Probability Surveys
Vol. 14 (2017) 1–52
ISSN: 1549-5787
DOI: 10.1214/16-PS278

Stein's method for comparison of univariate distributions

Christophe Ley

Ghent University
Department of Applied Mathematics
Computer Science and Statistics
Krijgslaan 281, S9 (2nd floor)
9000 Gent, Belgium
e-mail: christophe.ley@ugent.be

Gesine Reinert

University of Oxford
Department of Statistics
1 South Parks Road, Oxford OX1 3TG, UK
e-mail: gesine.reinert@keble.ox.ac.uk

and

Yvik Swan

Université de Liège
Département de Mathématique
12 allée de la découverte
B-4000 Liège, Belgium
e-mail: yswan@ulg.ac.be

Abstract: We propose a new general version of Stein's method for univariate distributions. In particular we propose a canonical definition of the *Stein operator* of a probability distribution which is based on a linear difference or differential-type operator. The resulting *Stein identity* highlights the unifying theme behind the literature on Stein's method (both for continuous and discrete distributions). Viewing the Stein operator as an operator acting on pairs of functions, we provide an extensive toolkit for distributional comparisons. Several abstract approximation theorems are provided. Our approach is illustrated for comparison of several pairs of distributions: normal vs normal, sums of independent Rademacher vs normal, normal vs Student, and maximum of random variables vs exponential, Fréchet and Gumbel.

MSC 2010 subject classifications: Primary 60B10; secondary 60E15, 60E05, 60F05.
Keywords and phrases: Density approach, Stein's method, comparison of distributions.

Received April 2016.

Contents

1. Introduction

Stein's method is a popular tool in applied and theoretical probability, widely used for Gaussian and Poisson approximation problems. The principal aim of the method is to provide quantitative assessments in distributional comparison statements of the form $W \approx Z$ where Z follows a known and well-understood probability law (typically normal or Poisson) and W is the object of interest. To this end, Charles Stein [90] in 1972 laid the foundation of what is now called "Stein's method". For Poisson approximation his student Louis Chen [22] adapted the method correspondingly, and hence for Poisson approximation the

method is often called "Stein-Chen method" or "Chen-Stein method". In recent years a third very fruitful area of application was born from Ivan Nourdin and Giovanni Peccati's pathbreaking idea to intertwine Stein's method and Malliavin calculus. First proposed in [70], this aspect of the method is now referred to as Malliavin-Stein (or Nourdin-Peccati) analysis. For an overview we refer to the monographs [91, 8, 71, 24] as well as Ivan Nourdin's dedicated webpage https://sites.google.com/site/malliavinstein.

Outside of the Gaussian and Poisson frameworks, for univariate distributions the method has now also been shown to be effective for: exponential approximation [19, 79], Gamma approximation [68, 81, 70], binomial approximation [31], Beta approximation [43, 28], the asymptotics of rank distributions [35], inverse and variance Gamma approximation [38, 37], Laplace approximation [82], negative binomial approximation [7] or semicircular approximation [46, 47]. It can also be tailored for specific problems such as preferential attachment graphs [80], the Curie-Weiss model [21], and other models from statistical mechanics [32, 33]. This list is by no means exhaustive and we refer the reader to the webpage https://sites.google.com/site/steinsmethod for an accurate overview of this rapidly moving field. For a target distribution for which Stein's method has not yet been developed, setting up the method can appear daunting. In this paper we give a straightforward yet very flexible framework which not only encompasses the known examples but is also able to cover any new distributions which can be given in explicit form.

Broadly speaking, Stein's method consists of two distinct components, namely

> Part A: a framework allowing to convert the problem of bounding the error in the approximation of W by Z into a problem of bounding the expectation of a certain functional of W.
> Part B: a collection of techniques to bound the expectation appearing in Part A; the details of these techniques are strongly dependent on the properties of W as well as on the form of the functional.

For a target probability distribution P with support $\mathcal{I}$, Part A of the method can be sketched as follows. First find a suitable operator $\mathcal{A}:=\mathcal{A}_{\mathrm{P}}=\mathcal{A}_Z$ (called *Stein operator*) and a wide class of functions $\mathcal{F}(\mathcal{A}) := \mathcal{F}(\mathcal{A}_{\mathrm{P}})=\mathcal{F}(\mathcal{A}_Z)$ (called *Stein class*) such that

$$Z \sim \mathrm{P} \text{ if and only if } \mathbb{E}[\mathcal{A}f(Z)] = 0 \text{ for all } f \in \mathcal{F}(\mathcal{A}) \tag{1}$$

(where $Z \sim \mathrm{P}$ means that Z has distribution P). This equivalence is called a *Stein characterization* of P. Next let $\mathcal{H}$ be a measure-determining class on $\mathcal{I}$. Suppose that for each $h \in \mathcal{H}$ one can find a solution $f = f_h \in \mathcal{F}(\mathcal{A})$ of the *Stein equation*

$$h(x) - \mathbb{E}[h(Z)] = \mathcal{A}f(x), \tag{2}$$

where $Z \sim \mathrm{P}$. Then, if taking expectations is permitted, we have

$$\mathbb{E}[h(W)] - \mathbb{E}[h(Z)] = \mathbb{E}\left[\mathcal{A}f(W)\right]. \tag{3}$$

There exist a number of probability distances (such as the Kolmogorov, the Wasserstein, and the Total Variation distance) which can be represented as *integral probability metrics* of the form

$$d_{\mathcal{H}}(W,Z) = \sup_{h\in\mathcal{H}} |\mathbb{E}[h(W)] - \mathbb{E}[h(Z)]|\,,$$

see [71, Appendix C] or [40, 83] for an overview. From (3) we get

$$d_{\mathcal{H}}(W,Z) \leq \sup_{f\in\mathcal{F}(\mathcal{H})} |\mathbb{E}\,[\mathcal{A}f(W)]| \tag{4}$$

where $\mathcal{F}(\mathcal{H}) = \{f_h \,|\, h\in\mathcal{H}\}$ is the collection of solutions of (2) for functions $h\in\mathcal{H}$.

When only certain features of W are known, for example that it is a sum of weakly dependent random variables, then (4) is the usual starting point for Part B of Stein's method. Now suppose that, furthermore, a Stein operator $\mathcal{A}_W$ (and a class $\mathcal{F}(\mathcal{A}_W)$) is available for W. Suppose also that $\mathcal{F}(\mathcal{A}_Z)\cap\mathcal{F}(\mathcal{A}_W)\neq\emptyset$ and choose $\mathcal{H}$ such that all solutions f of the Stein equation (2) for $\mathcal{A}_Z$ and $\mathcal{A}_W$ belong to this intersection. Then

$$\begin{aligned}\mathbb{E}[h(W)] - \mathbb{E}[h(Z)] &= \mathbb{E}[\mathcal{A}_Z f(W)]\\ &= \mathbb{E}[\mathcal{A}_Z f(W)] - \mathbb{E}[\mathcal{A}_W f(W)]\end{aligned}$$

(because $\mathbb{E}[\mathcal{A}_W f(W)] = 0$) and

$$d_{\mathcal{H}}(W,Z) \leq \sup_{f\in\mathcal{F}(\mathcal{A}_Z)\cap\mathcal{F}(\mathcal{A}_W)} |\mathbb{E}[\mathcal{A}_W f(W) - \mathcal{A}_Z f(W)]|. \tag{5}$$

Stein [90] discovered the magical relation that the r. h. s. of (4) or (5) provides a handle to assess the proximity between the laws of W and Z; this is precisely the object of Part B of Stein's method.

In many cases, not only are the functions f_h well-defined, but also they possess smoothness properties which render them particularly amenable to computations. Also there exist many ways by which one can evaluate $\mathbb{E}\,[\mathcal{A}f(W)]$ or $\mathbb{E}[\mathcal{A}_W f(W) - \mathcal{A}_Z f(W)]$ (even under unfavorable assumptions on W) including exchangeable pairs (as for example in [91, 52, 85, 21, 19, 28]), biasing mechanisms (as in [4, 44, 41, 79, 37]), and other couplings (as in [22, 9]); see [84, 87, 18] for overviews. Nourdin and Peccati [70] paved the way for many elegant results in the context of Malliavin calculus, for an overview see [71]. See also [52, 34, 35, 43, 28, 36] for examples where direct comparison (using the explicit distribution of W) via (5) is used.

