Statistical Science
2023, Vol. 38, No. 1, 120–139
https://doi.org/10.1214/22-STS863

Stein's Method Meets Computational Statistics: A Review of Some Recent Developments

Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E. Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, Lester Mackey, Chris J. Oates, Gesine Reinert and Yvik Swan

Abstract. Stein's method compares probability distributions through the study of a class of linear operators called Stein operators. While mainly studied in probability and used to underpin theoretical statistics, Stein's method has led to significant advances in computational statistics in recent years. The goal of this survey is to bring together some of these recent developments, and in doing so, to stimulate further research into the successful field of Stein's method and statistics. The topics we discuss include tools to benchmark and compare sampling methods such as approximate Markov chain Monte Carlo, deterministic alternatives to sampling methods, control variate techniques, parameter estimation and goodness-of-fit testing.

Key words and phrases: Stein's method, sample quality, approximate Markov chain Monte Carlo, variational inference, control variates, goodness-of-fit testing, maximum likelihood estimator, likelihood ratio, prior sensitivity.

1. INTRODUCTION

Stein's method was introduced by Charles Stein in the early 1970s [150] for distributional comparisons to the normal distribution. At the foundation of Stein's method lies a characterizing equation for the normal distribution. This equation is also a cornerstone in Stein's unbiased estimator of risk [153] and James–Stein shrinkage estimators [89, 149]; see [54] for a joined-up view. The latter paper also exploited these connections with Stein's method to propose and analyze new estimators in a non-Gaussian setting. Here, we concentrate on Stein's method for distributional comparisons.

Originally developed for normal approximation, the method was extended first to Poisson approximation by

Andreas Anastasiou is a Lecturer, Department of Mathematics and Statistics, University of Cyprus, P.O. Box: 20537, 1678 Nicosia, Cyprus (e-mail: anastasiou.andreas@ucy.ac.cy). Alessandro Barp is a Research Associate, University of Cambridge, Engineering Dept, Trumpington St, Cambridge CB2 1PZ, UK (e-mail: ab2286@cam.ac.uk). François-Xavier Briol is a Lecturer, University College London, 1-19 Torrington Place, London WC1E 7HB, UK (e-mail: f.briol@ucl.ac.uk). Bruno Ebner is a Lecturer, Institute of Stochastics, Karlsruhe Institute of Technology (KIT), Englerstr. 2, 76128 Karlsruhe, Germany (e-mail: Bruno.Ebner@kit.edu). Robert E. Gaunt is a Lecturer, The University of Manchester, Oxford Road, Manchester M13 9PL, UK (e-mail: robert.gaunt@manchester.ac.uk). Fatemeh Ghaderinezhad is a Data Analyst, amfori, The Gradient Building, Avenue de Tervueren 270, 1150 Brussels, Belgium (e-mail: Fatemeh.Ghaderinezhad@amfori.org). Jackson Gorham is a Data Scientist, Whisper.ai, Inc., USA (e-mail: jacksongorham@gmail.com). Arthur Gretton is Professor with the Gatsby Computational Neuroscience Unit, University College London, Sainsbury Wellcome Centre, 25 Howland Street, London W1T 4JG, UK (e-mail: arthur.gretton@gmail.com). Christophe Ley is Associate Professor, University of Luxembourg, Maison du Nombre, 6 Avenue de la Fonte, L-4364 Esch-sur-Alzette, Luxembourg (e-mail: christophe.ley@uni.lu). Qiang Liu is Assistant Professor, The University of Texas at Austin, Austin, Texas 78712, USA (e-mail: lqiang@utexas.edu). Lester Mackey is Principal Researcher, Microsoft Research New England, 1 Memorial Drive, Cambridge, Massachusetts 02142, USA (e-mail: lmackey@microsoft.com). Chris J. Oates is Professor, Newcastle University, UK (e-mail: Chris.Oates@newcastle.ac.uk). Gesine Reinert is Professor, University of Oxford, Department of Statistics, 24-29 St Giles', Oxford OX1 3LB, UK (e-mail: gesine.reinert@keble.ox.ac.uk). Yvik Swan is Professor, Université Libre de Bruxelles, Department of Mathematics—CP 210, Boulevard du Triomphe, 1050 Brussels, Belgium (e-mail: Yvik.Swan@ulb.be).

[36], then by a growing community to a growing collection of approximation problems including beta, binomial, gamma, Kummer-U, multinomial, variance-gamma, Wishart and many more. Stein's method has proved powerful in particular for deriving explicit bounds on distributional distances even when the underlying random elements are structures with dependence. Moreover, it thrives when the target distribution is known only up to a normalizing constant. Comprehensive introductions to the theory and its applications are available in the monographs [10, 16, 37, 124, 151]. We also refer to the lecture notes of [47] and the surveys of [15, 33, 98, 138]. The websites https://sites.google.com/site/malliavinstein and https://sites.google.com/site/steinsmethod provide regularly updated lists of references.

Over the past few decades, Stein's method has had substantial interactions with other mathematical fields, such as Malliavin calculus, information theory, functional analysis, dynamical systems and stochastic geometry. Some examples of applications of Stein's method in theoretical statistics are as follows. Stein et al. [152] employed the method for the analysis of sample quality in simulations, [85] developed a bootstrap method for network data which is analyzed via empirical processes, [141] obtained a Berry–Esseen bound for Student's t-statistic. Applications to self-normalized limit theorems and false discovery rates in simultaneous tests are surveyed in [142]. In [143], an overview on the use of randomized concentration inequalities in Stein's method for nonlinear statistics is provided. Lippert, Huang and Waterman [102] and [135] utilized the method to prove that there were flaws in then commonly used statistics for alignment-free sequence comparison, and subsequently introduced two new sequence comparison statistics, which avoid these flaws. This list is by no means exhaustive, but has the goal to give the reader a first taste of the versatile usage of Stein's method in statistics.

Starting around 2015, these early and ongoing successes of Stein's method in theoretical statistics have attracted the attention of researchers from computational statistics and machine learning. Indeed, the fact that target distributions only need to be known up to a normalizing constant for Stein's method to apply has sparked considerable interest in these domains. Here, ingredients from Stein's method such as so-called *Stein discrepancies* have been used to develop new methodological procedures based on Stein operators. The aim of this paper is to cover various (clearly not all) developments that took place in computational statistics and machine learning since around 2015; the choice of topics is biased by the research interests of the contributors. Related developments in applications of Stein's method in theoretical statistics are also included. By this survey, we wish to bring Stein's method and its different ingredients to the attention of the broad statistical community in order to further foster this fertile research domain.

This paper starts with a succinct introductory section on Stein's method (Section 2), followed by Section 3, which discusses the practical issue of computing Stein discrepancies. Section 4 presents various new statistical and machine learning procedures for assessing sample quality as well as constructing sample approximations and improving Monte Carlo integration, which are obtained by means of Stein method ingredients. Section 5 details new developments for testing goodness-of-fit, which are based on Stein's method, and discusses novel insights into existing inferential procedures such as the quality of asymptotic approximation of estimators and test statistics as well as the impact of the prior choice in Bayesian statistics. Section 6 then provides some summarizing conclusions.

2. THE BASIC INGREDIENTS OF STEIN'S METHOD

Stein's method provides a collection of tools permitting to quantify the dissimilarity between probability distributions. The method has many components, not all of which are pertinent to the present survey. The purpose of this introductory section is to provide a succinct overview of the basic ingredients, which shall be of use in the rest of the paper.

First, we fix some notation. The distribution of a random quantity X is denoted by $\mathcal{L}(X)$. Expectations with respect to a probability distribution Q are denoted by $\mathbb{E}_{X\sim Q}$; sometimes the subscript is omitted when the context is clear. The space $L^p(Q)$ denotes the set of functions such that $\mathbb{E}_{X\sim Q}[|f^p(X)|]$ is finite.

The function $\mathbb{I}_A(x)$ is the indicator function of $x \in A$, taking the value 1 if $x \in A$ and 0 otherwise. For $\mathbb{R}^d$-valued functions f and g, the notation $\langle f, g\rangle$ denotes the inner product; if f and g are matrix-valued, it denotes the Hilbert–Schmidt inner product. The notation $C^k(\mathbb{R}^d)$ denotes functions in $\mathbb{R}^d$ that are k times continuously differentiable. The norm $|\cdot|$ is the absolute value, $\|\cdot\|_2$ the Euclidean norm and $\|\cdot\|_\infty$ denotes the supremum norm. The operator ∇ denotes the gradient operator; the gradient of a smooth function $v : \mathbb{R}^d \to \mathbb{R}$ is the vector valued function ∇v with entries $(\nabla v)_i = \partial_i v$, $i = 1, \ldots, d$, by convention viewed as column vector. For a d-vector-valued function $\mathbf{v} : \mathbb{R}^d \to \mathbb{R}^d$ with components v_j, $j = 1, \ldots, d$, the divergence is $\operatorname{div}(v) = \nabla^\intercal v = \sum_{i=1}^d \partial_i v_i(x)$. For a vector or a matrix, the superscript $\intercal$ stands for the transpose; this also applies for vector- or matrix-valued operators. Finally, by convention, $0/0 = 0$.

2.1 Stein Operators, Stein Discrepancies and Stein Equations

The starting point of Stein's method for a target probability distribution P on some set $\mathcal{X}$ consists in identifying a linear operator $\mathcal{T}$ acting on a set $\mathcal{G}(\mathcal{T})$ of functions on

$\mathcal{X}$ such that, for any other probability measure Q on $\mathcal{X}$, it holds that

$$Q = P \quad \text{iff} \quad \mathbb{E}_{X\sim Q}\big[(\mathcal{T}g)(X)\big] = 0 \quad \forall g \in \mathcal{G}(\mathcal{T}). \tag{1}$$

Such an operator $\mathcal{T}$ is called a *Stein operator*, the collection $\mathcal{G}(\mathcal{T})$ of functions for which $\mathbb{E}_{X\sim P}[(\mathcal{T}g)(X)] = 0$ is called a *Stein class*, and equivalence (1) is called a *Stein characterization*. In many cases, the characterizing nature of the operator is superfluous, and we only need to require that a *Stein identity* for P is satisfied, namely that $\mathbb{E}_{X\sim P}[(\mathcal{T}g)(X)] = 0$ for all $g \in \mathcal{G}(\mathcal{T})$. Through a Stein identity, we only have a guarantee that the expectations taken against P vanish, but they could also be zero when taken against some $Q \neq P$.

We will discuss the topic of choosing Stein operators in Section 2.2. At this stage, let us suppose that we are given a characterizing Stein operator $\mathcal{T}$ with Stein class $\mathcal{G}(\mathcal{T})$. Then, for any *Stein set* $\mathcal{G} \subset \mathcal{G}(\mathcal{T})$, one may define a dissimilarity measure as

$$\mathcal{S}(Q, \mathcal{T}, \mathcal{G}) = \sup_{g\in\mathcal{G}} \big\|\mathbb{E}_{X\sim Q}\big[(\mathcal{T}g)(X)\big]\big\|^* \tag{2}$$

for some appropriate norm $\|\cdot\|^*$. By construction, if $\mathcal{S}(Q, \mathcal{T}, \mathcal{G}) \neq 0$, then $Q \neq P$ and, if $\mathcal{G}$ is sufficiently large, then $\mathcal{S}(Q, \mathcal{T}, \mathcal{G}) = 0$ also implies $Q = P$. Gorham and Mackey [69] call the quantity (2) a *Stein discrepancy* (in contrast to the use of the term in [96]). If the Stein operator $\mathcal{T}$ and the Stein set $\mathcal{G} \subset \mathcal{G}(\mathcal{T})$ are well chosen, the Stein discrepancy $\mathcal{S}(Q, \mathcal{T}, \mathcal{G})$ ought to capture some aspect of the dissimilarity between P and Q. Part of the magic of Stein's method lies in the fact that there are numerous combinations of target distribution P and approximating distribution Q for which one can identify operators $\mathcal{T}$ and sets $\mathcal{G}$ ensuring that the quantity $\mathcal{S}(Q, \mathcal{T}, \mathcal{G})$ is both tractable *and* relevant.

As an illustration, we now give an example of Stein discrepancy for quantifying the dissimilarity between any probability distribution Q on $\mathbb{R}^d$ and the normal distribution.

EXAMPLE 1 (Stein operator and discrepancy for the multivariate normal distribution). Let Σ be a $d \times d$ positive definite matrix; denote by $\mathrm{N}_d(0, \Sigma)$ the centered multivariate normal with covariance Σ. Let $g : \mathbb{R}^d \to \mathbb{R}$ be almost differentiable, that is, possess a gradient $\nabla g : \mathbb{R}^d \to \mathbb{R}^d$ such that, for all $z \in \mathbb{R}^d$, $g(x + z) - g(x) = \int_0^1 \langle z, \nabla g(x + tz)\rangle \,\mathrm{d}t$ for almost all $x \in \mathbb{R}^d$. Suppose furthermore that $\nabla g \in L^1(\mathrm{N}_d(0, \Sigma))$. Then

$$\mathbb{E}_{X\sim \mathrm{N}_d(0,\Sigma)}\big[\Sigma\nabla g(X) - Xg(X)\big] = 0;$$

see, for example, [153] (for Σ the identity matrix). We deduce that the first-order differential operator

$$(\mathcal{T}g)(x) = \Sigma\nabla g(x) - xg(x) \tag{3}$$

is a Stein operator for $\mathrm{N}_d(0, \Sigma)$ acting on the Stein class $\mathcal{G}(\mathcal{T})$ of all almost differentiable functions with (almost everywhere) gradient $\nabla g \in L^1(N_d(0, \Sigma))$. This leads to

$$\mathcal{S}(Q, \mathcal{T}, \mathcal{G}) = \sup_{g\in\mathcal{G}} \big\|\mathbb{E}_{X\sim Q}\big[\Sigma\nabla g(X) - Xg(X)\big]\big\|_2, \tag{4}$$

for any $\mathcal{G} \subset \mathcal{G}(\mathcal{T})$.

Of course it remains to ensure that the dissimilarity measures herewith obtained actually capture relevant aspects of the dissimilarity between P and Q. Classically, there are many ways to determine discrepancies between probability measures; see, for example, [63, 131]. In this survey, and in much of the literature on Stein's method, the focus is on distances known as *integral probability metrics* (IPMs, for short) [123, 173], which are defined as

$$d_{\mathcal{H}}(P, Q) := \sup_{h\in\mathcal{H}} \big|\mathbb{E}_{X\sim P}\big[h(X)\big] - \mathbb{E}_{X\sim Q}\big[h(X)\big]\big| \tag{5}$$

for some class of real-valued measurable test functions $\mathcal{H} \subset L^1(P) \cap L^1(Q)$. When $d_{\mathcal{H}}$ is a distance on the set of probability measures on $\mathcal{X}$ then $\mathcal{H}$ is called *measure determining*.

