A Stein–Papangelou Goodness-of-Fit Test for Point Processes

Jiasen Yang
Department of Statistics
Purdue University
West Lafayette, IN
jiaseny@purdue.edu

Vinayak Rao
Department of Statistics
Purdue University
West Lafayette, IN
varao@purdue.edu

Jennifer Neville
Computer Science and Statistics
Purdue University
West Lafayette, IN
neville@purdue.edu

Abstract

Point processes provide a powerful framework for modeling the distribution and interactions of events in time or space. Their flexibility has given rise to a variety of sophisticated models in statistics and machine learning, yet model diagnostic and criticism techniques remain underdeveloped. In this work, we propose a general Stein operator for point processes based on the Papangelou conditional intensity function. We then establish a kernel goodness-of-fit test by defining a Stein discrepancy measure for general point processes. Notably, our test also applies to non-Poisson point processes whose intensity functions contain intractable normalization constants due to the presence of complex interactions among points. We apply our proposed test to several point process models, and show that it outperforms a two-sample test based on the maximum mean discrepancy.

1 INTRODUCTION

Point pattern data, consisting of the locations of objects in some ambient space, occur widely in the physical, biological, and social sciences. Point process models have been applied to describe stars and galaxies [2], trees in a forest [15], earthquakes and aftershocks [33], neurons in the brain [29], and the dynamics of crime [28]. Point processes have also been the subject of much recent activity in statistics and machine learning, and a spate of sophisticated probabilistic and deep neural network models have been developed [16, 28, 34, 42, 44].

While the complexity of such point process models has grown at a rapid pace, corresponding tools for model diagnostics, evaluation, and criticism have lagged behind, restricted mostly to the spatial statistics literature. Beyond Poisson-type processes [9, 13], and residual-based analysis and diagnostic plots for some spatial processes [4], rigorous statistical tests to assess how well a point process model fits the observed data remains an important and under-studied topic [12].

In this work, we investigate an important class of statistical tests—the *goodness-of-fit test*—for point processes. Goodness-of-fit testing is a fundamental topic in statistics [27], but for point processes, well-established goodness-of-fit tests are only available under the simplest scenarios—such as when the null model is a Poisson process. For more general point processes, the construction of such tests typically rely on pseudo-likelihood approximations [40] which introduce biases and errors that are hard to quantify, or heuristic summary statistics (such as Ripley's K-function [35]) which could only capture certain aspects of the observed data and may lead to a considerable loss of statistical power.

A major hurdle preventing the construction of rigorous statistical tests (such as those based on the likelihood-ratio statistic) for more sophisticated point processes is the presence of *intractable normalization constants* in the density/intensity functions. For many widely used models that capture pairwise or higher-order dependencies between points, these functions can often be evaluated only up to a normalization constant, because summing over all possible configurations leads to an intractable infinite-dimensional integral. This precludes the use of classical tests (such as the likelihood-ratio test) which require the fully specified model density.

Recently, much progress have been made in developing nonparametric statistical tests which work directly with *unnormalized* probability distributions [11, 18, 19, 24, 30, 31, 43]. Central to these tests is a *Stein operator* [39] $\mathcal{A}_p$ such that, for functions f in some family, the expectation $\mathbb{E}\left[\mathcal{A}_p f\right]$ equals zero only under the distribution of interest p. All the aforementioned works have considered distributions over *fixed-length* (d-dimensional) vectors residing in a space $\mathbb{X}$ that is either the Euclidean space $\mathbb{R}^d$ (for continuous distributions) or $\mathcal{X}^d$ where $\mathcal{X}$ is a finite set (for discrete distributions). These works have shown how to con-

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

struct Stein operators (and goodness-of-fit tests) which only require unnormalized probability densities. On the other hand, a realization of a point process is a *set* containing an arbitrary number of points, and forms an element of an infinite-dimensional space. Constructing a Stein operator for this setting does not follow easily from previous work, and requires a new set of tools.

A primary contribution of our work is in identifying a suitable Stein operator for general point processes. While such constructions have been well-studied for Poisson process approximations in the probability literature [7, 8], constructions for general point processes have been largely unexplored. Our key technical tool in constructing a general Stein operator is the *Papangelou conditional intensity* of a point process (see Section 2). Importantly, any (intractable) normalization constant in the density or intensity function of the point process cancels out when evaluating the Papangelou conditional intensity. Using our proposed Stein operator, along with a suitable kernel function on the space of point configurations, we proceed to define a *kernelized Stein discrepancy* measure between distributions, following a similar strategy pioneered by [11, 30]. This allows us to develop a computationally feasible, nonparametric goodness-of-fit test for general point processes, including those with intractable normalization constants (*e.g.*, the Gibbs process). We apply our proposed goodness-of-fit test to the Poisson process, as well as two processes with inter-point interactions: the *Hawkes process* [21] exhibiting self-excitation, and the *Strauss process* [41] featuring repulsion. Our experiments show that the proposed test outperforms a two-sample test based on the *maximum mean discrepancy* [20] in terms of power while maintaining control on false-positive rate.

2 POINT PROCESSES

Notation. Let $\mathbb{X}$ be a locally compact metric space with $\mathcal{B}_{\mathbb{X}}$ its Borel σ-algebra. We will refer to $\mathbb{X}$ as the *ground space*, and consider point processes with points lying in this space. In practice, $\mathbb{X}$ is usually a compact subset of the d-dimensional Euclidean space $\mathbb{R}^d$.

A *configuration* or *realization* of a point process on $\mathbb{X}$ is a locally finite counting measure on $(\mathbb{X}, \mathcal{B}_{\mathbb{X}})$. We shall be primarily concerned with finite configurations in this work; these form finite integer-valued measures on $(\mathbb{X}, \mathcal{B}_{\mathbb{X}})$. Let us denote the space of finite configurations on $\mathbb{X}$ by $\mathcal{N}_{\mathbb{X}}$. While a configuration is formally defined as a counting measure, we shall also identify it as a (locally) finite set of points, and describe it using set-theoretic notations. Conversely, any (locally) finite set of points $\phi \subseteq \mathbb{X}$ also defines the configuration with set A having measure $\phi(A) := |\phi \cap A|$, $\forall A \in \mathcal{B}_{\mathbb{X}}$, where $|\cdot|$ denotes the cardinality of a set. For a point $x \in \mathbb{X}$, let δ_x denote the Dirac measure centered at x. Given a point configuration $\phi \in \mathcal{N}_{\mathbb{X}}$, the configurations $\phi + \delta_x$ and $\phi - \delta_x$ correspond to the point-sets $\phi \cup \{x\}$ and $\phi \backslash \{x\}$, respectively, and we shall use the measure-theoretic and set-theoretic notations interchangeably.

