Goodness-of-Fit Testing for Discrete Distributions via Stein Discrepancy

Jiasen Yang [1] **Qiang Liu** [*2] **Vinayak Rao** [*1] **Jennifer Neville** [1 3]

Abstract

Recent work has combined Stein's method with reproducing kernel Hilbert space theory to develop nonparametric goodness-of-fit tests for unnormalized probability distributions. However, the currently available tests apply exclusively to distributions with smooth density functions. In this work, we introduce a kernelized Stein discrepancy measure for discrete spaces, and develop a nonparametric goodness-of-fit test for discrete distributions with intractable normalization constants. Furthermore, we propose a general characterization of Stein operators that encompasses both discrete and continuous distributions, providing a recipe for constructing new Stein operators. We apply the proposed goodness-of-fit test to three statistical models involving discrete distributions, and our experiments show that the proposed test typically outperforms a two-sample test based on the maximum mean discrepancy.

1. Introduction

Goodness-of-fit testing is a central problem in statistics, measuring how well a model distribution $p(\mathbf{x})$ fits observed data $\{\mathbf{x}_i\}_{i=1}^n \subseteq \mathcal{X}^d$, for some domain $\mathcal{X}$ (*e.g.*, $\mathcal{X} \subseteq \mathbb{R}$ for continuous data or $\mathcal{X} \subseteq \mathbb{N}$ for discrete data). Examples of classical goodness-of-fit tests include the χ^2 test (Pearson, 1900), the Kolmogorov-Smirnov test (Kolmogorov, 1933; Smirnov, 1948), and the Anderson-Darling test (Anderson & Darling, 1954). These tests typically assume that the model distribution $p(\mathbf{x})$ is fully specified and is easy to evaluate. In modern statistical and machine learning applications, however, $p(\mathbf{x})$ is often specified only up to an intractable normalization constant; examples include large-scale graphical models, latent variable models, and statistical models for network data. While a variety of approximate inference techniques such as pseudo-likelihood estimation, Markov chain Monte Carlo (MCMC), and variational methods have been studied to allow learning and inference in these models, it is usually hard to quantify the approximation errors involved, making it difficult to establish statistical tests with calibrated uncertainty estimates.

Recently, a new line of research (Gorham & Mackey, 2015; Oates et al., 2017; Chwialkowski et al., 2016; Liu et al., 2016; Gorham & Mackey, 2017; Jitkrittum et al., 2017) has developed goodness-of-fit tests which work directly with un-normalized model distributions. Central to these tests is the notion of a *Stein operator*, originating from Stein's method (Stein, 1986) for characterizing convergence in distribution. Given a distribution $p(\mathbf{x})$ on $\mathcal{X}^d$ and a class of test functions $f \in \mathcal{F}$ on $\mathcal{X}^d$, a Stein operator $\mathcal{A}_p$ satisfies $\mathbb{E}_{\mathbf{x}\sim p}\left[\mathcal{A}_p f(\mathbf{x})\right] = 0$, so that when $\mathcal{A}_p$ is applied to any test function f, the resulting function $\mathcal{A}_p f$ has zero-expectation under p. Additionally, the expectation under any other distribution $q \neq p$ should be non-zero for at least some function f in $\mathcal{F}$. When $\mathcal{F}$ is sufficiently rich, the maximum value $\sup_{f\in\mathcal{F}} \mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_p f(\mathbf{x})\right]$ serves as a discrepancy measure, called *Stein discrepancy*, between distributions p and q.

The properties of the Stein discrepancy measure depends on two objects: the Stein operator $\mathcal{A}_p$, and the set $\mathcal{F}$. Different authors have studied different choices of $\mathcal{F}$: Gorham & Mackey (2015) considered test functions in the $\mathcal{W}^{2,\infty}$ Sobolev space, and the resulting test statistic requires solving a linear program under certain smoothness constraints. On the other hand, Oates et al. (2017); Chwialkowski et al. (2016); Liu et al. (2016) proposed taking $\mathcal{F}$ to be the unit ball of a reproducing kernel Hilbert space (RKHS), which leads to test statistics that can be computed in closed form and with time quadratic in n, the number of samples. Jitkrittum et al. (2017) further proposed a linear-time adaptive test that constructs test features by optimizing test power.

Regarding the choice of the Stein operator $\mathcal{A}_p$, all the aforementioned works consider the case when $\mathcal{X} \subseteq \mathbb{R}$ is a continuous domain, $p(\mathbf{x})$ is a smooth density on $\mathcal{X}^d$, and the Stein operator is defined in terms of the *score function* of p, $\mathbf{s}_p(\mathbf{x}) = \nabla \log p(\mathbf{x}) = {}^{\nabla p(\mathbf{x})}/_{p(\mathbf{x})}$, where ∇ is the gradient operator. Observe that any normalization constant in p cancels out in the score function, so that if the Stein operator $\mathcal{A}_p$

[*]Equal contribution [1]Department of Statistics, Purdue University, West Lafayette, IN [2]Department of Computer Science, The University of Texas at Austin, Austin, TX [3]Department of Computer Science, Purdue University, West Lafayette, IN. Correspondence to: Jiasen Yang <jiaseny@purdue.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

depends on p only through $\mathbf{s}_p$, then the discrepancy measure $\sup_{f \in \mathcal{F}} \mathbb{E}_{\mathbf{x} \sim q} [\mathcal{A}_p f(\mathbf{x})]$ can still be computed when p is unnormalized. However, constructing the Stein operator using the gradient becomes restrictive when one moves beyond distributions with smooth densities. For discrete distributions, even in the simple case of Bernoulli random variables, none of the aforementioned tests apply, since the probability mass function is no longer differentiable. This motivates more general constructions of tests based on Stein's method that would also be applicable to discrete domains.

In this work, we focus on the case where $\mathcal{X}$ is a finite set. The model distribution $p(\mathbf{x})$ is a probability mass function (pmf), whose normalization constant is computationally intractable. We note that examples of such intractable discrete distributions abound in statistics and machine learning, including the *Ising model* (Ising, 1924) in physics, the (Bernoulli) *restricted Boltzmann machine* (RBM) (Hinton & Salakhutdinov, 2006) for dimensionality reduction, and the *exponential random graph model* (ERGM) (Holland & Leinhardt, 1981) in statistical network analysis.

Our primary contribution is in establishing a kernelized Stein discrepancy measure between discrete distributions, using an appropriate choice of Stein operators for discrete spaces. Then, adopting a similar strategy as Chwialkowski et al. (2016); Liu et al. (2016), we develop a nonparametric goodness-of-fit test for discrete distributions. Notably, the proposed test also applies to discrete distributions that were previously not amenable to classical tests due to the presence of intractable normalization constants. Furthermore, we propose a general characterization of Stein operators that encompasses both discrete and continuous distributions, providing a recipe for constructing new Stein operators. For any Stein operator constructed as such, we could then define a kernelized Stein discrepancy measure to establish a valid goodness-of-fit test. Finally, we apply our proposed goodness-of-fit test to the Ising model, the Bernoulli RBM, and the ERGM, and our experiments show that the proposed test typically outperforms a two-sample test based on the maximum mean discrepancy (Gretton et al., 2012) in terms of power while maintaining control on false-positive rate.

Outline. Section 2 introduces notation and preliminaries. We construct and characterize discrete Stein operators in Section 3, establish the kernelized discrete Stein discrepancy measure in Section 4, and describe the goodness-of-fit testing procedure in Section 5. We apply the proposed test in experiments on several statistical models in Section 7, discuss related work in Section 6, and conclude in Section 8. All omitted results and proofs can be found in the Appendix.

2. Notation and Preliminaries

We primarily focus on domains $\mathcal{X}$ of finite cardinality $|\mathcal{X}|$. A probability mass function (pmf) p supported on $\mathcal{X}^d$ is said to be *positive* if $p(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^d$. A symmetric function $k(\cdot, \cdot)$ is a positive definite kernel on $\mathcal{X}^d$ if the Gram matrix $\mathbf{K} = [k(\mathbf{x}_i, \mathbf{x}_j)]_{i,j=1}^n$ is positive semidefinite for any $n \in \mathbb{N}$ and $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subseteq \mathcal{X}^d$. The kernel is *strictly* positive definite if $\mathbf{K}$ is positive definite. By the Moore-Aronszajin theorem, every such kernel k has a unique reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ of functions $f : \mathcal{X}^d \to \mathbb{R}$ satisfying the *reproducing property*: for any $f \in \mathcal{H}$, $f(\mathbf{x}) = \langle f(\cdot), k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}}$ (and in particular, $k(\mathbf{x}, \mathbf{x}') = \langle k(\cdot, \mathbf{x}), k(\cdot, \mathbf{x}') \rangle_{\mathcal{H}}$). More generally, let $\mathcal{H}^m = \mathcal{H} \times \mathcal{H} \cdots \times \mathcal{H}$ denote the Hilbert space of vector-valued functions $\mathbf{f} = \{f_\ell : f_\ell \in \mathcal{H}\}_{\ell=1}^m$, endowed with the inner-product $\langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{H}^m} = \sum_{\ell=1}^m \langle f_\ell, g_\ell \rangle_{\mathcal{H}}$ for $\mathbf{f} = \{f_\ell\}_{\ell=1}^m$ and $\mathbf{g} = \{g_\ell\}_{\ell=1}^m$, and norm $\|\mathbf{f}\|_{\mathcal{H}^m} = \sqrt{\sum_{\ell=1}^m \|f_\ell\|_{\mathcal{H}}^2}$.

3. Discrete Stein Operators

We first propose a simple Stein operator for discrete distributions, and then provide a general characterization of Stein operators for both the discrete and continuous cases. In particular, we draw upon ideas in the literature on *score-matching* methods (Hyvärinen, 2005; 2007; Lyu, 2009; Amari, 2016), which we elaborate on further in Section 6.

3.1. Difference Stein Operator

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

Thus, starting with any element of x, repeated application of the $\neg$ operator generates the set $\mathcal{X}$: $\mathcal{X} = \{x, \neg x, \dots, \neg^{(|\mathcal{X}|-1)} x\}$. In the simplest case, when $\mathcal{X}$ is a binary set, one can take $\mathcal{X} = \{\pm 1\}$ and define $\neg x = -x$.

The *inverse permutation* of $\neg$ is an operator $\neg : \mathcal{X} \to \mathcal{X}$ that satisfies $\neg(\neg x) = \neg(\neg x) = x$ for any $x \in \mathcal{X}$. Under the ordering of Definition 1, we have $\neg x^{[i]} = x^{[(i-1) \bmod |\mathcal{X}|]}$. It is easy to verify that $\neg$ is also a cyclic permutation on $\mathcal{X}$. When $\mathcal{X}$ is a binary set, the inverse of $\neg$ is itself: $\neg = \neg$.

