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

Andreas Anastasiou[1], Alessandro Barp[2], François-Xavier Briol[3], Bruno Ebner[4], Robert E. Gaunt[5], Fatemeh Ghaderinezhad[6], Jackson Gorham[7], Arthur Gretton[3], Christophe Ley[8],*, Qiang Liu[9], Lester Mackey[10], Chris. J. Oates[11], Gesine Reinert[12], Yvik Swan[13]

[1] *University of Cyprus, Cyprus;* [2] *University of Cambridge, UK;* [3] *University College London, UK;* [4] *Karlsruhe Institute of Technology, Germany;* [5] *The University of Manchester, UK;* [6] *Ghent University, Belgium;* [7] *Whisper.ai, Inc., US;* [8] *University of Luxembourg, Luxembourg;* [9] *The University of Texas at Austin, US* [10] *Microsoft Research New England, US;* [11] *Newcastle University, UK;* [12] *University of Oxford, UK;* [13] *Université Libre de Bruxelles, Belgium*

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: 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

MSC2020 subject classifications: Primary 62-02; Secondary 62-E99, 62-08

1 Introduction

Stein's method was introduced by Charles Stein in the early 1970s (Stein, 1972) 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 (Stein, 1981) and James-Stein

---

*Correspondance should be addressed to `christophe.ley@ugent.be`.

shrinkage estimators (Stein, 1956; James and Stein, 1961); see Fathi et al. (2020) for a joined-up view. The latter paper also exploited these connections with Stein's method to propose and analyse 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 Chen (1975), 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 Stein (1986); Barbour et al. (1992); Chen et al. (2010); Nourdin and Peccati (2012); Arras and Houdré (2019). We also refer to the lecture notes of Diaconis and Holmes (2004) and the surveys of Ross (2011); Chatterjee (2014); Barbour and Chen (2014); Ley et al. (2017b). 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. (2004) employed the method for the analysis of sample quality in simulations, Holmes and Reinert (2004) developed a bootstrap method for network data which is analysed via empirical processes, Shao (2005) 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 Shao (2010). In Shao et al. (2016), an overview on the use of randomised concentration inequalities in Stein's method for nonlinear statistics is provided. Lippert et al. (2002); Reinert et al. (2009) utilized the method to prove that there were flaws in then commonly used statistics for alignment-free sequence comparison, and subsequently introduce 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 $\text{div}(v) = \nabla^\mathsf{T} v = \sum_{i=1}^d \partial_i v_i(x)$. For a vector or a matrix, the superscript T 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

$$Q = P \text{ if and only if } \mathbb{E}_{X \sim Q}[(\mathcal{T}g)(X)] = 0 \text{ for all } g \in \mathcal{G}(\mathcal{T}). \qquad (1)$$

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

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

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 (2015) call the quantity (2) a *Stein discrepancy* (in contrast to the use of the term in Ledoux et al. (2015)). 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 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, i.e. 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*

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

see for example Stein (1981) (for Σ the identity matrix). We deduce that the first order differential operator

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

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

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

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 Gibbs and Su (2002); Rachev et al. (2013). 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) (Müller, 1997; Zolotarev, 1984), which are defined as

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

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 (Berlinet and Thomas-Agnan, 2004) 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

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

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

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

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 (Barbour et al., 1992; Reinert, 1998; Chen et al., 2010; Ross, 2011), the Malliavin-Stein method (Nourdin and Peccati, 2012) and comparison of Stein operators (Holmes, 2004; Ley et al., 2017b; Mijoule et al., 2021); 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 Courtade et al. (2019); Mijoule et al. (2021)). 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 Mackey and Gorham (2016); Fang et al. (2019) where bounds are obtained for operators given in the setting of Example 2 under assumptions of log-concavity, and Gorham et al. (2019), where Stein factors are obtained under the weaker assumption of *integrable Wasserstein decay*. An overview for continuous distributions is given in Mijoule et al. (2021).

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 Barbour (1988, 1990) and Götze (1991)) and the density approach (which dates back to Stein et al. (2004)). These are not the only available approaches (see for example Reinert (2005)) 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 (Oksendal, 2013, Lemma 8.1.4)) with invariant measure P, the *infinitesimal generator* $\mathcal{A}$ of the process given by

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

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 (1988, 1990) and Götze (1991) 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 analyse multivariate distributions via Stein's method.

Gorham et al. (2019) detailed the generator approach for a wide range of distributions of interest by using operators induced by *Itô diffusions*. An Itô diffusion (Oksendal, 2013, Def. 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

$$\mathrm{d}Z_{t,x} = b(Z_{t,x})\,\mathrm{d}t + \sigma(Z_{t,x})\,\mathrm{d}W_t \text{ with } Z_{0,x} = x \in \mathbb{R}^d, \tag{7}$$

where $(W_t)_{t \geq 0}$ is a m-dimensional Brownian motion. It is known (see, e.g., Gorham et al. (2019, Thm. 2) and Barp (2020, Thm. 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. Gorham et al. (2019) proposed the first order *diffusion Stein operator*

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

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, Gorham and Mackey (2015); Mackey and Gorham (2016) 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 Stein et al. (2004) for univariate distributions, and has since then been generalized in multiple directions, see for example Yang et al. (2018); Mijoule et al. (2021). 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 $p(x) = \lambda e^{-\lambda x} \mathbb{I}_{[0,\infty)}(x)$, 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*

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

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 $\lim_{x \to \infty} xg_0(x)e^{-\lambda x} = 0$ 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

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

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

$$(\mathcal{T}_2 g)(x) = \frac{\frac{\mathrm{d}}{\mathrm{d}\lambda}(\lambda g(\lambda x)e^{-\lambda x})}{(\lambda e^{-\lambda x})} = xg'(\lambda x) + (\tfrac{1}{\lambda} - x)g(\lambda x),$$

with no boundary assumptions on g since $\mathbb{E}_{X \sim \mathrm{Exp}(\lambda)}[(\mathcal{T}_2 g)(X)] = \frac{\mathrm{d}}{\mathrm{d}\lambda} \left(\int_0^\infty g(u)e^{-u} \, \mathrm{d}u \right) = 0$ 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 Kattumannil (2009) and Goldstein and Reinert (2013).

