neg x^{[i]} = x^{[(i+1) mod |X|]}, \forall i = 1, 2, \ldots, |X|$.*

Definition 2. *[Definition 2 [Yang et al., 2018]] Given a cyclic permutation $\neg$ on X , for any d-dimensional vector $x = (x_1, ..., x_d)^\top \in X^d$, write $\neg_i x := (x_1, \ldots, x_{i-1}, \neg x^i, x_{i+1}, \ldots, x_d)^\top$. For any function $f : X^d \to \mathbb{R}$, denote the (partial) difference operator as*

$$\Delta_{x^i} f(x) := f(x) - f(\neg_i x), \qquad i = 1, \ldots, d$$

and introduce the difference operator:

$$\Delta_\neg f(x) := (\Delta_{x_1} f(x), \ldots, \Delta_{x_d} f(x))^\top.$$

Here we use the notation $\Delta_\neg$ to distinguish it from the notation in the main text, where we used $\Delta_s h(x) = h(x^{(s,1)}) - h(x^{(s,0)})$ and $||\Delta h|| = \sup_{s\in[N]} |\Delta_s h(x)|$.

For discrete distributions q, [Yang et al., 2018] propose the following discrete Stein operator, which is based on the difference operator $\Delta_\neg$ constructed from a cyclic permutation:

$$\mathcal{A}_q^D f(x) = f(x)\frac{\Delta_\neg q(x)}{q(x)} - \Delta_\neg^* f(x), \tag{35}$$

where $\Delta_\neg^*$ denotes the adjoint operator of $\Delta_\neg$.

In Yang et al. [2018], the generalisation that better characterises the density q is stated in the following form,

$$\mathcal{A}_{q,\mathcal{L}}^D f(x) = f(x)\frac{\mathcal{L} q(x)}{q(x)} - \mathcal{L}^* f(x), \tag{36}$$

where $\mathcal{L}^*$ is the adjoint operator of $\mathcal{L}$.

D.2 Stochastic Stein Discrepancy

Gorham et al. [2020] proposed stochastic Stein discrepancy (SSD) via the following subset operators.

Given prior π_0, likelihood $\pi(\cdot|x)$ and samples $y_1, \ldots, y_L$, the posterior density $q(x) \propto \pi_0(x)\Pi_{l=1}^L \pi(y_l|x)$. With uniformly sampled index set $\sigma \subset [L]$ with $|\sigma| = m$, the stochastic Stein operator is defined as

$$\mathcal{T}_\sigma \mathbf{f}(x) = \frac{L}{m}\mathbf{f}(x)^\top \nabla \log q_\sigma(x) + \nabla \cdot \mathbf{f}(x) \tag{37}$$

for test function $\mathbf{f}$ and $q_\sigma(x) := \pi_0(x)^{m/L}\Pi_{l\in\sigma}\pi(y_l|x)$. The stochastic Stein variational gradient descent (SSVGD) is then developed for sampling procedures and nice convergence properties has been shown in Gorham et al. [2020].

where the latent samples z_j are functionally analogous to the observations y_l in SSD. The key difference is that for every single x in SSD, multiple y_l acts on it; while for latent-variable Stein operator, only one z_j acts on it at a time.

E Additional Generalisation via Standardisation-functions

The choice of the Langevin-diffusion type of Stein operator in Section 2.1 is not unique and many other Stein operators can characterise the same distribution. A particular method to generalise the Stein operator in Eq. (1), is via an appropriate linear operator $\mathcal{L}$ acting on the test function f, i.e.

$$\mathcal{T}_{q,\mathcal{L}} f = \mathcal{T}_q(\mathcal{L} f) = \mathcal{L} f^\top \nabla \log q + \nabla \cdot \mathcal{L} f. \tag{38}$$

Some related ideas involving $\mathcal{L}$ to generalise learning objectives for unnormalised model have been discussed in the context of score matching [Lyu, 2009][13]. Yang et al. [2018][14] suggests similar formulation for characterising discrete KSD, while not yet investigated.