Point process. Formally, a *point process* Φ on $\mathbb{X}$ is a random point configuration on $\mathbb{X}$. Define its *intensity measure* μ as

$$\mu(A) := \mathbb{E}\left[\Phi(A)\right], \quad \forall A \in \mathcal{B}_{\mathbb{X}}.$$

When $\mathbb{X} \subseteq \mathbb{R}^d$, the intensity measure is typically given in terms of a positive function $\lambda(\cdot)$ on $\mathbb{X}$, called the *rate* or *intensity function*: $\mu(A) = \int_A \lambda(x)\,\mathrm{d}x$.

When $d = 1$, the underlying space $\mathbb{X} \subseteq \mathbb{R}_+$ typically indexes time, and the process is called a *temporal* point process. When $d > 1$, the process is often termed a *spatial* point process (one typically considers $d = 2$ or $d = 3$ in applications). A crucial distinction between $d = 1$ and $d > 1$ is the existence of a natural ordering among the elements in $\mathbb{R}$, which is absent in $\mathbb{R}^d$ $(d > 1)$.

We now describe a few point processes and introduce some important theoretical tools along the way.

Poisson process. A point process Φ with intensity measure μ is called a Poisson process if (i) the counting measure Φ is *completely random*, *i.e.*, for any disjoint measurable subsets $A_1, A_2, \ldots, A_k \in \mathcal{B}_{\mathbb{X}}$, the point counts $\Phi(A_1), \Phi(A_2), \ldots, \Phi(A_k)$ are independent random variables; and (ii) for any set $A \in \mathcal{B}_{\mathbb{X}}$, $\Phi(A)$ follows a Poisson distribution with mean $\mu(A)$. A Poisson process is said to be *homogeneous* if its intensity function λ is constant, and *inhomogeneous* otherwise.

The following result, known as the *Mecke formula*, characterizes the Poisson process through the expectation of integrals (sums) with respect to a Poisson process, where the integrand depends on both the point process and a location in the ground space.

Theorem 1 (Mecke formula [26])**.** *Let μ be an s-finite measure and Φ be a point process on $\mathbb{X}$. Then Φ is a Poisson process with intensity measure μ if and only if*

$$\mathbb{E}\left[\int_{\mathbb{X}} h(x, \Phi)\Phi(\mathrm{d}x)\right] = \int_{\mathbb{X}} \mathbb{E}\left[h(x, \Phi + \delta_x)\right]\mu(\mathrm{d}x).$$

for all measurable functions $h : \mathbb{X} \times \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$.

More complicated point processes relax the assumption of complete randomness. We consider two examples, the *Hawkes process* [21] and the *Gibbs processes* [13].

Hawkes process. Consider a temporal point process Φ defined on the non-negative real line $\mathbb{R}_+$. For any $t \geq 0$, let $N(t) := \Phi([0, t))$ denote the number of points in the time interval $[0, t)$. Let $\mathcal{H}_t := \{N(s)\}_{s<t}$ denote the history of the point process prior to time t. We define the *conditional intensity* function as the

instantaneous arrival rate of the point process given the history $\mathcal{H}_t$. Formally, a Hawkes process is a temporal point process with conditional intensity

$$\lambda(t|\mathcal{H}_t) = \gamma + \int_0^t g(t-s)\,\Phi(\mathrm{d}s), \tag{1}$$

where γ is the base-rate, and $g(\cdot)$ is a *triggering function* that characterizes the excitatory effect that a past event has on the current event rate. For example, one could set $g(t) := \beta e^{-t/\tau}$, $t \geq 0$, implying that an event has an excitatory boost of magnitude $\beta \geq 0$, which decays exponentially with a time-scale $\tau > 0$.

Gibbs processes. These are a general class of point processes that model inter-point interactions in higher-dimensional spaces. The probability *density* of a Gibbs process (with respect to the unit-rate Poisson process on $\mathbb{X}$) takes the form [5, 36]:

$$f(\phi) = \frac{1}{Z}\exp\left\{-\sum_{k=1}^{|\phi|}\sum_{\omega\subseteq\phi,\,|\omega|=k} v_k(\omega)\right\},$$

where $v_k : \mathbb{X} \to \mathbb{R}$ is called the k-th order interaction potential, and Z is a normalization constant. Note that this normalization constant involves summing over all possible configurations $\phi \in \mathcal{N}_{\mathbb{X}}$, an *infinite-dimensional* integral which is intractable in all but the simplest situations (*e.g.*, the Poisson process, which is a Gibbs process with $v_k \equiv 0$, $\forall k > 1$).

Papangelou conditional intensity. The key challenge in generalizing the conditional intensity to spatial point processes is the lack of a natural ordering in $\mathbb{R}^d$ when $d > 1$: the 'history' of the process is not defined. For a point process Φ with density f, we follow [4] and define its *Papangelou conditional intensity* as

$$\rho(x|\phi) = \begin{cases} f(\phi\cup\{x\})\,/\,f(\phi)\,, & \text{if } x \notin \phi\,; \\ f(\phi)\,/\,f(\phi\backslash\{x\})\,, & \text{if } x \in \phi\,, \end{cases} \tag{2}$$

for $x \in \mathbb{X}$ and $\phi \in \mathcal{N}_{\mathbb{X}}$. We set $\rho(x|\phi) = 0$ if $f(\phi) = 0$. Informally, $\rho(x|\phi)\,\mathrm{d}x$ represents the relative probability of there being a point of Φ lying within an infinitesimal region of area $\mathrm{d}x$ containing x, given that the rest of the point process Φ coincides with ϕ [4]. Thus, the Papangelou conditional intensity provides an intuitive characterization of a point process.

For a Poisson process, its complete randomness ensures that its Papangelou conditional intensity is equivalent to its intensity: $\rho(x|\phi) \equiv \lambda(x)$, $\forall x \in \mathbb{X}$, $\phi \in \mathcal{N}_{\mathbb{X}}$.

For a Hawkes process, the density function is given by

$$f(\{t_i\}_{i=1}^n) = e^{-\Lambda(0,T)}\prod_{i=1}^n \lambda(t_i|\mathcal{H}_{t_i}),$$

where $\Lambda(t) := \int_0^T \lambda(t)\,\mathrm{d}t$ is the *integrated intensity*. Thus, we have the Papangelou conditional intensity:

$$\rho(x|\{t_i\}_{i=1}^n) = e^{-\int_0^{T-x} g(s)\,\mathrm{d}s}\cdot\left[\gamma + \sum_{k:\,t_k<x} g(x-t_k)\right] \times \prod_{i:\,t_i>x} \frac{\gamma + \sum_{k:\,t_k<t_i} g(t_i-t_k) + g(t_i-x)}{\gamma + \sum_{k:\,t_k<t_i} g(t_i-t_k)}\,.$$

Notice that the Papangelou conditional intensity is different from the conditional intensity function $\lambda(t|\mathcal{H}_t)$ which conditions only on events prior to t.