REMARK 1. Different choices of $\mathcal{H}$ give rise to different IPMs, including:

1. the *Kolmogorov distance*: $d_{\mathrm{Kol}}(P, Q)$, which is the IPM induced by the set of test functions $\mathcal{H}_{\mathrm{Kol}} = \{\mathbb{I}_{(-\infty,x]}(\cdot) : x \in \mathbb{R}^d\}$ (indicators of bottom left quadrants);
2. the *L^1-Wasserstein distance* (also known as the Kantorovich–Rubinstein or earth-mover's distance): $d_{\mathrm{W}}(P, Q)$, which is the IPM induced by the set of test functions $\mathcal{H}_{\mathrm{W}} = \{h : \mathbb{R}^d \to \mathbb{R} : \sup_{x\neq y\in\mathbb{R}^d} |h(x) - h(y)|/\|x - y\|_2 \leq 1\}$ (functions with Lipschitz constant at most 1);
3. the *bounded Wasserstein distance* (also known as the Dudley or bounded Lipschitz metric): $d_{\mathrm{bW}}(P, Q)$, which is the IPM induced by the set of test functions $\mathcal{H}_{\mathrm{bW}}$, which collects the bounded functions in $\mathcal{H}_{\mathrm{W}}$;
4. the *maximum mean discrepancy*: $d_k(P, Q)$, which is the IPM induced by the set of test functions $\mathcal{H}_k$, the unit-ball of some reproducing kernel Hilbert space [26] associated with kernel k. This case will be discussed extensively in Section 3.

To see the connection between IPMs $d_{\mathcal{H}}$ and Stein discrepancies $\mathcal{S}$, an additional ingredient enters the picture: the *Stein equation*. Given P the target distribution with Stein operator $\mathcal{T}$ and Stein class $\mathcal{G}(\mathcal{T})$, and given $\mathcal{H} \subset L^1(P)$ a measure-determining class of test functions, the *Stein equation* for $h \in \mathcal{H}$ is the functional equation

$$(\mathcal{T}g)(x) = h(x) - \mathbb{E}_{X\sim P}\big[h(X)\big] \tag{6}$$

evaluated over $x \in \mathcal{X}$, with solution $g = g(h) := \mathcal{L}h \in \mathcal{G}$, if it exists. Assuming that this solution exists for all

$h \in \mathcal{H}$, it follows that $\mathbb{E}_{X\sim Q}[h(X)] - \mathbb{E}_{X\sim P}[h(X)] = \mathbb{E}_{X\sim Q}[(\mathcal{T}(\mathcal{L}h))(X)]$ so

$$d_{\mathcal{H}}(P, Q) = \sup_{h\in\mathcal{H}} \big|\mathbb{E}_{X\sim P}\big[h(X)\big] - \mathbb{E}_{X\sim Q}\big[h(X)\big]\big| = \mathcal{S}(Q, \mathcal{T}, \mathcal{L}\mathcal{H}),$$

with $\mathcal{L}\mathcal{H}$ the Stein set collecting all solutions $\mathcal{L}h$ to the Stein equation (6) with $h \in \mathcal{H}$. Existence of a solution to the Stein equation depends on the properties of the target measure P, of the Stein operator $\mathcal{T}$, and of the Stein class $\mathcal{G}(\mathcal{T})$. In many cases, existence of these solutions is guaranteed and the IPMs listed in Remark 1 can be rewritten as Stein discrepancies whose underlying Stein set $\mathcal{L}\mathcal{H}$ depends on the measure P characterized by $\mathcal{T}$ through (6).

Often, bounding $\mathbb{E}_{X\sim Q}[h(X)] - \mathbb{E}_{X\sim P}[h(X)]$ through bounding $\mathbb{E}_{X\sim Q}[(\mathcal{T}(\mathcal{L}h))(X)]$ is advantageous as the latter only requires integrating under Q; the properties of P have been encoded in the Stein operator and Stein class. Commonly used approaches for bounding Stein discrepancies are coupling techniques [16, 37, 133, 138], the Malliavin–Stein method [124] and comparison of Stein operators [84, 98, 121]; here, the references only serve as pointers and the list is certainly not complete. In the context of theoretical statistics, IPM-based Stein discrepancies have been used for investigating finite-sample performance of statistical estimators with intractable exact distribution and known asymptotic behavior (here, thus Q is the exact distribution of some statistical procedure, and P its asymptotic distribution). An overview of some of these applications is provided in Section 5.

In order to bound $\mathbb{E}_{X\sim Q}[(\mathcal{T}(\mathcal{L}h))(X)]$, suitable bounds on the solutions $\mathcal{L}h$ of the Stein equation, as well as certain lower-order derivatives or differences of the solution, are usually required (although sometimes weak solutions of an appropriate equation suffice; see [43, 121]). Bounds on the solution are often referred to as *Stein factors*. Determining Stein factors has attracted attention in recent years. Of the many available references, we single out [53, 116] where bounds are obtained for operators given in the setting of Example 2 under assumptions of log-concavity, and [68], where Stein factors are obtained under the weaker assumption of *integrable Wasserstein decay*. An overview for continuous distributions is given in [121].

In this section, we have kept $\mathcal{H}$, or equivalently $d_{\mathcal{H}}$, mainly general, so that the task of deriving a Stein equation and bounds on Stein factors can be presented in a form which applies to any of the IPMs in Remark 1.

2.2 Choosing Stein Operators

When tackling Stein's method for a general target via a Stein discrepancy $\mathcal{S}(Q, \mathcal{T}, \mathcal{G})$, it is important to first choose $\mathcal{T}$ and $\mathcal{G}$ in a way which ensures relevance and tractability of the resulting metric or Stein discrepancy. For many target distributions, such useful Stein operators and Stein sets are readily available from the literature. One of the advantages of Stein's method, however, is that for a given P there is in principle full freedom of choice in the operator $\mathcal{T}$ and Stein set $\mathcal{G}$, and in particular no need to restrict to the operators from the literature nor Stein sets obtained from Stein equations.

Here, we shall mainly concentrate on two approaches for choosing a Stein operator, called the generator approach (which dates back to [13, 14] and [72]) and the density approach (which dates back to [152]). These are not the only available approaches (see, e.g., [134]) and we conclude the section with a brief pointer to other techniques.

2.2.1 *Stein operators via the generator approach.* We first describe the *generator approach*, which we present for a given target P on $\mathcal{X} = \mathbb{R}^d$. Given a Markov process with sufficient regularity $(Z_t)_{t\geq 0}$ (namely, a Feller process [129], Lemma 8.1.4) with invariant measure P, the *infinitesimal generator* $\mathcal{A}$ of the process given by

$$(\mathcal{A}u)(x) = \lim_{t\to 0} \frac{1}{t}\big(\mathbb{E}\big[u(Z_t) \mid Z_0 = x\big] - u(x)\big)$$

satisfies the property that $\mathbb{E}_{Z\sim P}[(\mathcal{A}u)(Z)] = 0$ for all $u : \mathbb{R}^d \to \mathbb{R}$ in the domain of $\mathcal{A}$. Barbour [13, 14] and [72] exploited this fact to provide both a Stein operator and a Stein class for all targets P that are invariant measures of sufficiently regular Markov processes, to analyze multivariate distributions via Stein's method.

Gorham et al. [68] detailed the generator approach for a wide range of distributions of interest by using operators induced by *Itô diffusions*. An Itô diffusion [129], Definition 7.1.1, with starting point $x \in \mathbb{R}^d$, Lipschitz *drift coefficient* $b : \mathbb{R}^d \to \mathbb{R}^d$, and Lipschitz *diffusion coefficient* $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times m}$ is a stochastic process $(Z_{t,x})_{t\geq 0}$ solving the Itô stochastic differential equation

$$\mathrm{d}Z_{t,x} = b(Z_{t,x})\,\mathrm{d}t + \sigma(Z_{t,x})\,\mathrm{d}W_t \tag{7}$$

with $Z_{0,x} = x \in \mathbb{R}^d$, where $(W_t)_{t\geq 0}$ is a m-dimensional Brownian motion. It is known (see, e.g., [68], Theorem 2, and [20], Theorem 19) that equation (7) will have invariant measure P with density p, which is positive and differentiable if and only if $b(x) = \langle \nabla, p(x)[\sigma(x)\sigma(x)^\intercal + c(x)]\rangle/2p(x)$ where the *stream coefficient* $c : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is some differentiable skew-symmetric valued function. [68] proposed the first-order *diffusion Stein operator*

$$(\mathcal{T}g)(x) = \frac{1}{p(x)}\langle \nabla, p(x)\big[(\sigma(x)\sigma(x)^\intercal + c(x)\big]g(x)\rangle, \tag{8}$$

based on the diffusion's second-order infinitesimal generator $\mathcal{A}u = \mathcal{T}(\nabla u/2)$. Under regularity conditions, the definition in equation (8) yields an infinite collection of Stein operators for a given target P, parametrized by the choice of σ and c.

EXAMPLE 2 (The Langevin–Stein operator on $\mathbb{R}^d$). As a concrete example, [69, 116] consider the case where $\sigma \equiv I_d$ and $c \equiv 0$, which corresponds to the overdamped Langevin diffusion. Assuming $\mathbb{E}_{X\sim P}[\|\nabla \log p(X)\|_2] < \infty$, this induces the Langevin–Stein operator

$$(\mathcal{T}g)(x) = \langle \nabla \log p(x), g(x)\rangle + \langle \nabla, g(x)\rangle. \tag{9}$$

The corresponding Stein discrepancies from equation (2) are often called *Langevin–Stein discrepancies*.

2.2.2 *Stein operators via the density approach*. The *density approach* was pioneered in [152] for univariate distributions, and has since then been generalized in multiple directions; see, for example, [121, 164]. Given a probability measure P on a set $\mathcal{X}$ with density function (with respect to some dominating measure) $p : \mathcal{X} \to \mathbb{R}^+$, consider operators of the form $g \mapsto \mathcal{D}(g(x)p(x))/p(x)$, where $\mathcal{D}$ is a linear operator with domain $\mathrm{dom}(\mathcal{D})$. Collecting into the class $\mathcal{G}$ all functions g on $\mathcal{X}$ such that $x \mapsto p(x)g(x) \in \mathrm{dom}(\mathcal{D})$ and $\int_{\mathcal{X}} \mathcal{D}(g(x)p(x))\,\mathrm{d}x = 0$, the $\mathcal{D}$- *density*, or for short, *density* Stein operator of the density approach for p is

$$g \mapsto (\mathcal{T}g)(x) = \frac{\mathcal{D}(g(x)p(x))}{p(x)}$$

with Stein class $\mathcal{G}(\mathcal{T}) = \mathcal{G}$. By construction, this operator satisfies $\mathbb{E}_{X\sim P}[(\mathcal{T}g)(X)] = 0$ for all $g \in \mathcal{G}(\mathcal{T})$. The following example illustrates the approach for univariate distributions with interval support.

EXAMPLE 3 (Density operators for the exponential distribution). Fix $d = 1$ and consider as target P the exponential distribution with density function $p(x) = \lambda \mathrm{e}^{-\lambda x}\mathbb{I}_{[0,\infty)}(x)$, for $\lambda > 0$. A natural choice of $\mathcal{D}$ is $\mathcal{D}f(x) = f'(x)$ the usual almost everywhere derivative. If $(gp)'$ is integrable on $\mathbb{R}^+$, then $\int_0^\infty (g(x)p(x))'\,\mathrm{d}x = \lim_{x\to\infty} g(x)p(x) - \lambda g(0)$. The corresponding density operator is therefore

$$(\mathcal{T}g)(x) = \frac{(g(x)p(x))'}{p(x)} = g'(x) - \lambda g(x), \quad x \in \mathbb{R}^+,$$

acting on the Stein class of functions g such that $(gp)'$ is integrable on $\mathbb{R}^+$ and $\lim_{x\to\infty} g(x)p(x) = \lambda g(0)$. Clearly, all functions $g(x) = xg_0(x)$ such that $\lim_{x\to\infty} xg_0(x)\mathrm{e}^{-\lambda x} = 0$ belong to $\mathcal{G}(\mathcal{T})$. Denoting $\tilde{\mathcal{G}}$ the collection of functions of this form, we reap a second operator for the exponential given by

$$(\mathcal{T}_1 g_0)(x) = \frac{(xg_0(x)\mathrm{e}^{-\lambda x})'}{\mathrm{e}^{-\lambda x}} = xg_0'(x) + (1 - \lambda x)g_0(x)$$

acting on the (restricted) Stein class $\tilde{\mathcal{G}}$. The advantage of the latter operator over the former is that it does not require any implicit boundary assumptions on the test functions.

Since the exponential density is also a parametric scale family in its parameter $\lambda > 0$, another natural derivative in this context is $\mathcal{D}f(x;\lambda) = \frac{\mathrm{d}}{\mathrm{d}\lambda} f(x;\lambda)$ for all functions $f(x;\lambda)$ of the form $f(x;\lambda) = \lambda f_0(\lambda x)$ for some f_0. This leads to

$$\begin{aligned}(\mathcal{T}_2 g)(x) &= \frac{\frac{\mathrm{d}}{\mathrm{d}\lambda}(\lambda g(\lambda x)\mathrm{e}^{-\lambda x})}{(\lambda \mathrm{e}^{-\lambda x})} \\ &= xg'(\lambda x) + \left(\frac{1}{\lambda} - x\right) g(\lambda x),\end{aligned}$$

with no boundary assumptions on g since

$$\mathbb{E}_{X\sim \mathrm{Exp}(\lambda)}[(\mathcal{T}_2 g)(X)] = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\int_0^\infty g(u)\mathrm{e}^{-u}\,\mathrm{d}u\right) = 0$$

for all $g \in L^1(\mathrm{Exp}(1))$.

Many choices of operator $\mathcal{D}$ lead to Stein operators. Moreover, using appropriate product rules, Stein operators can be tailored for the specifics of the problem at hand. This process is called *standardizing the Stein operator*; see [95] and [65].

The density approach and the generator approach are by no means the only methods for obtaining Stein operators. Other popular approaches include couplings ([35]), orthogonal polynomials ([64]), a perturbation approach ([17]), an ODE approach ([58]) and characteristic functions ([10, 156]).

2.2.3 *Some general remarks on Stein operators*. A Stein operator can often be found even when the density of the target distribution is not available in closed form, which will be particularly useful for applications in statistics. In this context, we highlight two classes of important problems:

2.2.3.1 *Bayesian computation*. In Bayesian statistics, usually the posterior distribution is known only in an unnormalized form. This is not a hindrance for Stein's method; see [99]. Take, for example, the Langevin–Stein operator of Example 2: $(\mathcal{T}g)(x) = \langle \nabla \log p(x), g(x)\rangle + \langle \nabla, g(x)\rangle$. Any function of the form $(\mathcal{T}g)$ can be evaluated pointwise provided that $\nabla \log p$ can be evaluated, which is often a reasonable requirement. In particular, this does not require knowledge of the normalizing constant of p, since if $p = \tilde{p}/C$ for $C > 0$, then $\nabla \log p = \nabla \log \tilde{p} - \nabla \log C = \nabla \log \tilde{p}$. In fact, $\nabla \log p$ is usually the basis of gradient-based Markov chain Monte Carlo algorithms to sample from posterior distributions. Illustrations of this principle can be found in [68] and [121], for instance.