In the presented work, we focus on a class of new Stein operators derived from Eq. (38) where the linear operator $\mathcal{L}$ is chosen as the element-wise product using a vector-valued function $\mathbf{g}$ that we call the *auxiliary function*. Even though this is a subclass of Stein operators in Eq. (38), we show that, with specific choice of $\mathbf{g}$, the class of Stein operators in this particular form is just enough to generalise the set of Stein operators introduced in Section 2.1.

Such formulation can be more general than the element-wise product cases developed in the main text. However, the elementwise product formulation is the simplest case to generalise existing Stein operators for goodness-of-fit test.

To see the interplay between Eq. (38) and Eq. (16). We use the generalisation of second order operator in Theorem 4 as an example. Multivariate notion utilises the ∇ notation as defined in the main text. In the $\mathbb{R}^d$ case, the metric tensor terms $[G^{-1}]_{ij} = \delta_{i=j}$ so that the We show here the more interesting scenario for the Riemannian manifold case, where the choice of generalised Stein operator corresponds to the second-order operator incorporating the Riemannian metric.

As $[G^{-1}]_{ij}$ may not vanish when $i \neq j$ in the Riemannian manifold scenario, it is not possible to generalise the second-order Stein operator for Riemannian manifold with the form of Eq. (16) using elementwise product between vector-valued functions; however, with the more general formulation in Eq. (38), we are able to show this.

Corollary 2. *For scalar test function $\tilde{f} : \mathcal{M} \to \mathbb{R}$, choosing $\mathcal{L}\tilde{f}(x) = \sum_{i,j}[G^{-1}]_{i,j}\frac{\partial}{\partial x^j}\tilde{f}(x)\partial x^i$, the Stein operator in the form of Eq. (38) recovers the second-order differential operator defined in Eq. (6) for Riemannian manifold, $\mathcal{T}_{q,\mathcal{L}}\tilde{f} = \mathcal{T}_q^{(2)}\tilde{f}$.*

Proof. By definition, we know that $\mathcal{L}\tilde{f} = \nabla\tilde{f}$ incorporating the Riemannian metric G. Thus,

$$\mathcal{T}_{q,\mathcal{L}}\tilde{f} = \mathcal{T}_q\nabla\tilde{f} = \nabla\tilde{f}^\top\nabla\log q(x) + \nabla\cdot\nabla\tilde{f} = \mathcal{T}_q^{(2)}\tilde{f}$$

by construction. And the ∇ and Δ notation is also w.r.t. the Riemannian manifold $\mathcal{M}$ given metric tensor G. □

We note from the Corollary 2 that, for each vector direction ∂x^i, it is summed over all possible differential operator, $\frac{\partial}{\partial x^j}$, acting on *scalar* test function $\tilde{f}$, instead of just using $\frac{\partial}{\partial x^i}$. Hence, the element-wise operation over vector-valued test function $\mathbf{f}$ in Eq. (16) does not generalise this form. An additional note to combine the element-wise product form of Eq. (16) and the second-order Stein operator on Riemmanian manifold can be the following.

F Additional Interpretations and Comparisons

F.1 Calculus of Variation for Optimality Conditions

We review the calculus of variation and apply techniques using Euler-Lagrangian equation that we used for optimality condition in Eq. (21). Consider some loss functional in the integral form

$$J[y] = \int_{\mathcal{X}} L(y(x), \dot{y}(x))dx.$$

In one dimension, $\mathcal{X}$ can be finite interval as well. The idea for calculus of variation is to perturb the test function y for a small amount and find the stationary point for the loss functional. Let $\eta(x)$ be any function vanishing at the boundary. We consider, $y + \varepsilon\eta$, and letting $\varepsilon \to 0$. Taking derivative w.r.t. ε we get

$$\frac{dL}{d\varepsilon} = \frac{\partial L}{\partial y}\frac{dy}{d\varepsilon} + \frac{\partial L}{\partial y'}\frac{dy'}{d\varepsilon} = \frac{\partial L}{\partial y}\eta + \frac{\partial L}{\partial y'}\eta'.$$

[13] In Lyu [2009], $\mathcal{L}$ acts on density q instead of the test function f here, where a common choice of $\mathcal{L}$ is *marginalization* operator.