Of course the devil is in the detail and the quest for suitable Stein operators which are tractable to deal with for the random variables in question is essential for the method to be effective. While no precise definition of what exactly a *Stein operator* is exists, most authors have used Stein operators which were differential operators (or difference operators in the case of discrete distributions) obtained through a suitable variation of one of the four following classical constructions:

- Stein's *density approach* pioneered in [91] relies on the target having an explicit density p (either continuous or discrete) and then using integration by parts and classical theory of ordinary differential (or difference) equations to characterize p (see [21, 28, 65, 75, 92] for the continuous case, [35, 64, 75] for the discrete case).
- Barbour and Götze's *generator approach* (see [5, 48]) is based on classical theory of Markov processes; this approach has the added advantage of also providing a probabilistic intuition to all the quantities at play. Some references detailing this approach for univariate distributions are [30, 34, 43, 52, 58, 59].
- Diaconis and Zabell's *orthogonal polynomial approach* (see [26]) uses Rodrigues type formulas, if available, for orthogonal polynomials associated with the target distribution. See also [88] as well as [2] and related references for an extensive study of Stein operators for the Pearson (or Ord) family of distributions.
- Probability transformations such as the size bias transformation [4] and the zero bias transformation [41] characterize a distribution through being the unique fixed point of a transformation. See also [42] and [80] for examples.

These four approaches are by no means hermetically separated: often the operators derived by one method are simple transformations of those derived by another one. See for instance [42] for a very general theory on the connection between Stein operators, probability transforms and orthogonal polynomials. Other methods of constructing Stein operators are available. In [93] Stein operators for discrete compound distributions are derived by exploiting properties of the moment generating function. In [3], both Fourier and Malliavin-based approaches are used to derive operators for targets which can be represented as linear combinations of independent chi-square random variables. An algebraic study of Stein operators is initiated in [37] with developments and applications in [39] and explicit bounds on some solutions provided in [29]. The parametric approach presented in [63, 66] laid the foundation to the current work.

Outline of the paper

In this paper we propose a generalization of Stein's density approach, in the spirit of [64, 66, 65] which leads to a canonical definition of "the" differential-type operator associated to any given density. The definition is canonical, or parsimonious, in the sense that, given a target p, we identify minimal conditions under which a Stein characterization of the form (1) can hold. Moreover we will show with a wealth of examples that all the "useful" operators mentioned in the Introduction can be derived as (sometimes not so straightforward) transformations of our operator.

In Section 2 we introduce our approach in the simplest setting: distributions with continuous probability density function. Two easy applications are provided. In Section 3 we establish the set-up and introduce our toolbox in full

generality. In Section 4 we discuss different important particular cases (which we call standardizations), hereby linking our approach with the classical literature on the topic. In Section 5 we provide abstract approximation theorems for comparing probability distributions. In Section 6 we illustrate the power of our approach by tackling applications to specific approximation problems.

2. The Stein operator for differentiable probability density functions

In this section we sketch our approach in the simplest setting: X has absolutely continuous probability density function (pdf) p with respect to the Lebesgue measure on $\mathbb{R}$. Furthermore we suppose that p has interval support $\mathcal{I}$ (i.e. $p(x) > 0$ for all $x \in \mathcal{I}$, some real interval which could be unbounded); we denote a, b the boundary points of $\mathcal{I}$.

2.1. The Stein operator

Definition 2.1. *The Stein class for p is the collection $\mathcal{F}(p)$ of functions $f : \mathbb{R} \to \mathbb{R}$ such that (i) $x \mapsto f(x)p(x)$ is differentiable, (ii) $x \mapsto (f(x)p(x))'$ is integrable and (iii) $\lim_{x\uparrow b} f(x)p(x) = \lim_{x\downarrow a} f(x)p(x) = 0$. The (differential) Stein operator for p is the differential operator $\mathcal{T}_p$ defined by*

$$f \mapsto \mathcal{T}_p f := \frac{(fp)'}{p} \tag{6}$$

with the convention that $\mathcal{T}_p f(x) = 0$ for x outside of $\mathcal{I}$.

Remark 2.2. *Condition (ii) in Definition 2.1 may easily be relaxed, e.g. by only imposing that $\int_a^b (f(x)p(x))'dx =: [f(x)p(x)]_a^b = 0$. This condition could also be dispensed with entirely, although this necessitates to re-define the operator as $\mathcal{T}_p f = (fp)'/p - [f(x)p(x)]_a^b$. See also Remark 3.10.*

Remark 2.3. *It should be stressed that the assumptions on $f \in \mathcal{F}(p)$ can be quite stringent, depending on the properties of p. There is, for instance, no guarantee a priori that constant functions $f \equiv 1$ belong to $\mathcal{F}(p)$, as this requires that p cancels at the edges of its support and is differentiable with integrable derivative; such assumptions are satisfied neither in the case of an exponential target nor in the case of a beta target.*

Obviously we can always expound the derivative in (6) (at least formally, because care must be taken with the implicit indicator functions) to obtain the equivalent expression

$$\mathcal{T}_p f(x) = f'(x) + \frac{p'(x)}{p(x)} f(x) \tag{7}$$

whose form is reminiscent of the operator advocated by [92, 21]. In our experience, however, operator (7) is a priori not necessarily useful in that form because most of the conditions inherited from p are still implicit in the properties of f, as illustrated in Examples 2.4 and 2.5.

Example 2.4. *If* $p(x) \propto (x(1-x))^{-1/2}\mathbb{I}[0,1]$ *then* $\mathcal{F}(p)$ *is the collection of functions* $f : \mathbb{R} \to \mathbb{R}$ *such that* $f(x)/\sqrt{x(1-x)}$ *is differentiable with integrable derivative and with the limiting behavior* $\lim_{x\to 0,1} f(x)/\sqrt{x(1-x)} = 0$. *Operator* (7) *becomes* $\mathcal{T}_p f(x) = f'(x) + (2x-1)/(2x(1-x))f(x)$. *The operator is cumbersome but nevertheless well defined at all points* $x \in [0,1]$ *thanks to the conditions on* $f \in \mathcal{F}(p)$. *In particular these conditions ensure that* $f(x)$ *cancels at 0 and 1 faster than* $p(x)$ *diverges. Taking functions of the form* $f(x) = (x(1-x))^{\alpha} f_0(x)$ *with* $\alpha > 1/2$ *suffices. For instance the choice* $\alpha = 1$ *yields the operator* $\mathcal{A}_p f_0(x) = \mathcal{T}_p f(x) = x(1-x)f_0'(x) + \left(\frac{1}{2} - x\right) f_0(x)$ *used in [43, 28] for Beta approximation.*

The pair $(\mathcal{T}_p, \mathcal{F}(p))$ is uniquely associated to p. By choosing to focus on different subclasses $\mathcal{F}(\mathcal{A}_p) \subset \mathcal{F}(p)$ one obtains different operators acting on different sets of functions.We call the passage from $(\mathcal{T}_p, \mathcal{F}(p))$ to $(\mathcal{A}_p, \mathcal{F}(A_p))$ a *parameterization* of the Stein operator. There remains full freedom in the choice of this explicit form and it remains necessary to further understand the properties of p in order to select those functions $f \in \mathcal{F}(p)$ for which (7) will assume the most tractable expression. In Example 2.4 this is achieved by a simple transformation of the test functions; in other cases the transformations are much more complex and the resulting operators are not even necessarily of first order.

Example 2.5 (Kummer-U distribution)**.** *Let* $U(a,b,z)$ *be the unique solution of the differential equation* $zd^2U/dz^2 + (b-z)dU/dz - aU = 0$. *Then* $U(a,b,z)$ *is the confluent hypergeometric function of the second kind (also known as the Kummer U function). A random variable* X *follows the Kummer-U distribution* K_s *if its density is*

$$\kappa_s(x) = \Gamma(s)\sqrt{\frac{2}{s\pi}}\exp\left(\frac{-x^2}{2s}\right) V_s(x)\mathbb{I}(x \in (0,\infty)), \quad s \geq 1/2,$$

with $\Gamma(s)$ *the Gamma function and* $V_s(x) = U\left(s-1, \frac{1}{2}, \frac{x^2}{2s}\right)$. *The class* $\mathcal{F}(\kappa_s)$ *contains all differentiable functions such that* $\lim_{x\to 0 \text{ or } \infty} f(x)\kappa_s(x) = 0$. *As noted in [80], the canonical Stein operator (as given in* (7)*) is cumbersome. One can show by direct computations that for differentiable* f_0 *we have*

$$\frac{\left(\kappa_s(x)\frac{(f_0(x)V_s(x))'}{V_s(x)}\right)'}{\kappa_s(x)} = sf_0''(x) - xf_0'(x) - 2(s-1)f_0(x) =: \mathcal{A}_0(f_0)(x)$$

for $x > 0$, *which suggests to consider functions* $f \in \mathcal{F}(\kappa_s)$ *of the form*

$$f(x) = \frac{(f_0(x)V_s(x))'}{V_s(x)},$$

hereby providing a new derivation of the second order *operator given in [80, Lemma 3.1, Lemma 3.2] where Stein's method was first set up for this distribution.*

2.2. The generalized Stein covariance identity

Given a function $f \in \mathcal{F}(p)$, we now introduce a *second* class of functions which contains all $g : \mathbb{R} \to \mathbb{R}$ which satisfy the integration by parts identity:

$$\int_a^b g(x)(f(x)p(x))'dx = -\int_a^b g'(x)(f(x)p(x))dx. \tag{8}$$

It is easy to deduce conditions under which (8) holds; these are summarized in the next definition.