For a Gibbs process, although its density f and intensity function λ are both intractable, the normalization constant Z cancels out when evaluating Eq. (2), and the Papangelou conditional intensity is fully available:

$$\rho(x|\phi) = \exp\left\{-\sum_{k=1}^{|\phi|}\sum_{\omega\subseteq\phi,\,|\omega|=k-1} v_k(\{x\}\cup\omega)\right\}. \tag{3}$$

An illustrative instance of Gibbs processes is the *Strauss process* [41], a popular *repulsive* point process.

Strauss process. The Strauss process is a spatial point process on $\mathbb{X} \subseteq \mathbb{R}^d$ with conditional intensity

$$\rho(x|\phi) = \beta\gamma^{t_r(x,\phi)}, \tag{4}$$

where $\beta > 0$, $\gamma \in [0,1]$, and $t_r(x,\phi) := \sum_{y\in\phi}\mathbb{I}\{\|x-y\|_2 \leq r\}$ counts the number of points in ϕ that lie within a distance $r > 0$ of the location x. Notice that Eq. (4) can be recovered from Eq. (3) by setting $v_1(\{x\}) \equiv -\beta$, $v_2(\{x,y\}) = -(\log\gamma)\cdot\mathbb{I}\{\|x-y\|_2 \leq r\}$, and $v_k(\omega) \equiv 0$, $\forall k > 2$. While the conditional intensity of the Strauss process takes the simple form of Eq. (4), we note that its density and intensity functions are generally computationally intractable for $d \geq 2$.

We conclude this section by reviewing an important identity that generalizes the Mecke formula (Theorem 1) to any finite point process. This identity will serve as an essential tool in our subsequent development of a Stein operator for general point processes.

Theorem 2 (Georgii–Nguyen–Zessin (GNZ) formula [13])**.** *Let Φ be a finite point process on $\mathbb{X}$ with Papangelou conditional intensity ρ. For any measurable function $h : \mathbb{X}\times\mathcal{N}_{\mathbb{X}} \to \mathbb{R}$,*

$$\mathbb{E}\left[\int_{\mathbb{X}} h(x, \Phi\backslash\{x\})\,\Phi(\mathrm{d}x)\right] = \mathbb{E}\left[\int_{\mathbb{X}} h(x,\Phi)\,\rho(x|\Phi)\,\mathrm{d}x\right].$$

3 STEIN OPERATORS FOR POINT PROCESSES

At a high level, Stein's method involves identifying an operator $\mathcal{A}$ that satisfies *Stein's identity* [39]: a random variable Φ is distributed according to the probability measure μ if and only if $\mathbb{E}_\mu[\mathcal{A}_\mu h(\Phi)] = 0$ for all functions h in some class $\mathcal{H}$. When Φ is real-valued, $\mathcal{A}$ can be characterized through a simple differential operator

(the *Langevin Stein operator* [18, 31]), with Stein's identity following easily from integration-by-parts. When ϕ is discrete-valued, an alternative Stein operator using partial differences was provided in [43]. However, when ϕ is a point process—a random variable taking values in an infinite-dimensional space $\mathcal{N}_{\mathbb{X}}$—we will require a new set of tools, based on the *generator method* of [7].

We begin by reviewing the Stein operator for the Poisson process, and then propose a general Stein operator for arbitrary finite point processes. Our proposed Stein operator can be easily evaluated for point processes whose intensity functions contain intractable normalization constants, such as the Gibbs process.

3.1 Stein Operator for the Poisson Process

Stein's method for Poisson process approximation was pioneered by [8], using the generator method of [7]. For a Poisson process Φ_μ on $\mathbb{X}$ with mean measure μ, [8] considered an immigration-death process on $\mathbb{X}$ with immigration intensity μ and unit per-capita death rate. This process has stationary distribution Φ_μ, and infinitesimal generator $\mathcal{A}_\mu$ given by

$$(\mathcal{A}_\mu h)(\phi) = \int_{\mathbb{X}} [h(\phi + \delta_x) - h(\phi)]\, \mu(\mathrm{d}x) + \int_{\mathbb{X}} [h(\phi - \delta_x) - h(\phi)]\, \phi(\mathrm{d}x) \qquad (5)$$

for any configuration $\phi \in \mathcal{N}_{\mathbb{X}}$. Notably, the infinitesimal generator $\mathcal{A}_\mu$ *characterizes* the Poisson process Φ_μ, as demonstrated by the following result:

Theorem 3 (Stein identity for the Poisson process; Barbour and Brown, 1992; see also Decreusefond and Vasseur, 2018)**.** *Let $\mathcal{A}_\mu$ be the infinitesimal generator defined in Eq. (5). A point process Φ on $\mathbb{X}$ is a Poisson process with intensity measure μ if and only if for any measurable and bounded function $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$,*

$$\mathbb{E}\left[\mathcal{A}_\mu h(\Phi)\right] = 0. \qquad (6)$$

In the literature on Stein's method [39], an operator $\mathcal{A}$ that characterizes the distribution of Φ is called a *Stein operator*, and Eq. (6) a *Stein identity*.

Although [7] derived the expression of $\mathcal{A}$ using the generator method, Theorem 3 can actually be viewed as a direct consequence of the Mecke formula (Theorem 1). This hints at a possible generalization of the Stein operator for Poisson processes in Eq. (5) to general (finite) point processes, which we discuss next.

3.2 The Stein–Papangelou Operator

Stein's method for Poisson process approximation has been extensively studied since [8], yet few works have considered more general point processes such as Hawkes processes and Gibbs processes (with the exceptions of [14, 38]). Here, we propose a generalization of the Stein operator in Eq. (5) to general (finite) point processes on $\mathbb{X}$. Our key insight is the analogy between the Mecke formula (Theorem 1) for Poisson processes and the GNZ formula (Theorem 2) for general point processes.

We begin by providing an interpretation of the right-hand side of Eq. (5). From the complete randomness of the Poisson process, $\mu(\mathrm{d}x) = \lambda(x)\,\mathrm{d}x$ gives the conditional intensity of an event at location x given the rest of the Poisson process realization ϕ. Then, the first integral equals the expected change in the value of the function h if a new event were added to the point process realization. Similarly, the second term gives the average change in h if one of the events were *removed* from ϕ. For a point process model with interactions, the conditional intensity at location x will depend on the rest of the point process realization; indeed, this is exactly the Papangelou conditional intensity $\rho(x|\phi)$. Thus, it is natural to consider substituting the intensity function $\lambda(x)$ with the Papangelou conditional intensity $\rho(x|\phi)$. Somewhat surprisingly, we can show that the resulting expression still gives a valid Stein operator for the associated point process.