The density approach and the generator approach are by no means the only methods for obtaining Stein operators. Other popular approaches include couplings (Chen and Röllin (2010)), orthogonal polynomials (Goldstein and Reinert (2005)), a perturbation approach (Barbour and Xia (1999)), an ODE approach (Gaunt (2017)), and characteristic functions (Tikhomirov (1980); Arras and Houdré (2019)).

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:

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 Ley et al. (2017a). 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 Gorham et al. (2019) and Mijoule et al. (2021), for instance.

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 Barp et al. (2019), Matsubara et al. (2021)).

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 Stein et al. (2004)). 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 (Gretton et al., 2006; Smola et al., 2007; Gretton et al., 2012). 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* (Aronszajn, 1950; Schwartz, 1964). This choice of Stein set was inspired by the zero mean reproducing kernel theory of Oates et al. (2017), used in Chwialkowski et al. (2016); Liu et al. (2016a); Gorham and Mackey (2017) and extended in Sun et al. (2021) 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 $\mathcal{G}_k := \{g = (g_1, \ldots, g_d) \mid \|v\|_2 \leq 1 \text{ for } v_j := \|g_j\|_k\}$:

$$\text{KSD}_k(Q) := \mathcal{S}(Q, \mathcal{T}, \mathcal{G}_k) = \sqrt{\mathbb{E}_{X, X' \sim Q}[k_P(X, X')]}, \qquad (10)$$

where the Stein reproducing kernel *is given by*

$$k_P(x, x') := \text{Trace}(\mathcal{T}_x \mathcal{T}_{x'} k(x, x')) = \langle \nabla_x, \nabla_{x'} k(x, x') \rangle + \langle \nabla_x k(x, x'), \nabla_{x'} \log p(x') \rangle$$
$$+ \langle \nabla_{x'} k(x, x'), \nabla_x \log p(x) \rangle + k(x, x') \langle \nabla_x \log p(x), \nabla_{x'} \log p(x') \rangle. \qquad (11)$$

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 Oates et al. (2017)). Whenever the approximating measure is* $Q_n = n^{-1} \sum_{i=1}^{n} \delta_{x_i}$, *the Langevin KSD has the simple closed form*

$$\text{KSD}_k(Q_n) = \mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_k) = \sqrt{\frac{1}{n^2} \sum_{i,j=1}^{n} k_P(x_i, x_j)}. \qquad (12)$$

The most common choice of kernel k is the inverse multi-quadric kernel $k(x, y) = (c^2 + \|x - y\|_2^2)^\beta$, $c > 0$, $\beta \in (-1, 0)$. This is because Gorham and Mackey (2017, Theorem 8) showed that, if $\nabla \log p$ is sufficiently regular, then Q_n converges weakly to P whenever $\text{KSD}_k(Q_n) \to 0$. We will return to the implications of this in Section 4.1.

Extensions of the Langevin KSD include Barp et al. (2019), who used the infinitesimal generator of general Itô diffusions to get a family of *diffusion kernel Stein discrepancies*; Yang et al. (2018) to discrete sets $\mathcal{X}$; Barp et al. (2018) 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 Gorham and Mackey (2015); Gorham et al. (2019).

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

$$\mathcal{G}_{\|\cdot\|, Q_n, E} = \left\{ g : \max \left(\|g(v)\|_\infty, \|\nabla g(v)\|_\infty, \frac{\|g(x)-g(y)\|_\infty}{\|x-y\|_1}, \frac{\|\nabla g(x)-\nabla g(y)\|_\infty}{\|x-y\|_1} \right) \le 1, \right.$$
$$\frac{\|g(x)-g(y)-\nabla g(x)(x-y)\|_\infty}{\frac{1}{2}\|x-y\|_1^2} \le 1, \frac{\|g(x)-g(y)-\nabla g(y)(x-y)\|_\infty}{\frac{1}{2}\|x-y\|_1^2} \le 1,$$
$$\left. \forall (x,y) \in E, v \in \mathrm{supp}(Q_n) \right\}$$

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 (Gorham and Mackey, 2015, Theorem 2, Proposition 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 Jitkrittum et al. (2017) 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 Huggins and Mackey (2018) 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 Gong et al. (2021) 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, Gorham et al. (2020) introduced *stochastic Stein discrepancies* (SSDs)

$$\mathcal{SS}(Q_n, \mathcal{T}, \mathcal{G}) \coloneqq \sup_{g \in \mathcal{G}} \left| \frac{L}{n} \sum_{i=1}^n (\mathcal{T}_{\sigma_i} g)(x_i) \right| \quad \text{for} \quad \sigma_i \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\{1, \dots, L\}). \tag{13}$$

They showed that SSDs inherit the convergence control properties of standard discrepancies with probability 1. In Xu and Reinert (2021), 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., Welling and Teh, 2011; Ahn et al., 2012; Korattikara et al., 2014). 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, Gorham and Mackey (2015, 2017); Huggins and Mackey (2018); Gorham et al. (2020) 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 $\{x_1, \ldots, x_n\}$.

Graph Stein discrepancies Gorham and Mackey (2015) 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, Mackey and Gorham (2016); Gorham et al. (2019); Erdogdu et al. (2018) bounded the GSD explicitly above and below by Wasserstein distances whenever the diffusion underlying the Stein operator couples quickly and has pseudo-Lipschitz drift.