Definition 2 (Partial difference operator and difference score function)**.** *Given a cyclic permutation $\neg$ on $\mathcal{X}$, for any vector $\mathbf{x} = (x_1, \dots, x_d)^\mathsf{T} \in \mathcal{X}^d$, write $\neg_i \mathbf{x} := (x_1, \dots, x_{i-1}, \neg x_i, x_{i+1}, \dots, x_d)^\mathsf{T}$. For any function $f : \mathcal{X}^d \to \mathbb{R}$, denote the* (partial) difference operator *as*

$$\Delta_{x_i} f(\mathbf{x}) := f(\mathbf{x}) - f(\neg_i \mathbf{x}), \quad i = 1, \dots, d,$$

and write $\Delta f(\mathbf{x}) = (\Delta_{x_1} f(\mathbf{x}), \dots, \Delta_{x_d} f(\mathbf{x}))^\mathsf{T}$. Define the (difference) score function *as $\mathbf{s}_p(\mathbf{x}) := {}^{\Delta p(\mathbf{x})}/_{p(\mathbf{x})}$, with*

$$(\mathbf{s}_p(\mathbf{x}))_i = \frac{\Delta_{x_i} p(\mathbf{x})}{p(\mathbf{x})} = 1 - \frac{p(\neg_i \mathbf{x})}{p(\mathbf{x})}, \quad i = 1, \dots, d. \quad (1)$$

We will also be interested in the difference operator defined

with respect to the inverse permutation $\multimap$. To avoid cluttering notation, we shall use Δ and $\mathbf{s}_p$ to denote the difference operator and score function defined with respect to $\neg$, and use Δ^* to denote the difference operator with respect to $\multimap$:

$$\Delta^*_{x_i} f(\mathbf{x}) := f(\mathbf{x}) - f(\multimap_i \mathbf{x}), \quad i = 1, \dots, d.$$

As in the continuous case, the score function $\mathbf{s}_p(\mathbf{x})$ can be easily computed even if p is only known up to a normalization constant: if $p(\mathbf{x}) = \tilde{p}(x)/Z$, then $\mathbf{s}_p(\mathbf{x}) = \Delta\tilde{p}(x)/\tilde{p}(x)$ does not depend on Z. For an exponential family distribution p with base measure $h(\mathbf{x})$, sufficient statistics $\boldsymbol{\phi}(\mathbf{x})$, and natural parameters $\boldsymbol{\theta}$: $p(\mathbf{x}) = \frac{1}{Z(\boldsymbol{\theta})} h(\mathbf{x}) \exp\{\boldsymbol{\theta}^\mathsf{T} \boldsymbol{\phi}(\mathbf{x})\}$, the (difference) score function is given by

$$(\mathbf{s}_p(\mathbf{x}))_i = 1 - \frac{h(\neg_i \mathbf{x})}{h(\mathbf{x})} \exp\{\boldsymbol{\theta}^\mathsf{T} (\boldsymbol{\phi}(\neg_i \mathbf{x}) - \boldsymbol{\phi}(\mathbf{x}))\}. \quad (2)$$

In the continuous case, it was obvious that two densities p and q are equal almost everywhere if and only if their score functions are equal almost everywhere. This still holds for the difference score function, but its proof is less trivial.

Theorem 1. *For any positive pmfs p and q on $\mathcal{X}^d$, we have that $\mathbf{s}_p(\mathbf{x}) = \mathbf{s}_q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^d$ if and only if $p = q$.*

Proof sketch. Clearly, $p = q$ implies that $\mathbf{s}_p(\mathbf{x}) = \mathbf{s}_q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^d$. For the converse, by Eq. (1), $\mathbf{s}_p(\mathbf{x}) = \mathbf{s}_q(\mathbf{x})$ implies that $p(\neg_i \mathbf{x})/p(\mathbf{x}) = q(\neg_i \mathbf{x})/q(\mathbf{x})$ for all $\mathbf{x}$ and i. Using the fact that $\neg$ is a cyclic permutation on $\mathcal{X}$, we can show that all the singleton conditional distributions of p and q must match, *i.e.*, $p(x_i|\mathbf{x}_{-i}) = q(x_i|\mathbf{x}_{-i})$ for all x_i and i, where $\mathbf{x}_{-i} := (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_d)$ (see the Appendix for details). By Brook's lemma (Brook, 1964; see Lemma 9 in the Appendix), the joint distribution is fully specified by the collection of singleton conditional distributions, and thus we must have $p(\mathbf{x}) = q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^d$. □

In the literature on score functions (Hyvärinen, 2007; Lyu, 2009), such results, showing that a score function $\mathbf{s}_p(\mathbf{x})$ uniquely determines a probability distribution, are called *completeness* results. For our purposes, such completeness results provide a basis for establishing statistical hypothesis tests to distinguish between two distributions. We first introduce the concept of a difference Stein operator.

Definition 3 (Difference Stein operator)**.** *Let $\neg$ be a cyclic permutation on $\mathcal{X}$ and let $\multimap$ be its inverse permutation. For any function $f : \mathcal{X}^d \to \mathbb{R}$ and pmf p on $\mathcal{X}^d$, define the* difference Stein operator *of p as*

$$\mathcal{A}_p f(\mathbf{x}) := \mathbf{s}_p(\mathbf{x}) f(\mathbf{x}) - \Delta^* f(\mathbf{x}), \quad (3)$$

where $\mathbf{s}_p(\mathbf{x}) = \Delta p(\mathbf{x})/p(\mathbf{x})$ is the difference score function defined w.r.t. $\neg$, and Δ^* is the difference operator w.r.t. $\multimap$.

We note that any intractable normalization constant in p cancels out in evaluating the Stein operator $\mathcal{A}_p$. The Stein operator satisfies an important identity:

Theorem 2 (Difference Stein's identity)**.** *For any function $f : \mathcal{X}^d \to \mathbb{R}$ and probability mass function p on $\mathcal{X}^d$,*

$$\mathbb{E}_{\mathbf{x}\sim p} [\mathcal{A}_p f(\mathbf{x})] = \mathbb{E}_{\mathbf{x}\sim p} [\mathbf{s}_p(\mathbf{x}) f(\mathbf{x}) - \Delta^* f(\mathbf{x})] = 0. \quad (4)$$

Proof. Notice that

$$\mathbb{E}_{\mathbf{x}\sim p} [\mathcal{A}_p f(\mathbf{x})] = \sum_{\mathbf{x}\in\mathcal{X}^d} [f(\mathbf{x}) \Delta p(\mathbf{x}) - p(\mathbf{x}) \Delta^* f(\mathbf{x})].$$

To complete the proof, simply note that for each i,

$$\sum_{\mathbf{x}\in\mathcal{X}^d} f(\mathbf{x}) \Delta_{x_i} p(\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{X}^d} f(\mathbf{x}) p(\mathbf{x}) - \sum_{\mathbf{x}\in\mathcal{X}^d} f(\mathbf{x}) p(\neg_i \mathbf{x}),$$
$$\sum_{\mathbf{x}\in\mathcal{X}^d} p(\mathbf{x}) \Delta^*_{x_i} f(\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{X}^d} p(\mathbf{x}) f(\mathbf{x}) - \sum_{\mathbf{x}\in\mathcal{X}^d} p(\mathbf{x}) f(\multimap_i \mathbf{x}).$$

The two equations are equal since $\neg$ and $\multimap$ are inverse cyclic permutations on $\mathcal{X}$, with $\neg_i(\multimap_i \mathbf{x}) = \multimap_i(\neg_i \mathbf{x}) = \mathbf{x}$. □

Finally, we can extend the definition of the difference Stein operator to vector-valued functions $\mathbf{f} : \mathcal{X}^d \to \mathbb{R}^m$. In this case, $\Delta\mathbf{f}$ is an $d \times m$ matrix with $(\Delta\mathbf{f})_{ij} = \Delta_{x_i} f_j(\mathbf{x})$, and the Stein operator takes the form

$$\mathcal{A}_p \mathbf{f}(\mathbf{x}) = \mathbf{s}_p(\mathbf{x})\, \mathbf{f}(\mathbf{x})^\mathsf{T} - \Delta^* \mathbf{f}(\mathbf{x}).$$

Similar to Theorem 2, one can show that for any function $\mathbf{f} : \mathcal{X}^d \to \mathbb{R}^m$ and positive pmf p on $\mathcal{X}^d$,

$$\mathbb{E}_{\mathbf{x}\sim p} [\mathcal{A}_p \mathbf{f}(\mathbf{x})] = \mathbb{E}_{\mathbf{x}\sim p} \left[\mathbf{s}_p(\mathbf{x})\, \mathbf{f}(\mathbf{x})^\mathsf{T} - \Delta^* \mathbf{f}(\mathbf{x})\right] = \mathbf{0}.$$

If $m = d$, taking the trace on both sides yields

$$\mathbb{E}_p [\mathrm{tr}\,(\mathcal{A}_p \mathbf{f}(\mathbf{x}))] = \mathbb{E}_p \left[\mathbf{s}_p(\mathbf{x})^\mathsf{T} \mathbf{f}(\mathbf{x}) - \mathrm{tr}\,(\Delta^* \mathbf{f}(\mathbf{x}))\right] = 0.$$

3.2. Characterization of Stein Operators

Generalizing our construction in the previous section, we can further identify a broad class of Stein operators which includes the difference Stein operator as a special case.

Let $\mathcal{L}$ be any operator defined on the space of functions $\mathcal{F} = \{f : \mathcal{X}^d \to \mathbb{R}\}$ that can be written in the form[1]

$$\mathcal{L} f(\mathbf{x}) = \sum_{\mathbf{x}'\in\mathcal{X}^d} g(\mathbf{x}, \mathbf{x}') f(\mathbf{x}'), \quad \forall f \in \mathcal{F} \quad (5)$$

for some bivariate (possibly vector-valued) function g on $\mathcal{X}^d \times \mathcal{X}^d$. Define a dual operator $\mathcal{L}^*$ via

$$\mathcal{L}^* f(\mathbf{x}) = \sum_{\mathbf{x}'\in\mathcal{X}^d} g(\mathbf{x}', \mathbf{x}) f(\mathbf{x}'), \quad \forall f \in \mathcal{F}. \quad (6)$$

[1] The notion can also be extended to vector-valued functions $\mathbf{f}$; we omit this generalization here for clarity.

In fact, when $\mathcal{X}$ is a finite set, any linear operator $\mathcal{L}$ on $\mathcal{F} = \{f : \mathcal{X}^d \to \mathbb{R}\}$ can be written in the form of Eq. (5). In this case, the operator $\mathcal{L}^*$ as defined in Eq. (6) is the *adjoint* operator of $\mathcal{L}$: $\langle \mathcal{L}f, g\rangle = \langle f, \mathcal{L}^* g\rangle$ for all $f, g \in \mathcal{F}$, where $\langle \cdot, \cdot \rangle$ is the appropriate inner-product on $\mathcal{X}^d$. If $g(\cdot, \cdot)$ is symmetric, then $\mathcal{L}$ is *self-adjoint*, *i.e.*, $\mathcal{L}^* = \mathcal{L}$.