To simplify presentation, let us define the 'inclusion' and 'exclusion' functionals $\mathcal{D}_x^+, \mathcal{D}_x^-$ at a point $x \in \mathbb{X}$ as

$$(\mathcal{D}_x^+ h)(\phi) := h(\phi + \delta_x) - h(\phi);$$
$$(\mathcal{D}_x^- h)(\phi) := h(\phi) - h(\phi - \delta_x),$$

for any measurable and bounded function $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$ and (finite) point configuration $\phi \in \mathcal{N}_{\mathbb{X}}$. Using these notations, we have the following definition:

Definition 1 (Stein–Papangelou operator for finite point processes)**.** *Let $\rho : \mathbb{X} \times \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$ be the Papangelou conditional intensity of a finite point process on $\mathbb{X}$. Define the* Stein–Papangelou operator $\mathcal{A}_\rho$ *via*

$$(\mathcal{A}_\rho h)(\phi) = \int_{\mathbb{X}} (\mathcal{D}_x^+ h)(\phi)\rho(x|\phi)\,\mathrm{d}x - \int_{\mathbb{X}} (\mathcal{D}_x^- h)(\phi)\phi(\mathrm{d}x) = \int_{\mathbb{X}} [h(\phi + \delta_x) - h(\phi)]\,\rho(x|\phi)\,\mathrm{d}x + \sum_{x \in \phi} [h(\phi - \delta_x) - h(\phi)] \qquad (7)$$

for any function $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$ and configuration $\phi \in \mathcal{N}_{\mathbb{X}}$.

Notice that Eq. (7) reduces to Eq. (5) for a Poisson process, since its Papangelou conditional intensity equals its intensity function: $\rho(x|\phi)\,\mathrm{d}x = \lambda(x)\,\mathrm{d}x = \mu(\mathrm{d}x)$. A crucial advantage of Eq. (7) is that the Stein operator $\mathcal{A}_\rho$ now depends only on the Papangelou conditional intensity ρ of the point process, which is usually easy to obtain even when the point process likelihood itself is computationally intractable.

We conclude this section by showing that Eq. (7) indeed defines a valid Stein operator for general (finite) point processes—*i.e.*, that it satisfies a Stein identity.

Theorem 4 (Stein identity for finite point processes)**.** *Let Φ be a finite point process on $\mathbb{X}$ with Papangelou conditional intensity $\rho : \mathbb{X} \times \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$, and let $\mathcal{A}_\rho$ be the operator defined via Eq. (7). Then, we have*

$$\mathbb{E}\left[\mathcal{A}_\rho h(\Phi)\right] = 0 \tag{8}$$

for all measurable and bounded functions $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$.

Proof. To prove Eq. (8), it suffices to show that

$$\mathbb{E}\left[\int_{\mathbb{X}} (\mathcal{D}_x^+ h)(\Phi)\, \rho(x|\Phi)\, \mathrm{d}x\right] = \mathbb{E}\left[\sum_{x\in\Phi} (\mathcal{D}_x^- h)(\Phi)\right]$$

for any function $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$ and configuration $\phi \in \mathcal{N}_{\mathbb{X}}$. Notice that for any $x \in \phi$, $(D_x^- h)(\phi) = h(\phi) - h(\phi - \delta_x) = h(\phi - \delta_x + \delta_x) - h(\phi - \delta_x) = (D_x^+ h)(\phi - \delta_x)$. Thus, applying the GNZ formula (Theorem 2) with $h(x, \Phi) := (D_x^+ h)(\Phi)$ gives the desired result. □

A similar idea, but under a different context, has also been proposed in the probability literature [38].

4 STEIN DISCREPANCY AND GOODNESS-OF-FIT TESTING

Equipped with a proper Stein operator, we are now ready to define a notion of *discrepancy* between two point processes with different intensity measures.

4.1 (Kernelized) Stein Discrepancy

Following a central observation made by [18] under the context of continuous distributions with smooth densities, we note that since the Stein identity of Eq. (8) holds when the point process Φ has Papangelou conditional intensity ρ (denoted $\Phi \sim \rho$), one could consider the *maximum violation* of Eq. (8) when $\Phi \sim \eta \neq \rho$ by choosing test functions within a function class $\mathcal{F}$. This leads to the following definition:[1]

Definition 2 (Stein discrepancy for point processes)**.** *Let Φ be a finite point process on $\mathbb{X}$ with Papangelou conditional intensity $\rho : \mathbb{X} \times \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$, and let $\mathcal{A}_\rho$ be the Stein operator defined via Eq. (7). For a family $\mathcal{F}$ of functions $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$, define the* Stein discrepancy *between Papangelou conditional intensities η and ρ as*

$$\mathbb{D}_{\mathcal{F}}(\eta \,\|\, \rho) := \sup_{h\in\mathcal{F}} \mathbb{E}_{\Phi\sim\eta}\left[\mathcal{A}_\rho h(\Phi)\right]. \tag{9}$$

Clearly, $\mathbb{D}_{\mathcal{F}}(\eta \,\|\, \rho) = 0$ when $\eta \equiv \rho$. While in principle the Stein discrepancy can be defined with respect to any family of functions $\mathcal{F}$, in practice we need to choose a function space that is both rich enough to ensure that the resulting Stein discrepancy has sufficient discriminative power, yet also suitably tractable such that Eq. (9) can be efficiently computed.

[1] As Eq. (7) reduces to Eq. (5) for Poisson processes, we present all results using the Stein–Papangelou operator.

Toward this end, we follow [11, 30] and take $\mathcal{F}$ to be the unit-ball in a *reproducing kernel Hilbert space* (RKHS). Specifically, let $k : \mathcal{N}_{\mathbb{X}} \times \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$ be a positive definite (p.d.) kernel on the space of finite point configurations $\mathcal{N}_{\mathbb{X}}$ (Section 4.3 discusses various choices of k), and let $\mathcal{H}_k$ be its associated RKHS (consisting of functions $h : \mathcal{N}_{\mathbb{X}} \to \mathbb{R}$). We have the following definition:

Definition 3. *The* kernelized Stein discrepancy *(KSD) between finite point processes with Papangelou conditional intensities η and ρ is*

$$\mathbb{D}_{\mathcal{H}_k}(\eta \,\|\, \rho) := \sup_{h\in\mathcal{H}_k,\, \|h\|_{\mathcal{H}_k}\leq 1} \mathbb{E}_{\Phi\sim\eta}\left[\mathcal{A}_\rho h(\Phi)\right], \tag{10}$$

where $\mathcal{H}_k$ is the RKHS of a p.d. kernel $k(\cdot,\cdot)$ on $\mathcal{N}_{\mathbb{X}}$.

Using the reproducing property of $\mathcal{H}_k$, our next result shows that Eq. (10) can actually be evaluated in closed-form. This follows directly from [30]; due to space constraints, we defer its proof to the Appendix.

Theorem 5. *The squared-KSD can be expressed as*

$$\mathbb{D}^2_{\mathcal{H}_k}(\eta \,\|\, \rho) = \mathbb{E}_{\Phi,\Psi\sim\eta}\left[\kappa_\rho(\Phi, \Psi)\right], \tag{11}$$

where $\kappa_\rho(\phi,\psi) := \mathcal{A}_\mu^\psi \mathcal{A}_\mu^\phi k(\phi,\psi)$ is a kernel function on $\mathcal{N}_{\mathbb{X}}$ obtained by applying the Stein operator $\mathcal{A}$ twice on each argument of the reproducing kernel $k(\cdot,\cdot)$ of $\mathcal{H}_k$. Its expression is shown in Eq. (12) on the next page.