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, Gorham and Mackey (2017) 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, Gorham and Mackey (2017) 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 Huggins and Mackey (2018); Chen et al. (2018b) for further developments).
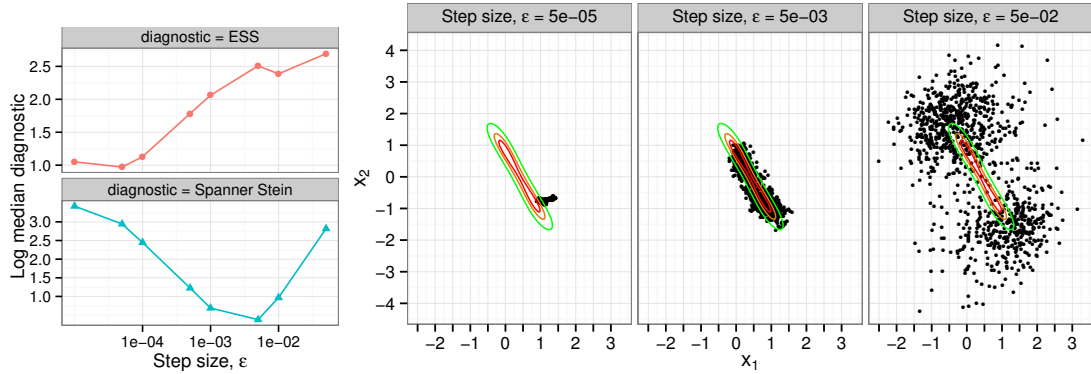
Figure 1: Selecting the step size ϵ for stochastic gradient Langevin dynamics (Welling and Teh, 2011), a popular approximate MCMC algorithm designed for scalability. Standard MCMC diagnostics like effective sample size (ESS) do not account for asymptotic bias and select overly large ϵ with greatly overdispersed 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 Gorham and Mackey (2015, Fig. 3).

These convergence-determining KSDs have been shown to deliver substantial speed-ups over the original GSDs in higher dimensions (Gorham and Mackey, 2017).

Random feature Stein discrepancies To identify a family of convergence-determining discrepancy measures that can be accurately and inexpensively approximated with random sampling, Huggins and Mackey (2018) 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). Huggins and Mackey (2018) 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 $\{x_1, \ldots, x_n\}$ 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 $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, which we refer to as *particles*, such that its empirical measure Q gives a close approximation to P. Stein variational gradient descent (SVGD) (Liu and Wang, 2016) 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 $\mathrm{KL}(Q \parallel P) < \infty$. We want to recursively "transport" the probability mass of Q with a deterministic map to move it closer to P in order to decrease $\mathrm{KL}(Q \parallel P)$ 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:

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

where the negative derivative $-\tfrac{\mathrm{d}}{\mathrm{d}\epsilon} \mathrm{KL}(\Phi_\sharp Q \parallel P) \, \big|_{\epsilon=0}$ 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*

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

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:

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

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):

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

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 $\{x_1, \ldots, x_n\}$ 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:

$$x_i \;\leftarrow\; x_i \;+\; \epsilon \tfrac{1}{n} \sum_{j=1}^n \left(\nabla \log p(x_j) k(x_j, x_i) + \nabla_{x_j} k(x_j, x_i) \right), \quad \forall i = 1, \ldots, n. \tag{17}$$

The two terms in (17) play intuitive roles. The term with the gradient $\nabla \log p$ pushes the particles towards 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 $\{x_1, \ldots, x_n\}$ interact and co-evolve to reach a desirable equilibrium. For understanding SVGD asymptotically, (Liu, 2017) 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 Lu et al. (2019); Liu et al. (2019a); Duncan et al. (2019); Chewi et al. (2020); Gorham et al. (2020); Nüsken and Renger (2021); Korba et al. (2020). In the non-asymptotic 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 (Liu and Wang, 2018).

SVGD has been extended and improved in various ways. For example, amortized SVGD (Feng et al., 2017) learns neural samplers in replacement of particle approximation; gradient-free SVGD (Han and Liu, 2018) provides an extension that requires no gradient information of the target distribution P; a number of other extensions and improvements can be found in, e.g., Wang et al. (2018); Zhuo et al. (2018); Wang et al. (2019); Liu and Zhu (2018); Detommaso et al. (2018); Chen et al. (2018a, 2019a); Li et al. (2020); Gorham et al. (2020); Gong et al. (2019); Wang and Liu (2019); Han and Liu (2017). SVGD has found applications in a variety of problems including in deep learning (e.g., Pu et al., 2017; Wang and Liu, 2016), reinforcement learning (e.g., Haarnoja et al.,
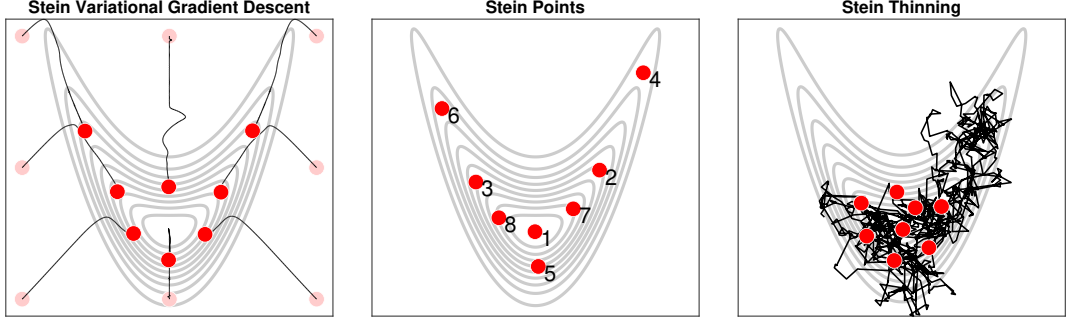
Figure 2: Sampling with Stein's method. Left: The initial (transparent red) and final (red) states of 8 particles, together with their trajectories (black) under the *Stein variational gradient descent* algorithm in (17). Middle: The first 8 states (red) of an extensible sequence produced by the *Stein points* algorithm in (18). The order in which the states are selected is indicated. Right: The first 8 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.)