Definition 2.6. *Let p be as above. To each $f \in \mathcal{F}(p)$ we associate $\mathcal{G}(p, f)$, the collection of functions such that*

(i) $x \mapsto |g(x)(f(x)p(x))'|$, $x \mapsto |g'(x)(f(x)p(x))|$ *are both integrable on $\mathcal{I}$;*
(ii) $[g(x)\, f(x)\, p(x)]_a^b = 0$.

We also define $\mathcal{G}(p) = \bigcap_{f \in \mathcal{F}(p)} \mathcal{G}(p, f)$, and call these functions the test functions for p.

If $\mathcal{F}(p)$ is not empty then neither are $\mathcal{G}(p, f)$ and $\mathcal{G}(p)$ because they must contain the constant function $g \equiv 1$. Rewriting identity (8) in terms of the Stein pair $(\mathcal{T}_p, \mathcal{F}(p))$ leads to the *generalized Stein covariance identity*

$$\mathbb{E}\left[g(X)\mathcal{T}_p f(X)\right] = -\mathbb{E}\left[g'(X)f(X)\right] \text{ for all } f \in \mathcal{F}(p) \text{ and } g \in \mathcal{G}(p, f). \tag{9}$$

This identity generalizes several fundamental probabilistic integration by parts formulas. For instance if, on the one hand, $f \equiv 1 \in \mathcal{F}(p)$ then $\mathcal{G}(p, 1)$ contains all $g : \mathbb{R} \to \mathbb{R}$ that are absolutely continuous with compact support and

$$\mathbb{E}\left[g(X)\rho(X)\right] = -\mathbb{E}\left[g'(X)\right] \text{ for all } g \in \mathcal{G}(p, 1),$$

with $\rho = \mathcal{T}_p 1$ the score function of X. If, on the other hand, $\mathbb{E}[X] = \mu$ is finite then choosing $f = \tau_p$ with

$$\tau_p(x) = \frac{1}{p(x)} \int_x^\infty (y - \mu)p(y)dy \tag{10}$$

the so-called *Stein kernel* of p and $\mathcal{G}$ the corresponding collection of functions leads to Stein's classical covariance identity

$$\mathbb{E}\left[(X - \mu)g(X)\right] = \mathbb{E}\left[\tau_p(X)g'(X)\right] \text{ for all } g \in \mathcal{G}(p, \tau_p)$$

(note how by definition $\mathcal{T}_p(\tau_p)(x) = \mu - x$); it is easy to see that it suffices that g be differentiable and bounded at the edges of the support of $\mathcal{I}$. This approach was first studied in [91] (see also [28, 58, 2]).

Remark 2.7. *Equation (9) also provides an alternative definition of the Stein operator (6) as some form of skew-adjoint operator to the derivative with respect to integration in pdx.*

2.3. Stein characterizations

In Section 3.5 we will show that, under reasonable assumptions on p, the classes $\mathcal{F}(p)$ and $\mathcal{G}(p)$ are sufficiently large to ensure that (9) also characterizes the distribution p. This realization leads to a collection of versions of the Stein characterization (1). For example, we shall prove that

$$\begin{aligned} &\text{for each } g \in \mathcal{G}(p), \\ Y \quad \sim \quad p \Longleftrightarrow \mathbb{E}\left[g(Y)\mathcal{T}_p f(Y)\right] &= -\mathbb{E}\left[g'(Y)f(Y)\right] \text{ for all } f \in \mathcal{F}(p); \end{aligned} \tag{11}$$

and

$$\begin{aligned} &\text{for each } f \in \mathcal{F}(p), \\ Y \quad \sim \quad p \Longleftrightarrow \mathbb{E}\left[g(Y)\mathcal{T}_p f(Y)\right] &= -\mathbb{E}\left[g'(Y)f(Y)\right] \text{ for all } g \in \mathcal{G}(p,f). \end{aligned} \tag{12}$$

We refer to Section 3.5 for more information as well as a precise statement of the conditions on p under which such characterizations hold.

The freedom of choice for test functions f and g implies that many different characterizations can be immediately deduced from (11), (12) or more generally from (9). For example taking $g = 1$ in (11) we obtain

$$Y \sim p \Longleftrightarrow \mathbb{E}\left[f'(X) + f(X)\frac{p'(X)}{p(X)}\right] = 0 \text{ for all } f \in \mathcal{F}(p) \tag{13}$$

with $\mathcal{F}(p)$ the functions such that $(fp)'$ is integrable with integral 0. If one is allowed to take $f = 1$ in (12) then we deduce the characterization

$$Y \sim p \Longleftrightarrow \mathbb{E}\left[g(Y)\frac{p'(Y)}{p(Y)}\right] = -\mathbb{E}\left[g'(Y)\right] \text{ for all } g \in \mathcal{G}(p,1), \tag{14}$$

with $\mathcal{G}(p,1)$ the functions such that gp' and $g'p$ are integrable and gp has integral 0. Although the difference between (13) and (14) may be subtle, the last characterization is more in line with the classical literature on the topic to be found e.g. in [21]'s general approach (the specific conditions outlined in [21] for their approach to work out guarantee that $1 \in \mathcal{F}(p)$).

2.4. Stein equations and Stein factors

The heuristic behind Stein's method outlined in the Introduction is that if $X \sim p$ is characterized by $\mathbb{E}\left[\mathcal{A}_X f(X)\right] = 0$ over the class $\mathcal{F}(\mathcal{A}_X)$ then $\Delta_f(Y,X) := |E\left[\mathcal{A}_X f(Y)\right]|$ ought to be a good measure of how far the law of Y is from that of X. Considering equations such as (3) leads to the conclusion that indeed $\sup_f \Delta_f(Y,X)$ provides a bound on all integral probability metrics such as (4).

A similar reasoning starting from the generalized Stein covariance identity (9) encourages us to consider, for some well chosen test function h, generalized Stein equations of the form

$$g(x)\mathcal{T}_p f(x) + g'(x)f(x) = h(x) - \mathbb{E}[h(X)] \tag{15}$$

(these are now equations in two unknown functions) and the corresponding quantities

$$\Delta_{f,g}(X,Y) = |\mathbb{E}\left[g(Y)\mathcal{T}_p f(Y) + g'(Y)f(Y)\right]| \tag{16}$$

for $f \in \mathcal{F}(p)$ and $g \in \mathcal{G}(p,f)$.

There are many ways to exploit the freedom of choice of test functions (f,g) in (16). A clear aim is to choose these functions in such a way that the expression is as manageable as possible and to this end it is natural to consider $f \in \mathcal{F}(p)$ such that

$$\mathcal{T}_p(f) = h \tag{17}$$

for some well-chosen h. Obviously for (17) to make sense it is necessary that h has mean 0 and, in this case, it is easy to solve this first order equation, at least formally. Introducing the class $\mathcal{F}^{(0)}(p)$ of functions with p-mean 0 we are now in a position to introduce the *inverse Stein operator*

$$\mathcal{T}_p^{-1} : \mathcal{F}^{(0)}(p) \mapsto \mathcal{F}(p) : h \mapsto \frac{1}{p(x)} \int_a^x h(u)p(u)du. \tag{18}$$

Similarly as with the differential Stein operator $\mathcal{T}_p$, the integral operator $\mathcal{T}_p^{-1}$ is uniquely associated to p.

Example 2.8. *The Stein kernel* (10) *is* $\mathcal{T}_p^{-1}h$ *with* h *the (recentered) identity function.*

In general one will choose f and g in such a way as to ensure that (i) both $\mathcal{T}_p f$ and f have agreeable expressions, and (ii) solutions to (15) have good properties, hereby ensuring that (16) is amenable to computations. We will show in Sections 5 and 6 that this is the case for a wide variety of target distributions. Given $\mathcal{H} \subset \mathcal{F}^{(0)}$, constants such as

$$\sup_{h\in\mathcal{H}} \left\|\mathcal{T}_p^{-1}h\right\|_\infty \qquad \text{and} \qquad \sup_{h\in\mathcal{H}} \left\|\left(\mathcal{T}_p^{-1}h\right)'\right\|_\infty \tag{19}$$

will play an important role in the success of the method. These are usually referred to as the *Stein factors* of p, and there is already a large body of literature dedicated to their study under various assumptions on p, see e.g. [12, 86, 7, 29].