[14] In Yang et al. [2018], $\mathcal{L}f(x) = \sum_{x'} l(x, x')f(x')$, for discrete variable x and bivariate function l. Different from Lyu [2009], $\mathcal{L}$ may act on both the probability mass function or the test function. For the particular form of discrete KSD studied in Yang et al. [2018] $\mathcal{L}$ is chosen as the partial difference operator, which is essentially different from the diffusion based generalisation in Eq. (38) for continuous variables. More details are included in Appendix D.1.

Then we consider $\int_{\mathcal{X}} \frac{dL}{d\varepsilon}|_{\varepsilon=0}$ and using integration by parts, we have

$$\frac{d}{d\varepsilon}|_{\varepsilon=0} \int_{\mathcal{X}} L dx = \int_{\mathcal{X}} \left(\frac{\partial L}{\partial y}\eta + \frac{\partial L}{\partial y'}\eta' \right) dx = \int_{\mathcal{X}} \frac{\partial L}{\partial y}\eta + \frac{\partial L}{\partial y'}\eta'|_{\partial X} - \int_{X} \eta \frac{d}{dx}\frac{\partial L}{\partial y'} dx = \int_{\mathcal{X}} \eta \left(\frac{\partial L}{\partial y} - \frac{d}{dx}\frac{\partial L}{\partial y'} \right) dx. \tag{39}$$

Setting the Eq. (39) above to 0 we get $\int_{\mathcal{X}} \eta \left(\frac{\partial L}{\partial y} - \frac{d}{dx}\frac{\partial L}{\partial y'} \right) dx = 0$. As η is chosen to be any perturbation function, we then have

$$\frac{\partial}{\partial y} L(y(x), \dot{y}(x)) - \frac{d}{dx}\frac{\partial}{\partial \dot{y}} L(y(x), \dot{y}(x)) = 0,$$

which is referred to as the Euler-Lagrangian equation. And we will see that such techniques apply to our multivariate case when we perturb our test function in each direction.

Proof of Theorem 6

Proof. Let $L = \mathbb{E}_p[\mathcal{T}_{q,\mathbf{g}}\mathbf{f}]$ and apply the Euler-Lagrangian equation in Eq. (21) perturbing f_i w.r.t. x^i, i.e. with y in Eq. (21) as f_i, we have condition

$$\frac{\partial L}{\partial f_i} - \frac{d}{dx^i}\frac{\partial L}{\partial \dot{f}_i} = 0,$$

for all $i \in [d]$. Referring back to the form of Eq. (16),

$$(\mathcal{T}_{q,\mathbf{g}}\mathbf{f}) = \sum_{i=1}^{d} g_i (f_i \frac{\partial}{\partial x^i} \log q + \frac{\partial}{\partial x^i} f_i) + f_i \frac{\partial}{\partial x^i} g_i,$$

we get $\frac{\partial L}{\partial f_i} = \mathbb{E}_p[g_i \frac{\partial}{\partial x^i} \log q + \frac{\partial g_i}{\partial x^i}]$ and $\frac{\partial L}{\partial \dot{f}_i} = \mathbb{E}_p[g_i]$. So the Euler-Lagrangian equation yields the optimality condition, where for all i,

$$\mathbb{E}_p[g_i \frac{\partial}{\partial x^i} \log q + \frac{\partial g_i}{\partial x^i} - \frac{\partial g_i}{\partial x^i}] = \mathbb{E}_p[g_i \frac{\partial}{\partial x^i} \log q] = 0. \tag{40}$$