2.2.3.2 *Intractable likelihood*. A second example includes models in which the likelihood itself is unnormalized, in which case the model is often called a Gibbs distribution. For these, $\ell(\theta; x) \propto \tilde{\ell}(\theta, x)$, where $\tilde{\ell}(\theta, x)$ can be pointwise evaluated. Once again, working with $\nabla_x \log \ell(\theta; x)$ may be practical even when the normalizing constant is an intractable integral. Furthermore, when the likelihood can be written as the density of a natural

exponential family model, $\nabla_x \log \ell(\theta; x)$ becomes linear in θ, which is particularly useful in the development of new statistical methodology based on the Langevin–Stein operator (see [21, 117]).

3. COMPUTABLE STEIN DISCREPANCIES

As mentioned in Section 2.1, many authors use Stein's method to assess IPMs between a target P and some approximating measure Q by using Stein discrepancies computed over sets $\mathcal{G}$ consisting of solutions to Stein equations. In this section, we will now show how Stein discrepancies may sometimes be computed *exactly* through a particular choice of Stein set (this issue was in fact already identified by [152]). Exact computation turns out to be possible when comparing an empirical measure $Q_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$, with data points $x_i \in \mathcal{X}$, δ_{x_i} giving all probability mass to x_i, to a given target distribution P. We will call any such discrepancy a *computable Stein discrepancy*.

The most common choice of computable discrepancies are the so-called *kernel Stein discrepancies* (KSD), which use the unit-ball of a reproducing kernel Hilbert space (RKHS) as a Stein set, and can therefore be considered the Stein discrepancy counterpart to the *maximum mean discrepancy* IPM [74, 75, 145]. An RKHS $\mathcal{H}_k$ is a Hilbert space (with norm $\|.\|_k$ and inner product $\langle\cdot,\cdot\rangle_k$) with an associated function $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ satisfying (i) symmetry; $k(x, y) = k(y, x)$ for all $x, y \in \mathcal{X}$, (ii) positive definiteness; $\sum_{i,j=1}^n c_i c_j k(x_i, x_j) \geq 0$ for all $c_i \in \mathbb{R}$, $x_i \in \mathcal{X}$ and (iii) the reproducing property $f(x) = \langle k(x,\cdot), f\rangle_k$ for all $f \in \mathcal{H}_k$, $x \in \mathcal{X}$. The function k is called a *reproducing kernel* [9, 139]. This choice of Stein set was inspired by the zero mean reproducing kernel theory of [127], used in [42, 70, 108] and extended in [154] to the case of matrix-valued kernels. The main advantage is that the supremum in (2) can be analytically computed in terms of the reproducing kernel:

EXAMPLE 4 (Langevin Kernel Stein discrepancies). The *Langevin KSD* on $\mathcal{X} = \mathbb{R}^d$ is obtained by combining the Langevin–Stein operator $\mathcal{T}$ from Example 2 with a *kernel Stein set* $\mathcal{G}_k := \{g = (g_1, \ldots, g_d) \mid \|v\|_2 \leq 1 \text{ for } v_j := \|g_j\|_k\}$:

$$\begin{aligned}\mathrm{KSD}_k(Q) &:= \mathcal{S}(Q, \mathcal{T}, \mathcal{G}_k)\\ &= \sqrt{\mathbb{E}_{X,X'\sim Q}\big[k_P(X, X')\big]},\end{aligned} \tag{10}$$

where the *Stein reproducing kernel* is given by

$$\begin{aligned}k_P(x, x') &:= \mathrm{Trace}\big(\mathcal{T}_x\mathcal{T}_{x'}k(x, x')\big)\\ &= \big\langle\nabla_x, \nabla_{x'}k(x, x')\big\rangle\\ &\quad + \big\langle\nabla_x k(x, x'), \nabla_{x'}\log p(x')\big\rangle\\ &\quad + \big\langle\nabla_{x'}k(x, x'), \nabla_x \log p(x)\big\rangle\\ &\quad + k(x, x')\big\langle\nabla_x \log p(x), \nabla_{x'}\log p(x')\big\rangle.\end{aligned} \tag{11}$$

Here, the subscript in $\mathcal{T}_x$ indicates that the input of $\mathcal{T}$ is seen as a function of x. Most notably, this Stein reproducing kernel satisfies $\mathbb{E}_{X\sim P}[k_P(X, x)] = 0$ for all $x \in \mathbb{R}^d$ under mild regularity conditions (see [127]). Whenever the approximating measure is $Q_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$, the Langevin KSD has the simple closed form

$$\begin{aligned}\mathrm{KSD}_k(Q_n) &= \mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_k)\\ &= \sqrt{\frac{1}{n^2}\sum_{i,j=1}^n k_P(x_i, x_j)}.\end{aligned} \tag{12}$$

The most common choice of kernel k is the inverse multiquadric kernel $k(x, y) = (c^2 + \|x - y\|_2^2)^\beta$, $c > 0$, $\beta \in (-1, 0)$. This is because [70], Theorem 8, showed that, if $\nabla \log p$ is sufficiently regular, then Q_n converges weakly to P whenever $\mathrm{KSD}_k(Q_n) \to 0$. We will return to the implications of this in Section 4.1.

Extensions of the Langevin KSD include [21], who used the infinitesimal generator of general Itô diffusions to get a family of *diffusion kernel Stein discrepancies*; [164] to discrete sets $\mathcal{X}$; [22] to the case where $\mathcal{X}$ is a Riemannian manifold, such as in directional statistics.

A second type of computational Stein discrepancies are the *graph Stein discrepancies* (GSDs) of [68, 69].

EXAMPLE 5 (Graph Stein discrepancies). The *graph Stein discrepancies* combine a diffusion Stein operator $\mathcal{T}$ as in (8) with a *graph Stein set*

$$\begin{aligned}&\mathcal{G}_{\|\cdot\|,Q_n,E}\\ &\quad= \Bigg\{g : \max\Bigg(\big\|g(v)\big\|_\infty, \big\|\nabla g(v)\big\|_\infty,\\ &\qquad\frac{\|g(x) - g(y)\|_\infty}{\|x - y\|_1}, \frac{\|\nabla g(x) - \nabla g(y)\|_\infty}{\|x - y\|_1}\Bigg) \leq 1,\\ &\qquad\frac{\|g(x) - g(y) - \nabla g(x)(x - y)\|_\infty}{\frac{1}{2}\|x - y\|_1^2} \leq 1,\\ &\qquad\frac{\|g(x) - g(y) - \nabla g(y)(x - y)\|_\infty}{\frac{1}{2}\|x - y\|_1^2} \leq 1,\\ &\qquad\forall (x, y) \in E, v \in \mathrm{supp}(Q_n)\Bigg\},\end{aligned}$$

where ∇g denotes the Jacobian matrix of g and E is a set of pairs of the form (x_i, x_j), which must be taken sufficiently large to ensure that the GSD has Wasserstein convergence control [69], Theorem 2, Propositions 5 and 6.

Once again, the Stein set is selected so that the discrepancy can be computed efficiently. The GSD is actually the solution of a finite-dimensional linear program, with the size of E as low as linear in n, implying that it can be efficiently computed.

While computable, both KSDs and GSDs suffer from a computational cost that grows at least quadratically in the

sample size n. There exist at least two practical options for large sample sizes. The *finite set Stein discrepancies* of [90] achieve a linear runtime by learning a small number of adaptive features based on Stein-transformed kernels, so as to distinguish P from Q samples with maximum test power. The *random feature Stein discrepancies* of [87] approximate a broad class of convergence-determining Stein discrepancies in near-linear time using importance sampling. To reduce the computational cost of Stein discrepancies in high dimensions, the *sliced Stein discrepancies* of [67] can be used.

Finally, the computation of a Stein discrepancy can also be prohibitive if the Stein operator is expensive to evaluate. This commonly occurs in Bayesian and probabilistic inference where $\mathcal{T} = \sum_{l=1}^{L} \mathcal{T}_l$ is a sum over likelihood terms or potentials which are each more easily evaluated than $\mathcal{T}$ itself. To address this deficiency, [71] introduced *stochastic Stein discrepancies* (SSDs)

$$\mathcal{SS}(Q_n, \mathcal{T}, \mathcal{G}) := \sup_{g \in \mathcal{G}} \left| \frac{L}{n} \sum_{i=1}^{n} (\mathcal{T}_{\sigma_i} g)(x_i) \right| \tag{13}$$

for $\sigma_i \overset{\text{i.i.d.}}{\sim} \text{Unif}(\{1, \dots, L\})$. They showed that SSDs inherit the convergence control properties of standard discrepancies with probability 1. In [163], for a special case of a stochastic Stein discrepancy, Stein's method is used to establish its asymptotic normality.

4. NEW STATISTICAL METHODS FOR ASSESSING SAMPLE QUALITY, CONSTRUCTING SAMPLE APPROXIMATIONS AND IMPROVING MONTE CARLO INTEGRATION

This section details how ingredients from Stein's method have been successfully used to uncover methodological tools and procedures, and discusses a range of recent applications of Stein's method in computational statistics and machine learning. Section 4.1 shows how computable Stein discrepancies can be employed to quantify the quality of approximate MCMC schemes. Section 4.2 introduces a variety of ways of using Stein's method to construct and improve a sample approximation, including Stein variational gradient descent (Section 4.2.1), Stein points (Section 4.2.2) and Stein thinning (Section 4.2.3). Section 4.3 describes Stein-based control variates for improved Monte Carlo integration, Section 4.4 presents statistical estimators, and Section 5 details goodness-of-fit tests.

4.1 Measuring Sample Quality

This section presents practical tools based on Stein's method for computing how well a given sample, represented as an empirical measure $Q_n = n^{-1} \sum_{i=1}^{n} \delta_{x_i}$, approximates a given target distribution P. This line of work was motivated by the approximate Markov chain Monte Carlo (MCMC) revolution in which practitioners have turned to asymptotically biased MCMC procedures that sacrifice asymptotic correctness for improved sampling speed (see, e.g., [1, 93, 161]). The reasoning is sound—the reduction in Monte Carlo variance from faster sampling can outweigh the bias introduced, but standard Monte Carlo diagnostics like effective sample size, asymptotic variance, trace and mean plots and pooled and within-chain variance diagnostics presume eventual convergence to the target distribution, and hence do not account for asymptotic bias. To address this deficiency, [69–71, 87] introduced the computable Stein discrepancies of Section 3 as measures of sample quality suitable for comparing asymptotically exact, asymptotically biased, and even deterministic sample sequences $\{x_1, \dots, x_n\}$.

4.1.1 *Graph Stein discrepancies*. [69] used the GSDs of Example 5 to select and tune approximate MCMC samplers, assess the empirical convergence rates of Monte Carlo and quasi-Monte Carlo procedures, and quantify bias-variance tradeoffs in posterior inference. An illustrative example is given in Figure 1. These applications were enabled by a series of analyses establishing that the GSD converges to 0 if and only if its empirical measure Q_n converges to P. Specifically, [52, 68, 116] bounded the GSD explicitly above and below by Wasserstein distances whenever the diffusion underlying the Stein operator couples quickly and has pseudo-Lipschitz drift.

4.1.2 *Kernel Stein discrepancies*. The closed form of the KSDs of Example 4 represents a significant practical advantage for sample quality measurement, as no linear program solvers are necessary, and the computation of the discrepancy can be easily parallelized. However, [70] showed that not all KSDs are suitable for measuring sample quality. In particular, in dimension $d \geq 3$, KSDs based on popular kernels like the Gaussian and Matérn kernels fail to detect when a sample is not converging to the target, even when the target is normal. To address this shortcoming, [70] developed a theory of weak convergence control for KSDs and designed a class of KSDs that provably control weak convergence for a large set of target distributions (see [40, 87] for further developments). These convergence-determining KSDs have been shown to deliver substantial speed-ups over the original GSDs in higher dimensions [70].

4.1.3 *Random feature Stein discrepancies*. To identify a family of convergence-determining discrepancy measures that can be accurately and inexpensively approximated with random sampling, [87] introduce a new domain for the Stein operator using a feature function, giving rise to a *feature Stein set* and a corresponding *feature Stein discrepancy*. The feature Stein discrepancy is then approximated using importance sampling, which results a *random feature Stein discrepancy* (RΦSD). Higgins and

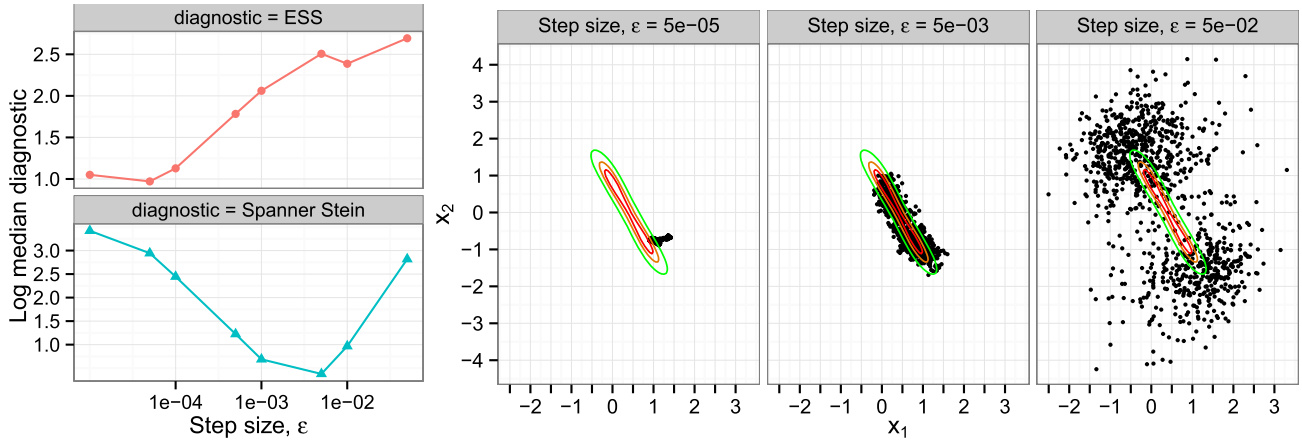

FIG. 1. *Selecting the step size ϵ for stochastic gradient Langevin dynamics* [161], *a popular approximate MCMC algorithm designed for scalability. Standard MCMC diagnostics like effective sample size (ESS) do not account for asymptotic bias and select overly large ϵ with greatly overdispersed samples (right panel). Overly small ϵ leads to slow mixing (left panel). The Stein discrepancy selects an intermediate value offering the best approximation (center panel). Figure reproduced from* [69], *Figure* 3.