2.5. Comparing probability densities by comparing Stein operators

Now let X_1 and X_2 have densities p_1, p_2 with supports $\mathcal{I}_1, \mathcal{I}_2$ and Stein pairs $(\mathcal{T}_1, \mathcal{F}_1)$ and $(\mathcal{T}_2, \mathcal{F}_2)$, respectively. Equation (15) leads to an ensemble of *Stein equations* for $X_i, i = 1, 2$ of the form

$$h(x) - \mathbb{E}[h(X_i)] = g'(x)f(x) + g(x)\mathcal{T}_i f(x) \tag{20}$$

whose solutions are now pairs $(f,g) \in \mathcal{F}(p_i) \times \mathcal{G}(p_i)$. Given a sufficiently regular function h then any pair $f_i, g_i \in \mathcal{F}(p_i) \times \mathcal{G}(p_i)$ satisfying

$$f_i(x)g_i(x) = \frac{1}{p_i(x)} \int_{a_i}^x p_i(u) \left(h(u) - \mathbb{E}[h(X_i)]\right) du \tag{21}$$

(with $a_i, i = 1, 2$ the lower edge of $\mathcal{I}_i$) is a solution to (20) for $i = 1, 2$. Functions such as the one on the rhs of (21) have been extensively studied, see e.g. [91, 58].

There are many starting points from here. For example taking differences between Equations (20) for $i = 1, 2$ leads to the unusual identity

$$\begin{aligned} &\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)] \\ &= (g_1'(x)f_1(x) - g_2'(x)f_2(x)) + (g_1(x)\mathcal{T}_1 f_1(x) - g_2(x)\mathcal{T}_2 f_2(x)) \end{aligned} \tag{22}$$

for all $x \in \mathcal{I}_1 \cap \mathcal{I}_2$ and all $(f_i, g_i) \in \mathcal{F}(p_i) \times \mathcal{G}(p_i)$ which satisfy (21). It is more usual to exploit (20) as follows. Pick a solution (f_1, g_1) of (21) and a pair $(f_2, g_2) \in \mathcal{F}(p_2) \times \mathcal{G}(p_2)$, which ensures that $\mathbb{E}[g_2'(X_2)f_2(X_2) + g_2(X_2)\mathcal{T}_2 f_2(X_2)] = 0$ while (f_2, g_2) is not required to satisfy (21). Taking expectations in X_2 on both sides of (20) yields

$$\begin{aligned} \mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)] &= \mathbb{E}\left[g_1'(X_2)f_1(X_2) + g_1(X_2)\mathcal{T}_1 f_1(X_2)\right] \\ &= \mathbb{E}\left[g_1'(X_2)f_1(X_2) - g_2'(X_2)f_2(X_2)\right] \\ &\quad -\mathbb{E}\left[g_1(X_2)\mathcal{T}_1 f_1(X_2) - g_2(X_2)\mathcal{T}_2 f_2(X_2)\right], \end{aligned} \tag{23}$$

under the additional assumption that all expectations exist. Identity (23) is a powerful starting point for stochastic approximation problems, as one can handpick the functions $f_i, i = 1, 2$ and $g_i, i = 1, 2$ best suited to the problem under study.

- Assume that $f_1 = f_2 = 1$ is permitted and that g_1, defined in (21), belongs to $\mathcal{G}(p_2)$. Then from (23) we deduce that

  $$\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)] = \mathbb{E}\left[g_1(X_2)\left(\rho_2(X_2) - \rho_1(X_2)\right)\right]$$

  where ρ_i is the score function of X_i. This identity (which holds as soon as $g_1 \in \mathcal{F}(p_2)$) in turn leads to the Fisher information inequalities studied, e.g., in [89, 54, 65].
- Assume that X_1, X_2 both have mean ν and pick f_1, f_2 such that $\mathcal{T}_1 f_1 = \mathcal{T}_2 f_2 = x - \nu$. Let g_1 be the corresponding function from (21) and assume that $g_1 \in \mathcal{G}(p_2)$. Then

  $$\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)] = \mathbb{E}\left[g_1'(X_2)\left(\tau_1(X_2) - \tau_2(X_2)\right)\right] \tag{24}$$

  where τ_i is the Stein kernel of X_i. From here one readily recovers the key inequalities from [17, 13]. This is also the starting point of the Nourdin-Peccati approach to Stein's method [71].

Many other identities can be obtained. We have recently applied this result to the computation of explicit bounds in a problem of Bayesian analysis, see [62]. Several applications will be provided in Sections 5 and 6. We conclude this section with two easy applications.

2.6. Application 1: Rates of convergence to the Fréchet distribution

Let X_α follow the Fréchet distribution with tail index α so that

$$P(X_\alpha \leq x) =: \Phi_\alpha(x) = \exp(-x^{-\alpha})\mathbb{I}(x \geq 0).$$

The Stein class $\mathcal{F}(\alpha)$ for the Fréchet is the collection of all differentiable functions f on $\mathbb{R}$ such that $\lim_{x\to+\infty} f(x)x^{-\alpha-1}e^{-x^{-\alpha}} = \lim_{x\to 0} f(x)x^{-\alpha-1}e^{-x^{-\alpha}} = 0$. We restrict our attention to functions of the form $f(x) = x^{\alpha+1}f_0(x)$. In this parameterization the differential Stein operator becomes

$$\mathcal{A}_\alpha f_0(x) = x^{\alpha+1}f_0'(x) + \alpha f_0(x). \tag{25}$$

The generalized Stein equation (20) with $g = 1$ reads $x^{\alpha+1}f_0'(x) + \alpha f_0(x) = h(x) - \mathbb{E}[h(X_\alpha)]$ and, given $h(x) = \mathbb{I}(x \leq z)$, has unique bounded solution

$$f_z(x) = \frac{1}{\alpha}\left(\Phi_\alpha(x \wedge z)/\Phi_\alpha(x) - \Phi_\alpha(z)\right). \tag{26}$$

This function is continuous and differentiable everywhere except at $x = z$; it satisfies $0 \leq \alpha f_z(x) \leq 1$ for all $x, z \geq 0$ as well as $\lim_{x\to+\infty} f_z(x) = 0$.

Next take $F(x) = (1 - x^{-\alpha})\mathbb{I}(x \geq 1)$ the Pareto distribution and for $n \geq 1$ consider the random variable $W_n = M_n/n^{1/\alpha}$. Its probability density function is $p_n(x) = \alpha x^{-\alpha-1}\left(1 - x^{-\alpha}/n\right)^{n-1}$ on $[n^{-1/\alpha}, +\infty)$. For each n the random variable W_n has a Stein pair $(\mathcal{T}_n, \mathcal{F}(n))$, say. In order to compare with the Fréchet distribution we consider the standardization

$$\mathcal{A}_n(f_0)(x) = \frac{(x^{\alpha+1}f_0(x)p_n(x))'}{p_n(x)} = x^{\alpha+1}f_0'(x) + \alpha\frac{n-1}{n}\left(1 - \frac{x^{-\alpha}}{n}\right)^{-1} f_0(x)$$

with f_0 an absolutely continuous function such that

$$\lim_{x\to+\infty} x^{\alpha+1}f_0(x)p_n(x) = \lim_{x\to n^{-1/\alpha}} x^{\alpha+1}f_0(x)p_n(x) = 0.$$

The function f_z given in (26) satisfies these two constraints. Hence $\mathbb{E}\left[\mathcal{A}_n(f_z)(W_n)\right] = 0$ and from (23) we get in this particular case

$$P(W_n \leq z) - \Phi_\alpha(z) = \alpha\mathbb{E}\left[f_z(W_n)\left(1 - \frac{n-1}{n}\left(1 - \frac{W_n^{-\alpha}}{n}\right)^{-1}\right)\right].$$

The function $x \mapsto 1 - \frac{n-1}{n}\left(1 - \frac{x^{-\alpha}}{n}\right)^{-1}$ is negative for all $x \geq n^{-1/\alpha}$. Also, it is easy to compute explicitly $E\left[\frac{n-1}{n}\left(1 - \frac{W_n^{-\alpha}}{n}\right)^{-1} - 1\right] = \frac{2}{n-1}\left(1 - \frac{1}{n}\right)^n$. We deduce the upper bound

$$\sup_{z\in\mathbb{R}} |P(W_n \leq z) - \Phi_\alpha(z)| \leq \frac{2e^{-1}}{n-1}.$$

More general bounds of the same form can be readily obtained for maxima of independent random variables satisfying adhoc tail assumptions.