□

Remarks It is interesting to know that for Sf-KSD case, the calculus of variation for f does not itself depends on f. That is also why we call it optimality condition rather than solving an optimisation problem. Moreover, as we can see from Eq. (16), the role of $\mathbf{f}$ and $\mathbf{g}$ are exchangeable, meaning that if we perturb g_i instead, we will have f_i to satisfy condition $\mathbb{E}_p[f_i \frac{\partial}{\partial x^i} \log q] = 0$ for g_i to be a stationary point. Hence, the role of $\mathbf{f}$ and $\mathbf{g}$ need to be pre-specified, where we use $\mathbf{f} \in \mathcal{H}^d$ to be our RKHS function and $\mathbf{g}$ to be the auxiliary function.

We also note that the solution to Eq.(40) is not unique. As such, the optimality condition is only sufficient condition for stationary point instead of seeking for a global optimal functional.

To further characterise $\mathbf{g}$ from optimality condition $\mathbb{E}_p[g_i \frac{\partial}{\partial x^i} \log q] = 0$, we may impose additional constraint for function $\mathbf{g}$. Connecting back to the infinitesimal generator, g can be viewed as the covariance of diffusion function to be constrained for unit norm diffusion regularisation $\mathbb{E}_p[g_i(x)] = 1$ and satisfy the optimality condition in Eq. (22).

F.2 Additional Interpretation via Density Ratio as Auxiliary Function

Density ratio function $g = \frac{q}{p}$ naturally satisfy $\mathbb{E}_p[g(x)] = 1$ and $\mathbb{E}_p[g(\log q)'] = 0$. However, such "optimal" function depends on the alternative distribution p which is unknown in closed form. Although not unique, it can act as a good guideline for choosing useful g function in practice.

Truncated Gaussian In the Gaussian distribution in truncated sphere in Eq. (33) (results shown in Fig.1), the density ratio between the null and the alternative $\frac{q(x)}{p(x)} = \exp\{\nu x^1 x^2\}$. Projecting to each dimension with $\exp\{\nu x^i\}, i = 1, 2$, the L_2 distance with $g_i^{(p)}(x) = 1 - \|x\|^p$ decreases monotonically when p increases for the cases presented, i.e. polynomial increase is slower than exponential increase. Hence, $g^{(p)}$ with larger p better distinguishes the alternative from the null, yielding higher test power as shown.

Dirichlet Distribution For Dirichlet distribution example in Eq. (34) (results shown in Fig.1), the density ration $\frac{q(x)}{p(x)} = (x^{(1)})^{\frac{1}{3}}$. The auxiliary functions used are the geometric mean on simplex $g^{(1)}(x) = (x^{(1)}x^{(2)}(1 - x^{(1)} - x^{(2)}))^{\frac{1}{3}}$ where the first coordinate exactly matches the density ratio, which is closer compared to the Euclidean distance to the nearest boundary which decays quadraticall from the center of the simplex.

F.3 Connection to Unnormalised Model Learning Objectives via Density Ratio

Beyond providing a unifying view for KSDs in various testing scenarios, Sf-KSD can also be related to learning objectives for unnormalised models such as score matching [Hyvärinen, 2005], i.e. score matching can be related to Sf-KSD form with specific choice of kernels and auxiliary functions. Recall the score matching objective [Hyvärinen, 2005],

$$J(p\|q) = \mathbb{E}_p \left[(\log p(x)' - \log q(x)')^2 \right]. \tag{41}$$

$J(p\|q) \geq 0$ and the equality holds if and only if $p = q$ under mild regularity conditions [Hyvärinen, 2005].

The discrepancy measure in score matching objective [Hyvärinen, 2005] is constructed via the squared difference between derivative of log densities, without the presence of test function, opposing to the KSD-type of discrepancy measure. Hence, instead of just specifying the auxiliary function g to recover various KSDs in the previous results, we also need specific kernel function here to make connections the score matching objective.