To evaluate $\kappa_\rho(\phi,\psi)$ for a pair of configurations (ϕ,ψ) using Eq. (12), we need to compute one double integral and two single integrals over the domain $\mathbb{X} \subseteq \mathbb{R}^d$ as well as summations over the points in both ϕ and ψ.[2] Evaluating these integrals could require numerical integration techniques, but observe that we have reduced the problem of evaluating a normalization constant for a distribution on $\mathcal{N}_{\mathbb{X}}$ (an infinite-dimensional integral) to a finite-dimensional one. For most applications, d is small ($d = 1$ for temporal point processes and typically $d = 2$ for spatial point processes), and standard numerical quadrature methods should suffice.

While Theorem 4 implies that $\mathbb{D}_{\mathcal{H}_k}(\eta \,\|\, \rho) = 0$ for $\eta \equiv \rho$, we note that for non-Poisson processes, $\mathbb{D}_{\mathcal{H}_k}(\eta \,\|\, \rho) = 0$ may not be sufficient to guarantee that $\eta \equiv \rho$. This is due to the fact that while the Mecke formula fully characterizes a Poisson process, the GNZ formula (which was crucial in establishing our Stein operator) provides only a necessary condition for a point process to have a specific Papangelou conditional intensity.

4.2 Goodness-of-Fit Testing via KSD

We now apply the kernelized Stein discrepancy measure of Definition 3 to construct a goodness-of-fit test for general (finite) point processes, including those with computationally intractable intensity functions.

[2] For concreteness, we provide example Python code for implementing Eq. (12) in the appendix.

$$\kappa_\rho(\phi,\psi) = \int_{\mathbb{X}}\int_{\mathbb{X}} \Big[k(\phi+\delta_u,\psi+\delta_v) - k(\phi,\psi+\delta_v) - k(\phi+\delta_u,\psi) + k(\phi,\psi)\Big]\rho(u|\phi)\,\rho(v|\psi)\,\mathrm{d}u\,\mathrm{d}v$$
$$+ \int_{\mathbb{X}}\Bigg[\sum_{x\in\phi}\big[k(\phi-\delta_x,\psi+\delta_v) - k(\phi-\delta_x,\psi)\big] - |\phi|\cdot\big[k(\phi,\psi+\delta_v) - k(\phi,\psi)\big]\Bigg]\rho(v|\psi)\,\mathrm{d}v$$
$$+ \int_{\mathbb{X}}\Bigg[\sum_{y\in\psi}\big[k(\phi+\delta_u,\psi-\delta_y) - k(\phi,\psi-\delta_y)\big] - |\psi|\cdot\big[k(\phi+\delta_u,\psi) - k(\phi,\psi)\big]\Bigg]\rho(u|\phi)\,\mathrm{d}u$$
$$+ \Bigg[\sum_{x\in\phi}\sum_{y\in\psi} k(\phi-\delta_x,\psi-\delta_y) - |\phi|\cdot\sum_{y\in\psi}k(\phi,\psi-\delta_y) - |\psi|\cdot\sum_{x\in\phi}k(\phi-\delta_x,\psi) + |\phi|\cdot|\psi|\cdot k(\phi,\psi)\Bigg]. \tag{12}$$

Suppose we observe samples $\{\mathcal{X}_i\}_{i=1}^m$ from a point process with *unknown* Papangelou conditional intensity η, where each $\mathcal{X}_i := \{x_k\}_{k=1}^{n_i} \subseteq \mathbb{X}$ is a collection of points in $\mathbb{X}$ (note that the cardinalities n_i would vary). Given a statistical model which posits that the observed samples arose from a point process with (known) Papangelou conditional intensity ρ, we would like to quantify the 'goodness-of-fit' of the model ρ to the data $\{\mathcal{X}_i\}_{i=1}^m$. (Often we have only a single realization $\mathcal{X}$ of a point process, rather than many realizations of the process. In this case, it suffices to partition the space into a collection of blocks with equal volume, and treat the restriction of $\mathcal{X}$ to block i as the i-th realization $\mathcal{X}_i$.)

Formally, we perform the hypothesis test $H_0 : \rho = \eta$ vs. $H_1 : \rho \neq \eta$ using kernelized Stein discrepancy (KSD). For convenience, we omit the dependency on $\mathcal{H}_k$ and denote $\mathbb{S}(\eta \,\|\, \rho) := \mathbb{D}^2_{\mathcal{H}_k}(\eta \,\|\, \rho)$. Given observed samples $\{\mathcal{X}_i\}_{i=1}^m$ from a point process with (unknown) Papangelou conditional intensity η, by Eq. (12) we can estimate $\mathbb{S}(\eta \,\|\, \rho)$ via a U-statistic [22] which gives a minimum-variance unbiased estimator:

$$\widehat{\mathbb{S}}(\eta \,\|\, \rho) = \frac{1}{m(m-1)}\sum_{i=1}^m\sum_{j\neq i}^m \kappa_\rho(\mathcal{X}_i,\mathcal{X}_j), \tag{13}$$

where the expression for $\kappa_\rho(\phi,\psi)$ is shown in Eq. (12).[3] By standard asymptotic results on U-statistics (see Theorem 7 in the Appendix for details), the U-statistic $\widehat{\mathbb{S}}(\eta \,\|\, \rho)$ is asymptotically normally distributed under the alternative hypothesis $H_1 : \rho \neq \eta$, but becomes degenerate under the null hypothesis $H_0 : \rho = \eta$.

Since the asymptotic distribution of $\widehat{\mathbb{S}}(\eta \,\|\, \rho)$ under H_0 is not available in closed-form, we follow [30] and adopt the generalized bootstrap method for degenerate U-statistics [1, 23] to approximate the distribution. To obtain a bootstrap sample, we draw random multinomial weights $w_1,\ldots,w_m \sim \mathrm{Mult}(m; 1/m,\ldots,1/m)$, set $\widetilde{w}_i = (w_i - 1)/m$, and compute

$$\widehat{\mathbb{S}}^*(\eta \,\|\, \rho) = \sum_{i=1}^m\sum_{j\neq i}^m \widetilde{w}_i\widetilde{w}_j\kappa_p(\mathcal{X}_i,\mathcal{X}_j). \tag{14}$$

Upon repeating this procedure $\widetilde{m}$ times, we calculate the critical value of the test by taking the $(1-\alpha)$-th quantile of the bootstrapped statistics $\{\widehat{\mathbb{S}}^*_b\}_{b=1}^{\widetilde{m}}$. We reject the null hypothesis H_0 if $\widehat{\mathbb{S}}(\eta \,\|\, \rho) > \gamma_{1-\alpha}$. The overall goodness-of-fit testing procedure is summarized in Algorithm 1 of the Appendix.