2.7. Application 2: A CLT for random variables with a Stein kernel

Let $X_1, \dots, X_n$ be independent centered continuous random variables with unit variance and Stein kernels $\tau_1, \dots, \tau_n$ as given by (10). Also let Z be a standard normal random variable independent of all else. The standard normal random variable (is characterized by the fact that it) has constant Stein kernel $\tau_Z(x) = 1$. Finally let $W = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$. We will prove in Section 5.3 that

$$\tau_W(w) = \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\tau_i(X_i) \,|\, W = w\right] \tag{27}$$

(see Proposition 5.12) which we can use in (24) (setting $X_2 = W$ and $X_1 = Z$) to deduce that

$$\begin{aligned} \mathbb{E}\left[h(W)\right] - \mathbb{E}\left[h(Z)\right] &= \frac{1}{n}\mathbb{E}\left[g_1'(W) \sum_{i=1}^n (1 - \tau_i(X_i))\right] \\ &\leq \frac{1}{n} \sqrt{\mathbb{E}\left[g_1'(W)^2\right] \mathbb{E}\left[\left(\sum_{i=1}^n (1 - \tau_i(X_i))\right)^2\right]}. \end{aligned}$$

Classical results on Gaussian Stein's method give that $\|g_1'\|_\infty \leq 1$ if $h(x) = \mathbb{I}(x \leq z)$, see [24, Lemma 2.3]. Also, using the fact that $\mathbb{E}[1 - \tau_i(X_i)] = 0$ for all $i = 1, \dots, n$ as well as $\mathbb{E}\left[(\tau_i(X_i) - 1)^2\right] = \mathbb{E}\left[\tau_i(X_i)^2\right] - 1$, we get

$$\mathbb{E}\left[\left(\sum_{i=1}^n (1 - \tau_i(X_i))\right)^2\right] = \mathrm{Var}\left(\sum_{i=1}^n (1 - \tau_i(X_i))\right) = \sum_{i=1}^n \left(\mathbb{E}\left[\tau_i(X_i)^2\right] - 1\right).$$

If the X_i are i.i.d. then we finally conclude that

$$\sup_z |P(W \leq z) - P(Z \leq z)| \leq \frac{1}{\sqrt{n}} \sqrt{\left(\mathbb{E}\left[\tau_1(X_1)^2\right] - 1\right)}. \tag{28}$$

Of course (28) is for illustrative purposes only because the requirement that the $X_i, i = 1, \dots, n$, possess a Stein kernel is very restrictive (even more restrictive than the existence of a fourth moment). In this application it is assumed that W has a continuous distribution; this assumption is not necessary because Stein kernels can be defined for any univariate distribution. We will provide a general version of (28) in Section 5.3.

3. The canonical Stein operator

In this section we lay down the foundations and set the framework for our general theory of canonical Stein operators.

3.1. The setup

Let $(\mathcal{X}, \mathcal{B}, \mu)$ be a measure space with $\mathcal{X} \subset \mathbb{R}$ (see Remark 3.7). Let $\mathcal{X}^\star$ be the set of real-valued functions on $\mathcal{X}$. We require the existence of a linear operator

$$\mathcal{D} : dom(\mathcal{D}) \subset \mathcal{X}^\star \to im(\mathcal{D})$$

such that $dom(\mathcal{D}) \setminus \{0\}$ is not empty. As is standard we define

$$\mathcal{D}^{-1} : im(\mathcal{D}) \to dom(\mathcal{D})$$

as the linear operator which sends any $h = \mathcal{D}f$ onto f. This operator is a right-inverse for $\mathcal{D}$ in the sense that $\mathcal{D}\left(\mathcal{D}^{-1}h\right) = h$ for all $h \in im(\mathcal{D})$ whereas, for $f \in dom(\mathcal{D})$, $\mathcal{D}^{-1}\left(\mathcal{D}f\right)$ is only defined up to addition with an element of $ker(\mathcal{D})$. We impose the following assumption.

Assumption 1. *There exists a linear operator $\mathcal{D}^\star : dom(\mathcal{D}^\star) \subset \mathcal{X}^\star \to im(\mathcal{D}^\star)$ and a constant $l := l_{\mathcal{X},\mathcal{D}}$ such that*

$$\mathcal{D}(f(x)g(x+l)) = g(x)\mathcal{D}f(x) + f(x)\mathcal{D}^\star g(x) \tag{29}$$

for all $(f, g) \in dom(\mathcal{D}) \times dom(\mathcal{D}^\star)$ and for all $x \in \mathcal{X}$.

Assumption 1 guarantees that operators $\mathcal{D}$ and $\mathcal{D}^\star$ are skew-adjoint in the sense that

$$\int_{\mathcal{X}} g\mathcal{D}f d\mu = -\int_{\mathcal{X}} f\mathcal{D}^\star g d\mu \tag{30}$$

for all $(f, g) \in dom(\mathcal{D}) \times dom(\mathcal{D}^\star)$ such that $g\mathcal{D}f \in L^1(\mu)$, or $f\mathcal{D}^\star g \in L^1(\mu)$, and in addition $\int_{\mathcal{X}} \mathcal{D}(f(\cdot)g(\cdot + l))d\mu = 0$.

Example 3.1 (Lebesgue measure)**.** *Let μ be the Lebesgue measure on $\mathcal{X} = \mathbb{R}$ and take $\mathcal{D}$ the usual strong derivative. Then*

$$\mathcal{D}^{-1}f(x) = \int_{\bullet}^{x} f(u)du$$

is the usual antiderivative. Assumption 1 is satisfied with $\mathcal{D}^\star = \mathcal{D}$ and $l = 0$.

Example 3.2 (Counting measure)**.** *Let μ be the counting measure on $\mathcal{X} = \mathbb{Z}$ and take $\mathcal{D} = \Delta^+$, the forward difference operator $\Delta^+ f(x) = f(x+1) - f(x)$. Then*

$$\mathcal{D}^{-1}f(x) = \sum_{k=\bullet}^{x-1} f(k).$$

Also we have the discrete product rule

$$\Delta^+(f(x)g(x-1)) = g(x)\Delta^+ f(x) + f(x)\Delta^- g(x)$$

for all $f, g \in \mathbb{Z}^\star$ and all $x \in \mathbb{Z}$. Hence Assumption 1 is satisfied with $\mathcal{D}^\star = \Delta^-$, the backward difference operator, and $l = -1$.

Example 3.3 (Counting measure on the grid)**.** *Let μ be the counting measure on $\mathcal{X} = \delta\mathbb{Z}$ with $\delta > 0$ and take $\mathcal{D} = \Delta_\delta^+$, the scaled forward difference operator $\Delta_\delta^+ f(x) = \delta^{-1}\left(f(x+\delta) - f(x)\right)$. The inverse $\mathcal{D}^{-1}$ is defined similarly as in the*

previous example. Setting $\Delta_\delta^- f(x) = \delta^{-1}(f(x) - f(x-\delta))$*, we have the discrete product rule*

$$\Delta_\delta^+(f(x)g(x-\delta)) = g(x)\Delta_\delta^+ f(x) + f(x)\Delta_\delta^- g(x)$$

for all $f, g \in \mathbb{Z}^\star$ *and all* $x \in \mathbb{R}$*. Hence Assumption 1 is satisfied with* $\mathcal{D}^\star = \Delta_\delta^-$ *and* $l = -\delta$*.*

Example 3.4 (Standard normal)**.** *Let* φ *be the standard normal density function so that* $\varphi'(x) = -x\varphi(x)$*. Let* $\mu(x)$ *be the standard normal measure on* $\mathbb{R}$ *and take* $\mathcal{D} = \mathcal{D}_\varphi$ *the differential operator defined by*

$$\mathcal{D}_\varphi f(x) = f'(x) - xf(x) = \frac{(f(x)\varphi(x))'}{\varphi(x)},$$

see e.g. [60]. Then

$$\mathcal{D}_\varphi^{-1} f(x) = \frac{1}{\varphi(x)} \int_\bullet^x f(y)\varphi(y)dy.$$

Also we have the product rule

$$\begin{aligned} \mathcal{D}_\varphi(gf)(x) &= (gf)'(x) - xg(x)f(x) \\ &= g(x)\mathcal{D}_\varphi f(x) + f(x)g'(x). \end{aligned}$$

Hence Assumption 1 is satisfied with $\mathcal{D}^\star g = g'$ *and* $l = 0$*.*

Example 3.5 (Poisson)**.** *Let* γ_λ *be the Poisson probability mass function with parameter* λ*. Let* $\mu(x)$ *be the corresponding Poisson measure on* $\mathbb{Z}^+$ *and take* $\mathcal{D} = \Delta_\lambda^+$ *the difference operator defined by*

$$\Delta_\lambda^+ f(x) = \lambda f(x+1) - xf(x) = \frac{\Delta^+(f(x)x\gamma_\lambda(x))}{\gamma_\lambda(x)}.$$