Under these definitions, we have the following result which characterizes the Stein operators on a discrete space $\mathcal{X}^d$.

Theorem 3. *Denote $\mathcal{F} = \{f : \mathcal{X}^d \to \mathbb{R}\}$. For any positive pmf p on $\mathcal{X}^d$, a linear operator $\mathcal{T}_p$ satisfies Stein's identity*

$$\mathbb{E}_{\mathbf{x}\sim p}\left[\mathcal{T}_p f(\mathbf{x})\right] = 0 \tag{7}$$

for all functions $f \in \mathcal{F}$ if and only if there exist linear operators $\mathcal{L}$ and $\mathcal{L}^*$ of the forms (5) and (6), such that

$$\mathcal{T}_p f(\mathbf{x}) = \frac{\mathcal{L}p(\mathbf{x})}{p(\mathbf{x})} f(\mathbf{x}) - \mathcal{L}^* f(\mathbf{x}) \tag{8}$$

holds for all $\mathbf{x} \in \mathcal{X}^d$ and functions $f \in \mathcal{F}$.

Proof. Sufficiency: Suppose the linear operators $\mathcal{L}$ and $\mathcal{L}^*$ take the forms of Eqs. (5) and (6) for some function g, we show that the operator $\mathcal{T}_p$ defined via Eq. (8) satisfies Stein's identity of Eq. (7). We can write

$$\begin{aligned}
\mathbb{E}_p\left[\mathcal{T}_p f(\mathbf{x})\right] &= \sum_{\mathbf{x}\in\mathcal{X}^d} \left[f(\mathbf{x})\mathcal{L}p(\mathbf{x}) - p(\mathbf{x})\mathcal{L}^* f(\mathbf{x})\right] \\
&= \sum_{\mathbf{x}\in\mathcal{X}^d}\sum_{\mathbf{x}'\in\mathcal{X}^d} f(\mathbf{x})g(\mathbf{x},\mathbf{x}')p(\mathbf{x}') \\
&\quad - \sum_{\mathbf{x}\in\mathcal{X}^d}\sum_{\mathbf{x}'\in\mathcal{X}^d} p(\mathbf{x})g(\mathbf{x}',\mathbf{x})f(\mathbf{x}')\,.
\end{aligned}$$

The two terms in the last line cancel out since the double-summations are invariant under a swapping of summation indices $\mathbf{x}$ and $\mathbf{x}'$, giving $\mathbb{E}_p\left[\mathcal{T}_p f(\mathbf{x})\right] = 0$.

Necessity: See the Appendix for the remaining proof. □

We note that the sufficiency part of Theorem 3 remains valid when $\mathcal{X}$ is a continuous space, p is a density, $\mathcal{F} \subseteq \{f : \mathcal{X}^d \to \mathbb{R}\}$ is some family of functions for which $\mathcal{T}_p f$ and $\mathcal{L}f$ are well-defined, and the summations in Eqs. (5) and (6) are replaced by integrations. However, the necessity part requires further conditions on the expressiveness of $\mathcal{F}$.

Theorem 3 essentially states that (for a fixed p) given any pair of adjoint operators $\mathcal{L}$ and $\mathcal{L}^*$, one can construct a linear operator $\mathcal{T}_p$ satisfying Stein's identity; conversely, any Stein operator $\mathcal{T}_p$ can be expressed using a pair of adjoint operators $\mathcal{L}$ and $\mathcal{L}^*$. This connection between adjoint operators and Stein operators enables us to unify different forms of Stein operators for discrete and continuous distributions (see also Ley et al. (2017) for related discussions).

Remark 4 (Continuous case)**.** *For a continuous space $\mathcal{X} \subseteq \mathbb{R}$, consider a smooth density p on $\mathcal{X}^d$. Take $\mathcal{L} = \nabla$ to be the gradient operator, and let $\mathcal{F}$ consist of smooth functions $f : \mathcal{X}^d \to \mathbb{R}$ for which $f(\mathbf{x})\,p(\mathbf{x})$ vanishes at the boundary $\partial\mathcal{X}$. Using integration-by-parts, it can be shown that the adjoint operator of $\mathcal{L}$ is $\mathcal{L}^* = -\nabla$. Then, applying Eq. (8) of Theorem 3 recovers the standard continuous Stein operator*

$$\mathcal{A}_p f(\mathbf{x}) = \nabla \log p(\mathbf{x}) f(\mathbf{x}) + \nabla f(\mathbf{x}).$$

Remark 5 (Discrete case)**.** *In Eqs. (5) and (6), define the vector-valued function $\mathbf{g} : \mathcal{X}^d \times \mathcal{X}^d \to \mathbb{R}^d$ with*

$$(\mathbf{g}(\mathbf{x},\mathbf{x}'))_i = \mathbb{I}\{\mathbf{x}' = \mathbf{x}\} - \mathbb{I}\{\mathbf{x}' = \neg_i \mathbf{x}\} \tag{9}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. Then, we have

$$(\mathcal{L}f(\mathbf{x}))_i = \sum_{\mathbf{x}\in\mathcal{X}^d} (\mathbf{g}(\mathbf{x},\mathbf{x}'))_i f(\mathbf{x}) = f(\mathbf{x}) - f(\neg_i \mathbf{x}),$$

which recovers the difference operator Δ. Similarly, define $\mathbf{g}^*$ by replacing $\neg$ with its inverse permutation $\neg$ in Eq. (9). Notice that $\mathbf{g}(\mathbf{x},\mathbf{x}') = \mathbf{g}^*(\mathbf{x}',\mathbf{x})$, and thus the adjoint of $\mathcal{L}$ is given by $\mathcal{L}^* = \Delta^*$. In this case, Eq. (8) boils down to the difference Stein operator defined in Eq. (3).

Note that if $\mathcal{X}$ is binary, then $\neg = \neg$, and $\mathcal{L}$ is self-adjoint. When $\mathcal{L}$ is self-adjoint, in addition to Stein's identity, the Stein operator defined via Eq. (8) also satisfies $\mathcal{T}_p\, p(\mathbf{x}) = 0$.

Graph-based discrete Stein operators. Extending the form of Eq. (9), we can obtain a more general recipe for constructing g, which, upon applying Theorem 3, gives rise to other Stein operators on $\mathcal{X}^d$. Specifically, suppose we have identified a simple graph $\mathcal{G} = (\mathcal{X}^d, \mathcal{E})$ on $|\mathcal{X}|^d$ vertices, with each vertex corresponding to a possible configuration $\mathbf{x} \in \mathcal{X}^d$. Then, it is natural to define g such that it respects the structure of $\mathcal{G}$, in the sense that $g(\mathbf{x},\mathbf{x}') = 0$ if $\mathbf{x}' \notin \mathcal{N}_\mathbf{x} \cup \{\mathbf{x}\}$, where $\mathcal{N}_\mathbf{x} := \{\mathbf{x}' : (\mathbf{x},\mathbf{x}') \in \mathcal{E}\}$ is the set of neighbors of $\mathbf{x}$ in $\mathcal{G}$. If $\mathcal{G}$ is undirected, one would also make g symmetric, in which case $\mathcal{L} \equiv \mathcal{L}^*$ is self-adjoint.

Revisiting the difference Stein operator in this light, notice that $\neg$ defines a d-dimensional (undirected) lattice graph $\mathcal{G}$ on $\mathcal{X}^d$, in which two vertices $\mathbf{x}$ and $\mathbf{x}'$ are connected if and only if $\mathbf{x}' = \neg_i \mathbf{x}$ for some $i \in \{1,\ldots,d\}$. In this case, every vertex $\mathbf{x}$ has exactly d neighbors in $\mathcal{G}$: $\mathcal{N}_\mathbf{x} = \{\neg_1\mathbf{x},\ldots,\neg_d\mathbf{x}\}$. We then set $\mathbf{g}(\mathbf{x},\neg_i\mathbf{x}) = -\mathbf{e}_i$ for each i, $\mathbf{g}(\mathbf{x},\mathbf{x}) = \mathbf{e}$, and $\mathbf{g}(\mathbf{x},\mathbf{x}') = \mathbf{0}$ for $\mathbf{x}' \notin \mathcal{N}_\mathbf{x} \cup \{\mathbf{x}\}$, where $\mathbf{e}_i \in \mathbb{R}^d$ is the i-th standard basis vector, and $\mathbf{e} \in \mathbb{R}^d$ is the all-ones vector. This recovers the form of $\mathbf{g}$ in Eq. (9).

As another example, one could take $g(\mathbf{x},\mathbf{x}') = -|\mathcal{N}_\mathbf{x}|^{-1}$ for $\mathbf{x}' \in \mathcal{N}_\mathbf{x}$ and set $g(\mathbf{x},\mathbf{x}') = \mathbb{I}[\mathbf{x} = \mathbf{x}']$ otherwise. Then, Eq. (5) becomes $\mathcal{L}f(\mathbf{x}) = \frac{1}{|\mathcal{N}_\mathbf{x}|}\sum_{\mathbf{x}'\in\mathcal{N}_\mathbf{x}} \left(f(\mathbf{x}) - f(\mathbf{x}')\right)$, which recovers the normalized Laplacian of $\mathcal{G}$ (see also Amari, 2016). Thus, by specifying an arbitrary graph structure $\mathcal{G}$ on $\mathcal{X}^d$, one could also utilize its Laplacian $\mathcal{L}$ to define a corresponding Stein operator $\mathcal{T}$ by applying Theorem 3.

4. Kernelized Discrete Stein Discrepancy

We can now proceed similarly as in the continuous case (Liu et al., 2016; Chwialkowski et al., 2016) to define the discrete Stein discrepancy and its kernelized counterpart. While all results in this section hold for the general Stein operators discussed in Section 3.2, for clarity we state them for the difference Stein operator described in Section 3.1.

Definition 4 (Discrete Stein discrepancy)**.** *Let $\mathcal{X}$ be a finite set. For a family $\mathcal{F}$ of functions $\mathbf{f} : \mathcal{X}^d \to \mathbb{R}^d$, define the* discrete Stein discrepancy *between two positive pmfs p, q as*

$$\mathbb{D}(q \,\|\, p) := \sup_{\mathbf{f} \in \mathcal{F}} \mathbb{E}_{\mathbf{x} \sim q} \left[\operatorname{tr} \left(\mathcal{A}_p \mathbf{f}(\mathbf{x}) \right) \right],$$

where $\mathcal{A}_p \mathbf{f}(\mathbf{x}) = \mathbf{s}_p(\mathbf{x})\, \mathbf{f}(\mathbf{x})^\mathsf{T} - \Delta^* \mathbf{f}(\mathbf{x})$ is the difference Stein operator w.r.t. p. Taking $\mathcal{F}$ to be the unit ball in an RKHS $\mathcal{H}^d$ of vector-valued functions $\mathbf{f} : \mathcal{X}^d \to \mathbb{R}^d$, we obtain the kernelized discrete Stein discrepancy *(KDSD):*

$$\mathbb{D}(q \,\|\, p) = \sup_{\mathbf{f} \in \mathcal{H}^d,\, \|\mathbf{f}\|_{\mathcal{H}^d} \le 1} \mathbb{E}_{\mathbf{x} \sim q} \left[\operatorname{tr} \left(\mathcal{A}_p \mathbf{f}(\mathbf{x}) \right) \right]. \tag{10}$$

Although Eq. (10) involves solving a variational problem, the next two results show that the kernelized discrete Stein discrepancy can actually be computed in closed-form. Due to space constraints, we defer their proofs to the Appendix.