As noted at the end of Section 4.1, $\mathbb{S}(\eta \,\|\, \rho) = 0$ may be insufficient to guarantee that $\eta \equiv \rho$. Thus, the KSD goodness-of-fit test may fail to reject H_0 even when the observed data arose from a point process with a Papangelou conditional intensity different from that specified by the null model, yielding Type-II errors. To the best of our knowledge, no necessary-and-sufficient condition for characterizing general (non-Poisson) point processes is known in the literature, and existing approaches [5, 12] also only guarantee Type-I error control, and suffer from the same loss of power.

Computational complexity. Calculating the test statistic $\widehat{\mathbb{S}}(\eta \,\|\, \rho)$ in Eq. (13) requires $\mathcal{O}(m^2)$ evaluations of $\kappa_\rho(\mathcal{X}_i,\mathcal{X}_j)$, where m is the number of data samples. Once the kernel matrix $[\kappa_\rho(\mathcal{X}_i,\mathcal{X}_j)]_{i,j=1}^m$ is cached, the bootstrapping procedure takes $\mathcal{O}(\widetilde{m}\cdot m^2)$ time, where $\widetilde{m}$ is the number of bootstrap samples.

To be more precise, recall that a sample $\mathcal{X}_i$ consists of $n_i := |\mathcal{X}_i|$ points in $\mathbb{X}$. Evaluating $\kappa_\rho(\mathcal{X}_i,\mathcal{X}_j)$ for each pair of samples $(\mathcal{X}_i,\mathcal{X}_j)$ using Eq. (12) requires numerical integration. Assuming q quadrature points per dimension, the time complexity for a single evaluation of $\kappa_\rho(\mathcal{X}_i,\mathcal{X}_j)$ is given by $\mathcal{O}((q^{2d} + 2\,q^d\bar{n} + |\bar{n}|^2)\cdot t_k + 2\,q^d t_\rho) = \mathcal{O}((q^d + |\bar{n}|)^2\cdot\bar{n}^2)$. Here, $\bar{n}$ is the average cardinality of the observed samples, t_k is the time required to evaluate the kernel function $k(\cdot,\cdot)$ for a pair of samples with size $\bar{n}$, and t_ρ is the time needed for a single evaluation of the Papangelou conditonal intensity (typically, $t_k, t_\rho = \mathcal{O}(\bar{n}^2)$ in the worst case). Putting everything together, the overall time complexity of Algorithm 1 is $\mathcal{O}(m^2\cdot(q^d + |\bar{n}|)^2\cdot\bar{n}^2 + \widetilde{m}\cdot m^2)$. Note that when d is large, one could apply Monte Carlo integration in lieu of numerical quadrature to avoid the curse of dimensionality, and the q^d term would be replaced by c, the number of Monte Carlo points.

4.3 Kernel Functions for Point Processes

Our theoretical development so far hold generally for any positive definite kernel on the space of finite counting measures $\mathcal{N}_{\mathbb{X}}$. There has been work on *set kernels* or *multi-instance kernels* [17], where the similarity of

[3] When evaluating Eq. (12), recall that $\phi+\delta_x$ and $\phi-\delta_x$ are equivalent to $\phi\cup\{x\}$ and $\phi\backslash\{x\}$, respectively.

two sets is measured by their average pairwise point similarities, as well as kernels which make parametric assumptions on the distributions of the points [3, 10, 25].

We argue that a proper kernel function $k(\mathcal{X}, \mathcal{Y})$ between two point configurations $\mathcal{X}, \mathcal{Y} \subseteq \mathbb{X}$ should capture their similarities with regard to their *extrinsic* and/or *intrinsic* characteristics as needed. Extrinsic characteristics refer to inhomogeneities in intensity, resulting in different expected counts for different point processes in the same parts of the $\mathbb{X}$-space. Intrinsic characteristics pertain to point interactions within a point process—*i.e.*, whether the points exhibit attraction or repulsiveness. Any prior knowledge regarding the nature of deviations from the null model could accordingly be incorporated into the kernel function. One simple approach would be to map each point configuration into a feature-vector, with components including *e.g.*, the number of points in different regions of the space, the number of points within some distance r of each other, the average distance from a point to their k-th nearest neighbor, etc.

As a flexible nonparametric alternative that takes both extrinsic and intrinsic features into consideration, we propose to use the *maximum mean discrepancy* (MMD) [20] between two counting measures to define a p.d. kernel. Specifically, we have the following:

Proposition 6. *Given a positive definite kernel $k_{\mathbb{X}}(\cdot,\cdot)$ on the ground space $\mathbb{X}$, define the $\mathcal{M}$-*kernel*:*

$$k_{\mathcal{M}}(\phi, \psi) := \exp\{-\widehat{d^2}(\phi, \psi)\}, \tag{15}$$

where $\widehat{d^2}(\phi, \psi)$ denotes the V-statistic estimate of the squared-MMD between configurations $\phi, \psi \in \mathcal{N}_{\mathbb{X}}$:

$$\widehat{d^2}(\phi, \psi) := \frac{1}{|\phi|^2} \sum_{x\in\phi}\sum_{x'\in\phi} k_{\mathbb{X}}(x, x') + \frac{1}{|\psi|^2} \sum_{y\in\psi}\sum_{y'\in\psi} k_{\mathbb{X}}(y, y') - \frac{2}{|\phi| \cdot |\psi|} \sum_{x\in\phi}\sum_{y\in\psi} k_{\mathbb{X}}(x, y). \tag{16}$$

Then, $k_{\mathcal{M}}(\cdot,\cdot)$ is a positive definite kernel on $\mathcal{N}_{\mathbb{X}}$.[4]

Proof. By [37], to prove that $k_{\mathcal{M}}(\cdot,\cdot)$ is a p.d. kernel, it suffices to show that $\widehat{d^2}$ is a conditionally negative definite function; see the Appendix for details. □

5 RELATED WORK

Classical diagnostic measures for point processes have largely been restricted to temporal point processes. For spatial point processes, traditional approaches [15] primarily rely on heuristic summary statistics (*e.g.*, the 'K-function' of [35]) to test for specific properties of the data, such as complete randomness or clustering.

Related to our work, and also motivated by the GNZ formula, [4] defined the *h-weighted residual measure* for a parametric model $\hat{\rho}$ fitted to an observed configuration ϕ on a bounded domain $B \subseteq \mathbb{X}$:

$$\gamma(B, h, \hat{\rho}) := \sum_{x\in\phi\cap B} h(x, \phi\backslash\{x\}) - \int_B h(u, \phi)\hat{\rho}(u|\phi)\,\mathrm{d}u,$$

where h is a user-specified weight function. Informally, our proposed KSD goodness-of-fit test statistic could be viewed as a *kernelization* of the h-weighted residuals, where we take the supremum over all test functions h in an RKHS. In doing so, we obtain a parsimonious and more powerful test capturing various aspects of the model intensity that would have been difficult for any specific h to fully cover. In addition, the KSD test allows users the flexibility to emphasize specific aspects of interest through the design of the kernel function.