Then

$$(\Delta_\lambda^+)^{-1} f(x) = \frac{1}{x\gamma_\lambda(x)} \sum_{k=\bullet}^{x-1} f(k)\gamma_\lambda(k)$$

which is ill-defined at $x = 0$ *(see, e.g., [6, 8]). We have the product rule*

$$\Delta_\lambda^+(g(x-1)f(x)) = g(x)\Delta_\lambda^+ f(x) + f(x)x\Delta^- g(x).$$

Hence Assumption 1 is satisfied with $\mathcal{D}^\star g(x) = x\Delta^- g(x)$ *and* $l = -1$*.*

Remark 3.6. *In all examples considered above the choice of* $\mathcal{D}$ *is, in a sense, arbitrary and other options are available. In the Lebesgue measure setting of Example 3.1 one could, for instance, use* $\mathcal{D}$ *the derivative in the sense of distributions, or even* $\mathcal{D}f(x) = \frac{\partial}{\partial t} f(P_t x)$ *for* $x \mapsto P_t x$ *some transformation of* $\mathcal{X}$*; see e.g. [66]. In the counting measure setting of Example 3.2 the roles of backward and forward difference operators can be exchanged; these operators can also be replaced by linear combinations as, e.g., in [51]. The discrete construction is also easily extended to general spacing* $\delta \neq 1$ *: if* $\mathcal{X} = \delta\mathbb{Z}$*, then we can take* $\mathcal{D} = \Delta_\delta^+$ *such that* $\mathcal{D}f(x) = f(x+\delta) - f(x)$*. In the Poisson example one could also consider*

$$\mathcal{D}f(x) = \frac{\lambda}{x+1}f(x+1) - f(x) = \frac{\Delta^+(f(x)\gamma_\lambda(x))}{\gamma_\lambda(x)}.$$

In all cases less conventional choices of $\mathcal{D}$ can be envisaged (even forward differences in the continuous setting).

Remark 3.7. *Nowhere is the restriction to dimension 1 necessary in this subsection and our approach could in principle be transcribed verbatim to arbitrary dimension. In fact, in the wake of pathbreaking references [5, 48], much has already been achieved via Stein's method for multivariate analysis, particularly in the context of Gaussian approximation as e.g. in [44, 20, 73, 85]. Recent notable contributions to a general multivariate theory are to be found in [45, 69], [11] and [10]. A multivariate version of our "canonical" theory requires much care and we choose to delay its exposition to a later publication (in preparation). We will not discuss this topic further here.*

3.2. Canonical Stein class and operator

Following [43] we say that a subset $\mathcal{I} \subset \mathcal{X}$ is a finite interval if $\mathcal{I} = \{a, b\} \cap \mathcal{X}$ for $a, b \in \mathbb{R}$ with $a \leq b$, and an infinite interval if either $\mathcal{I} = (-\infty, b\} \cap \mathcal{X}$ or $\mathcal{I} = \{a, \infty) \cap \mathcal{X}$ or $\mathcal{I} = \mathcal{X}$ (provided, of course, that $\mathcal{X}$ itself has infinite length). Here $\{$ is used as shorthand for either (or [, and similarly $\}$ is either) or]. In the sequel we consistently denote intervals by $\mathcal{I} = \{a, b\}$ where $-\infty \leq a \leq b \leq +\infty$ (we omit the intersection with $\mathcal{X}$ unless necessary).

Now consider a real-valued random variable X on $\mathcal{X}$ such that $P_X(A) = \mathrm{P}(X \in A)$ for $A \in \mathcal{B}$ is absolutely continuous w.r.t. μ. Let $p = dP_X/d\mu$ be the Radon-Nikodym derivative of P_X; throughout we call p the *density* of X (even if X is not a continuous random variable). In the sequel, we only consider random variables such that $p \in dom(\mathcal{D})$ and whose support $supp(p) = \{x \in \mathcal{X} \,|\, p(x) > 0\} =: \mathcal{I}$ is an interval of $\mathcal{X}$. For any real-valued function h on $\mathcal{X}$ we write

$$\mathbb{E}_p h = \mathbb{E}[h(X)] = \int_{\mathcal{X}} hpd\mu = \int_{\mathcal{I}} hpd\mu;$$

this expectation exists for all functions $h : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}_p|h| < \infty$; we denote this set of functions by $L^1_\mu(p) \equiv L^1_\mu(X)$.

Definition 3.8. *The* canonical $\mathcal{D}$-Stein class $\mathcal{F}(p) \equiv \mathcal{F}(X) (= \mathcal{F}_\mu(p))$ *for X is the collection of functions $f \in L^1_\mu(p)$ such that (i) $fp \in dom(\mathcal{D})$, (ii) $\mathcal{D}(fp) \in L^1(\mu)$ and (iii) $\int_{\mathcal{I}} \mathcal{D}(fp)d\mu = 0$. The* canonical $\mathcal{D}$-Stein operator $\mathcal{T}_p \equiv \mathcal{T}_X$ *for p is the linear operator on $\mathcal{F}(X)$ defined as*

$$\mathcal{T}_X f : \mathcal{F}(X) \to L^1_\mu(p) : f \mapsto \frac{\mathcal{D}(fp)}{p}, \tag{31}$$

with the convention that $\mathcal{T}_X f = 0$ outside of $\mathcal{I}$. We call the construction $(\mathcal{T}_X, \mathcal{F}(X)) = (\mathcal{T}_p, \mathcal{F}(p))$ a $\mathcal{D}$-Stein pair for X.

Remark 3.9. *In the sequel we shall generally drop the reference to the dominating differential $\mathcal{D}$.*

To avoid triviality we from hereon assume that $\mathcal{F}(X) \setminus \{0\} \neq \emptyset$. Note that $\mathcal{F}(X)$ is closed under multiplication by constants. By definition, $\mathcal{T}_X f \in L^1_\mu(p)$ for all $f \in \mathcal{F}(X)$, and

$$\mathbb{E}[\mathcal{T}_X f(X)] = \int_{\mathcal{I}} \frac{\mathcal{D}(fp)(x)}{p(x)} p(x) d\mu(x) = \int_{\mathcal{I}} \mathcal{D}(fp)(x) d\mu(x) = 0,$$

so that $\mathcal{T}_X$ satisfies Equation (1), qualifying it as a Stein operator.

Remark 3.10. *The assumption for $\mathcal{F}(X)$ that $\int_{\mathcal{I}} \mathcal{D}(fp) d\mu = 0$ is made for convenience of calculation but it is not essential. Indeed sometimes it may be more natural not to impose this restriction. For example if μ is the continuous uniform measure on $[0,1]$ and $p = 1$, with $\mathcal{D}$ the usual derivative, then imposing that $\int_0^1 f'(x)dx = f(1) - f(0) = 0$ may not be natural. The price to pay for relaxing the assumption is that in the definition of $\mathcal{T}_X f(X)$ we would have to subtract this integral, as in [92], to assure that $\mathbb{E}[\mathcal{T}_X f(X)] = 0$.*

The canonical Stein operator (31) bears an intuitive interpretation in terms of the linear operator $\mathcal{D}$.

Proposition 3.11. *For all $f \in \mathcal{F}(X)$ define the class of functions*

$$dom(\mathcal{D}, X, f) = \{g \in dom(\mathcal{D}^\star) : g(\cdot + l)f(\cdot) \in \mathcal{F}(X), \\ \mathbb{E}|f(X)\mathcal{D}^\star(g)(X)| < \infty \text{ or } \mathbb{E}|g(X)\mathcal{T}_X f(X)| < \infty\}. \quad (32)$$

Then

$$\mathbb{E}\left[f(X)\mathcal{D}^\star(g)(X)\right] = -\mathbb{E}\left[g(X)\mathcal{T}_X f(X)\right] \quad (33)$$

for all $f \in \mathcal{F}(X)$ and all $g \in dom(\mathcal{D}, X, f)$.