Mackey [87] showed that RΦSDs upper bound standard discrepancy measures with high probability. This translates into high-probability convergence control whenever the approximating sample sequence is uniformly integrable.

4.2 Constructing and Improving Sample Approximation

Popular stochastic Monte Carlo methods such as MCMC provide a standard approach for constructing and improving a sample-based approximation of the form $Q_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$ for an intractable distribution P of interest. In this section, we explain how Stein's method can be used to develop a suit of *optimization-based* alternatives to Monte Carlo methods. We demonstrate this with three examples: Section 4.2.1 introduces Stein variational gradient descent, a gradient based algorithm that iteratively updates the location of the particles $\{x_1, \dots, x_n\}$ to improve the approximation quality w.r.t P. Section 4.2.2 introduces Stein Points, a greedy algorithm that constructs the approximation by sequentially adding the particles to minimize KSD. Section 4.2.3 introduces Stein Thinning, which compresses an existing approximation using KSD.

4.2.1 *Sampling with Stein variational gradient.* Let P be a distribution with a continuously differentiable density function p supported on $\mathcal{X}$. We want to find a set of points $\{x_1, \dots, x_n\} \subset \mathcal{X}$, which we refer to as *particles*, such that its empirical measure Q gives a close approximation to P. Stein variational gradient descent (SVGD) [111] achieves this by iteratively updating the particles to minimize the KL divergence between Q and P, which is made possible by exploiting an intrinsic connection between KL divergence and Stein's method, as follows.

For the purpose of derivation, we assume for now that Q is a continuous distribution with a finite KL divergence $\mathrm{KL}(Q\|P) < \infty$. We want to recursively "transport" the probability mass of Q with a deterministic map to move it closer to P in order to decrease $\mathrm{KL}(Q\|P)$ as fast as possible. Specifically, we consider mappings of the form

$$\Phi(x) = x + \epsilon g(x),$$

where ϵ is a small positive scalar that serves as a step size, and $g\colon \mathcal{X} \to \mathcal{X}$ is a one-to-one mapping that serves as the velocity field. Denote by $\Phi_\sharp Q$ the distribution of $\Phi(X)$ when $X \sim Q$; this is also called the *pushforward measure*.

The key challenge is to optimally choose g for each given Q, so that the KL divergence between $\Phi_\sharp Q$ and P is decreased as much as possible. Assuming ϵ is infinitesimal, the optimal choice of g can be framed into a functional optimization problem:

$$\max_{g\in\mathcal{G}} \left\{ -\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{KL}(\Phi_\sharp Q\|P)|_{\epsilon=0} \right\}, \tag{14}$$

where the negative derivative $-\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{KL}(\Phi_\sharp Q\|P)|_{\epsilon=0}$ measures the decreasing rate of KL divergence under the transport map Φ as we increase the step size ϵ starting from zero, and $\mathcal{G}$ is a function space that specifies the candidate set of g. The key observation is that the objective in (14) is in fact equivalent to the expectation $\mathbb{E}_Q[(\mathcal{T}g)(X)]$ of the Langevin–Stein operator.

THEOREM 1. *Assume P and Q have positive densities on $\mathcal{X} = \mathbb{R}^d$, and the density p of P is in $C^1(\mathbb{R}^d)$. Let $\Phi(x) = x + \epsilon g(x)$, where $\epsilon \in \mathbb{R}$ and $g\colon \mathbb{R}^d \to \mathbb{R}^d$ is a C^1 map with $\sup_{x\in\mathbb{R}^d}\|\nabla g(x)\|_2 < \infty$, where $\|\cdot\|_2$ denotes the spectral norm. We have*

$$-\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{KL}(\Phi_\sharp Q\|P)|_{\epsilon=0} = \mathbb{E}_{X\sim Q}[(\mathcal{T}g)(X)],$$

where $(\mathcal{T}g)(x) = \langle \nabla \log p(x), g(x)\rangle + \langle \nabla, g(x)\rangle$.

Theorem 1 draws an intriguing connection between Stein's method, the KL divergence and optimal transport. It shows that (14) is equivalent to the optimization in Langevin KSD:

$$\begin{aligned}\mathrm{KSD}_k(Q) &= \max_{g\in\mathcal{G}}\{\mathbb{E}_{X\sim Q}[(\mathcal{T}g)(X)]\}\\ &= \max_{g\in\mathcal{G}}\left\{-\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{KL}(\Phi_\sharp Q\|P)|_{\epsilon=0}\right\}.\end{aligned}\tag{15}$$

Therefore, the Langevin KSD can be interpreted as the maximum decreasing rate of KL divergence between Q and P under the best transport map in $\mathcal{G}$. Taking $\mathcal{G}$ to be the unit ball of the RKHS with kernel k, we can solve equation (15) in closed form (see Example 4):

$$g^*_{Q,P}(\cdot)\propto\mathbb{E}_{X\sim Q}[\nabla\log p(X)k(X,\cdot)+\nabla_x k(X,\cdot)].\tag{16}$$

This yields the best update direction for "transporting" particles from Q to P under KL divergence. In practice, we take $Q=n^{-1}\sum_{i=1}^n\delta_{x_i}$ to be the empirical measure of the particles while iteratively updating $\{x_1,\dots,x_n\}$ by using the optimal transport map found above, $\Phi^*_{Q,P}(x)=x+\epsilon g^*_{Q,P}(x)$. This yields the following simple update rule on the particles, which is illustrated in the left panel of Figure 2:

$$\begin{aligned}x_i\leftarrow x_i+\frac{\epsilon}{n}\sum_{j=1}^n&\big(\nabla\log p(x_j)k(x_j,x_i)\\&+\nabla_{x_j}k(x_j,x_i)\big),\end{aligned}\tag{17}$$

for all $i=1,\dots,n$. The two terms in (17) play intuitive roles. The term with the gradient $\nabla\log p$ pushes the particles toward the high probability regions of P, while the term with $\nabla_x k$ can be viewed as a repulsive force to enforce the diversity between the particles if k is a stationary kernel of form $k(x,x')=\phi(x-x')$: in this case, performing $x_i'\leftarrow x_i+\epsilon\nabla_{x_j}k(x_j,x_i)$ would decrease $k(x_i,x_j)$, which measures the similarity between x_i and x_j, when ϵ is sufficiently small. If there is no repulsive force, or when there is only a single particle (and the kernel satisfies $\nabla_x k(x,x')=0$ for $x=x'$), the solution would collapse to the local optima of $\log p$, reducing to the maximum a posteriori (MAP) point. Therefore, by using different particle sizes, SVGD provides an interpolation between MAP to a full particle-based approximation.

SVGD defines a *deterministic interacting particle* system in which $\{x_1,\dots,x_n\}$ interact and coevolve to reach a desirable equilibrium. For understanding SVGD asymptotically, [107] considers the limit of large particle size ($n\to\infty$) and continuous time ($\epsilon\to 0$), and interprets SVGD as a *gradient flow* of KL divergence induced by a kernel-Wasserstein geometric structure on the infinite dimensional space of distributions; a set of theoretical studies along this line can be found in [41, 49, 71, 94, 105, 115, 125]. In the nonasymptotic regime of a finite number n of particles, SVGD acts like a *numerical quadrature* method in which the particles are arranged to exactly estimate the true expectation of a set of special basis functions determined by the Stein operator and kernel function [112].

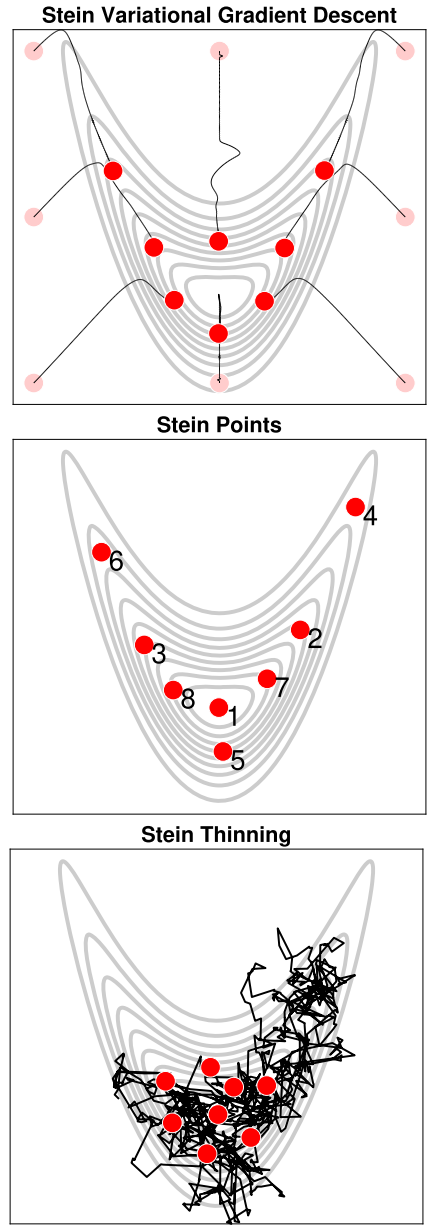

FIG. 2. *Sampling with Stein's method. Top*: *The initial* (*transparent red*) *and final* (*red*) *states of eight particles, together with their trajectories* (*black*) *under the Stein variational gradient descent algorithm in* (17). *Middle*: *The first* 8 *states* (*red*) *of an extensible sequence produced by the Stein points algorithm in* (18). *The order in which the states are selected is indicated. Bottom*: *The first eight representative states* (*red*) *selected from a Markov chain sample path* (*black*), *according to* (19). (*Grey contours are shown for the distributional target, which in each case the red states are intended to represent.*)

SVGD has been extended and improved in various ways. For example, amortized SVGD [55] learns neural samplers in replacement of particle approximation;

gradient-free SVGD [78] provides an extension that requires no gradient information of the target distribution P; a number of other extensions and improvements can be found in, for example, [34, 38, 45, 66, 71, 77, 101, 104, 158–160, 172]. SVGD has found applications in a variety of problems including in deep learning (e.g., [130, 157]), reinforcement learning (e.g., [76, 103, 114]), meta learning (e.g., [55, 92]) and uncertainty quantification in science and engineering (e.g., [167–170]).

4.2.2 *Sampling with Stein points*. The *Stein points* [39, 40] approach progressively constructs a set of points $\{x_1, \dots, x_n\} \subset \mathcal{X}$ to approximate P by minimizing a Stein discrepancy. For example, the KSD can be minimized in a sequential greedy manner: $x_1 \in \operatorname{argmin}_{x\in\mathcal{X}} \mathrm{KSD}_k(\{x\})$ and

$$x_n \in \operatorname*{argmin}_{x\in\mathcal{X}} \mathrm{KSD}_k\big(\{x_1, \dots, x_{n-1}, x\}\big) \quad \text{for } n > 1, \tag{18}$$

where $\mathrm{KSD}_k(\{x_1, \dots, x_n\}) = \mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_k)$ and $\mathcal{G}_k$ is a kernel Stein set; then the set $\{x_1, \dots, x_n\}$ is selected as to approximately minimize this KSD. A typical sequence obtained in this way is presented in the middle panel of Figure 2.

Finding the global minima in equation (18) may be difficult. However, [40], Theorem 2, showed that even imperfect optimization methods can lead to a fast decrease of the KSD. More precisely, if k_P in equation (11) satisfies $\mathbb{P}_{X\sim P}(k_P(X, X) \geq t) \leq b_1 \mathrm{e}^{-b_2 t}$ for some constants $b_1, b_2 > 0$ and all $t \geq 0$, then there exist constants $c_1, c_2 > 0$ depending only on k_P and P such that any $n \in \mathbb{N}$ and $\{x_1, \dots, x_n\} \subset \mathcal{X}$ satisfying

$$\mathrm{KSD}_k\big(\{x_1, \dots, x_j\}\big)^2 \leq \frac{\delta}{n^2} + \min_{x\in\mathcal{X}: k_P(x,x) \leq \frac{2\log(j)}{c_2}} \mathrm{KSD}_k\big(\{x_1, \dots, x_{j-1}, x\}\big)^2$$

for all $j = 1, \dots, n$, lead to an upper bound on the KSD of the form

$$\mathrm{KSD}_k\big(\{x_1, \dots, x_n\}\big) \leq \mathrm{e}^{\pi/2} \sqrt{\frac{2\log(n)}{c_2 n} + \frac{c_1}{n} + \frac{\delta}{n}}.$$

Thus, KSD can be used to transform the sampling problem of approximating P into an optimization problem that admits a provably convergent numerical method.

4.2.3 *Stein thinning*. [137] use KSD in a post-processing approach to select states from a large predetermined candidate set, with application to debiasing MCMC output. Their approach can be summarized as

$$\begin{aligned} x_1 &\in \operatorname*{argmin}_{x\in\{X_1,\dots,X_N\}} \mathrm{KSD}_k\big(\{x\}\big), \\ x_n &\in \operatorname*{argmin}_{x\in\{X_1,\dots,X_N\}} \mathrm{KSD}_k\big(\{x_1, \dots, x_{n-1}, x\}\big) \end{aligned} \tag{19}$$

for $n > 1$, where $(X_i)_{i=1,\dots,N}$ is a Q-invariant Markov chain; Q and P need not be equal. A typical sequence obtained in this way is presented in the right panel of Figure 2. These authors extended earlier convergence results to prove almost sure weak convergence of $Q_n = n^{-1}\sum_{i=1}^n \delta_{x_i}$ to P in the limit as $N \geq n \to \infty$. Indeed, provided that the Markov chain is V-uniformly ergodic with $V(x) \geq \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\sqrt{k_P(x,x)}$ and that certain moments of the chain are finite, [137], Theorem 3, showed that $\mathrm{KSD}_k(\{x_1, \dots, x_n\}) \to 0$ almost surely as $n \to \infty$.

Thus, Stein discrepancies may be used to post-process MCMC output, which can have the benefits of improving approximation quality, mitigating sampler bias and providing a compressed representation of P. The closed form of KSD renders such post-processing straightforward. Extensions of Stein thinning, to allow for nonmyopic optimization and for mini-batching, were recently studied in [155]. In related work, [83, 109] proposed to use Stein discrepancies to reweight Markov chain output, as opposed to selecting a smaller subset.

4.3 Improving Monte Carlo Integration

As already mentioned, the problem of approximating expectations $\mathbb{E}_{X\sim P}[f(X)]$, where $f : \mathcal{X} \to \mathbb{R}$ is a test function of interest, is at the heart of Stein's method, see [151]. In Bayesian statistics, it is most common for expectations to be approximated using ergodic averages from MCMC, though of course the algorithms described in Sections 4.2.1 and 4.2 can also be used. The convergence of estimators based on MCMC is characterized by the central limit theorem, whose asymptotic variance will depend on the variance of f along the sample path of the Markov chain (see Chapter 17 of [119]). In [152], auxiliary variables are constructed for such variance reduction in a particular setting. A recent approach to reducing the asymptotic variance is to use so-called *control variates*. This consists of designing a function $h : \mathcal{X} \to \mathbb{R}$ such that, if we rewrite the expectation as

$$\mathbb{E}_{X\sim P}\big[f(X)\big] = \mathbb{E}_{X\sim P}\big[h(X)\big] + \mathbb{E}_{X\sim P}\big[f(X) - h(X)\big],$$

then the first term on the right-hand side is known analytically (by some auxiliary argument) and the second integrand, $f - h$, should have smaller variance than f along the sample path of the Markov chain. In this way, estimation of the original expectation is reduced to estimation of an alternative expectation, which is more amenable to MCMC. Indeed, in an ideal situation we would pick h such that $f - h$ is constant along the sample path of the Markov chain, so that the ergodic average is exact after just one iteration of the chain has been performed [122].