Theorem 6. *The kernelized discrete Stein discrepancy as defined in Eq. (10) admits an equivalent representation:*

$$\mathbb{D}(q \,\|\, p)^2 = \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim q} \left[\boldsymbol{\delta}_{p,q}(\mathbf{x})^\mathsf{T} k(\mathbf{x}, \mathbf{x}')\, \boldsymbol{\delta}_{p,q}(\mathbf{x}') \right], \tag{11}$$

where $\boldsymbol{\delta}_{p,q}(\mathbf{x}) := \mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})$ is the score-difference between p and q.

Theorem 7. *Define the kernel function*

$$\begin{aligned} \kappa_p(\mathbf{x}, \mathbf{x}') = {} & \mathbf{s}_p(\mathbf{x})^\mathsf{T} k(\mathbf{x}, \mathbf{x}')\, \mathbf{s}_p(\mathbf{x}') - \mathbf{s}_p(\mathbf{x})^\mathsf{T} \Delta^*_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}') \\ & - \Delta^*_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}')^\mathsf{T} \mathbf{s}_p(\mathbf{x}') + \operatorname{tr} \left(\Delta^*_{\mathbf{x}, \mathbf{x}'} k(\mathbf{x}, \mathbf{x}') \right), \end{aligned} \tag{12}$$

then

$$\mathbb{D}(q \,\|\, p)^2 = \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim q} \left[\kappa_p(\mathbf{x}, \mathbf{x}') \right]. \tag{13}$$

The next result justifies $\mathbb{D}(q \,\|\, p)$ as a divergence measure.

Lemma 8. *For a finite set $\mathcal{X}$, let p and q be positive pmfs on $\mathcal{X}^d$. Let $\mathcal{H}$ be an RKHS on $\mathcal{X}^d$ with kernel $k(\cdot, \cdot)$, and let $\mathbb{D}(q \,\|\, p)$ be defined as in Eq. (10). Assume that the Gram matrix $\mathbf{K} = [k(\mathbf{x}, \mathbf{x}')]_{\mathbf{x}, \mathbf{x}' \in \mathcal{X}^d}$ is strictly positive definite, then $\mathbb{D}(q \,\|\, p) = 0$ if and only if $p = q$.*

Proof. By Theorem 6, we have

$$\begin{aligned} \mathbb{D}(q \,\|\, p)^2 &= \mathbb{E}_{\mathbf{x}, \mathbf{x}' \sim q} \left[\boldsymbol{\delta}_{p,q}(\mathbf{x})^\mathsf{T} k(\mathbf{x}, \mathbf{x}')\, \boldsymbol{\delta}_{p,q}(\mathbf{x}') \right] \\ &= \sum_{\mathbf{x} \in \mathcal{X}^d} \sum_{\mathbf{x}' \in \mathcal{X}^d} q(\mathbf{x}) \boldsymbol{\delta}_{p,q}(\mathbf{x})^\mathsf{T} k(\mathbf{x}, \mathbf{x}')\, \boldsymbol{\delta}_{p,q}(\mathbf{x}') q(\mathbf{x}'), \end{aligned}$$

where $\boldsymbol{\delta}_{p,q}(\mathbf{x}) = \mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x}) \in \mathbb{R}^d$. Denote the ℓ-th element of $\boldsymbol{\delta}_{p,q}$ by $\delta^\ell_{p,q}$, and write $\mathbf{g}_\ell := [q(\mathbf{x}) \delta^\ell_{p,q}(\mathbf{x})]_{\mathbf{x} \in \mathcal{X}^d}$ for $\ell = 1, \ldots, d$. Then, $\mathbb{D}(q \,\|\, p)^2 = \sum_{\ell=1}^d \mathbf{g}_\ell^\mathsf{T} \mathbf{K} \mathbf{g}_\ell$. Since $\mathbf{K}$ is strictly positive-definite, $\mathbb{D}(q \,\|\, p)^2 = 0$ if and only if $\mathbf{g}_\ell = \mathbf{0}$ for all ℓ. Therefore, $\boldsymbol{\delta}_{p,q}(\mathbf{x}) = \mathbf{0}$ for all $\mathbf{x} \in \mathcal{X}^d$. By Theorem 1, this holds if and only if $p = q$. □

5. Goodness-of-Fit Testing via KDSD

Given a (possibly un-normalized) model distribution p and *i.i.d.* samples $\{\mathbf{x}_i\}_{i=1}^n$ from an unknown data distribution q on $\mathcal{X}^d$, we would like to measure the goodness-of-fit of the model distribution p to the observed data $\{\mathbf{x}_i\}_{i=1}^n$. To this end, we perform the hypothesis test $H_0 : p = q$ vs. $H_1 : p \neq q$ using the kernelized discrete Stein discrepancy (KDSD) measure. Denote $\mathbb{S}(q \,\|\, p) := \mathbb{D}(q \,\|\, p)^2$; we can estimate $\mathbb{S}(q \,\|\, p)$ via a U-statistic (Hoeffding, 1948) which provides a minimum-variance unbiased estimator:

$$\widehat{\mathbb{S}}(q \,\|\, p) = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n \kappa_p(\mathbf{x}_i, \mathbf{x}_j). \tag{14}$$

As in the continuous case (Liu et al., 2016), the U-statistic $\widehat{\mathbb{S}}(q \,\|\, p)$ is asymptotically Normal under the alternative hypothesis $H_1 : p \neq q$, $\sqrt{n} \left(\widehat{\mathbb{S}}(q \,\|\, p) - \mathbb{S}(q \,\|\, p) \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \operatorname{Var}_{\mathbf{x} \sim q}(\mathbb{E}_{\mathbf{x}' \sim q} [\kappa_p(\mathbf{x}, \mathbf{x}')]) > 0$, but becomes degenerate ($\sigma^2 = 0$) under the null hypothesis $H_0 : p = q$ (see Theorem 11 in the Appendix for a precise statement).

Since the asymptotic distribution of $\widehat{\mathbb{S}}(q \,\|\, p)$ under the null hypothesis cannot be easily calculated, we follow Liu et al. (2016) and adopt the bootstrap method for degenerate U-statistics (Arcones & Gine, 1992; Huskova & Janssen, 1993) to draw samples from the null distribution of the test statistic. Specifically, to obtain a bootstrap sample, we draw random multinomial weights $w_1, \ldots, w_n \sim \text{Mult}(n; 1/n, \ldots, 1/n)$, set $\widetilde{w}_i = (w_i - 1)/n$, and compute

$$\widehat{\mathbb{S}}^*(q \,\|\, p) = \sum_{i=1}^n \sum_{j \neq i}^n \widetilde{w}_i \widetilde{w}_j \kappa_p(\mathbf{x}_i, \mathbf{x}_j). \tag{15}$$

Upon repeating this procedure m times, we calculate the critical value of the test by taking the $(1 - \alpha)$-th quantile of the bootstrapped statistics $\{\widehat{\mathbb{S}}^*_b\}_{b=1}^m$.

The overall goodness-of-fit testing procedure is summarized in Algorithm 1. Computing the test statistic in Eq. (14) takes $\mathcal{O}(n^2)$ time, where n is the number of observations, and the bootstrapping procedure takes $\mathcal{O}(mn^2)$ time, where m is the number of bootstrap samples used.

Kernel choice. A practical question that arises when performing the KDSD test is the choice of the kernel function $k(\cdot, \cdot)$ on $\mathcal{X}^d$. For continuous spaces, the RBF kernel might be a natural choice; Gorham & Mackey (2017) also provide further recommendations. For discrete spaces, a naive choice is the δ-kernel, $k(\mathbf{x}, \mathbf{x}') = \mathbb{I}\{\mathbf{x} = \mathbf{x}'\}$, which suffers from the curse of dimensionality. A more sensible choice is

Algorithm 1 Goodness-of-fit testing via KDSD

Input: Difference score function $\mathbf{s}_p$ of p, data samples $\{\mathbf{x}_i\}_{i=1}^n \sim q$, kernel function $k(\cdot,\cdot)$, bootstrap sample size m, significance level α.
Objective: Test $H_0 : p = q$ vs. $H_1 : p \neq q$.
Compute test statistic $\widehat{\mathbb{S}}(q \,\|\, p)$ via Eq. (14).
for $b = 1, \ldots, m$ **do**
  Compute bootstrap test statistic $\widehat{\mathbb{S}}_b^*$ via Eq. (15).
end for
Compute critical value $\gamma_{1-\alpha}$ by taking the $(1-\alpha)$-th quantile of the bootstrap test statistics $\{\widehat{\mathbb{S}}_b^*\}_{b=1}^m$.
Output: Reject H_0 if test statistic $\widehat{\mathbb{S}}(q \,\|\, p) > \gamma_{1-\alpha}$, otherwise do not reject H_0.

the *exponentiated Hamming kernel*:

$$k(\mathbf{x}, \mathbf{x}') = \exp\{-H(\mathbf{x}, \mathbf{x}')\}, \tag{16}$$

where $H(\mathbf{x}, \mathbf{x}') := \frac{1}{d}\sum_{i=1}^d \mathbb{I}\{x_i \neq x_i'\}$ is the normalized Hamming distance. Lemma 12 in the Appendix shows that Eq. (16) defines a positive definite kernel.

When the inputs $\mathbf{x}$ and $\mathbf{x}'$ encode additional structure about $\mathcal{X}^d$, the Hamming distance may no longer be appropriate. For instance, when $\mathbf{x} \in \{0,1\}^{\binom{d}{2}}$ represents the (flattened) adjacency matrix of an undirected and unweighted graph on d vertices, two graphs $\mathbf{x}$ and $\mathbf{x}'$ may be isomorphic yet have non-zero Hamming distance. In this case, we can resort to the literature on graph kernels (Vishwanathan et al., 2010). Section 7 gives an example of using the *Weisfeiler-Lehman graph kernel* of Shervashidze et al. (2011) to test whether a set of graphs $\{\mathbf{x}_i\}_{i=1}^n$ comes from a specific distribution.

6. Related Work and Discussion

Stein's method. In probability theory, Stein's method has become an important tool for deriving approximations to probability distributions and characterizing convergence rates (see *e.g.*, Barbour & Chen (2014) for an overview). Related to our characterization via adjoint operators, Ley et al. (2017) also proposed the notion of a canonical Stein operator. Recently, Bresler & Nagaraj (2017); Reinert & Ross (2017) applied Stein's method to bound the distance between two stationary distributions of irreducible Markov chains in terms of their Glauber dynamics. Notably, they also make use of a difference operator for the binary case, and it is interesting to investigate whether their analysis techniques could be adopted for goodness-of-fit testing.