6 EMPIRICAL EVALUATION

We apply the kernelized Stein discrepancy (KSD) test to the point process models described in Section 2. We also compare with a test based on the maximum mean discrepancy (MMD) [20], which draws samples from the null model, and performs a two-sample test between the drawn samples and the observed data. Note that here we are computing the MMD test statistic between two *collections* of *point configurations* in $\mathcal{N}_{\mathbb{X}}$, as opposed to Eq. (16) which estimates the MMD between two sets of *points* in $\mathbb{X}$. Given samples $\{\mathcal{X}_i\}_{i=1}^m, \{\mathcal{Y}_j\}_{j=1}^m$ from two point processes ρ and η, we compute the U-statistic estimate of $\mathrm{MMD}^2(\rho, \eta)$:

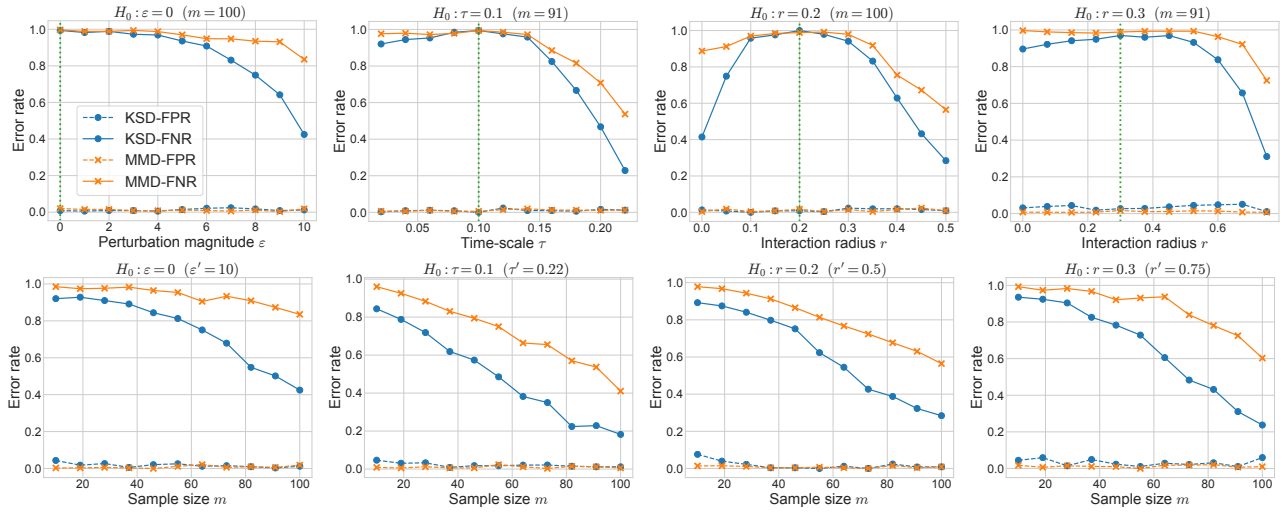
$$\widehat{\mathrm{MMD}^2}(\rho, \eta) := \frac{1}{m(m-1)} \sum_{i=1}^{m}\sum_{j\neq i}^{m} k(\mathcal{X}_i, \mathcal{X}_j) + \frac{1}{m(m-1)} \sum_{i=1}^{m}\sum_{j\neq i}^{m} k(\mathcal{Y}_i, \mathcal{Y}_j) - \frac{2}{m^2} \sum_{i=1}^{m}\sum_{j=1}^{n} k(\mathcal{X}_i, \mathcal{Y}_j).$$

The critical value of the MMD test is calculated by bootstrapping on the aggregated data.

Setup. We adopt a common experiment setup used in [30, 43]. Denote the Papangelou intensities for the null and the alternative point process models by ρ and η, respectively. For KSD, we draw m *i.i.d.* samples (point configurations) from η; for MMD, we draw m samples from η and another m samples from ρ. For the kernel function $k(\cdot,\cdot)$ on $\mathcal{N}_{\mathbb{X}}$ (used in both KSD and MMD), we utilize the $\mathcal{M}$-kernel defined via Eqs. (15) and (16), where the *ground kernel* $k_{\mathbb{X}}(\cdot,\cdot)$ in Eq. (16) is set to a Gaussian RBF kernel. To ensure fair comparison, we set the bandwidth of the RBF kernel for both KSD and MMD to the median pairwise distance [20] of the aggregated points in the samples drawn from η. We use $\widetilde{m} = 10,000$ bootstrap samples for both methods.

For each model, we choose a single parameter, fix its value for the null model ρ, and draw samples for η under different values of that parameter. For each value of the chosen parameter and sample size m, we conduct 500 independent trials. In each trial, we flip a fair coin to decide whether the alternative model η will be set to the same as ρ or with a different value

[4] If either ϕ or ψ is an empty configuration, we define $k(\phi, \psi) = 1$ if both are empty and $k(\phi, \psi) = 0$ otherwise.



(a) Poisson process (b) Hawkes process (c) Strauss process ($d = 1$) (d) Strauss process ($d = 2$)

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

of the chosen parameter (in the former case, the null hypothesis $H_0 : \rho = \eta$ should not be rejected, and in the latter case it should be). We conduct the hypothesis test $H_0 : \rho = \eta$ vs. $H_1 : \rho \neq \eta$ under significance level $\alpha = 0.01$, and evaluate the performance of KSD and MMD in terms of their false-positive rate (FPR; Type-I error) and false-negative rate (FNR; Type-II error).

Poisson process. We consider a Poisson process on the unit-square $[0, 1]^2$ with intensity function $\lambda(x) = \gamma + \varepsilon \sin(2\pi(x + y))$, where γ is a base-rate, and ε represents the perturbation magnitude. We fix $\gamma = 50$ throughout, vary the perturbation magnitude ε, and test the hypotheses $H_0 : \varepsilon = 0$ vs. $H_1 : \varepsilon \neq 0$.

Hawkes process. We consider a Hawkes process on $[0, 1]$ with intensity function given in Eq. (1) and set $g(t) = \beta e^{-t/\tau}$. We fix $\gamma = 20$ and $\beta = 2$ throughout, vary the time-scale parameter τ, and test the hypotheses $H_0 : \tau = 0.1$ vs. $H_1 : \tau \neq 0.1$. To simulate from a Hawkes process, we employ Ogata's thinning algorithm [32].

Strauss process. We consider Strauss processes on $[0, 1]^d$ ($d = 1$ or 2) with conditional intensity given in Eq. (4). We fix $\beta = 20$ and $\gamma = 0.8$ ($d = 1$) or 0.9 ($d = 2$), vary the interaction radius r and test the hypotheses $H_0 : r = r_0$ vs. $H_1 : r \neq r_0$ with $r_0 = 0.2$ or 0.3. To simulate from a 1-D Strauss process, we apply rejection sampling to realizations of a Poisson process with intensity β. To simulate from a 2-D Strauss process, we use the MCMC sampler provided in the R package `spatstat` [5, 6].