The principal limitation to the successful application of control variates is the identification of a set of candidates for h that (a) is sufficiently rich to approximate f and (b) for which the expectations $\mathbb{E}_{X\sim P}[h(X)]$ can be evaluated. Several authors have developed bespoke solutions

that are specific to a particular MCMC algorithm, including [8, 44, 79, 120, 152]. It was pointed out in [127] that the image of a Stein operator adapted to P can serve as such a set in general. In concrete terms, one may identify a Stein operator $\mathcal{T}$ and a Stein set $\mathcal{G}$ that are adapted to P and then attempt to pick an element $g \in \mathcal{G}$ for which $f - h \approx$ constant along the Markov chain sample path, where $h = \mathcal{T}g$. This problem is closely related to numerical solution of the Stein equation (6).

In [11, 122, 128], the authors selected g from the set of all polynomials of a fixed maximum degree, minimizing the squared error $J_n(g) = \sum_{i=1}^n (f(x_i) - \mathcal{T}g(x_i))^2$ along the Markov chain sample path $\{x_1, \ldots, x_n\}$, with no complexity penalty used. In [148], the authors used an ℓ_1 or ℓ_2 penalty on the polynomial coefficients and recommended cross-validation as a means to select an appropriate polynomial degree. Kernel methods with a minimum norm penalty were proposed in [21, 126, 127, 154]. In [147], the authors showed how polynomials and reproducing kernels can be combined in a manner that leads to polynomial exactness of the control variate estimator in the Bernstein–von Mises limit. The use of neural networks for g was empirically assessed in [144, 171]. If one specializes to particular MCMC algorithms then it may be possible to consistently estimate the asymptotic variance under the Markov chain, which can be used to construct a more appropriate functional J_n. This approach is exemplified in [23–25]. [106] provides a detailed application of Stein control variates to policy optimization in reinforcement learning.

The diverse set of approaches for constructing control variates based on Stein operators supports the view that no single method will be universally optimal for all real-world computational problems and, to some extent, the estimation of a suitable control variate remains as much an "art" as the design of an efficient MCMC method.

4.4 Statistical Estimators Based on Stein Discrepancies

Computable Stein discrepancies have also been used for parameter estimation. Let $\mathscr{P}_\Theta = \{P_\vartheta : \vartheta \in \Theta\}$ denote a parametric family of distributions, and assume we would like to recover the element of this family, which generated some data $x_1, \ldots, x_n$ (represented by $Q_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$). Barp et al. [21] proposed *minimum Stein discrepancy estimators*, which are a general class of estimators of the form

$$\hat{\vartheta}_n := \arg\inf_{\vartheta \in \Theta} \mathcal{S}(Q_n, \mathcal{T}_\vartheta, \mathcal{G}), \tag{20}$$

where $\mathcal{T}_\vartheta$ is a Stein operator characterising P_ϑ, and showed that a number of machine learning algorithms including score-matching [88], contrastive divergence [82] and minimum probability flow [146] are specific instances of this framework. have also been proposed. Barp et al. [21] studied the special case of minimum diffusion KSD estimators and showed that these enjoy desirable robustness properties under regularity conditions on the kernel. This was then studied further in the context of discrete models by [12], while [73] considered a Stein discrepancy where the Stein space is indexed by a neural network. Relatedly, [30] studied minimum L^q distance estimators based on Stein operators, and [113] considered a minimum distance estimator based on likelihood ratios estimated through Stein operators.

Most notably, estimators of the form in (20) are useful for unnormalized likelihood models, since Stein operators usually rely on unnormalized densities. When the parametric family is in some exponential family, the Langevin Stein discrepancies become quadratic forms in ϑ, which implies that the optimizer can be obtained in closed-form. Matsubara [117] built on this idea to propose a fully conjugate generalized Bayesian approach for unnormalized densities. This was latter extended to discrete data settings by [118].

5. NEW METHODS FOR AND INSIGHTS IN STATISTICAL INFERENCE VIA STEIN OPERATORS AND STEIN DISCREPANCIES

In this section, we focus on statistical inference and show how tools from Stein's method have been put to use to build new powerful tools as well as to gain novel insights in long-existing procedures. Section 5.1 is concerned with new goodness-of-fit tests obtained from Stein discrepancies, while Section 5.2 deals with composite goodness-of-fit tests based on Stein operators. These goodness-of-fit tests lend themselves very naturally to a further Stein-based analysis, namely to quantify the distance, at a given finite sample size n, between the asymptotic distribution and the unknown exact distribution, hereby getting an idea of how good the asymptotic approximation actually is. More generally, recently Stein's method has been used to quantify the asymptotic behaviour of statistical estimators and hypothesis tests, which is the topic of Section 5.3. In a similar vein, Section 5.4 deals with the Bayesian setting, hereby showing a new way to quantify the finite-sample effect of the prior choice.

5.1 Goodness-of-Fit Tests from Stein Discrepancies

Suppose we would like to test for the null hypothesis $H_0 : Q = P$ based on realizations $\{x_1, \ldots, x_n\}$ from Q (which may or may not be independent). Chwialkowski, Strathmann and Gretton and Lin, Lee and Jordan [42, 110] proposed to use a KSD as test statistic, which is particularly powerful for a distribution P whose density is known up to a normalizing constant.

These tests are motivated by the general approach of using IPMs within a hypothesis testing framework. In particular, an influential line of work in machine learning has

been to use IPMs with a kernel-based underlying function class, leading to the so-called MMD hypothesis tests [74, 75]. This approach has previously been used to test for a range of hypotheses, including two-sample tests and independence tests. Their popularity can be explained through their generality: they only rely on the choice of a kernel and samples from both P and Q, and can hence be implemented for a wide range of problems.

In the goodness-of-fit setting, when P has a density known up to normalizing, sampling from P may introduce unnecessary variance to our test statistic. The test is also somewhat suboptimal since it does not use any specific properties of P. It is therefore natural to consider the use of Stein operators in this setting. This can be achieved by selecting an IPM whose underlying function class is of the form $\mathcal{T}g$ for g in some Stein set $\mathcal{G}$. When using a Langevin–Stein operator and kernel Stein set, this leads to the Langevin KSD of Example 4, which is the case most often considered in this literature. Recalling the expression for the population Langevin KSD given in equation (10), an unbiased estimate of the squared KSD takes the convenient form of a U-statistic:

$$\widehat{\mathrm{KSD}}_k^2(Q) = \frac{2}{n(n-1)} \sum_{i<j} k_P(x_i, x_j).$$

This estimate can be used as a test statistic. It is degenerate under the null hypothesis that $Q = P$, and nondegenerate under the alternative. As a result, when the sample is i.i.d. the asymptotic behavior of the statistic is obtained via standard results [140]. Unfortunately, the asymptotic distribution under the null is a function of the eigenvalues of k_P with respect to Q, which are rarely computable in closed form. Nonetheless, a test threshold of asymptotic level α may be obtained using a wild-bootstrap procedure on a V-statistic approximation to the KSD. The wild bootstrap may also be adapted to the case where the sample from Q is not i.i.d., but satisfies a τ-mixing condition [97]. This is especially helpful when the goodness-of-fit test is used for bias quantification of approximate MCMC procedures since these are not i.i.d. [42], Section 4.

In order to guarantee consistency of the tests, it is of interest to establish when the KSD uniquely determines whether Q and P correspond. We refer to [42], Theorem 2.2: if k is C_0-universal ([32], Definition 4.1), and if $\mathbb{E}_{X\sim Q}[\|\nabla(\log(p(X)/q(X)))\|_2^2] < \infty$, then $\mathrm{KSD}_k(q) = 0$ if and only if $P = Q$. Many popular kernels, including the exponentiated quadratic (Gaussian) kernel $k(x, y) = \exp(-\|x - y\|_2^2/l^2)$ $(l > 0)$, are C_0-universal. We however recall the result of [70] that stronger conditions on the kernel are required when one wishes to control *weak convergence* to a target using the KSD.

Apart from U-statistic based tests, alternative tests exist, which can be computed in linear time, using adaptive kernel Stein features that indicate where the data distribution Q differs from the model P [90], or importance sampling approaches [87]. In the former case, the features are learned on a held-out sample from Q, so as to maximize the power of the resulting test.

Stein goodness-of-fit tests may also be defined for right-censored time-to-event data. Indeed, [56] defined three Stein operators for this setting, which exploit well-known identities in survival analysis that arise from the underlying structure of the data. The first is the *Survival Stein Operator*, which arises from a direct application of the Langevin–Stein operator to the density function; the second, the *martingale Stein operator*, applies a well-known martingale equality in a similar fashion as for log-rank statistics; and the third, the *Proportional Stein Operator*, applies the Langevin–Stein operator to the hazard function. The resulting Stein tests were used to validate models of survival times in real-world medical studies of leukemia, chronic granulotamous disease, ovarian cancer and lung cancer.

For discrete distributions, KSD tests include the work of [164], which derives a discrepancy for discrete data and that of [165] that focuses on point processes. For exponential random graph models when only one network observation is available, [163] use the Stein operator for exponential random graph models from [136] as basis for a kernelized Stein discrepancy test.

5.2 Composite Goodness-of-Fit Tests from Stein Operators

Consider the classical problem of testing the composite null hypothesis $H_0 : Q \in \mathscr{P}_\Theta = \{P_\vartheta : \vartheta \in \Theta\}$, where $\Theta \subset \mathbb{R}^s$, $s \in \mathbb{N}$, is an open parameter space, and P_ϑ is the unique distribution corresponding to $\vartheta \in \Theta$ in the parametric family $\mathscr{P}_\Theta$. This hypothesis is to be tested based on an i.i.d. sample $\{x_1, \ldots, x_n\}$ from Q. For example, tests for normality fall into this category.

For this problem, test statistics based on parametric families of Stein operators as in [100] have been developed as follows. Let $\{\mathcal{T}_\vartheta : \vartheta \in \Theta\}$ be a family of Stein operators characterizing the family $\mathscr{P}_\Theta$. By the Stein characterization, we have $\mathbb{E}_{X\sim P_\vartheta}[(\mathcal{T}_\vartheta g)(X)] = 0$ for all $g \in \mathcal{G}(\mathcal{T}_\vartheta)$ and $\vartheta \in \Theta$. A natural extension of the KSD framework to the composite hypothesis was proposed by [91] and built on the minimum Stein discrepancy estimators of [21]. However, we will focus this section on an alternative test for the composite hypothesis based on a suitable set of test functions $\mathcal{G} = \{g_t(x) : t \in M\}$, $M \subset \mathbb{R}^d$ given by the weighted L^2 statistic

$$T_n = n \int_M \left\| \frac{1}{n} \sum_{i=1}^n (\mathcal{T}_{\hat{\vartheta}_n} g_t)(x_i) - \mathbb{E}_{X\sim P_{\hat{\vartheta}_n}}\left[(\mathcal{T}_{\hat{\vartheta}_n} g_t)(X)\right] \right\|^2 \omega(t)\,\mathrm{d}t \tag{21}$$

$$= n \int_M \left\| \frac{1}{n} \sum_{i=1}^{n} (\mathcal{T}_{\widehat{\vartheta}_n} g_t)(x_i) \right\|^2 \omega(t)\, \mathrm{d}t,$$

where $\widehat{\vartheta}_n$ is a consistent estimator of ϑ, $\|\cdot\|$ is a suitable norm and $\omega : M \to [0, \infty)$ is a positive weight function satisfying some weak integrability conditions. Heuristically, T_n should be close to 0 if and only if the data stems from $\mathscr{P}_\Theta$, and we hence reject H_0 for large values of T_n.

Henze and Visagie [81] implicitly used such a test for multivariate normality based on the classical Stein operator $\mathcal{T}$ from Example 1. An alternative test of univariate normality based on $\mathcal{T}$ from Example 1 is proposed in [50], but in this case test functions of the form $\{g_t(x) = \exp(\mathrm{i}tx) : t \in \mathbb{R}\}$ (i.e., related to characteristic functions) are used. Dörr, Ebner and Henze [48] also introduce a test of multivariate normality, based on $(\mathcal{T}g)(x) = -\Delta g(x) + (\|x\|_2^2 - d)g(x)$ (where Δ denotes the Laplacian), and the class of test functions $\{g_t(x) = \exp(\mathrm{i}t^\intercal x) : t \in \mathbb{R}^d\}$. There are considerable differences in power against specific alternatives between the tests, especially w.r.t. the choice of test functions. For a comparative Monte Carlo simulation study, see [51].

In a similar vein, [29] and [31] provide new characterizations of continuous and discrete parametric families of distributions through the density approach for novel tests for univariate normality [28], the gamma family [27] and the inverse Gaussian law [2]. Note that other test statistics of type (21) based on Stein operators are implicitly proposed in tests for parametric families, although originally motivated by characterizing (partial) differential equations for integral transforms; see, for instance, [18] for a test of exponentiality, [19] for a test of Poissonity and [80] for a test of the gamma law.

The expression in (21) can be thought of as a weighted L^2-difference between the expectation of $\mathcal{T}_{\widehat{\vartheta}_n} g_t$ under $P_{\widehat{\vartheta}_n}$ and $Q_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$. This is in contrast with the IPMs, such as the KSD of the previous section, which measure worst-case types of differences (recall equation (5) which considers the supremum instead of an average). As a result, although the tests in Sections 5.1 and 5.2 are both based on Stein operators, they use these in rather different manners. The tests in Section 5.1 use an RKHS setting, which allows for a rich set of alternative distributions. For the tests in Section 5.2, the benefit of considering the structure of a L^2-Hilbert space lies in the fact that the central limit theorem for Hilbert-space valued random elements can be exploited to derive limit distributions under H_0, as well as fixed and contiguous alternatives.