Goodness-of-fit tests. Closely related to our work is the kernelized Stein discrepancy test proposed independently by Chwialkowski et al. (2016); Liu et al. (2016) for smooth densities on continuous spaces. Our work further identifies and characterizes Stein operators for discrete domains, unifying them via Theorem 3 under a general framework for constructing Stein operators from adjoint operators. Under this framework, any Stein operator can be directly used to establish a KDSD test (under completeness conditions).

In addition to kernel-based tests, other forms of goodness-of-fit tests have also been examined for discrete distributions. Some recent examples include Valiant & Valiant (2016); Martín del Campo et al. (2017); Daskalakis et al. (2018). However, these tests are often model-specific, and typically assume that the normalization constant is easy to evaluate. In contrast, the KDSD test we propose is fully nonparametric, and applies to any un-normalized statistical model.

Score-matching methods. Proposed by Hyvärinen (2005), score-matching methods make use of score functions to perform parameter estimation in un-normalized models. Suppose we observe data $\{\mathbf{x}\}_{i=1}^n$ from some unknown density $q(\mathbf{x})$ which we would like to approximate using a parameterized model density $p(\mathbf{x}; \boldsymbol{\theta})$. To estimate the parameters $\boldsymbol{\theta}$, score-matching methods minimize the *Fisher divergence*:

$$J(\boldsymbol{\theta}) = \int_{\boldsymbol{\xi} \in \mathbb{R}^d} q(\boldsymbol{\xi}) \, \|\nabla_{\boldsymbol{\xi}} \log p(\boldsymbol{\xi}; \boldsymbol{\theta}) - \nabla_{\boldsymbol{\xi}} \log q(\boldsymbol{\xi})\|_2^2 \, \mathrm{d}\boldsymbol{\xi}.$$

Similar to the continuous KSD (Liu et al., 2016), if we set $k(\mathbf{x}, \mathbf{x}') = \mathbb{I}\{\mathbf{x} = \mathbf{x}'\}/\sqrt{q(\mathbf{x})\, q(\mathbf{x}')}$ and apply Theorem 6, the KDSD statistic can be written as $\mathbb{D}(q \,\|\, p)^2 = \mathbb{E}_{\mathbf{x}\sim q}\left[\|\mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})\|_2^2\right]$, which takes the same form as $J(\boldsymbol{\theta})$ with the continuous score function $\nabla \log p(\mathbf{x})$ replaced by the difference score function $\mathbf{s}_p(\mathbf{x})$.

Extensions of score-matching to discrete data have also been considered in Hyvärinen (2007); Lyu (2009); Amari (2016), and our work draws insights from these in the design of score functions for Stein operators. In particular, Lyu (2009) examined the connections between adjoint operators and Fisher divergence, and Amari (2016) discussed score functions for data from a graphical model. However, the connections to Stein operators and kernel-based hypothesis testing have not appeared in the score-matching literature.

Two-sample tests. Complementing goodness-of-fit tests (or one-sample tests) are two-sample tests, where we test if two collections of samples come from the same distribution. A well-known kernel two-sample test statistic is the *maximum mean discrepancy* (MMD) of Gretton et al. (2012). Given *i.i.d.* samples $\{\mathbf{x}_i\}_{i=1}^n \sim p$ and $\{\mathbf{y}_j\}_{j=1}^{n'} \sim q$, one could compute a U-statistic estimate of $\mathrm{MMD}(p, q)$ in $\mathcal{O}(nn')$ time. The critical value of the test is calculated by bootstrapping on the aggregated data.

Two-sample tests can also be used as goodness-of-fit tests by comparing observed data with samples from the null model. For distributions with intractable normalization constants, obtaining exact samples from p could become very difficult or expensive. Further, approximate samples may introduce bias and/or correlation among the samples, violating the test assumptions, and leading to unpredictable test errors.

7. Applications

We apply the proposed KDSD goodness-of-fit test to three statistical models involving discrete distributions. We describe the models and derive their difference score functions in Section 7.1, and present experiments in Section 7.2.

7.1. Statistical Models

Ising model. The Ising model (Ising, 1924) is a canonical example of a Markov random field (MRF). Consider an (undirected) graph $G = (V, E)$, where each vertex $i \in V$ is associated with a binary *spin*. The collection of spins form a random vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, whose components x_i and x_j ($i \neq j$) interact directly only if $(i, j) \in E$. The pmf is $p_{\boldsymbol{\Theta}}(\mathbf{x}) = \frac{1}{Z(\boldsymbol{\Theta})} \exp\{\sum_{(i,j)\in E} \theta_{ij} x_i x_j\}$, where θ_{ij} are the edge potentials and $Z(\boldsymbol{\Theta})$ is the partition function which is prohibitive to compute when d is high. Recognizing the pmf as an exponential family distribution, we can apply Eq. (2) to obtain the difference score function: $(\mathbf{s}_p(\mathbf{x}))_i = 1 - \exp\{-2x_i \sum_{j\in\mathcal{N}_i} \theta_{ij} x_j\}$, where $\mathcal{N}_i := \{j : (i, j) \in E\}$ denotes the set of vertices adjacent to node i in graph G.

Bernoulli restricted Boltzmann machine (RBM). The RBM (Hinton, 2002) is an undirected graphical model consisting of a bipartite graph between visible units $\mathbf{v}$ and hidden units $\mathbf{h}$. In a Bernoulli RBM, both $\mathbf{v}$ and $\mathbf{h}$ are Bernoulli-distributed; $\mathcal{X} = \{0, 1\}$. The joint pmf of an RBM with M visible units and K hidden units is given by $p(\mathbf{h}, \mathbf{v}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\{-E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})\}$, with energy function $E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta}) = -(\mathbf{v}^\mathsf{T}\mathbf{W}\mathbf{h} + \mathbf{v}^\mathsf{T}\mathbf{b} + \mathbf{h}^\mathsf{T}\mathbf{c})$, where $\mathbf{W} \in \mathbb{R}^{M\times K}$ are the weights, $\mathbf{b} \in \mathbb{R}^M$ and $\mathbf{c} \in \mathbb{R}^K$ are the bias terms, $\boldsymbol{\theta} := (\mathbf{W}, \mathbf{b}, \mathbf{c})$, and $Z(\boldsymbol{\theta}) = \sum_{\mathbf{v}} \sum_{\mathbf{h}} \exp\{-E(\mathbf{v}, \mathbf{h}; \boldsymbol{\theta})\}$ is the partition function.

Marginalizing out the hidden variables $\mathbf{h}$, the pmf of $\mathbf{v}$ is given by $p(\mathbf{v}|\boldsymbol{\theta}) = \frac{1}{Z'(\boldsymbol{\theta})} \exp\{-F(\mathbf{v}; \boldsymbol{\theta})\}$, with free energy $F(\mathbf{v}; \boldsymbol{\theta}) = -\mathbf{v}^\mathsf{T}\mathbf{b} - \sum_{k=1}^K \log(1 + \exp\{\mathbf{v}^\mathsf{T}\mathbf{W}_{*k} + c_k\})$. Here, $\mathbf{W}_{*k}$ denotes the k-th column of $\mathbf{W}$, and $Z'(\boldsymbol{\theta}) = \sum_{\mathbf{v}} \exp\{-F(\mathbf{v}; \boldsymbol{\theta})\}$ is another normalization constant. Thus, we can write down the (difference) score function as $(\mathbf{s}_p(\mathbf{v}; \boldsymbol{\theta}))_i = 1 - e^{\tilde{v}_i b_i} \prod_{k=1}^K \frac{1+\exp\{\mathbf{v}^\mathsf{T}\mathbf{W}_{*k} + \tilde{v}_i w_{ik} + c_k\}}{1+\exp\{\mathbf{v}^\mathsf{T}\mathbf{W}_{*k} + c_k\}}$, where $\tilde{v}_i = \neg v_i - v_i$. Note that $\mathbf{s}_p(\mathbf{v}; \boldsymbol{\theta})$ is again free of normalization constants and can be easily evaluated.

Exponential random graph model (ERGM). The ERGM is a well-studied statistical model for network data (Holland & Leinhardt, 1981). In a typical ERGM, the probability of observing an adjacency matrix $\mathbf{y} \in \{0, 1\}^{n\times n}$ is $p(\mathbf{y}) = \frac{1}{Z(\boldsymbol{\theta},\tau)} \exp\left\{\sum_{k=1}^{n-1} \theta_k S_k(\mathbf{y}) + \tau T(\mathbf{y})\right\}$. Here, $S_k(\cdot)$ counts the number of edges ($k = 1$) or k-stars ($k \geq 2$), $T(\cdot)$ counts triangles, and $Z(\boldsymbol{\theta}, \tau)$ is the normalization constant.

We consider an ERGM distribution of undirected graphs $\mathbf{y}$ with three sufficient statistics: $S_1(\mathbf{y})$, the number of edges (1-stars); $S_2(\mathbf{y})$, the number of wedges (2-stars); and $T(\mathbf{y})$, the number of triangles.[2] The parameters for these sufficient statistics are θ_1, θ_2, and τ, respectively. The score function can be written as $(\mathbf{s}_p(\mathbf{y}))_{ij} = 1 - \exp\{\theta_1\delta_1(\mathbf{y}) + \theta_2\delta_2(\mathbf{y}) + \tau\delta_3(\mathbf{y})\}$, with the *change statistics* given by $\delta_1(\mathbf{y}) := [S_1(\neg_{ij}\mathbf{y}) - S_1(\mathbf{y})] = (-1)^{y_{ij}}$, $\delta_2(\mathbf{y}) := [S_2(\neg_{ij}\mathbf{y}) - S_2(\mathbf{y})] = (-1)^{y_{ij}}(|\mathcal{N}_i^{\setminus j}| + |\mathcal{N}_j^{\setminus i}|)$, and $\delta_3(\mathbf{y}) := [T(\neg_{ij}\mathbf{y}) - T(\mathbf{y})] = (-1)^{y_{ij}}|\mathcal{N}_i \cap \mathcal{N}_j|$, where $\mathcal{N}_i$ denotes the neighbor-set of node i, and $\mathcal{N}_i^{\setminus j} := \mathcal{N}_i \setminus \{j\}$.