Proof. Assumption 1 assures us that

$$\mathcal{D}(g(\cdot + l)f(\cdot)p(\cdot))(x) = g(x)\mathcal{D}(fp)(x) + f(x)p(x)\mathcal{D}^\star g(x)$$

for all $f \in \mathcal{F}(X)$ and all $g \in dom(\mathcal{D}^\star)$. If $g \in dom(\mathcal{D}, X, f)$ then $\int_{\mathcal{X}} \mathcal{D}(g(x + l)f(x)p(x))d\mu(x) = 0$ and

$$\begin{aligned}\mathbb{E}\left[g(X)\frac{\mathcal{D}(fp)}{p}(X)\right] &= \int_{\mathcal{I}} g\mathcal{D}(fp)d\mu \\ &= -\int_{\mathcal{I}} fp\mathcal{D}^\star(g)d\mu \\ &= -\mathbb{E}[f(X)\mathcal{D}^\star(g)(X)],\end{aligned}$$

with both integrals being finite. This yields (33). □

As anticipated in the Introduction, the relationship (33) shows that if $\mathcal{D}$ is skew-adjoint with respect to $\mathcal{D}^\star$ under integration in μ then the canonical Stein operator is skew-adjoint to $\mathcal{D}^\star$ under integration in the measure P_X. This motivates the use of the terminology "canonical" in Definition 3.8; we will further elaborate on this topic in Section 3.5.

Example 3.12 (Example 3.1, continued)**.** *Let X be a random variable with absolutely continuous density p with support $\mathcal{I} = \{a, b\}$. Then $\mathcal{F}(X)$ is the collection of functions $f : \mathbb{R} \to \mathbb{R}$ such that $fp \in W^{1,1}$ the Sobolev space of order 1 on $L^1(dx)$ and $\lim_{x \searrow a} f(x)p(x) = \lim_{x \nearrow b} f(x)p(x)$; the canonical Stein operator is*

$$\mathcal{T}_X f = \frac{(fp)'}{p}$$

which we set to 0 outside of $\mathcal{I}$. Also, for $f \in \mathcal{F}(X)$, $dom((\cdot)', X, f)$ is the class of differentiable functions $g : \mathbb{R} \to \mathbb{R}$ such that $\int (gfp)' dx = 0$, $\int |g'fp| dx < \infty$ or $\int |g(fp)'| dx < \infty$. (Note that the first requirement implicitly requires $\int |(gfp)'| dx < \infty$.) In particular all constant functions are in $dom((\cdot)', X, f)$.

In the case that p itself is differentiable (and not only the function $x \mapsto f(x)p(x)$ is) we can write

$$\mathcal{T}_X f(x) = \left(f'(x) + f(x) \frac{p'(x)}{p(x)} \right) \mathbb{I}(x \in \mathcal{I}), \tag{34}$$

with $\mathbb{I}(\cdot)$ the usual indicator function. This is operator (7) from Stein's density approach. Note that, in many cases, the constant functions may not belong to $\mathcal{F}(X)$. Operator (34) was also discussed (under slightly different – more restrictive – assumptions) in [21]. See also [65] for a similar construction.

Example 3.13 (Example 3.2, continued)**.** *Recall $\mathcal{D} = \Delta^+$ and consider X some discrete random variable whose density p has interval support $\mathcal{I} = [a, b]$ (with, for simplicity, $a > -\infty$). The associated (forward) Stein operator is*

$$\mathcal{T}_X f = \frac{\Delta^+(fp)}{p},$$

which we set to 0 outside of $\mathcal{I}$. We divide the example in two parts.

1. *If $b < +\infty$: the associated (forward) canonical Stein class $\mathcal{F}(X)$ is the collection of functions $f : \mathbb{Z} \to \mathbb{R}$ such that $f(a) = 0$, and, for $f \in \mathcal{F}(X)$, $dom(\Delta^+, X, f)$ is the collection of functions $g : \mathbb{Z} \to \mathbb{R}$.*
2. *If $b = +\infty$: the (forward) canonical Stein class $\mathcal{F}(X)$ is the collection of functions $f : \mathbb{Z} \to \mathbb{R}$ such that $f(a) = 0$ and $\sum_{n=a}^{\infty} |f(n)|p(n) < +\infty$, and for $f \in \mathcal{F}(X)$, $dom(\Delta^+, X, f)$ is the collection of functions $g : \mathbb{Z} \to \mathbb{R}$ such that $\lim_{n \to \infty} g(n-1)f(n)p(n) = 0$ and, either $\sum_{k=a}^{\infty} p(k) |f(k)\Delta^+ g(k)| < \infty$ or $\sum_{k=a}^{\infty} p(k) |g(k)\mathcal{T}_X f(k)| < \infty$. In particular all bounded functions g are in $dom(\Delta^+, X, f)$.*

If p itself is in $\mathcal{F}(X)$ then we have

$$\mathcal{T}_X f(x) = f(x+1) \frac{p(x+1)}{p(x)} - f(x).$$

Similarly it is straightforward to define a backward Stein class and operator.

Example 3.14 (Example 3.4, continued)**.** *Let X be a random variable with density p with support $\mathcal{I} = \{a, b\}$ with respect to $\varphi(x)dx$ the Gaussian measure. Recall $\mathcal{D}_\varphi f(x) = f'(x) - xf(x)$ and $\mathcal{D}^\star g(x) = g'(x)$. Then $\mathcal{F}(X)$ is the collec-*

tion of functions $f : \mathbb{R} \to \mathbb{R}$ such that $fp \in L^1(\varphi)$ is absolutely continuous, $\int_{\mathbb{R}} |\mathcal{D}_\varphi(fp)|\varphi(x)dx < \infty$ and $\lim_{x \searrow a} f(x)p(x)\varphi(x) = \lim_{x \nearrow b} f(x)p(x)\varphi(x)$; the canonical Stein operator is

$$\mathcal{T}_X f = \frac{\mathcal{D}_\varphi(fp)}{p} = \frac{(fp\varphi)'}{p\varphi}$$

which we set to 0 outside of $\mathcal{I}$. Also, for $f \in \mathcal{F}(X)$, $dom(\mathcal{D}_\varphi, X, f)$ contains all differentiable functions $g : \mathbb{R} \to \mathbb{R}$ such that $gf \in L^1_\mu(p)$ (or, equivalently, $gfp \in L^1(\varphi)$), $\int (gfp\varphi)' dx = 0$, and either $\int |g'f|p\varphi dx < \infty$ or $\int |g(fp\varphi)'|dx < \infty$. In particular all constant functions are in $dom(\mathcal{D}_\varphi, X, f)$. The above construction can also be obtained directly by replacing p with $p\varphi$ in Example 3.12.

Example 3.15 (Example 3.5, continued)**.** *Recall $\mathcal{D} = \Delta^+_\lambda$ and consider X some discrete random variable whose density p has interval support $\mathcal{I} = [0, b]$. The (forward) Stein operator is*

$$\mathcal{T}_X f = \frac{\Delta^+_\lambda(fp)}{p} = \frac{\Delta^+(f(x)xp(x)\gamma_\lambda(x))}{p(x)\gamma_\lambda(x)},$$

which we set to 0 outside of $\mathcal{I}$. Then, as in the previous example, we simply recover the construction of Example 3.13 with $f(x)$ replaced by $xf(x)$ (and thus no condition on $f(0)$) and $p(x)$ replaced by $p(x)\gamma_\lambda(x)$.

Remark 3.16. *As noted already in the classic paper [26], the abstract theory of Stein operators is closely connected to Sturm-Liouville theory. This connection is quite easy to see from our notations and framework; it remains however outside of the scope of the present paper and will be explored in future publications.*

3.3. The canonical inverse Stein operator

The Stein operator being defined (in terms of $\mathcal{D}$), we now define its inverse (in terms of $\mathcal{D}^{-1}$). To this end first note that if $\mathcal{D}(fp) = hp$ for $f \in \mathcal{F}(X)$ then $\mathcal{T}_X(f) = h$. As $\mathcal{D}(fp + \chi) = hp$ for any $\chi \in ker(\mathcal{D})$, to define a unique right-inverse of $\mathcal{T}_X$ we make the following assumption.

Assumption 2. $ker(\mathcal{D}) \cap L^1(\mu) = \{0\}$.

This assumption ensures that the only μ-integrable χ is 0 and thus $\mathcal{T}_X$ (as an operator acting on $\mathcal{F}(X) \subset L^1(\mu)$) possesses a *bona fide* inverse, and also that $ker(\mathcal{D}) \cap L^1_\mu(p) = \{0\}$.

Definition 3.17. *Let X have density p with support $\mathcal{I}$. The* canonical inverse Stein operator $\mathcal{T}_p^{-1} \equiv \mathcal{T}_X^{-1}$ *for X is defined for all h such that $hp \in im(\mathcal{D})$ as the unique function $f \in \mathcal{F}(X)$ such that $\mathcal{D}(fp) = hp$.*

We will use the shorthand

$$\mathcal{T}_X^{-1} h = \frac{\mathcal{D}^{-1}(hp)}{p}$$

with the convention that $\mathcal{T}_X^{-1} h = 0$ outside of $\mathcal{I}$.