5.3 Maximum Likelihood Estimation and Chi-Square Tests

With Stein's method it is possible to give explicit bounds at finite sample size n to the asymptotic approximation of estimators and test statistics. The arguably most famous example is the asymptotic normal distribution for maximum likelihood estimators (MLEs) under fairly general conditions. For example, in the simple case of $X_1, X_2, \ldots, X_n$ being i.i.d. random variables from a single-parameter distribution, then for $Z \sim \mathrm{N}(0, 1)$, and under classical regularity conditions,

$$\sqrt{n i(\theta_0)}\big(\hat{\theta}_n(\boldsymbol{X}) - \theta_0\big) \to_d Z, \quad \text{as } n \to \infty, \tag{22}$$

where $\to_d$ denotes convergence in distribution. Starting with the single-parameter case, under some natural regularity assumptions, which we do not detail here, [6] obtain general bounds w.r.t. the bounded Wasserstein distance as follows. Let $W_n := \sqrt{n i(\theta_0)}(\hat{\theta}_n(X) - \theta_0)$. Then the interest is to find upper bounds on $|\mathbb{E}[h(W_n)] - \mathbb{E}[h(Z)]|$, where $h \in \mathcal{H}_{\mathrm{bW}}$ as in Remark 1. The general idea is to represent the standardized MLE in such a way that it contains a quantity which is a sum of independent random variables plus a term that can be controlled. The part involving the sum is handled via a classical use of Stein's method. While the underlying random sample $X_1, \ldots, X_n$ are assumed i.i.d. in [6], they are locally dependent in [3].

As an illustration, consider the exponential distribution in its canonical form. The probability density function is $f(x|\theta) = \theta \exp\{-\theta x\}$ for $x > 0$ and the unique MLE for θ is $\hat{\theta}_n(X) = 1/\overline{X}$, the inverse of the sample average. Then [6] established that

$$d_{\mathrm{bW}}\big(\mathcal{L}(W_n), \mathcal{L}(Z)\big) \leq \frac{4.41456}{\sqrt{n}} + \frac{8(n+2)(1+\sqrt{n})}{(n-1)(n-2)},$$

with $Z \sim \mathrm{N}(0, 1)$ and $W_n := \sqrt{n i(\theta_0)}(\hat{\theta}_n(X) - \theta_0)$; here, $i(\theta_0)$ is the expected Fisher information for one variable. This bound is explicit and of the order $n^{-1/2}$. Using the delta method combined with Stein's method, [5] give an explicit bound for MLEs, which are a smooth function of a sum of independent terms. This result is generalized to the multivariate case in [4].

Since the MLE can be used as a basis for likelihood ratio tests, which under regularity assumptions follow approximately a chi-square distribution, it is natural to measure the finite-sample approximation error of such tests. An explicit general bound of order $O(n^{-1/2})$ is obtained in [7] using Stein's method.

Explicit bounds on chi-square approximations for Pearson's chi-square test for goodness-of-fit of categorical data are obtained in [60], and more generally the power divergence family of statistics in [59]. Gaunt and Reinert [61] provided explicit bounds of the order r/n to quantify the chi-square approximation with $r - 1$ degrees of freedom to Friedman's statistic.

5.4 The Effect of Prior Choice on the Posterior in Bayesian Statistics

In Bayesian statistics, [46] proved that, under certain regularity conditions and for large sample sizes, the

choice of a prior distribution gets irrelevant for posterior inference. With the help of Stein's method, [62, 99] complemented this result by estimating prior sensitivity for fixed (and often small) sample sizes by quantifying the Wasserstein distance between posterior distributions arising from two distinct priors in the one-dimensional one-parameter setting. The argument was extended to the multivariate setting in [121].

Let us start by fixing the notation. Suppose that the observations $X_1, \dots, X_n$ are i.i.d. from a parametric model with scalar parameter of interest, which we model as some random variable Θ. Now, assume we have two distinct (possibly improper) prior densities $p_1(\theta)$ and $p_2(\theta)$ for the random quantity Θ. The resulting posterior densities for Θ can be expressed as

$$p_i(\theta; x) = \kappa_i(x) p_i(\theta) \ell(\theta; x), \quad i = 1, 2, \tag{23}$$

where κ_1 and κ_2 are normalizing constants. Denote by (Θ_1, P_1) and (Θ_2, P_2) pairs of random variables and cumulative distribution functions, which correspond to the densities $p_1(\theta; x)$ and $p_2(\theta; x)$, respectively. We assume that the densities $p_1(\theta; x)$ and $p_2(\theta; x)$ are *nested*, so that the support of one is included in the support of the other. We suppose $I_2 \subseteq I_1$, which allows us to write $p_2(\theta; x) = \frac{\kappa_2(x)}{\kappa_1(x)} \rho(\theta) p_1(\theta; x)$, where $\rho(\theta) = p_2(\theta)/p_1(\theta)$ is the ratio of prior densities. The key idea relies on the elementary identity

$$\begin{aligned} &\frac{\frac{\mathrm{d}}{\mathrm{d}\theta}(p_2(\theta; x) f(\theta))}{p_2(\theta; x)} \\ &\quad = \frac{\frac{\mathrm{d}}{\mathrm{d}\theta}(p_1(\theta; x) f(\theta))}{p_1(\theta; x)} + \left(\frac{\mathrm{d}}{\mathrm{d}\theta} \log(\rho(\theta)) \right) f(\theta), \end{aligned}$$

which is an immediate consequence of (23) and the nestedness of the densities. This identity no longer involves the normalizing constants and it relates the density operators of $p_1(\cdot; x)$ and $p_2(\cdot; x)$ in such a way that, with f_h a solution to the Stein equation $h(x) - \mathbb{E}_{X_1 \sim P_1} h(X_1) = \mathcal{T}_1 f_h(x)$, we get (writing shorthand $\mathbb{E}_{P_j}$ for $\mathbb{E}_{\Theta_j \sim P_j}$, $j = 1, 2$)

$$\begin{aligned} &\mathbb{E}_{P_2}[h(\Theta_2)] - \mathbb{E}_{P_1}[h(\Theta_1)] \\ &\quad = \mathbb{E}_{P_2}[\mathcal{T}_1 f_h(\Theta_2)] \\ &\quad = \mathbb{E}_{P_2}[(\mathcal{T}_1 - \mathcal{T}_2) f_h(\Theta_2)] \\ &\quad = \mathbb{E}_{P_2}\left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log(\rho(\Theta_2)) f_h(\Theta_2) \right] \end{aligned}$$

(the second equality holds because, by definition, $\mathbb{E}_{P_2}[\mathcal{T}_2 f_h(\Theta_2)] = 0$). Thus, bounding an IPM generated by some class $\mathcal{H}$ between Θ_2 and Θ_1 can be achieved by bounding $\mathbb{E}_{P_2}[\frac{\mathrm{d}}{\mathrm{d}\theta} \log(\rho(\theta))|_{\theta=\Theta_2} f_h(\Theta_2)]$ over all $h \in \mathcal{H}$.

For the sake of illustration, consider normal data with fixed variance σ^2, and the mean the parameter of interest. Ley, Reinert and Swan [99] compare a normal $\mathrm{N}(\mu, \delta^2)$ prior for the location parameter (the conjugate prior in this situation) with a uniform prior. They bounded the Wasserstein distance between the resulting posteriors P_1 and P_2 by

$$\begin{aligned} \frac{\sigma^2}{n\delta^2 + \sigma^2} |\overline{x} - \mu| &\le d_{\mathrm{W}}(P_1, P_2) \\ &\le \frac{\sigma^2}{n\delta^2 + \sigma^2} |\overline{x} - \mu| \\ &\quad + \frac{\sqrt{2}}{\sqrt{\pi}} \frac{\sigma^3}{n\delta\sqrt{n\delta^2 + \sigma^2}} \end{aligned}$$

with $\overline{x} = n^{-1} \sum_{i=1}^n x_i$ the sample average. Both bounds are of the order of $O(n^{-1})$ and are easily interpreted: the better the initial guess of the prior, meaning here of the location, the smaller the bounds, and hence the smaller the influence of the prior.

6. CONCLUSION

The goal of this paper is to highlight some recent developments in computational statistics that have been accomplished via tools inherited from Stein's method. Moreover, this paper illustrates that there is considerable scope for more interplay between the research strand on how to set up Stein operators and that of devising computable Stein discrepancies and related algorithms. For example, for a given target distribution, it is mostly an open problem which Stein operator and class to choose so as to obtain a computable Stein discrepancy, which is most useful for the problem at hand. This answer may differ depending on whether we want to construct a hypothesis test, develop a sampling method, or measure sample quality; a step in this direction is taken in [162]. Section 5 highlights how Stein's method can be brought to fruition not only to devise estimators but also to quantify their behavior. There is plenty of scope for analyzing the procedures and estimators from Sections 4.4 and 5.1 to obtain quantitative bounds on their performance.

The list of results given in this paper are but a mere sample of the ongoing activity in this newly established area of research at the boundary between probability, functional analysis, data science and computational statistics. For instance, Stein's method has been used for designing sampling-based algorithms for nonconvex optimization [52], or for learning semiparametric multiindex models in high dimensions [166]. In Bayesian statistics, Stein discrepancies have been used as variational objectives for posterior approximation (e.g., [57, 86, 132]).

A complete exhaustive description of all recent developments in this area is an impossible task within the constrained space of a review paper such as this one. Yet, we hope that the range of problems which are addressed in this paper show the versatility of Stein's method, and the promise that it holds for further exciting developments.

ACKNOWLEDGMENTS

The authors thank the Editor, Associate Editor and two anonymous reviewers for helpful comments that led to a clear improvement of the presentation of this paper.

Christophe Ley is the corresponding author.

FUNDING

AA was supported by a start-up grant from the University of Cyprus. AB was supported by the UK Defence Science and Technology Laboratory (Dstl) and Engineering and Physical Research Council (EPSRC) under the grant EP/R018413/2. FXB and CJO were supported by the Lloyds Register Foundation Programme on Data-Centric Engineering and The Alan Turing Institute under the EPSRC grant EP/N510129/1. AG was supported by the Gatsby Charitable Foundation. RG was supported by a Dame Kathleen Ollerenshaw Research Fellowship. FG and CL were supported by a BOF Starting Grant of Ghent University. QL was supported in part by NSF CAREER No. 1846421. GR was supported in part by EP/T018445/1 and EP/R018472/1. YS was supported in part by CDR/OL J.0197.20 from FRS-FNRS.

REFERENCES

[1] AHN, S., KORATTIKARA, A. and WELLING, M. (2012). Bayesian posterior sampling via stochastic gradient Fisher scoring. In *International Conference on Machine Learning* (*ICML*) 1591–1598.

[2] ALLISON, J. S., BETSCH, S., EBNER, B. and VISAGIE, I. J. H. (2022). On testing the adequacy of the inverse Gaussian distribution. *Mathematics* **10** 350.

[3] ANASTASIOU, A. (2017). Bounds for the normal approximation of the maximum likelihood estimator from m-dependent random variables. *Statist. Probab. Lett.* **129** 171–181. MR3688530 https://doi.org/10.1016/j.spl.2017.04.022

[4] ANASTASIOU, A. and GAUNT, R. E. (2021). Wasserstein distance error bounds for the multivariate normal approximation of the maximum likelihood estimator. *Electron. J. Stat.* **15** 5758–5810. MR4355697 https://doi.org/10.1214/21-ejs1920

[5] ANASTASIOU, A. and LEY, C. (2017). Bounds for the asymptotic normality of the maximum likelihood estimator using the delta method. *ALEA Lat. Am. J. Probab. Math. Stat.* **14** 153–171. MR3622464

[6] ANASTASIOU, A. and REINERT, G. (2017). Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli* **23** 191–218. MR3556771 https://doi.org/10.3150/15-BEJ741

[7] ANASTASIOU, A. and REINERT, G. (2020). Bounds for the asymptotic distribution of the likelihood ratio. *Ann. Appl. Probab.* **30** 608–643. MR4108117 https://doi.org/10.1214/19-AAP1510

[8] ANDRADÓTTIR, S., HEYMAN, D. P. and OTT, T. J. (1993). Variance reduction through smoothing and control variates for Markov chain simulations. *ACM Trans. Model. Comput. Simul.* **3** 167–189.

[9] ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.* **68** 337–404. MR0051437 https://doi.org/10.2307/1990404

[10] ARRAS, B. and HOUDRÉ, C. (2019). *On Stein's Method for Infinitely Divisible Laws with Finite First Moment*. *SpringerBriefs in Probability and Mathematical Statistics*. Springer, Cham. MR3931309

[11] ASSARAF, R. and CAFFAREL, M. (1999). Zero-variance principle for Monte Carlo algorithms. *Phys. Rev. Lett.* **83** 4682.

[12] BANERJEE, T., LIU, Q., MUKHERJEE, G. and SUN, W. (2021). A general framework for empirical Bayes estimation in discrete linear exponential family. *J. Mach. Learn. Res.* **22** 67. MR4253760

[13] BARBOUR, A. D. (1988). Stein's method and Poisson process convergence. *J. Appl. Probab.* **25A** 175–184.

[14] BARBOUR, A. D. (1990). Stein's method for diffusion approximations. *Probab. Theory Related Fields* **84** 297–322. MR1035659 https://doi.org/10.1007/BF01197887

[15] BARBOUR, A. D. and CHEN, L. H. Y. (2014). Stein's (magic) method. ArXiv preprint. Available at arXiv:1411.1179.

[16] BARBOUR, A. D., HOLST, L. and JANSON, S. (1992). *Poisson Approximation*. *Oxford Studies in Probability* **2**. The Clarendon Press, New York. MR1163825

[17] BARBOUR, A. D. and XIA, A. (1999). Poisson perturbations. *ESAIM Probab. Stat.* **3** 131–150. MR1716120 https://doi.org/10.1051/ps:1999106

[18] BARINGHAUS, L. and HENZE, N. (1991). A class of consistent tests for exponentiality based on the empirical Laplace transform. *Ann. Inst. Statist. Math.* **43** 551–564. MR1143640 https://doi.org/10.1007/BF00053372

[19] BARINGHAUS, L. and HENZE, N. (1992). A goodness of fit test for the Poisson distribution based on the empirical generating function. *Statist. Probab. Lett.* **13** 269–274. MR1160747 https://doi.org/10.1016/0167-7152(92)90033-2

[20] BARP, A. A. (2020). The Bracket Geometry of Statistics Ph.D. thesis Imperial College London.

[21] BARP, A. A., BRIOL, F. X., DUNCAN, A. B., GIROLAMI, M. and MACKEY, L. (2019). Minimum Stein discrepancy estimators. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 12964–12976.

[22] BARP, A. A., OATES, C., PORCU, E. and GIROLAMI, M. (2018). A Riemannian-Stein kernel method. ArXiv preprint. Available at arXiv:1810.04946.

[23] BELOMESTNY, D., IOSIPOI, L., MOULINES, E., NAUMOV, A. and SAMSONOV, S. (2020). Variance reduction for Markov chains with application to MCMC. *Stat. Comput.* **30** 973–997. MR4108687 https://doi.org/10.1007/s11222-020-09931-z

[24] BELOMESTNY, D., IOSIPOI, L. and ZHIVOTOVSKIY, N. (2017). Variance reduction via empirical variance minimization: Convergence and complexity. ArXiv preprint. Available at arXiv:1712.04667.