7.2. Experiments

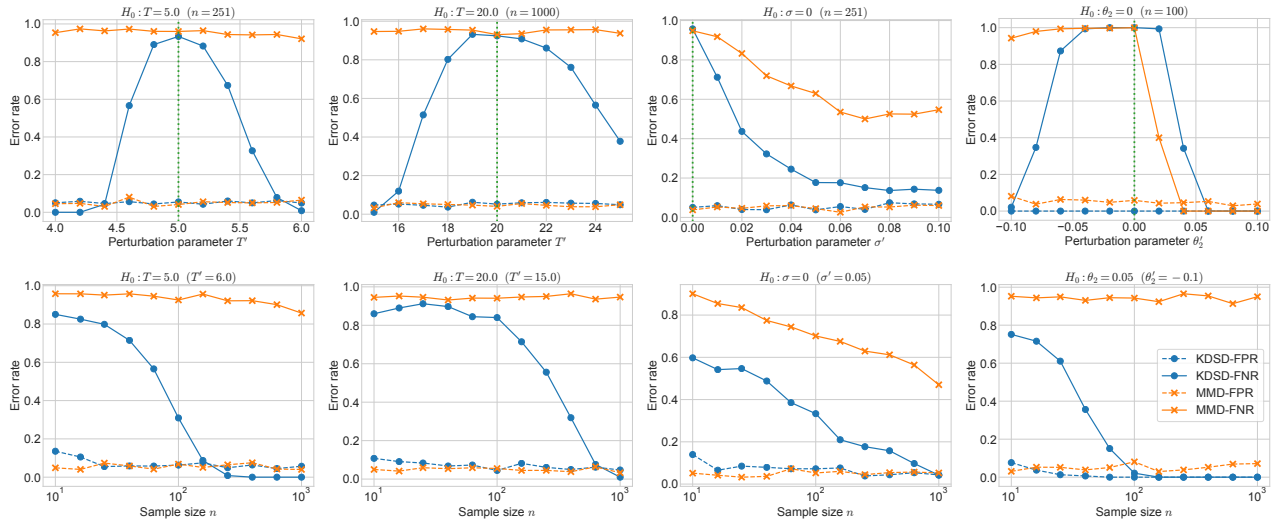
We apply the kernelized discrete Stein discrepancy (KDSD) test to the statistical models described in Sections 7.1. In the absence of established baselines, we compare with a two-sample test based on the maximum mean discrepancy (MMD) (see Section 6). For both KDSD and MMD, we utilize the exponentiated Hamming kernel (Eq. (16)) for the Ising model and RBM, and the Weisfeiler-Lehman graph kernel (Shervashidze et al., 2011) for the ERGM.

Setup. Denote the null model distribution by p and the alternative distribution by q. For each distribution, we draw exact *i.i.d.* samples by running n independent Markov chains with different random initializations, each for 10^5 iterations, and collecting only the last sample of each chain. For KDSD, we draw n samples from q; for MMD, we draw n samples from q and another n samples from p. Under this setup, both KDSD and MMD takes time $\mathcal{O}(mn^2)$, where m is the number of bootstrap samples used to determine the critical threshold. We set $m = 5000$ for both methods throughout.

For each model, we choose a "perturbation parameter" and fix its value for the null distribution p, while drawing data samples under various values of the perturbation parameter. We also vary the sample size n to examine the performance of the test as n increases. For each value of the perturbation parameter and each sample size n, we conduct 500 independent trials. In each trial, we first randomly flip a fair coin to decide whether to set the alternative distribution q to be the same as p or with a different value of the perturbation parameter. (In the former case, the null hypothesis $H_0 : p = q$ should not be rejected, and in the latter case it should be.) Then, we draw n independent samples from q (for KDSD) or both p and q (for MMD) and perform the hypothesis test $H_0 : p = q$ vs. $H_1 : p \neq q$ under significance level $\alpha = 0.05$. We evaluate the performance of the KDSD and MMD tests in terms of their false-positive rate (FPR; Type-I error) and false-negative rate (FNR; Type-II error), and report the results across 500 independent trials.

Ising model. We consider a periodic 10-by-10 lattice, with $d = 100$ random variables. We focus on the ferromagnetic

[2] Notice that the sufficient statistics are not independent: *e.g.*, $S_2(\mathbf{y}) > T(\mathbf{y})$ since every triangle contains three 2-stars.



(a) Ising model ($T_0 = 5$) (b) Ising model ($T_0 = 20$) (c) Bernoulli RBM (d) ERGM

Figure 1. Top row: KDSD and MMD testing error rate vs. perturbation parameter (the vertical dotted lines indicate the value of the perturbation parameter under H_0). *Bottom row:* KDSD and MMD testing error rate vs. sample size.

setting and set $\theta_{ij} = 1/T$, where T is the temperature of the system. For $T_0 \in \{5, 20\}$ and various values of T', we test the hypotheses $H_0 : T = T_0$ vs. $H_1 : T \neq T_0$ using data samples drawn from the model under $T = T'$. To draw samples from the Ising model, we apply the Metropolis algorithm: in each iteration, we propose to flip the spin of a randomly chosen variable x_i, and adopt this proposal with probability $\min(1, \exp\{-2x_i \sum_{j \in \mathcal{N}_i} \theta_{ij} x_j\})$.

Bernoulli RBM. We use $M = 50$ visible units and $K = 25$ hidden units. We draw the entries of the weight matrix $\mathbf{W}$ *i.i.d.* from a Normal distribution with mean zero and standard deviation $1/M$, and the entries of the bias terms $\mathbf{b}$ and $\mathbf{c}$ *i.i.d.* from the standard Normal distribution. We corrupt the weights in $\mathbf{W}$ by adding *i.i.d.* Gaussian noise with mean zero and standard deviation σ, and test the hypotheses $H_0 : \sigma = 0$ (no-corruption) vs. $H_1 : \sigma \neq 0$ using data samples drawn under $\sigma = \sigma'$ for various values of σ'. To draw samples from the RBM, we perform block Gibbs sampling by exploiting the bipartite structure of the graphical model.

ERGM. We consider an ERGM distribution for undirected graphs on 20 nodes, with the dimension of each sample $d = \binom{20}{2} = 190$. We fix $\theta_1 = -2$ and $\tau = 0.01$, For various values of the 2-star parameter θ_2', we test the hypotheses $H_0 : \theta_2 = 0$ vs. $H_1 : \theta_2 \neq 0$ using data samples drawn under $\theta_2 = \theta_2'$. To draw MCMC samples from the ERGM, we utilize the `ergm` R package (Handcock et al., 2017).

Results. In Figure 1, the top row plots the testing error rate vs. different values of the perturbation parameter in H_1, for a fixed H_0 and sample size; while the bottom row plots the error rate vs. sample size n for a fixed pair of H_0 and H_1. We observe that both KDSD and MMD maintain a false-positive rate (Type-I error) around or below the significance level $\alpha = 0.05$. In addition, KDSD consistently achieves lower false-negative rate (Type-II error) than MMD in most cases, indicating that KDSD, by utilizing the score function information of p, leads to a more powerful test.

It is interesting to note that in the ERGM example, MMD exhibits higher power than KDSD when the data samples were drawn from an ERGM distribution with $\theta_2' \in (0, 0.05)$ (roughly). We hypothesize that this may correspond to a regime in which a small change in θ_2 causes a subtle change in the *global* graph structure that can be more easily detected by MMD, while the difference Stein operator of Section 3.1 may be more adapt in detecting *local* differences. Thus, the performance of the KDSD test could be improved by constructing Stein operators (using the characterization of Section 3.2) that exploit higher-order structure in the graph samples, and we plan to investigate this in future work.

8. Conclusion

We have introduced a kernelized Stein discrepancy measure for discrete probability distributions, which enabled us to establish a nonparametric goodness-of-fit test for discrete distributions with intractable normalization constants. Furthermore, we have proposed a general characterization of Stein operators that encompasses both discrete and continuous distributions, providing a recipe for constructing new Stein operators. We have applied the proposed goodness-of-fit test to three statistical models involving discrete distributions, and shown that it typically outperforms a two-sample test based on the maximum mean discrepancy.

Acknowledgements. We thank the anonymous reviewers for their helpful comments. This research is supported by NSF under contract numbers IIS-1149789, IIS-1618690, IIS-1546488, and CCF-0939370.

References

Amari, S.-i. *Information Geometry and Its Applications*. Springer, 2016.

Anderson, T. W. and Darling, D. A. A test of goodness of fit. *Journal of the American Statistical Association*, 49 (268):765–769, 1954.

Arcones, M. A. and Gine, E. On the bootstrap of U and V statistics. *The Annals of Statistics*, 20(2):655–674, 1992.

Barbour, A. D. and Chen, L. H. Y. Stein's (magic) method. *arXiv:1411.1179*, 2014.

Bresler, G. and Nagaraj, D. Stein's method for stationary distributions of markov chains and application to Ising models. *arXiv:1712.05736*, 2017.

Brook, D. On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems. *Biometrika*, 51 (3/4):481–483, 1964.

Chwialkowski, K., Strathmann, H., and Gretton, A. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, pp. 2606–2615, 2016.

Daskalakis, C., Dikkala, N., and Kamath, G. Testing Ising models. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1989–2007, 2018.

Gorham, J. and Mackey, L. Measuring sample quality with Stein's method. In *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS)*, pp. 226–234, 2015.

Gorham, J. and Mackey, L. W. Measuring sample quality with kernels. In *Proceedings of The 34th International Conference on Machine Learning (ICML)*, pp. 1292–1301, 2017.

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

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Krivitsky, P. N., and Morris, M. *ergm: Fit, Simulate and Diagnose Exponential-Family Models for Networks*. The Statnet Project (http://www.statnet.org), 2017. URL https://CRAN.R-project.org/package=ergm. R package version 3.8.0.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786):504–507, 2006.

Hoeffding, W. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Holland, P. W. and Leinhardt, S. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

Huskova, M. and Janssen, P. Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics*, 21(4):1811–1823, 1993.

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

Hyvärinen, A. Some extensions of score matching. *Computational Statistics & Data Analysis*, 51(5):2499–2512, 2007.

Ising, E. *Beitrag zur Theorie des Ferro- und Paramagnetismus*. PhD thesis, 1924.

Jitkrittum, W., Xu, W., Szabo, Z., Fukumizu, K., and Gretton, A. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems 30*, pp. 261–270. 2017.

Kolmogorov, A. N. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:83–91, 1933.

Ley, C. and Swan, Y. Stein's density approach and information inequalities. *Electronic Communications in Probability*, 18:14 pp., 2013.

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

Liu, Q., Lee, J. D., and Jordan, M. I. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, 2016.

Lyu, S. Interpretation and generalization of score matching. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 359–366, 2009.

Martín del Campo, A., Cepeda, S., and Uhler, C. Exact goodness-of-fit testing for the Ising model. *Scandinavian Journal of Statistics*, 44(2):285–306, 2017.

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

Pearson, K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.

Reinert, G. and Ross, N. Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs. *arXiv:1712.05743*, 2017.

Shervashidze, N., Schweitzer, P., van Leeuwen, E. J., Mehlhorn, K., and Borgwardt, K. M. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561, 2011.

Smirnov, N. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948.

Stein, C. Approximate computation of expectations. *Institute of Mathematical Statistics Lecture Notes–Monograph Series*, 7:i–164, 1986.

Valiant, G. and Valiant, P. Instance optimal learning of discrete distributions. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 142–155, 2016.

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. Graph kernels. *Journal of Machine Learning Research*, 11:1201–1242, 2010.