2017; Liu et al., 2017, 2020), meta learning (e.g., Kim et al., 2018; Feng et al., 2017), and uncertainty quantification in science and engineering (e.g., Zhu and Zabaras, 2018; Zhang and Curtis, 2019, 2020; Zhang, 2019).

4.2.2 Sampling with Stein Points

The *Stein points* (Chen et al., 2018b, 2019b) approach progressively constructs a set of points $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ 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

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

where $\mathrm{KSD}_k(\{x_1, \ldots, x_n\}) = \mathcal{S}(Q_n, \mathcal{T}, \mathcal{G}_k)$ and $\mathcal{G}_k$ is a kernel Stein set; then the set $\{x_1, \ldots, x_n\}$ 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, Chen et al. (2018b, 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 $\{x_1, \ldots, x_n\} \subset \mathcal{X}$ satisfying

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

16

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

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

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

Riabiz et al. (2020) use KSD in a post-processing approach to select states from a large pre-determined candidate set, with application to de-biasing MCMC output. Their approach can be summarized as:

$$x_1 \in \text{argmin}_{x \in \{X_1, \ldots, X_N\}} \text{KSD}_k(\{x\}),$$
$$x_n \in \text{argmin}_{x \in \{X_1, \ldots, X_N\}} \text{KSD}_k(\{x_1, \ldots, x_{n-1}, x\}) \text{ for } n > 1, \tag{19}$$

where $(X_i)_{i=1,\ldots,N}$ 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 $V(x) \geq \frac{dP}{dQ}(x)\sqrt{k_P(x,x)}$ and that certain moments of the chain are finite, Riabiz et al. (2020, Theorem 3) showed that $\text{KSD}_k(\{x_1, \ldots, x_n\}) \to 0$ 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 non-myopic optimization and for mini-batching, were recently studied in Teymur et al. (2021). In related work, Liu and Lee (2017); Hodgkinson et al. (2020) proposed to use Stein discrepancies to re-weight 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 Stein (1986). 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 Meyn and Tweedie, 2012). In Stein et al. (2004), 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 re-write the expectation as

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

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 (Mira et al., 2013).

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 Andradóttir et al. (1993); Stein et al. (2004); Henderson and Simon (2004); Dellaportas and Kontoyiannis (2012); Mijatović and Vogrinc (2018). It was pointed out in Oates et al. (2017) 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 Assaraf and Caffarel (1999); Mira et al. (2013); Oates et al. (2016), 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 South et al. (2018), 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 Oates et al. (2017, 2019); Barp et al. (2019); Sun et al. (2021). In South et al. (2020), 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 Zhu et al. (2018); Si et al. (2020). 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 Belomestny et al. (2017, 2019, 2020). Liu et al. (2018) 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 $x_1, \ldots, x_n$ (repre-

sented by $Q_n = n^{-1} \sum_{i=1}^{n} \delta_{x_i}$). Barp et al. (2019) 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 (Hyvärinen, 2005), contrastive divergence (Hinton, 2002), and minimum probability flow (Sohl-Dickstein et al., 2011) are specific instances of this framework. have also been proposed. Barp et al. (2019) 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 Banerjee et al. (2021), whilst Grathwohl et al. (2020) considered a Stein discrepancy where the Stein space is indexed by a neural network. Relatedly, Betsch et al. (2020) studied minimum L^q distance estimators based on Stein operators, and Liu et al. (2019b) considered a minimum distance estimator based on likelihood ratios estimated through Stein operators.

Most notably, estimators of the form in (20) are useful for unnormalised likelihood models, since Stein operators usually rely on unnormalised densities. When the parametric family is in some exponential family, the Langevin Stein discrepancies become quadratic forms in ϑ, which implies that the optimiser can be obtained in closed-form. Matsubara et al. (2021) built on this idea to propose a fully conjugate generalised Bayesian approach for unnormalised densities. This was latter extended to discrete data settings by Matsubara et al. (2022).

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 et al.

(2016); Liu et al. (2016b) 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 (Gretton et al., 2006, 2012). 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 sub-optimal 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:

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

This estimate can be used as a test statistic. It is degenerate under the null hypothesis that $Q = P$, and non-degenerate under the alternative. As a result, when the sample is i.i.d. the asymptotic behaviour of the statistic is obtained via standard results (Serfling, 2009). 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 (Leucht and Neumann, 2013). 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. (Chwialkowski et al., 2016, 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 Chwialkowski et al. (2016, Theorem 2.2): if k is C_0-universal (Carmeli et al., 2010, 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 Gorham and Mackey (2017) 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 (Jitkrittum et al., 2017), or importance sampling approaches

(Huggins and Mackey, 2018). 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, Fernández et al. (2020) 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 Yang et al. (2018) which derives a discrepancy for discrete data and that of Yang et al. (2019) which focuses on point processes. For exponential random graph models when only one network observation is available, Xu and Reinert (2021) use the Stein operator for exponential random graph models from Reinert and Ross (2019) as basis for a kernelised 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 Ley and Swan (2016) 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 Key et al. (2021) and built on the minimum Stein discrepancy estimators of Barp et al. (2019). 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

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

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 (2020) 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 Ebner (2021), 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 et al. (2021) 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 $\{g_t(x) = \exp(\mathrm{i}t^\mathsf{T}x) : t \in \mathbb{R}^d\}$. 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 Ebner and Henze (2020).