[25] BELOMESTNY, D., MOULINES, E., SHAGADATOV, N. and URUSOV, M. (2019). Variance reduction for MCMC methods via martingale representations. ArXiv preprint. Available at arXiv:1903.07373.

[26] BERLINET, A. and THOMAS-AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic, Boston, MA. MR2239907 https://doi.org/10.1007/978-1-4419-9096-9

[27] BETSCH, S. and EBNER, B. (2019). A new characterization of the Gamma distribution and associated goodness-of-fit tests. *Metrika* **82** 779–806. MR4008662 https://doi.org/10.1007/s00184-019-00708-7

[28] BETSCH, S. and EBNER, B. (2020). Testing normality via a distributional fixed point property in the Stein characterization. *TEST* **29** 105–138. MR4063385 https://doi.org/10.1007/s11749-019-00630-0

[29] BETSCH, S. and EBNER, B. (2021). Fixed point characterizations of continuous univariate probability distributions and their applications. *Ann. Inst. Statist. Math.* **73** 31–59. MR4205241 https://doi.org/10.1007/s10463-019-00735-1

[30] BETSCH, S., EBNER, B. and KLAR, B. (2021). Minimum L^q-distance estimators for non-normalized parametric models. *Canad. J. Statist.* **49** 514–548. MR4267931 https://doi.org/10.1002/cjs.11574

[31] BETSCH, S., EBNER, B. and NESTMANN, F. (2022). Characterizations of non-normalized discrete probability distributions and their application in statistics. *Electron. J. Stat.* **16** 1303–1329. MR4381061 https://doi.org/10.1214/22-ejs1983

[32] CARMELI, C., DE VITO, E., TOIGO, A. and UMANITÀ, V. (2010). Vector valued reproducing kernel Hilbert spaces and universality. *Anal. Appl. (Singap.)* **8** 19–61. MR2603770 https://doi.org/10.1142/S0219530510001503

[33] CHATTERJEE, S. (2014). A short survey of Stein's method. In *Proceedings of the International Congress of Mathematicians—Seoul* 2014. *Vol. IV* 1–24. Kyung Moon Sa, Seoul. MR3727600

[34] CHEN, C., ZHANG, R., WANG, W., LI, B. and CHEN, L. (2018). A unified particle-optimization framework for scalable Bayesian sampling. In *Uncertainty in Artificial Intelligence* (*UAI*).

[35] CHEN, L. H. and RÖLLIN, A. (2010). Stein couplings for normal approximation. ArXiv preprint. Available at arXiv:1003.6039.

[36] CHEN, L. H. Y. (1975). Poisson approximation for dependent trials. *Ann. Probab.* **3** 534–545. MR0428387 https://doi.org/10.1214/aop/1176996359

[37] CHEN, L. H. Y., GOLDSTEIN, L. and SHAO, Q.-M. (2011). *Normal Approximation by Stein's Method*. *Probability and Its Applications* (*New York*). Springer, Heidelberg. MR2732624 https://doi.org/10.1007/978-3-642-15007-4

[38] CHEN, P., WU, K., CHEN, J., O'LEARY-ROSEBERRY, T. and GHATTAS, O. (2019). Projected Stein variational Newton: A fast and scalable Bayesian inference method in high dimensions. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 15130–15139.

[39] CHEN, W. Y., BARP, A. A., BRIOL, F.-X., GORHAM, J., GIROLAMI, M., MACKEY, L. and OATES, C. J. (2019). Stein point Markov chain Monte Carlo. In *International Conference on Machine Learning* (*ICML*) 1011–1021.

[40] CHEN, W. Y., MACKEY, L., GORHAM, J., BRIOL, F.-X. and OATES, C. J. (2018). Stein points. In *International Conference on Machine Learning* (*ICML*) 844–853.

[41] CHEWI, S., GOUIC, T. L., LU, C., MAUNU, T. and RIGOLLET, P. (2020). SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence. In *Advances on Neural Information Processing Systems* (*NeurIPS*).

[42] CHWIALKOWSKI, K., STRATHMANN, H. and GRETTON, A. (2016). A kernel test of goodness of fit. In *International Conference on Machine Learning* (*ICML*) 2606–2615.

[43] COURTADE, T. A., FATHI, M. and PANANJADY, A. (2019). Existence of Stein kernels under a spectral gap, and discrepancy bounds. *Ann. Inst. Henri Poincaré Probab. Stat.* **55** 777–790. MR3949953 https://doi.org/10.1214/18-aihp898

[44] DELLAPORTAS, P. and KONTOYIANNIS, I. (2012). Control variates for estimation based on reversible Markov chain Monte Carlo samplers. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 133–161. MR2885843 https://doi.org/10.1111/j.1467-9868.2011.01000.x

[45] DETOMMASO, G., CUI, T., MARZOUK, Y., SCHEICHL, R. and SPANTINI, A. (2018). A Stein variational Newton method. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 9169–9179.

[46] DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates (with a discussion and a rejoinder by the authors). *Ann. Statist.* **14** 1–67. MR0829555 https://doi.org/10.1214/aos/1176349830

[47] DIACONIS, P. and HOLMES, S., eds. (2004). *Stein's Method*: *Expository Lectures and Applications*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **46**.

[48] DÖRR, P., EBNER, B. and HENZE, N. (2021). A new test of multivariate normality by a double estimation in a characterizing PDE. *Metrika* **84** 401–427. MR4233599 https://doi.org/10.1007/s00184-020-00795-x

[49] DUNCAN, A., NÜSKEN, N. and SZPRUCH, L. (2019). On the geometry of Stein variational gradient descent. ArXiv preprint. Available at arXiv:1912.00894.

[50] EBNER, B. (2021). On combining the zero bias transform and the empirical characteristic function to test normality. *ALEA Lat. Am. J. Probab. Math. Stat.* **18** 1029–1045. MR4282180 https://doi.org/10.30757/alea.v18-38

[51] EBNER, B. and HENZE, N. (2020). Tests for multivariate normality—a critical review with emphasis on weighted L^2-statistics. *TEST* **29** 845–892. MR4182841 https://doi.org/10.1007/s11749-020-00740-0

[52] ERDOGDU, M. A., MACKEY, L. and SHAMIR, O. (2018). Global non-convex optimization with discretized diffusions. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 9694–9703.

[53] FANG, X., SHAO, Q.-M. and XU, L. (2019). Multivariate approximations in Wasserstein distance by Stein's method and Bismut's formula. *Probab. Theory Related Fields* **174** 945–979. MR3980309 https://doi.org/10.1007/s00440-018-0874-5

[54] FATHI, M., GOLDSTEIN, L., REINERT, G. and SAUMARD, A. (2020). Relaxing the Gaussian assumption in shrinkage and SURE in high dimension. ArXiv preprint. Available at arXiv:2004.01378.

[55] FENG, Y., WANG, D. and LIU, Q. (2017). Learning to draw samples with amortized Stein variational gradient descent. In *Uncertainty in Artificial Intelligence* (*UAI*).

[56] FERNÁNDEZ, T., RIVERA, N., XU, W. and GRETTON, A. (2020). Kernelized Stein discrepancy tests of goodness-of-fit for time-to-event data. In *International Conference on Machine Learning* (*ICML*).

[57] FISHER, M. A., NOLAN, T. H., GRAHAM, M. M., PRANGLE, D. and OATES, C. J. (2021). Measure transport with kernel Stein discrepancy. In *International Conference on Artificial Intelligence and Statistics* (*AISTATS*).

[58] GAUNT, R. E. (2017). On Stein's method for products of normal random variables and zero bias couplings. *Bernoulli* **23** 3311–3345. MR3654808 https://doi.org/10.3150/16-BEJ848

[59] GAUNT, R. E. (2022). Bounds for the chi-square approximation of the power divergence family of statistics. *J. Appl. Probab.*

[60] GAUNT, R. E., PICKETT, A. M. and REINERT, G. (2017). Chi-square approximation by Stein's method with application to Pearson's statistic. *Ann. Appl. Probab.* **27** 720–756. MR3655852 https://doi.org/10.1214/16-AAP1213

[61] GAUNT, R. E. and REINERT, G. (2021). Bounds for the chi-square approximation of Friedman's statistic by Stein's method. ArXiv preprint. Available at arXiv:2111.00949.

[62] GHADERINEZHAD, F. and LEY, C. (2019). Quantification of the impact of priors in Bayesian statistics via Stein's method. *Statist. Probab. Lett.* **146** 206–212. MR3884714 https://doi.org/10.1016/j.spl.2018.11.012

[63] GIBBS, A. L. and SU, F. E. (2002). On choosing and bounding probability metrics. *Int. Stat. Rev.* **70** 419–435.

[64] GOLDSTEIN, L. and REINERT, G. (2005). Distributional transformations, orthogonal polynomials, and Stein characterizations. *J. Theoret. Probab.* **18** 237–260. MR2132278 https://doi.org/10.1007/s10959-004-2602-6

[65] GOLDSTEIN, L. and REINERT, G. (2013). Stein's method for the beta distribution and the Pólya-Eggenberger urn. *J. Appl. Probab.* **50** 1187–1205. MR3161381 https://doi.org/10.1239/jap/1389370107

[66] GONG, C., PENG, J. and LIU, Q. (2019). Quantile Stein variational gradient descent for parallel Bayesian optimization. In *International Conference on Machine Learning* (*ICML*) 2347–2356.

[67] GONG, W., LI, Y. and HERNÁNDEZ-LOBATO, J. M. (2021). Sliced kernelized Stein discrepancy. In *International Conference on Learning Representations* (*ICLR*).

[68] GORHAM, J., DUNCAN, A. B., VOLLMER, S. J. and MACKEY, L. (2019). Measuring sample quality with diffusions. *Ann. Appl. Probab.* **29** 2884–2928. MR4019878 https://doi.org/10.1214/19-AAP1467

[69] GORHAM, J. and MACKEY, L. (2015). Measuring sample quality with Stein's method. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 226–234. Curran Associates, Red Hook.

[70] GORHAM, J. and MACKEY, L. (2017). Measuring sample quality with kernels. In *International Conference on Machine Learning* (*ICML*) 1292–1301.

[71] GORHAM, J., RAJ, A. and MACKEY, L. (2020). Stochastic Stein discrepancies. In *Advances on Neural Information Processing Systems* (*NeurIPS*).

[72] GÖTZE, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19** 724–739. MR1106283

[73] GRATHWOHL, W., WANG, K. C., JACOBSEN, J. H., DUVENAUD, D. and ZEMEL, R. (2020). Learning the Stein discrepancy for training and evaluating energy-based models without sampling. In *International Conference on Machine Learning* 9485–9499.

[74] GRETTON, A., BORGWARDT, K. M., RASCH, M., SCHÖLKOPF, B. and SMOLA, A. J. (2006). A kernel method for the two-sample-problem. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 513–520.

[75] GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

[76] HAARNOJA, T., TANG, H., ABBEEL, P. and LEVINE, S. (2017). Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning* (*ICML*) 1352–1361.

[77] HAN, J. and LIU, Q. (2017). Stein variational adaptive importance sampling. In *Uncertainty in Artificial Intelligence* (*UAI*).

[78] HAN, J. and LIU, Q. (2018). Stein variational gradient descent without gradient. In *International Conference on Machine Learning* (*ICML*) 1900–1908.

[79] HENDERSON, S. G. and SIMON, B. (2004). Adaptive simulation using perfect control variates. *J. Appl. Probab.* **41** 859–876. MR2074828 https://doi.org/10.1017/s0021900200020593

[80] HENZE, N., MEINTANIS, S. G. and EBNER, B. (2012). Goodness-of-fit tests for the gamma distribution based on the empirical Laplace transform. *Comm. Statist. Theory Methods* **41** 1543–1556. MR3003807 https://doi.org/10.1080/03610926.2010.542851

[81] HENZE, N. and VISAGIE, J. (2020). Testing for normality in any dimension based on a partial differential equation involving the moment generating function. *Ann. Inst. Statist. Math.* **72** 1109–1136. MR4137748 https://doi.org/10.1007/s10463-019-00720-8

[82] HINTON, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14** 1771–1800.

[83] HODGKINSON, L., SALOMONE, R. and ROOSTA, F. (2020). The reproducing Stein kernel approach for post-hoc corrected sampling. ArXiv preprint. Available at arXiv:2001.09266.

[84] HOLMES, S. (2004). Stein's method for birth and death chains. In *Stein's Method*: *Expository Lectures and Applications*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **46** 45–67. IMS, Beachwood, OH. MR2118602

[85] HOLMES, S. and REINERT, G. (2004). Stein's method for the bootstrap. In *Stein's Method*: *Expository Lectures and Applications*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **46** 95–136. IMS, Beachwood, OH. MR2118605

[86] HU, T., CHEN, Z., SUN, H., BAI, J., YE, M. and CHENG, G. (2018). Stein neural sampler. ArXiv preprint. Available at arXiv:1810.03545.

[87] HUGGINS, J. H. and MACKEY, L. (2018). Random feature Stein discrepancies. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 1899–1909.

[88] HYVÄRINEN, A. (2005). Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.* **6** 695–709. MR2249836

[89] JAMES, W. and STEIN, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. I* 361–379. Univ. California Press, Berkeley, CA. MR0133191

[90] JITKRITTUM, W., XU, W., SZABO, Z., FUKUMIZU, K. and GRETTON, A. (2017). A linear-time kernel goodness-of-fit test. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 261–270.

[91] KEY, O., FERNANDEZ, T., GRETTON, A. and BRIOL, F.-X. (2021). Composite goodness-of-fit tests with kernels. In *NeurIPS* 2021 *Workshop Your Model Is Wrong*: *Robustness and Misspecification in Probabilistic Modeling*. Available at arXiv:2111.10275.

[92] KIM, T., YOON, J., DIA, O., KIM, S., BENGIO, Y. and AHN, S. (2018). Bayesian model-agnostic meta-learning. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 7332–7342.

[93] KORATTIKARA, A., CHEN, Y. and WELLING, M. (2014). Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proceedings of International Conference on Machine Learning* (*ICML*). *ICML*'14.

[94] KORBA, A., SALIM, A., ARBEL, M., LUISE, G. and GRETTON, A. (2020). A non-asymptotic analysis for Stein variational gradient descent. In *Advances in Neural Information Processing Systems* (*NeurIPS*) **33**.