Appendix to
Goodness-of-Fit Testing for Discrete Distributions via Stein Discrepancy

Proof of Theorem 1. Clearly, $p = q$ implies that $\mathbf{s}_p(\mathbf{x}) = \mathbf{s}_q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^d$. It remains to be shown that the converse is true. By Eq. (1), $\mathbf{s}_p(\mathbf{x}) = \mathbf{s}_q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^d$ implies that $p(\neg_i \mathbf{x})/p(\mathbf{x}) = q(\neg_i \mathbf{x})/q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^d$ and all $i = 1, \ldots, d$. We show that the latter implies that all the singleton conditional distributions of p and q must match, *i.e.*, $p(x_i|\mathbf{x}_{-i}) = q(x_i|\mathbf{x}_{-i})$ for all $x_i \in \mathcal{X}$ and for all $i = 1, \ldots, d$, where $\mathbf{x}_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$.

Specifically, using the fact that $\neg$ is a cyclic permutation on $\mathcal{X}$, we can write

$$\begin{aligned}
\frac{1}{p(x_i|\mathbf{x}_{-i})} &= \frac{\sum_{\xi_i \in \mathcal{X}} p(x_1, \ldots, x_{i-1}, \xi_i, x_{i+1}, \ldots, x_d)}{p(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)} = \sum_{\xi_i \in \mathcal{X}} \frac{p(x_1, \ldots, x_{i-1}, \xi_i, x_{i+1}, \ldots, x_d)}{p(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)} \\
&= \sum_{\ell=1}^{|\mathcal{X}|} \frac{p(x_1, \ldots, x_{i-1}, \neg^{(\ell)} x_i, x_{i+1}, \ldots, x_d)}{p(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_d)} \\
&= \sum_{\ell=1}^{|\mathcal{X}|} \frac{p(\neg_i^{(\ell)} \mathbf{x})}{p(\mathbf{x})} = \sum_{\ell=1}^{|\mathcal{X}|} \prod_{j=0}^{\ell-1} \frac{p(\neg_i^{(j+1)} \mathbf{x})}{p(\neg_i^{(j)} \mathbf{x})} = \sum_{\ell=1}^{|\mathcal{X}|} \prod_{j=0}^{\ell-1} \frac{p(\neg_i \mathbf{y}_{ij})}{p(\mathbf{y}_{ij})}\,,
\end{aligned} \tag{17}$$

where we adopted the convention that $\neg^{(0)} \mathbf{x} = \mathbf{x}$ and written $\mathbf{y}_{ij} := \neg_i^{(j)} \mathbf{x}$ in the last term. By Eq. (1), all the terms on the right-hand-side of Eq. (17) will be determined by the score function $\mathbf{s}_p(\mathbf{x})$, and thus $\mathbf{s}_p(\mathbf{x}) = \mathbf{s}_q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^d$ implies that all the singleton conditional distributions must match: $p(x_i|\mathbf{x}_{-i}) = q(x_i|\mathbf{x}_{-i})$, $\forall \mathbf{x} \in \mathcal{X}^d$. By Brook's lemma (Brook, 1964; see Lemma 9 for a self-contained proof), the joint probability distribution is fully specified by the collection of singleton conditional distributions, and thus we must have $p(\mathbf{x}) = q(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}^d$. □

Lemma 9 (Brook, 1964)**.** *Assume that $p(\mathbf{x}) > 0$ for all $\mathbf{x} \in \mathcal{X}^d$. The joint distribution $p(\mathbf{x})$ is completely determined by the collection of singleton conditional distributions $p(x_i|\mathbf{x}_{-i})$, where $\mathbf{x}_{-i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$, $i = 1, \ldots, d$.*

Proof. Let $p(x_1, \ldots, x_d)$ and $p(y_1, \ldots, y_d)$ denote the joint densities (pmfs or pdfs) for $(x_1, \ldots, x_d)$ and $(y_1, \ldots, y_d)$, respectively. We can write

$$\begin{aligned}
\frac{p(x_1, x_2, \ldots, x_d)}{p(y_1, y_2, \ldots, y_d)} &= \frac{p(x_1, x_2, \ldots, x_d)}{p(y_1, x_2, \ldots, x_d)} \cdot \frac{p(y_1, x_2, \ldots, x_d)}{p(y_1, y_2, \ldots, x_d)} \cdots \frac{p(y_1, y_2, \ldots, y_{d-1}, x_d)}{p(y_1, y_2, \ldots, y_{d-1}, y_d)} \\
&= \frac{p(x_1|x_2, \ldots, x_d)}{p(y_1|x_2, \ldots, x_d)} \cdot \frac{p(x_2|y_1, x_3, \ldots, x_d)}{p(y_2|y_1, x_3, \ldots, x_d)} \cdots \frac{p(x_d|y_1, \ldots, y_{d-1})}{p(y_d|y_1, \ldots, y_{d-1})}\,.
\end{aligned}$$

Thus, the collection of all singleton conditional distributions completely determine the ratios of joint probability densities, which in turn completely determine the joint densities themselves, since they have to sum to one. □

The following result provides more convenient expressions for evaluating $\mathbb{E}_{\mathbf{x}\sim p}\left[\mathcal{A}_p \mathbf{f}(\mathbf{x})\right]$ and $\mathbb{E}_{\mathbf{x}\sim p}\left[\mathrm{tr}\left(\mathcal{A}_p \mathbf{f}(\mathbf{x})\right)\right]$.

Lemma 10 (See also Ley & Swan (2013))**.** *For positive pmfs p, q and any function $\mathbf{f} : \mathcal{X}^d \to \mathbb{R}^d$, we have*

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_p \mathbf{f}(\mathbf{x})\right] &= \mathbb{E}_{\mathbf{x}\sim q}\left[\left(\mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})\right) \mathbf{f}(\mathbf{x})^\mathsf{T}\right], \\
\mathbb{E}_{\mathbf{x}\sim q}\left[\mathrm{tr}\left(\mathcal{A}_p \mathbf{f}(\mathbf{x})\right)\right] &= \mathbb{E}_{\mathbf{x}\sim q}\left[\left(\mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})\right)^\mathsf{T} \mathbf{f}(\mathbf{x})\right].
\end{aligned}$$

Proof. Theorem 2 states that $\mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_q \mathbf{f}(\mathbf{x})\right] = 0$. Thus, writing $\mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_p \mathbf{f}(\mathbf{x})\right] = \mathbb{E}_{\mathbf{x}\sim q}\left[\mathcal{A}_p \mathbf{f}(\mathbf{x}) - \mathcal{A}_q \mathbf{f}(\mathbf{x})\right] = \mathbb{E}_{\mathbf{x}\sim q}[\left(\mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})\right) \mathbf{f}(\mathbf{x})^\mathsf{T}]$ and taking the trace on both sides completes the proof. □

Proof of Theorem 3 (Continued)**.** *Necessity:* Assume that a linear operator $\mathcal{T}$ satisfies Eq. (7); we show that it can be written in the form of Eq. (8) for some linear operators $\mathcal{L}$ and $\mathcal{L}^*$ of the forms (5) and (6). Recall that for a finite set $\mathcal{X}$, any function $f : \mathcal{X}^d \to \mathbb{R}$ can be represented by a vector $\mathbf{f} \in \mathbb{R}^{|\mathcal{X}|^d}$, and any linear operator $\mathcal{T}$ on the set of functions f can be represented via a matrix $\mathbf{T} \in \mathbb{R}^{|\mathcal{X}|^d \times |\mathcal{X}|^d}$ under the standard basis of $\mathbb{R}^{|\mathcal{X}|^d}$. Under these notations, $\mathcal{T}f$ can be represented by $\mathbf{T}\mathbf{f}$, and Eq. (7) can be rewritten in matrix form as

$$\mathbb{E}_{\mathbf{x}\sim p}\left[\mathcal{T}_p f(\mathbf{x})\right] = \sum_{\mathbf{x}\in\mathcal{X}^d} p(\mathbf{x})\mathcal{T}_p f(\mathbf{x}) = \mathbf{p}^\mathsf{T}(\mathbf{T}_p\mathbf{f}) = 0\,,$$

which holds for any function f (*i.e.*, for any vector $\mathbf{f}$) if and only if $\mathbf{p}^\mathsf{T}\mathbf{T}_p = \mathbf{0}$. We can always find a diagonal matrix $\mathbf{D}$ and a matrix $\mathbf{L}$ such that $\mathbf{T}_p = \mathbf{D} - \mathbf{L}$. Observe that $\mathbf{p}^\mathsf{T}\mathbf{T}_p = \mathbf{0}$, *i.e.*, $\mathbf{p}^\mathsf{T}\mathbf{D} = \mathbf{p}^\mathsf{T}\mathbf{L}$ if and only if $d_{ii} = \mathbf{p}^\mathsf{T}\mathbf{L}_{*i}/p_i$ for all i, where d_{ii} is the i-th diagonal element of $\mathbf{D}$ and $\mathbf{L}_{*i}$ is the i-th column of $\mathbf{L}$. Thus, Eq. (7) holds if and only if

$$\mathbf{T}_p = \operatorname{diag}\{\mathbf{p}\}^{-1}\operatorname{diag}\left\{\mathbf{L}^\mathsf{T}\mathbf{p}\right\} - \mathbf{L}$$

for some matrix $\mathbf{L}$, where $\operatorname{diag}\{\mathbf{p}\}$ denotes the diagonal matrix whose i-th diagonal entry equals p_i. Rewriting, we have

$$\operatorname{diag}\{\mathbf{p}\}\,\mathbf{T}_p = \operatorname{diag}\left\{\mathbf{L}^\mathsf{T}\mathbf{p}\right\} - \operatorname{diag}\{\mathbf{p}\}\,\mathbf{L}\,.$$

Right-multiplying both sides by an arbitrary vector $\mathbf{f} \in \mathbb{R}^{|\mathcal{X}|^d}$, we obtain

$$\mathbf{p}\odot(\mathbf{T}_p\mathbf{f}) = (\mathbf{L}^\mathsf{T}\mathbf{p})\odot\mathbf{f} - \mathbf{p}\odot(\mathbf{L}^\mathsf{T}\mathbf{f})\,, \tag{18}$$

where $\odot$ denotes the Hadamard product. Let $\mathcal{L}$ and $\mathcal{L}^*$ be the linear operators with matrices $\mathbf{L}^\mathsf{T}$ and $\mathbf{L}$ under the standard basis, Eq. (18) can be re-written as

$$p(\mathbf{x})\mathcal{T}_p f(\mathbf{x}) = \mathcal{L}p(\mathbf{x})f(\mathbf{x}) - p(\mathbf{x})\mathcal{L}^* f(\mathbf{x})$$

for all $\mathbf{x} \in \mathcal{X}^d$. Finally, dividing by $p(\mathbf{x})$ on both sides yields Eq. (8). □