In a similar vein, Betsch and Ebner (2021) and Betsch et al. (2022) provide new characterizations of continuous and discrete parametric families of distributions through the density approach for novel tests for univariate normality (Betsch and Ebner, 2020), the gamma family (Betsch and Ebner, 2019) and the inverse Gaussian law (Allison et al., 2022). 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, Baringhaus and Henze (1991) for a test of exponentiality, Baringhaus and Henze (1992) for a test of Poissonity, and Henze et al. (2012) 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 chisquare 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,

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

where $\to_d$ denotes convergence in distribution. Starting with the single-parameter case, under some natural regularity assumptions which we do not detail here, Anastasiou and Reinert (2017) 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 Anastasiou and Reinert (2017), they are locally dependent in Anastasiou (2017).

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, Anastasiou and Reinert (2017) established that

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

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, Anastasiou and Ley (2017) give an explicit bound for MLEs which are a smooth function of a sum of independent terms. This result is generalised to the multivariate case in Anastasiou and Gaunt (2021).

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 Anastasiou and Reinert (2020) using Stein's method.

Explicit bounds on chisquare approximations for Pearson's chi-square test for goodness-of-fit of categorical data are obtained in Gaunt et al. (2017), and more generally the power divergence family of statistics in Gaunt (2022). Gaunt and Reinert (2021) 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, Diaconis and Freedman (1986) 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, Ley et al. (2017a); Ghaderinezhad and Ley (2019) 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 Mijoule et al. (2021).

Let us start by fixing the notation. Suppose that the observations $X_1, \ldots, X_n$ 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

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

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

$$\mathbb{E}_{\Theta_2 \sim P_2}[h(\Theta_2)] - \mathbb{E}_{\Theta_1 \sim P_1}[h(\Theta_1)] = \mathbb{E}_{\Theta_2 \sim P_2}[\mathcal{T}_1 f_h(\Theta_2)] = \mathbb{E}_{\Theta_2 \sim P_2}[(\mathcal{T}_1 - \mathcal{T}_2) f_h(\Theta_2)]$$

$$= \mathbb{E}_{\Theta_2 \sim P_2}\left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log(\rho(\theta))\big|_{\theta=\Theta_2} f_h(\Theta_2) \right]$$

(the second equality holds because, by definition, $\mathbb{E}_{\Theta_2 \sim P_2}[\mathcal{T}_2 f_h(\Theta_2)] = 0$). Thus, bounding an IPM generated by some class $\mathcal{H}$ between Θ_2 and Θ_1 can be achieved by bounding $\mathbb{E}_{\Theta_2 \sim P_2}\left[\frac{\mathrm{d}}{\mathrm{d}\theta} \log(\rho(\theta))\big|_{\theta=\Theta_2} f_h(\Theta_2) \right]$ 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 et al. (2017a) 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

$$\frac{\sigma^2}{n\delta^2+\sigma^2}\left|\overline{x} - \mu\right| \leq d_{\mathrm{W}}(P_1, P_2) \leq \frac{\sigma^2}{n\delta^2+\sigma^2}\left|\overline{x} - \mu\right| + \frac{\sqrt{2}}{\sqrt{\pi}} \frac{\sigma^3}{n\delta\sqrt{n\delta^2+\sigma^2}}$$

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 Xu (2022). Section 5 highlights how Stein's method can be brought to fruition not only to devise estimators but also to quantify their behaviour. There is plenty of scope for analysing 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 non-convex optimization (Erdogdu et al., 2018), or for learning semi-parametric multi-index models in high dimensions (Yang et al., 2017). In Bayesian statistics, Stein discrepancies have been used as variational objectives for posterior approximation (e.g., Ranganath et al., 2016; Hu et al., 2018; Fisher et al., 2021).

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.

References

Ahn, S., Korattikara, A., and Welling, M. (2012). Bayesian posterior sampling via stochastic gradient Fisher scoring. In *International Conference on Machine Learning (ICML)*, pages 1591–1598.

Allison, J., Betsch, S., Ebner, B., and Visagie, I. (2022). On testing the adequacy of the Inverse Gaussian distribution. *Mathematics*, 10:350.

Anastasiou, A. (2017). Bounds for the normal approximation of the maximum likelihood estimator from M-dependent random variables. *Statist. Probab. Lett.*, 129:171–181.

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.

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.

Anastasiou, A. and Reinert, G. (2017). Bounds for the normal approximation of the maximum likelihood estimator. *Bernoulli*, 23(1):191–218.

Anastasiou, A. and Reinert, G. (2020). Bounds for the asymptotic distribution of the likelihood ratio. *Ann. Appl. Probab*, 30(2):608–643.

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. Simu.*, 3(3):167–189.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68(3):337–404.

Arras, B. and Houdré, C. (2019). *On Stein's Method for Infinitely Divisible Laws with Finite First Moment*. Springer.

Assaraf, R. and Caffarel, M. (1999). Zero-variance principle for Monte Carlo algorithms. *Phys. Rev. Lett.*, 83(23):4682.

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.*, 22:1–46.

Barbour, A. D. (1988). Stein's method and Poisson process convergence. *J. Appl. Probab.*, 25A:175–184.

Barbour, A. D. (1990). Stein's method for diffusion approximations. *Probab. Theory Related Fields*, 84(3):297–322.

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

Barbour, A. D., Holst, L., and Janson, S. (1992). *Poisson Approximation*, volume 2. The Clarendon Press Oxford University Press.

Barbour, A. D. and Xia, A. (1999). Poisson perturbations. *ESAIM Probab. Stat*, 3:131–150.

Baringhaus, L. and Henze, N. (1991). A class of consistent tests for exponentiality based on the empirical Laplace transform. *Ann. Inst. Statist. Math.*, 43(3):551–564.

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(4):269–274.