[95] KUMAR KATTUMANNIL, S. (2009). On Stein's identity and its application. *Statist. Probab. Lett.* **79** 1444–1449. MR2536504 https://doi.org/10.1016/j.spl.2009.03.021

[96] LEDOUX, M., NOURDIN, I. and PECCATI, G. (2015). Stein's method, logarithmic Sobolev and transport inequalities. *Geom. Funct. Anal.* **25** 256–306. MR3320893 https://doi.org/10.1007/s00039-015-0312-0

[97] LEUCHT, A. and NEUMANN, M. H. (2013). Dependent wild bootstrap for degenerate U- and V-statistics. *J. Multivariate Anal.* **117** 257–280. MR3053547 https://doi.org/10.1016/j.jmva.2013.03.003

[98] LEY, C., REINERT, G. and SWAN, Y. (2017). Stein's method for comparison of univariate distributions. *Probab. Surv.* **14** 1–52. MR3595350 https://doi.org/10.1214/16-PS278

[99] LEY, C., REINERT, G. and SWAN, Y. (2017). Distances between nested densities and a measure of the impact of the

prior in Bayesian statistics. *Ann. Appl. Probab.* **27** 216–241. MR3619787 https://doi.org/10.1214/16-AAP1202

[100] LEY, C. and SWAN, Y. (2016). Parametric Stein operators and variance bounds. *Braz. J. Probab. Stat.* **30** 171–195. MR3481100 https://doi.org/10.1214/14-BJPS271

[101] LI, L., LI, Y., LIU, J.-G., LIU, Z. and LU, J. (2020). A stochastic version of Stein variational gradient descent for efficient sampling. *Commun. Appl. Math. Comput. Sci.* **15** 37–63. MR4113783 https://doi.org/10.2140/camcos.2020.15.37

[102] LIPPERT, R. A., HUANG, H. and WATERMAN, M. S. (2002). Distributional regimes for the number of k-word matches between two random sequences. *Proc. Natl. Acad. Sci. USA* **99** 13980–13989. MR1944413 https://doi.org/10.1073/pnas.202468099

[103] LIU, A., LIANG, Y. and VAN DEN BROECK, G. (2020). Off-policy deep reinforcement learning with analogous disentangled exploration. In *International Conference on Autonomous Agents and Multiagent Systems* (*AAMAS*).

[104] LIU, C. and ZHU, J. (2018). Riemannian Stein variational gradient descent for Bayesian inference. In *AAAI Conference on Artificial Intelligence* 3627–3634.

[105] LIU, C., ZHUO, J., CHENG, P., ZHANG, R. and ZHU, J. (2019). Understanding and accelerating particle-based variational inference. In *International Conference on Machine Learning* (*ICML*) 4082–4092.

[106] LIU, H., FENG, Y., MAO, Y., ZHOU, D., PENG, J. and LIU, Q. (2018). Action-dependent control variates for policy optimization via Stein's identity. In *International Conference on Learning Representations* (*ICLR*).

[107] LIU, Q. (2017). Stein variational gradient descent as gradient flow. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 3115–3123.

[108] LIU, Q., LEE, J. and JORDAN, M. (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning* (*ICML*) 276–284.

[109] LIU, Q. and LEE, J. D. (2017). Black-box importance sampling. In *International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 952–961.

[110] LIU, Q., LEE, J. D. and JORDAN, M. I. (2016). A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. In *International Conference on Machine Learning* (*ICML*) 276–284.

[111] LIU, Q. and WANG, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 2370–2378.

[112] LIU, Q. and WANG, D. (2018). Stein variational gradient descent as moment matching. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 8854–8863.

[113] LIU, S., KANAMORI, T., JITKRITTUM, W. and CHEN, Y. (2019). Fisher efficient inference of intractable models. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 8793–8803.

[114] LIU, Y., RAMACHANDRAN, P., LIU, Q. and PENG, J. (2017). Stein variational policy gradient. In *Uncertainty in Artificial Intelligence* (*UAI*).

[115] LU, J., LU, Y. and NOLEN, J. (2019). Scaling limit of the Stein variational gradient descent: The mean field regime. *SIAM J. Math. Anal.* **51** 648–671. MR3919409 https://doi.org/10.1137/18M1187611

[116] MACKEY, L. and GORHAM, J. (2016). Multivariate Stein factors for a class of strongly log-concave distributions. *Electron. Commun. Probab.* **21** 56. MR3548768 https://doi.org/10.1214/16-ecp15

[117] MATSUBARA, T., KNOBLAUCH, J., BRIOL, F. X. and OATES, C. J. (2021). Robust generalised Bayesian inference for intractable likelihoods. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*. To appear. Available at arXiv:2104.07359.

[118] MATSUBARA, T., KNOBLAUCH, J., BRIOL, F. X. and OATES, C. J. (2022). Generalised Bayesian inference for discrete intractable likelihood. Available at arXiv:2206.08420.

[119] MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. *Communications and Control Engineering Series*. Springer, London. MR1287609 https://doi.org/10.1007/978-1-4471-3267-7

[120] MIJATOVIĆ, A. and VOGRINC, J. (2018). On the Poisson equation for Metropolis-Hastings chains. *Bernoulli* **24** 2401–2428. MR3757533 https://doi.org/10.3150/17-BEJ932

[121] MIJOULE, G., REINERT, G. and SWAN, Y. (2021). Stein's density method for multivariate continuous distributions. ArXiv preprint. Available at arXiv:2101.05079.

[122] MIRA, A., SOLGI, R. and IMPARATO, D. (2013). Zero variance Markov chain Monte Carlo for Bayesian estimators. *Stat. Comput.* **23** 653–662. MR3094805 https://doi.org/10.1007/s11222-012-9344-6

[123] MÜLLER, A. (1997). Integral probability metrics and their generating classes of functions. *Adv. in Appl. Probab.* **29** 429–443. MR1450938 https://doi.org/10.2307/1428011

[124] NOURDIN, I. and PECCATI, G. (2012). *Normal Approximations with Malliavin Calculus*: *From Stein's Method to Universality*. *Cambridge Tracts in Mathematics* **192**. Cambridge Univ. Press, Cambridge. MR2962301 https://doi.org/10.1017/CBO9781139084659

[125] NÜSKEN, N. and RENGER, D. (2021). Stein variational gradient descent: Many-particle and long-time asymptotics. ArXiv preprint. Available at arXiv:2102.12956.

[126] OATES, C. J., COCKAYNE, J., BRIOL, F.-X. and GIROLAMI, M. (2019). Convergence rates for a class of estimators based on Stein's method. *Bernoulli* **25** 1141–1159. MR3920368 https://doi.org/10.3150/17-bej1016

[127] OATES, C. J., GIROLAMI, M. and CHOPIN, N. (2017). Control functionals for Monte Carlo integration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79** 695–718. MR3641403 https://doi.org/10.1111/rssb.12185

[128] OATES, C. J., PAPAMARKOU, T. and GIROLAMI, M. (2016). The controlled thermodynamic integral for Bayesian model evidence evaluation. *J. Amer. Statist. Assoc.* **111** 634–645. MR3538693 https://doi.org/10.1080/01621459.2015.1021006

[129] OKSENDAL, B. (2013). *Stochastic Differential Equations*: *An Introduction with Applications*, 6th ed. Springer, Berlin.

[130] PU, Y., GAN, Z., HENAO, R., LI, C., HAN, S. and CARIN, L. (2017). VAE learning via Stein variational gradient descent. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 4236–4245.

[131] RACHEV, S. T., KLEBANOV, L. B., STOYANOV, S. V. and FABOZZI, F. J. (2013). *The Methods of Distances in the Theory of Probability and Statistics*. Springer, New York. MR3024835 https://doi.org/10.1007/978-1-4614-4869-3

[132] RANGANATH, R., TRAN, D., ALTOSAAR, J. and BLEI, D. (2016). Operator variational inference. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 496–504.

[133] REINERT, G. (1998). Couplings for normal approximations with Stein's method. In *Microsurveys in Discrete Probability* (*Princeton, NJ*, 1997). *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* **41** 193–207. Amer. Math. Soc., Providence, RI. MR1630415 https://doi.org/10.1089/cmb.1998.5.223

[134] REINERT, G. (2005). Three general approaches to Stein's method. In *An Introduction to Stein's Method. Lect. Notes*

Ser. Inst. Math. Sci. Natl. Univ. Singap. **4** 183–221. Singapore Univ. Press, Singapore. MR2235451 https://doi.org/10.1142/9789812567680_0004

[135] REINERT, G., CHEW, D., SUN, F. and WATERMAN, M. S. (2009). Alignment-free sequence comparison. I. Statistics and power. *J. Comput. Biol.* **16** 1615–1634. MR2578699 https://doi.org/10.1089/cmb.2009.0198

[136] REINERT, G. and ROSS, N. (2019). Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs. *Ann. Appl. Probab.* **29** 3201–3229. MR4019886 https://doi.org/10.1214/19-AAP1478

[137] RIABIZ, M., CHEN, W., COCKAYNE, J., SWIETACH, P., NIEDERER, S. A., MACKEY, L. and OATES, C. (2020). Optimal thinning of MCMC output. ArXiv preprint. Available at arXiv:2005.03952.

[138] ROSS, N. (2011). Fundamentals of Stein's method. *Probab. Surv.* **8** 210–293. MR2861132 https://doi.org/10.1214/11-PS182

[139] SCHWARTZ, L. (1964). Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *J. Anal. Math.* **13** 115–256. MR0179587 https://doi.org/10.1007/BF02786620

[140] SERFLING, R. J. (2009). *Approximation Theorems of Mathematical Statistics* **162**. Wiley, New York.

[141] SHAO, Q.-M. (2005). An explicit Berry-Esseen bound for Student's t-statistic via Stein's method. In *Stein's Method and Applications. Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap.* **5** 143–155. Singapore Univ. Press, Singapore. MR2205333 https://doi.org/10.1142/9789812567673_0009

[142] SHAO, Q.-M. (2010). Stein's method, self-normalized limit theory and applications. In *Proceedings of the International Congress of Mathematicians. Volume IV* 2325–2350. Hindustan Book Agency, New Delhi. MR2827974

[143] SHAO, Q.-M., ZHANG, K. and ZHOU, W.-X. (2016). Stein's method for nonlinear statistics: A brief survey and recent progress. *J. Statist. Plann. Inference* **168** 68–89. MR3412222 https://doi.org/10.1016/j.jspi.2015.06.008

[144] SI, S., OATES, C. J., DUNCAN, A. B., CARIN, L. and BRIOL, F.-X. (2020). Scalable control variates for Monte Carlo methods via stochastic optimization. ArXiv preprint. Available at arXiv:2006.07487.

[145] SMOLA, A., GRETTON, A., SONG, L. and SCHÖLKOPF, B. (2007). A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory* 13–31.

[146] SOHL-DICKSTEIN, J., BATTAGLINO, P. and DEWEESE, M. R. (2011). Minimum probability flow learning. In *International Conference on Machine Learning* 905–912.

[147] SOUTH, L. F., KARVONEN, T., NEMETH, C., GIROLAMI, M. and OATES, C. (2020). Semi-exact control functionals from sard's method. ArXiv preprint. Available at arXiv:2002.00033.

[148] SOUTH, L. F., OATES, C. J., MIRA, A. and DROVANDI, C. (2018). Regularised zero-variance control variates for high-dimensional variance reduction. ArXiv preprint. Available at arXiv:1811.05073.

[149] STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 197–206. Univ. California Press, Berkeley-Los Angeles, CA. MR0084922

[150] STEIN, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. II*: *Probability Theory* 583–602. MR0402873

[151] STEIN, C. (1986). *Approximate Computation of Expectations. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **7**. IMS, Hayward, CA. MR0882007

[152] STEIN, C., DIACONIS, P., HOLMES, S. and REINERT, G. (2004). Use of exchangeable pairs in the analysis of simulations. In *Stein's Method*: *Expository Lectures and Applications. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **46** 1–26. IMS, Beachwood, OH. MR2118600 https://doi.org/10.1214/lnms/1196283797

[153] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

[154] SUN, Z., BARP, A. and BRIOL, F.-X. (2021). Vector-valued control variates. Available at arXiv:2109.08944.

[155] TEYMUR, O., GORHAM, J., RIABIZ, M. and OATES, C. (2021). Optimal quantisation of probability measures using maximum mean discrepancy. In *International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 1027–1035.

[156] TIHOMIROV, A. N. (1980). Convergence rate in the central limit theorem for weakly dependent random variables. *Teor. Veroyatn. Primen.* **25** 800–818. MR0595140

[157] WANG, D. and LIU, Q. (2016). Learning to draw samples: With application to amortized MLE for generative adversarial learning. ArXiv preprint. Available at arXiv:1611.01722.

[158] WANG, D. and LIU, Q. (2019). Nonlinear Stein variational gradient descent for learning diversified mixture models. In *International Conference on Machine Learning* (*ICML*) 6576–6585.

[159] WANG, D., TANG, Z., BAJAJ, C. and LIU, Q. (2019). Stein variational gradient descent with matrix-valued kernels. In *Advances on Neural Information Processing Systems* (*NeurIPS*) 7834–7844.

[160] WANG, D., ZENG, Z. and LIU, Q. (2018). Stein variational message passing for continuous graphical models. In *International Conference on Machine Learning* (*ICML*) 5219–5227.

[161] WELLING, M. and TEH, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning* (*ICML*) 681–688.

[162] XU, W. (2022). Standardisation-function kernel Stein discrepancy: A unifying view on kernel Stein discrepancy tests for goodness-of-fit. In *International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 1575–1597.

[163] XU, W. and REINERT, G. (2021). A Stein goodness-of-fit test for exponential random graph models. In *International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 415–423.

[164] YANG, J., LIU, Q., RAO, V. and NEVILLE, J. (2018). Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *International Conference on Machine Learning* (*ICML*) 5561–5570.

[165] YANG, J., RAO, V. and NEVILLE, J. (2019). A Stein–papangelou goodness-of-fit test for point processes. In *International Conference on Artificial Intelligence and Statistics* (*AISTATS*) 226–235.

[166] YANG, Z., BALASUBRAMANIAN, K., WANG, Z. and LIU, H. (2017). Learning non-Gaussian multi-index model via second-order Stein's method. In *Advances in Neural Information Processing Systems* (*NeurIPS*) **30** 6097–6106.

[167] ZHANG, X. and CURTIS, A. (2019). Seismic tomography using variational inference methods. *J. Geophys. Res., Solid Earth* **125** e2019JB018589.

[168] ZHANG, X. and CURTIS, A. (2020). Variational full-waveform inversion. *Geophys. J. Int.* **222** 406–411.

[169] ZHANG, Y. and LEE, A. A. (2019). Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.* **10** 8154–8163. https://doi.org/10.1039/c9sc00616h

[170] ZHU, Y. and ZABARAS, N. (2018). Bayesian deep convolutional encoder-decoder networks for surrogate modeling and uncertainty quantification. *J. Comput. Phys.* **366** 415–447. MR3800689 https://doi.org/10.1016/j.jcp.2018.04.018

[171] ZHU, Z., WAN, R. and ZHONG, M. (2018). Neural control variates for variance reduction. ArXiv preprint. Available at arXiv:1806.00159.

[172] ZHUO, J., LIU, C., SHI, J., ZHU, J., CHEN, N. and ZHANG, B. (2018). Message passing Stein variational gradient descent. In *International Conference on Machine Learning* (*ICML*) 6013–6022.

[173] ZOLOTAREV, V. M. (1984). Probability metrics. *Theory Probab. Appl.* **28** 278–302.