Proof of Theorem 6. Observe that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}\sim q}\left[\operatorname{tr}\left(\mathcal{A}_p\mathbf{f}(\mathbf{x})\right)\right] &= \sum_{\ell=1}^{d}\mathbb{E}_{\mathbf{x}\sim q}\left[s_p^\ell(\mathbf{x})\,f_\ell(\mathbf{x}) - \Delta^*_{x_\ell}f_\ell(\mathbf{x})\right] \\
&= \sum_{\ell=1}^{d}\mathbb{E}_{\mathbf{x}\sim q}\left[s_p^\ell(\mathbf{x})\,\langle f_\ell, k(\cdot,\mathbf{x})\rangle_{\mathcal{H}} - \left\langle f_\ell, \Delta^*_{x_\ell}k(\cdot,\mathbf{x})\right\rangle_{\mathcal{H}}\right] \\
&= \sum_{\ell=1}^{d}\left\langle f_\ell, \mathbb{E}_{\mathbf{x}\sim q}\left[s_p^\ell(\mathbf{x})\,k(\cdot,\mathbf{x}) - \Delta^*_{x_\ell}k(\cdot,\mathbf{x})\right]\right\rangle_{\mathcal{H}},
\end{aligned}$$

where we used the reproducing property $\langle f_\ell, k(\cdot,\mathbf{x})\rangle_{\mathcal{H}} = f_\ell(\mathbf{x})$ and the fact that

$$\Delta^*_{x_j}f_i(\mathbf{x}) = f_i(\mathbf{x}) - f_i(\neg_j\mathbf{x}) = \langle f_i, k(\cdot,\mathbf{x})\rangle - \langle f_i, k(\cdot,\neg_j\mathbf{x})\rangle = \langle f_i, k(\cdot,\mathbf{x}) - k(\cdot,\neg_j\mathbf{x})\rangle = \left\langle f_j, \Delta^*_{x_j}k(\cdot,\mathbf{x})\right\rangle.$$

Denoting $\boldsymbol{\beta}(\cdot) := \mathbb{E}_{\mathbf{x}\sim q}\left[\mathbf{s}_p(\mathbf{x})k(\cdot,\mathbf{x}) - \Delta^*k(\cdot,\mathbf{x})\right] \in \mathcal{H}^m$, we have

$$\mathbb{E}_{\mathbf{x}\sim q}\left[\operatorname{tr}\left(\mathcal{A}_p\mathbf{f}(\mathbf{x})\right)\right] = \sum_{\ell=1}^{d}\langle f_\ell, \beta_\ell\rangle_{\mathcal{H}} = \langle\mathbf{f},\boldsymbol{\beta}\rangle_{\mathcal{H}^m}\,.$$

Thus, we can rewrite the kernelized discrete Stein discrepancy as

$$\mathbb{D}(q\,\|\,p) = \sup_{\mathbf{f}\in\mathcal{H}^m,\,\|\mathbf{f}\|_{\mathcal{H}^m}\le 1}\langle\mathbf{f},\boldsymbol{\beta}\rangle_{\mathcal{H}^m}\,,$$

which immediately implies that $\mathbb{D}(q\,\|\,p) = \|\boldsymbol{\beta}\|_{\mathcal{H}^m}$ since the supremum will be attained by $\mathbf{f} = \boldsymbol{\beta}/\|\boldsymbol{\beta}\|_{\mathcal{H}^m}$.

By Lemma 10, we have

$$\boldsymbol{\beta}(\cdot) = \mathbb{E}_{\mathbf{x}\sim q}\left[\mathbf{s}_p(\mathbf{x})k(\cdot,\mathbf{x}) - \Delta^* k(\cdot,\mathbf{x})\right] = \mathbb{E}_{\mathbf{x}\sim q}\left[(\mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x}))k(\cdot,\mathbf{x})\right].$$

Writing $\boldsymbol{\delta}_{p,q}(\mathbf{x}) := \mathbf{s}_p(\mathbf{x}) - \mathbf{s}_q(\mathbf{x})$, we have

$$\begin{aligned}
\mathbb{D}(q\,\|\,p)^2 = \|\boldsymbol{\beta}\|^2_{\mathcal{H}^m} &= \sum_{\ell=1}^d \langle \beta_\ell, \beta_\ell\rangle_{\mathcal{H}} = \sum_{\ell=1}^d \left\langle \mathbb{E}_{\mathbf{x}\sim q}\left[\delta^\ell_{p,q}(\mathbf{x})\, k(\cdot,\mathbf{x})\right], \mathbb{E}_{\mathbf{x}'\sim q}\left[\delta^\ell_{p,q}(\mathbf{x}')\, k(\cdot,\mathbf{x}')\right]\right\rangle_{\mathcal{H}} \\
&= \sum_{\ell=1}^d \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\delta^\ell_{p,q}(\mathbf{x})\, \langle k(\cdot,\mathbf{x}), k(\cdot,\mathbf{x}')\rangle_{\mathcal{H}}\, \delta^\ell_{p,q}(\mathbf{x}')\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\boldsymbol{\delta}_{p,q}(\mathbf{x})^\mathsf{T}\, \langle k(\cdot,\mathbf{x}), k(\cdot,\mathbf{x}')\rangle_{\mathcal{H}}\, \boldsymbol{\delta}_{p,q}(\mathbf{x}')\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\boldsymbol{\delta}_{p,q}(\mathbf{x})^\mathsf{T} k(\mathbf{x},\mathbf{x}')\, \boldsymbol{\delta}_{p,q}(\mathbf{x}')\right],
\end{aligned}$$

where we used the reproducing property, $k(\mathbf{x},\mathbf{x}') = \langle k(\cdot,\mathbf{x}), k(\cdot,\mathbf{x}')\rangle_{\mathcal{H}}$. This concludes the proof. □

Proof of Theorem 7. Expanding the expression for $\boldsymbol{\delta}_{p,q}(x)$ and applying Lemma 10 twice, we obtain

$$\begin{aligned}
\mathbb{D}(q\,\|\,p)^2 &= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\boldsymbol{\delta}_{p,q}(\mathbf{x})^\mathsf{T} k(\mathbf{x},\mathbf{x}')\boldsymbol{\delta}_{p,q}(\mathbf{x}')\right] \\
&= \mathbb{E}_{\mathbf{x}\sim q}\left[\boldsymbol{\delta}_{p,q}(\mathbf{x})^\mathsf{T}\mathbb{E}_{\mathbf{x}'\sim q}\left[k(\mathbf{x},\mathbf{x}')\boldsymbol{\delta}_{p,q}(\mathbf{x}')\right]\right] \\
&= \mathbb{E}_{\mathbf{x}\sim q}\left[\boldsymbol{\delta}_{p,q}(\mathbf{x})^\mathsf{T}\mathbb{E}_{\mathbf{x}'\sim q}\left[k(\mathbf{x},\mathbf{x}')\mathbf{s}_p(\mathbf{x}') - \Delta^*_{\mathbf{x}'}k(\mathbf{x},\mathbf{x}')\right]\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\mathbf{s}_p(\mathbf{x})^\mathsf{T} k(\mathbf{x},\mathbf{x}')\,\mathbf{s}_p(\mathbf{x}') - \mathbf{s}_p(\mathbf{x})^\mathsf{T}\Delta^*_{\mathbf{x}'}k(\mathbf{x},\mathbf{x}') - \Delta^*_{\mathbf{x}}k(\mathbf{x},\mathbf{x}')^\mathsf{T}\mathbf{s}_p(\mathbf{x}') + \operatorname{tr}\left(\Delta^*_{\mathbf{x},\mathbf{x}'}k(\mathbf{x},\mathbf{x}')\right)\right] \\
&= \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\kappa_p(\mathbf{x},\mathbf{x}')\right],
\end{aligned}$$

which completes the proof. □

Theorem 11 (Adapted from Liu et al., 2016)**.** *Let $k(x,x')$ be a strictly positive definite kernel on $\mathcal{X}^d$, and assume that $\mathbb{E}_{\mathbf{x},\mathbf{x}'\sim q}\left[\kappa_p(\mathbf{x},\mathbf{x}')^2\right] < \infty$. We have the following two cases:*

(i) If $q \neq p$, then $\widehat{\mathbb{S}}(q\,\|\,p)$ is asymptotically Normal:

$$\sqrt{n}\left(\widehat{\mathbb{S}}(q\,\|\,p) - \mathbb{S}(q\,\|\,p)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2),$$

where $\sigma^2 = \operatorname{Var}_{\mathbf{x}\sim q}(\mathbb{E}_{\mathbf{x}'\sim q}\left[\kappa_p(\mathbf{x},\mathbf{x}')\right]) > 0$.

(ii) If $q = p$, then $\sigma^2 = 0$, and the U-statistic is degenerate:

$$n\,\widehat{\mathbb{S}}(q\,\|\,p) \xrightarrow{\mathcal{D}} \sum_j c_j(Z_j^2 - 1),$$

where $\{Z_j\} \overset{iid}{\sim} \mathcal{N}(0,1)$ and $\{c_j\}$ are the eigenvalues of the kernel $\kappa_p(\cdot,\cdot)$ under q.

Lemma 12. *The exponentiated Hamming kernel*

$$k(\mathbf{x},\mathbf{x}') = \exp\{-H(\mathbf{x},\mathbf{x}')\},$$

where $H(\mathbf{x},\mathbf{x}') := \frac{1}{d}\sum_{i=1}^d \mathbb{I}\{x_i \neq x'_i\}$ is the normalized Hamming distance, is positive definite.

Proof. Without loss of generality, assume that $\mathcal{X} = \{0,1\}$ is a binary set; the general case can be easily accommodated by modifying the feature map to be described next. Define the feature map $\boldsymbol{\phi} : \mathcal{X}^d \to \mathcal{X}^{2d}$, $\mathbf{x} \mapsto \widetilde{\mathbf{x}}$, where $\widetilde{x}_{2i-1} = \mathbb{I}\{x_i = 0\}$ and $\widetilde{x}_{2i} = \mathbb{I}\{x_i = 1\}$ for $i = 1,\ldots,d$. Then, the normalized Hamming distance can be expressed as

$$H(\mathbf{x},\mathbf{x}') = 1 - \frac{1}{d}\sum_{i=1}^d \mathbb{I}\{x_i = x'_i\} = 1 - \frac{1}{2d}\sum_{j=1}^{2d}\widetilde{x}_j\widetilde{x}'_j = 1 - \frac{1}{2d}\widetilde{\mathbf{x}}^\mathsf{T}\widetilde{\mathbf{x}}' = 1 - \frac{1}{2d}\boldsymbol{\phi}(\mathbf{x})^\mathsf{T}\boldsymbol{\phi}(\mathbf{x}').$$

Thus, $1 - H(\mathbf{x},\mathbf{x}')$ is a positive definite kernel. By Taylor expansion, $\exp\{1 - H(\mathbf{x},\mathbf{x}')\}$ (and hence $\exp\{-H(\mathbf{x},\mathbf{x}')\}$) also constitutes a positive definite kernel on $\mathcal{X}^d$. □