Barp, A. A. (2020). *The Bracket Geometry of Statistics*. PhD thesis, Imperial College London.

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)*, pages 12964–12976.

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

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.

Belomestny, D., Iosipoi, L., and Zhivotovskiy, N. (2017). Variance reduction via empirical variance minimization: convergence and complexity. *arXiv preprint arXiv:1712.04667*.

Belomestny, D., Moulines, E., Shagadatov, N., and Urusov, M. (2019). Variance reduction for MCMC methods via martingale representations. *arXiv preprint arXiv:1903.07373*.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer Science+Business Media.

Betsch, S. and Ebner, B. (2019). A new characterization of the gamma distribution and associated goodness-of-fit tests. *Metrika*, 82(7):779–806.

Betsch, S. and Ebner, B. (2020). Testing normality via a distributional fixed point property in the Stein characterization. *TEST*, 29(1):105–138.

Betsch, S. and Ebner, B. (2021). Fixed point characterizations of continuous univariate probability distributions and their applications. *Ann. Inst. Statist. Math.*, 73:31–59.

Betsch, S., Ebner, B., and Klar, B. (2020). Minimum L^q-distance estimators for non-normalized parametric models. *Can. J. Stat.*, 49(2):514–548.

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.

Carmeli, C., De Vito, E., Toigo, A., and Umanità, V. (2010). Vector valued reproducing kernel Hilbert spaces and universality. *Anal. Appl.*, 8(1):19–61.

Chatterjee, S. (2014). A short survey of Stein's method. In *International Congress of Mathematicians (ICM)*, pages 1–24.

Chen, C., Zhang, R., Wang, W., Li, B., and Chen, L. (2018a). A unified particle-optimization framework for scalable Bayesian sampling. In *Uncertainty in Artificial Intelligence (UAI)*.

Chen, L. H. (1975). Poisson approximation for dependent trials. *Ann. Probab.*, 3:534–545.

Chen, L. H., Goldstein, L., and Shao, Q.-M. (2010). *Normal Approximation by Stein's method*. Springer Science & Business Media.

Chen, L. H. and Röllin, A. (2010). Stein couplings for normal approximation. *arXiv preprint arXiv:1003.6039*.

Chen, P., Wu, K., Chen, J., O'Leary-Roseberry, T., and Ghattas, O. (2019a). Projected Stein variational Newton: A fast and scalable Bayesian inference method in high dimensions. In *Advances on Neural Information Processing Systems (NeurIPS)*, pages 15130–15139.

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

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

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)*.

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

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(2):777–790.

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(1):133–161.

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)*, pages 9169–9179.

Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates (with discussion and rejoinder by the authors). *Ann. Statist.*, 14:1–67.

Diaconis, P. and Holmes, S., editors (2004). *Stein's Method: Expository Lectures and Applications*. Institute of Mathematical Statistics Lecture Notes - Monograph Series, Volume 46.

Dörr, P., Ebner, B., and Henze, N. (2021). A new test of multivariate normality by a double estimation in a characterizing PDE. *Metrika*, 84(3):401–427.

Duncan, A., Nüsken, N., and Szpruch, L. (2019). On the geometry of Stein variational gradient descent. *arXiv preprint arXiv:1912.00894*.

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.

Ebner, B. and Henze, N. (2020). Tests for multivariate normality – a critical review with emphasis on weighted L^2-statistics. *TEST*, 29:845–892.

Erdogdu, M. A., Mackey, L., and Shamir, O. (2018). Global non-convex optimization with discretized diffusions. In *Advances on Neural Information Processing Systems (NeurIPS)*, pages 9694–9703.

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(3-4):945–979.

Fathi, M., Goldstein, L., Reinert, G., and Saumard, A. (2020). Relaxing the Gaussian assumption in shrinkage and SURE in high dimension. *arXiv preprint arXiv:2004.01378*.

Feng, Y., Wang, D., and Liu, Q. (2017). Learning to draw samples with amortized Stein variational gradient descent. In *Uncertainty in Artificial Intelligence (UAI)*.

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)*.

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)*.

Gaunt, R. E. (2017). On Stein's method for products of normal random variables and zero bias couplings. *Bernoulli*, 23(4B):3311–3345.

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

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(2):720–756.

Gaunt, R. E. and Reinert, G. (2021). Bounds for the chi-square approximation of Friedman's statistic by stein's method. *arXiv preprint arXiv:2111.00949*.

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.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *Int. Stat. Rev.*, 70(3):419–435.

Goldstein, L. and Reinert, G. (2005). Distributional transformations, orthogonal polynomials, and Stein characterizations. *J. Theor. Probab.*, 18(1):237–260.

Goldstein, L. and Reinert, G. (2013). Stein's method for the beta distribution and the Pólya-Eggenberger urn. *J. Appl. Probab.*, 50(4):1187–1205.

Gong, C., Peng, J., and Liu, Q. (2019). Quantile Stein Variational Gradient Descent for parallel Bayesian optimization. In *International Conference on Machine Learning (ICML)*, pages 2347–2356.

Gong, W., Li, Y., and Hernández-Lobato, J. M. (2021). Sliced kernelized Stein discrepancy. In *International Conference on Learning Representations (ICLR)*.

Gorham, J., Duncan, A. B., Vollmer, S. J., and Mackey, L. (2019). Measuring sample quality with diffusions. *Ann. Appl. Probab.*, 29(5):2884–2928.

Gorham, J. and Mackey, L. (2015). Measuring sample quality with Stein's method. In *Advances on Neural Information Processing Systems (NeurIPS)*, pages 226–234. Curran Associates, Inc.

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

Gorham, J., Raj, A., and Mackey, L. (2020). Stochastic Stein discrepancies. In *Advances on Neural Information Processing Systems (NeurIPS)*.