Sf-KSD Asymptotically Connects to Score Matching

It was shown that score matching objective cna be viewed as the limit of the diffusion based KSD [Barp et al., 2019, Theorem 10]. Here, we derive the result, making explicit connections to density ratio form in Section 4.

Theorem 9. *Let f, g be scalar functions. Choosing $g(x) = 1/\sqrt{p(x)}$, and exponential quadratic kernel with bandwidth σ $k_\sigma(x, x') = \exp\{-\frac{1}{\sigma}\|x - \tilde{x}\|^2\}$, Sf-KSD in the form of Eq. (16) converges the score matching objective as $\sigma \to 0$.*

Proof. We rewrite g in the form involving density ratio between q and p, such that

$$g = \frac{q}{p} \cdot \underbrace{\frac{\sqrt{p}}{q}}_{\xi}.$$

Then we can write the Sf-KSD of the following form,

$$\begin{aligned}
\text{Sf-KSD}_g(q\|p) &= \mathbb{E}_p[((\log q(x))'(f\xi)(x) + (f\xi)'(x))\frac{q}{p}(x)] + \mathbb{E}_p[(f\xi)(x)(\frac{q}{p})'(x)] &(42)\\
&= \mathbb{E}_q[((\log q(x))'(f\xi)(x) + (f\xi)'(x))\frac{q}{p}(x)] + \mathbb{E}_p[(f\xi)(x)(\frac{q}{p})'(x)] &(43)\\
&= \mathbb{E}_p[\frac{f(x)}{\sqrt{p(x)}}((\log q(x))' - (\log p(x))')]. &(44)
\end{aligned}$$

The first expectation is 0 under q from Stein's identity of operator in Eq. (16) and the second expectation follows from

$$\left(\frac{q(x)}{p(x)}\right)' = \frac{q'}{p}(x) - \frac{qp'}{p^2}(x) = \frac{q}{p}((\log q(x))' - (\log p(x))')$$

For derivation, we first consider the δ-type function, that we call $k(x, \tilde{x}) = \delta_{x=\tilde{x}}$, we recover the original

score-matching objective in Hyvärinen [2005] since

$$\sup_{f\in\mathcal{H}} \mathbb{E}_p[\frac{f(x)}{\sqrt{p(x)}}((\log p(x))' - (\log q(x))')] = \left\| \mathbb{E}_p \left[k(x,\cdot) \frac{(\log p(x))' - (\log q(x))'}{\sqrt{p(x)}} \right] \right\|^2$$
$$= \mathbb{E}_{x,\tilde{x}\sim p} \left[\frac{k(x,\tilde{x})}{\sqrt{p(x)p(\tilde{x})}} ((\log p(x))' - (\log q(x))')((\log p(\tilde{x}))' - (\log q(\tilde{x}))') \right]$$
$$= \int\int \frac{\delta_{x=\tilde{x}}}{\sqrt{p(x)p(\tilde{x})}} ((\log p(x))' - (\log q(x))')((\log p(\tilde{x}))' - (\log q(\tilde{x}))')p(x)p(\tilde{x})dxd\tilde{x}$$
$$= \int \frac{1}{p(x)} ((\log p(x))' - (\log q(x))')^2 p(x)^2 dx = \mathbb{E}_p[((\log p(x))' - (\log q(x))')^2].$$

Using $k_\sigma(x,\tilde{x}) \to \delta_{x=\tilde{x}}$, as $\sigma \to 0$, the result follows. □

We note that the delta function is not bounded even tough integrally bounded by probability density. As such, the result appears in the form of limit with vanishing bandwidth. As we mentioned in the density ratio argument for optimality conditions, the choice of auxiliary function g here depends on the data density p, which is unknown in practice. Score matching relied on the following result for estimating the empirical version of the objective [Hyvärinen, 2005],

$$J(p\|q) = \mathbb{E}_p \left[(\log p(x)' - \log q(x)')^2 \right] = \mathbb{E}_p \left[\log q(x)' + \frac{1}{2} \log q(x)'' \right]. \tag{45}$$