Results. In Figure 1, the top row plots the testing error rate vs. different values of the parameter we chose to vary for η, under a given sample size. The bottom row plots the error rate vs. sample size for a specific value of the chosen parameter. We observe that both methods generally maintain a false-positive rate (Type-I error) around the significance level, while KSD consistently achieves lower false-negative rate (Type-II error) than MMD across different parameter settings as well as sample sizes.[5] This indicates that KSD, by utilizing information from the Papangelou conditional intensity ρ of the null model, gives rise to a more powerful test. We emphasize that the MMD two-sample test requires generating exact samples from the null model, which could be computationally costly or intractable. Finally, we note that the statistical power of both methods could be improved by using more sophisticated constructions of kernel functions on the space of counting measures, which we leave for future work.

7 CONCLUSION

We have introduced a general Stein operator based on the Papangelou conditional intensity for point processes which can be evaluated even when the intensity function contains an intractable normalization constant. Using the proposed Stein operator, we have developed a kernelized Stein discrepancy test for measuring the goodness-of-fit of a point process model. We have applied the proposed test to several point process models, and showed that it outperforms a two-sample test based on the maximum mean discrepancy, which assumes the availability of exact samples from the null model.

Acknowledgements. We thank Qiang Liu and the anonymous reviewers for their helpful feedback. This research is supported by NSF under contract numbers IIS-1618690, IIS-1546488, CCF-0939370, IIS-1816499 and DMS-1812197.

[5] In Figure 1d, the Type-I error for KSD appears slightly higher than the nominal significance level 0.01. We found that this was due to numerical quadrature error involved in evaluating Eq. (12) under limited computational budget (since the double-integral over $\mathbb{X}$ is now four-dimensional). This issue could be alleviated using Monte Carlo integration techniques, which shall be investigated in future work.

References

[1] Miguel A. Arcones and Evarist Gine. On the bootstrap of U and V statistics. *The Annals of Statistics*, 20(2):655–674, 1992.

[2] G. J. Babu and E. D. Feigelson. *Astrostatistics*. Chapman and Hall, 1996.

[3] Francis R. Bach. Graph kernels between point clouds. In *Proceedings of the 25th International Conference on Machine Learning*, pages 25–32, 2008.

[4] A. Baddeley, R. Turner, J. Møller, and M. Hazelton. Residual analysis for spatial point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):617–666, 2005.

[5] Adrian Baddeley and Rolf Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12(6):1–42, 2005.

[6] Adrian Baddeley, Ege Rubak, and Rolf Turner. *Spatial Point Patterns: Methodology and Applications with R*. Chapman and Hall/CRC Press, London, 2015.

[7] A. D. Barbour. Stein's method and Poisson process convergence. *Journal of Applied Probability*, 25:175–184, 1988.

[8] A.D. Barbour and T.C. Brown. Stein's method and point process approximation. *Stochastic Processes and their Applications*, 43(1):9 – 31, 1992.

[9] Emery N Brown, Riccardo Barbieri, Valérie Ventura, Robert E Kass, and Loren M Frank. The time-rescaling theorem and its application to neural spike train data analysis. *Neural computation*, 14(2):325–346, 2002.

[10] Mathieu Carrière, Marco Cuturi, and Steve Oudot. Sliced Wasserstein kernel for persistence diagrams. In *Proceedings of the 34th International Conference on Machine Learning*, pages 664–673, 2017.

[11] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2606–2615, 2016.

[12] Jean-François Coeurjolly and Frédéric Lavancier. Residuals and goodness-of-fit tests for stationary marked Gibbs point processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):247–276, 2013.

[13] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes (Vol. II)*. Springer, second edition, 2008.

[14] Laurent Decreusefond and Aurélien Vasseur. Stein's method and Papangelou intensity for Poisson or Cox process approximation. *arXiv:1807.02453*, 2018.

[15] Peter J. Diggle. *Statistical analysis of spatial point patterns*. Edward Arnold, 2003.

[16] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1555–1564, 2016.

[17] Thomas Gärtner, Peter A. Flach, Adam Kowalczyk, and Alex J. Smola. Multi-instance kernels. In *Proceedings of the International Conference on Machine Learning*, pages 179–186, 2002.

[18] Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems 28*, pages 226–234, 2015.

[19] Jackson Gorham and Lester W. Mackey. Measuring sample quality with kernels. In *Proceedings of The 34th International Conference on Machine Learning*, pages 1292–1301, 2017.

[20] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

[21] Alan G. Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58 (1):83–90, 1971.

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

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

[24] Wittawat Jitkrittum, Wenkai Xu, Zoltan Szabo, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *Advances in Neural Information Processing Systems 30*, pages 261–270. 2017.

[25] Risi Kondor and Tony Jebara. A kernel between sets of vectors. In *Proceedings of the 20th International Conference on Machine Learning*, pages 361–368, 2003.

[26] Günter Last and Mathew Penrose. *Lectures on the Poisson Process*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2017.

[27] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, third edition, 2005.

[28] Scott Linderman and Ryan Adams. Discovering latent network structure in point process data. In *Proceedings of the 31st International Conference on Machine Learning*, pages 1413–1421, 2014.

[29] Scott Linderman, Christopher Stock, and Ryan Adams. A framework for studying synaptic plasticity with neural spike train data. In *Advances in Neural Information Processing Systems 27*, pages 2330–2338. 2014.

[30] Qiang Liu, Jason D. Lee, and Michael I. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016.

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

[32] Y. Ogata. On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory*, 27(1):23–31, 1981.

[33] Yosihiko Ogata. Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27, 1988.

[34] Alex Reinhart. A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318, 2018.

[35] B. D. Ripley. The second-order analysis of stationary point processes. *Journal of Applied Probability*, 13(2):255–266, 1976.

[36] B. D. Ripley and F. P. Kelly. Markov point processes. *Journal of the London Mathematical Society*, s2-15(1):188–192, 1977.

[37] I. J. Schoenberg. Metric spaces and positive definite functions. *Transactions of the American Mathematical Society*, 44(3):522–536, 1938.

[38] Dominic Schuhmacher and Kaspar Stucki. Gibbs point process approximation: Total variation bounds using Stein's method. *Ann. Probab.*, 42 (5):1911–1951, 09 2014.

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

[40] David Strauss and Michael Ikeda. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212, 1990.

[41] David J. Strauss. A model for clustering. *Biometrika*, 62(2):467–475, 1975.

[42] Shuai Xiao, Mehrdad Farajtabar, Xiaojing Ye, Junchi Yan, Le Song, and Hongyuan Zha. Wasserstein learning of deep generative point process models. In *Advances in Neural Information Processing Systems 30*, pages 3247–3257. 2017.

[43] Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5561–5570, 2018.

[44] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan R Salakhutdinov, and Alexander J Smola. Deep sets. In *Advances in Neural Information Processing Systems 30*, pages 3391–3401. 2017.