Götze, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.*, 19(2):724–739.

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*, pages 9485–9499.

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)*, pages 513–520.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.*, 13(1):723–773.

Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning (ICML)*, pages 1352–1361.

Han, J. and Liu, Q. (2017). Stein variational adaptive importance sampling. In *Uncertainty in Artificial Intelligence (UAI)*.

Han, J. and Liu, Q. (2018). Stein variational gradient descent without gradient. In *International Conference on Machine Learning (ICML)*, pages 1900–1908.

Henderson, S. G. and Simon, B. (2004). Adaptive simulation using perfect control variates. *J. Appl. Probab.*, 41(3):859–876.

Henze, N., Meintanis, S. G., and Ebner, B. (2012). Goodness-of-fit tests for the gamma distribution based on the empirical Laplace transform. *Commun. Stat. Theory Methods*, 41(9):1543–1556.

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(5):1109–1136.

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

Hodgkinson, L., Salomone, R., and Roosta, F. (2020). The reproducing Stein kernel approach for post-hoc corrected sampling. *arXiv preprint arXiv:2001.09266*.

Holmes, S. (2004). Stein's method for birth and death chains. In *Stein's Method: Expository Lectures and Applications*, volume 46, pages 42–65. Institute of Mathematical Statistics.

Holmes, S. and Reinert, G. (2004). Stein's method for the bootstrap. In *Stein's Method: Expository Lectures and Applications*, volume 46, pages 93–132. Institute of Mathematical Statistics.

Hu, T., Chen, Z., Sun, H., Bai, J., Ye, M., and Cheng, G. (2018). Stein neural sampler. *arXiv preprint arXiv:1810.03545*.

Huggins, J. H. and Mackey, L. (2018). Random Feature Stein Discrepancies. In *Advances on Neural Information Processing Systems (NeurIPS)*, pages 1899–1909.

Hyvärinen, A. (2005). Estimation of non-normalized statistical models by score matching. In *J. Mach. Learn. Res.*, pages 695–709.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

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)*, pages 261–270.

Kattumannil, S. K. (2009). On Stein's identity and its applications. *Statist. Probab. Lett.*, 79(12):1444–1449.

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. arXiv:2111.10275*.

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)*, pages 7332–7342.

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.

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)*, volume 33.

Ledoux, M., Nourdin, I., and Peccati, G. (2015). Stein's method, logarithmic Sobolev and transport inequalities. *Geom. Funct. Anal.*, 25(1):256–306.

Leucht, A. and Neumann, M. H. (2013). Dependent wild bootstrap for degenerate U- and V-statistics. *J. Multivariate Anal.*, 117:257–280.

Ley, C., Reinert, G., and Swan, Y. (2017a). Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. *Ann. Appl. Probab.*, 27:216–241.

Ley, C., Reinert, G., and Swan, Y. (2017b). Stein's method for comparison of univariate distributions. *Probab. Surv.*, 14:1–52.

Ley, C. and Swan, Y. (2016). Parametric Stein operators and variance bounds. *Braz. J. Probab. Stat.*, 30(2):171–195.

Li, L., Li, Y., Liu, J.-G., Liu, Z., and Lu, J. (2020). A stochastic version of Stein variational gradient descent for efficient sampling. *Comm. App. Math. Comp. Sci.*, 15:37–63.

Lippert, R. A., Huang, H., and Waterman, M. S. (2002). Distributional regime for the number of k-word matches between two random sequences. *Proc. Natl. Acad. Sci. USA USA*, 99:13980–13989.

Liu, A., Liang, Y., and Broeck, G. V. d. (2020). Off-policy deep reinforcement learning with analogous disentangled exploration. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Liu, C. and Zhu, J. (2018). Riemannian Stein variational gradient descent for Bayesian inference. In *AAAI Conference on Artificial Intelligence*, pages 3627–3634.

Liu, C., Zhuo, J., Cheng, P., Zhang, R., and Zhu, J. (2019a). Understanding and accelerating particle-based variational inference. In *International Conference on Machine Learning (ICML)*, pages 4082–4092.

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)*.

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

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

Liu, Q. and Lee, J. D. (2017). Black-box importance sampling. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 952–961.

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

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances on Neural Information Processing Systems (NeurIPS)*, pages 2370–2378.

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

Liu, S., Kanamori, T., Jitkrittum, W., and Chen, Y. (2019b). Fisher efficient inference of intractable models. In *Advances on Neural Information Processing Systems (NeurIPS)*, pages 8793–8803.

Liu, Y., Ramachandran, P., Liu, Q., and Peng, J. (2017). Stein variational policy gradient. In *Uncertainty in Artificial Intelligence (UAI)*.

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(2):648–671.

Mackey, L. and Gorham, J. (2016). Multivariate Stein factors for a class of strongly log-concave distributions. *Electron. Commun. Probab.*, 21:1–14.

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). arXiv:2104.07359*.

Matsubara, T., Knoblauch, J., Briol, F.-X., and Oates, C. J. (2022). Generalised Bayesian inference for discrete intractable likelihood. *arXiv:2206.08420*.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov Chains and Stochastic Stability*. Springer Science & Business Media.

Mijatović, A. and Vogrinc, J. (2018). On the Poisson equation for Metropolis–Hastings chains. *Bernoulli*, 24(3):2401–2428.

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

Mira, A., Solgi, R., and Imparato, D. (2013). Zero variance Markov chain Monte Carlo for Bayesian estimators. *Stat. Comput.*, 23(5):653–662.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Ann. Appl. Probab.*, 29(2):pp. 429–443.

Nourdin, I. and Peccati, G. (2012). *Normal Approximations with Malliavin Calculus: from Stein's Method to Universality*, volume 192. Cambridge University Press.

Nüsken, N. and Renger, D. (2021). Stein variational gradient descent: many-particle and long-time asymptotics. *arXiv preprint arXiv:2102.12956*.

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

Oates, C. J., Girolami, M., and Chopin, N. (2017). Control functionals for Monte Carlo integration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):695–718.

Oates, C. J., Papamarkou, T., and Girolami, M. (2016). The controlled thermodynamic integral for Bayesian model comparison. *J. Amer. Statist. Assoc.*, 111(514):634–645.

Oksendal, B. (2013). *Stochastic Differential Equations: an Introduction with Applications*. Springer Science & Business Media, 6th edition.

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)*, pages 4236–4245.

Rachev, S. T., Klebanov, L., Stoyanov, S. V., and Fabozzi, F. (2013). *The Methods of Distances in the Theory of Probability and Statistics*. Springer Science & Business Media.

Ranganath, R., Tran, D., Altosaar, J., and Blei, D. (2016). Operator variational inference. In *Advances on Neural Information Processing Systems (NeurIPS)*, pages 496–504.

Reinert, G. (1998). Couplings for normal approximations with Stein's method. *Microsurveys in Discrete Probability. D. Aldous, J. Propp eds., Dimacs series. AMS*, pages 193–207.

Reinert, G. (2005). Three general approaches to stein's method. In *An Introduction to Stein's Method*, pages 183–221. World Scientific.

Reinert, G., Chew, D., Sun, F., and Waterman, M. S. (2009). Alignment-free sequence comparison (I): Statistics and power. *J. Comput. Biol.*, 16(12):1615–1634.

Reinert, G. and Ross, N. (2019). Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs. *Ann. Appl. Probab.*, 29(5):3201–3229.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., and Oates, C. (2020). Optimal thinning of MCMC output. *arXiv preprint arXiv:2005.03952*.

Ross, N. (2011). Fundamentals of Stein's method. *Probab. Surv.*, 8:210–293.

Schwartz, L. (1964). Sous-espaces hilbertiens d'espaces vectoriels topologiques et noyaux associés (noyaux reproduisants). *J. Anal. Math.*, 13(1):115–256.

Serfling, R. J. (2009). *Approximation Theorems of Mathematical Statistics*, volume 162. John Wiley & Sons.

Shao, Q.-M. (2005). An explicit Berry–Esseen bound for Student's t-statistic via Stein's method. In Chen, L. H. Y. and Barbour, A., editors, *Stein's Method And Applications*, pages 143–155. World Scientific.

Shao, Q.-M. (2010). Stein's method, self-normalized limit theory and applications. In *International Congress of Mathematicians (ICM)*, pages 2325–2350.

Shao, Q.-M., Zhang, K., and Zhou, W.-X. (2016). Stein's method for nonlinear statistics: A brief survey and recent progress. *J. Statist. Plan. Infer.*, 168:68–89.

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 arXiv:2006.07487*.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007). A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31.

Sohl-Dickstein, J., Battaglino, P., and DeWeese, M. R. (2011). Minimum probability flow learning. In *International Conference on Machine Learning*, pages 905–912.

South, L. F., Karvonen, T., Nemeth, C., Girolami, M., and Oates, C. (2020). Semi-exact control functionals from Sard's method. *arXiv preprint arXiv:2002.00033*.

South, L. F., Oates, C. J., Mira, A., and Drovandi, C. (2018). Regularised zero-variance control variates for high-dimensional variance reduction. *arXiv preprint arXiv:1811.05073*.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Technical report, Stanford University, Stanford, United States.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings 6th Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 583–602. Univ. California Press, Berkeley, Calif.

Stein, C. (1986). Approximate computation of expectations. *Lecture Notes-Monograph Series*, 7:1–164.

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*, pages 1–25. Institute of Mathematical Statistics.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, 9:1135–1151.

Sun, Z., Barp, A., and Briol, F.-X. (2021). Vector-valued control variates. *arXiv:2109.08944*.

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)*, pages 1027–1035.

Tikhomirov, A. N. (1980). On the rate of convergence in the central limit theorem for weakly dependent random variables. *Teoriya Veroyatnostei i ee Primeneniya*, 25(4):800–818.

Wang, D. and Liu, Q. (2016). Learning to draw samples: With application to amortized MLE for generative adversarial learning. *arXiv preprint arXiv:1611.01722*.

Wang, D. and Liu, Q. (2019). Nonlinear Stein variational gradient descent for learning diversified mixture models. In *International Conference on Machine Learning (ICML)*, pages 6576–6585.

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)*, pages 7834–7844.

Wang, D., Zeng, Z., and Liu, Q. (2018). Stein variational message passing for continuous graphical models. In *International Conference on Machine Learning (ICML)*, pages 5219–5227.

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning (ICML)*, pages 681–688.

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)*, pages 1575–1597.

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)*, pages 415–423.

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)*, pages 5561–5570.

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)*, pages 226–235.

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)*, volume 30, pages 6097—6106.

Zhang, X. and Curtis, A. (2019). Seismic tomography using variational inference methods. *J. Geophys. Res. Solid Earth*, 125:e2019JB018589.

Zhang, X. and Curtis, A. (2020). Variational full-waveform inversion. *Geophys. J. Int.*, 222(1):406–411.

Zhang, Y. (2019). Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.*, 10(35):8154–8163.

Zhu, Y. and Zabaras, N. (2018). Bayesian deep convolutional encoder–decoder networks for surrogate modeling and uncertainty quantification. *J. Comput. Phys.*, 366:415–447.

Zhu, Z., Wan, R., and Zhong, M. (2018). Neural control variates for variance reduction. *arXiv preprint arXiv:1806.00159*.

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)*, pages 6013–6022.

Zolotarev, V. M. (1984). Probability metrics. *Theory Probab. Appl.*, 28(2):278–302.