We state the counterpart of Proposition 3.11 for the inverse Stein operator.

Proposition 3.18. *Define the class of functions*

$$\mathcal{F}^{(0)}(X) = \{h \in im(\mathcal{T}_X) : hp = \mathcal{D}(fp) \text{ with } f \in \mathcal{F}(X)\}.$$

Then

$$\mathbb{E}\left[\mathcal{T}_X^{-1}h(X)\mathcal{D}^\star g(X)\right] = -\mathbb{E}\left[g(X)h(X)\right] \tag{35}$$

for all $h \in \mathcal{F}^{(0)}(X)$ *and all* $g \in dom(\mathcal{D}, X, \mathcal{T}_X^{-1}h)$.

Example 3.19 (Example 3.1, continued)**.** *Let* X *have support* $\mathcal{I} = \{a, b\}$ *with Stein class* $\mathcal{F}(X)$ *and Stein operator* $\mathcal{T}_X(f) = (fp)'/p$. *Then*

$$\mathcal{T}_X^{-1}h(x) = \frac{1}{p(x)}\int_a^x h(u)p(u)du = -\frac{1}{p(x)}\int_x^b h(u)p(u)du$$

for all $h \in \mathcal{F}^{(0)}(X)$ *the collection of functions* $h \in L^1_\mu(p)$ *such that* $\mathbb{E}_p h = 0$.

Example 3.20 (Example 3.2, continued)**.** *Let* X *have support* $\mathcal{I} = [a, b]$ *with Stein class* $\mathcal{F}(X)$ *and Stein operator* $\mathcal{T}_X(f) = \Delta^+(fp)/p$. *Then*

$$\mathcal{T}_X^{-1}h(x) = \frac{1}{p(x)}\sum_{k=a}^{x} h(k)p(k) = -\frac{1}{p(x)}\sum_{k=x+1}^{b} h(k)p(k)$$

for all $h \in \mathcal{F}^{(0)}(X)$ *the collection of functions* h *such that* $\mathbb{E}_p h = 0$.

The inverse operator and corresponding sets in Example 3.4 (resp., Example 3.5) are simply obtained by replacing p with φp (resp., with $\gamma_\lambda p$) in Example 3.19 (resp., in Example 3.20).

3.4. Stein differentiation and the product rule

Define the new class of functions

$$dom(\mathcal{D}, X) := \bigcap_{f \in \mathcal{F}(X)} dom(\mathcal{D}, X, f)$$

with $dom(\mathcal{D}, X, f)$ as in (32). Then the following holds.

Lemma 3.21. *If the constant function* 1 *belongs to* $dom(\mathcal{D}) \cap dom(\mathcal{D}^\star)$, *then all constant functions are in* $ker(\mathcal{D}^\star)$ *and in* $dom(\mathcal{D}, X)$.

Proof. Taking $g \equiv 1$ in (29) we see that $\mathcal{D}f(x) = \mathcal{D}f(x) + f(x)\mathcal{D}^\star 1(x)$ for all $f \in dom(\mathcal{D})$. Taking $f \equiv 1$ ensures the first claim. The second claim then follows immediately. □

From here onwards we make the following assumption.

Assumption 3. $1 \in dom(\mathcal{D}) \cap dom(\mathcal{D}^\star)$.

Starting from the product rule (29) we also obtain the following differentiation rules for $\mathcal{D}$ and $\mathcal{D}^\star$.

Lemma 3.22. *Under Assumptions 1 and 3 we have*

1. $\mathcal{D}g(\cdot + l) = g\mathcal{D}1 + \mathcal{D}^\star g$
2. $\mathcal{D}^\star(fg) = g\mathcal{D}^\star f + f(\cdot + l)\mathcal{D}^\star g$

for all $f, g \in dom(\mathcal{D}) \cap dom(\mathcal{D}^\star)$.

Proof. Claim 1. is immediate. To see 2., using Assumption 1 we write

$$\begin{aligned}\mathcal{D}^\star(fg) &= -fg\mathcal{D}1 + \mathcal{D}(f(\cdot + l)g(\cdot + l)) \\ &= -fg\mathcal{D}1 + g\mathcal{D}f(\cdot + l) + f(\cdot + l)\mathcal{D}^\star g.\end{aligned}$$

Applying Claim 1. to the second summand we then get

$$\begin{aligned}\mathcal{D}^\star(fg) &= -fg\mathcal{D}1 + fg\mathcal{D}1 + g\mathcal{D}^\star f + f(\cdot + l)\mathcal{D}^\star g \\ &= g\mathcal{D}^\star f + f(\cdot + l)\mathcal{D}^\star g.\end{aligned}$$

□

Remark 3.23. *From Point 1. in Lemma 3.22 we see that if* $l = 0$ *and* $1 \in ker(\mathcal{D})$ *then* $\mathcal{D} = \mathcal{D}^\star$ *on* $dom(\mathcal{D}) \cap dom(\mathcal{D}^*)$. *Neither of these assumptions are always satisfied (see Examples 3.2 and 3.4).*

The following result is the basis of what we call "Stein differentiation". It is also the key to the standardizations leading to the different Stein operators that will be discussed in Section 4.

Theorem 3.24 (Stein product rule)**.** *The Stein triple* $(\mathcal{T}_X, \mathcal{F}(X), dom(\mathcal{D}, X, \cdot))$ *satisfies the product rule*

$$f(x)\mathcal{D}^\star(g)(x) + g(x)\mathcal{T}_X f(x) = \mathcal{T}_X(f(\cdot)g(\cdot + l))(x) \tag{36}$$

for $f \in \mathcal{F}(X)$ *and* $g \in dom(\mathcal{D}, X, f)$.

Proof. Use Assumption 1 to deduce

$$\begin{aligned}f(x)\mathcal{D}^\star(g)(x) + g(x)\mathcal{T}_X f(x) &= f(x)\mathcal{D}^\star(g)(x) + g(x)\frac{\mathcal{D}(fp)(x)}{p(x)} \\ &= \frac{1}{p(x)}\mathcal{D}(f(\cdot)p(\cdot)g(\cdot + l))(x),\end{aligned}$$

which is the claim. □

To see how (36) can be put to use, let $h \in L^1_\mu(X)$ and consider the equation

$$h(x) - \mathbb{E}[h(X)] = f(x)\mathcal{D}^\star(g)(x) + g(x)\mathcal{T}_X f(x), \quad x \in \mathcal{I}. \tag{37}$$

As discussed in the Introduction, Equation (37) is indeed a *Stein equation for the target* X in the sense of (2), although the solutions of (37) are now pairs of functions (f, g) with $f \in \mathcal{F}(X)$ and $g \in dom(\mathcal{D}, X, f)$ which satisfy the relationship

$$f(\cdot)g(\cdot + l) = \mathcal{T}_X^{-1}(h - \mathbb{E}_p h). \tag{38}$$

We stress that although fg is uniquely defined by (38), the individual f and g are not (just consider multiplication by constants).

Equation (37) and its solutions (38) are not equivalent to Equation (2) and its solutions already available from the literature, but rather contain them, as illustrated in the following example.

Example 3.25 (Example 3.1, continued)**.** *Taking $g = 1$ (this is always permitted by Lemma 3.21) and p differentiable we get the equation*

$$h(x) - \mathbb{E}[h(X)] = f'(x) + \frac{p'(x)}{p(x)} f(x), \quad x \in \mathcal{I}, \tag{39}$$

whose solution is some function $f \in \mathcal{F}(X)$, as in e.g. [65]. If the constant function $f \equiv 1$ is in $\mathcal{F}(X)$ then keeping instead g variable but taking $f \equiv 1$ yields the equation

$$h(x) - \mathbb{E}[h(X)] = g'(x) + \frac{p'(x)}{p(x)} g(x), \quad x \in \mathcal{I}, \tag{40}$$

whose solution is any function in $dom(\mathcal{D}, X, 1)$ the collection of functions $g \in \mathcal{F}(X)$ such that $gp'/p \in L^1_\mu(X)$, a family of equations considered e.g. in [92]. Similar considerations hold in the settings of Examples 3.2, 3.4 and 3.5. We stress the fact that the difference between (39) and (40) lies in the space of solutions.

3.5. Stein characterizations

Pursuing the tradition in the literature on Stein's method, we provide a general family of *Stein characterizations for X*. In addition to Assumptions 1, 2 and 3 we will further need the following two assumptions to hold.

Assumption 4. *$f \in ker(\mathcal{D}^\star)$ if $f \equiv \alpha$ for some $\alpha \in \mathbb{R}$.*

Assumption 5. *$\mathcal{D}f/f = \mathcal{D}g/g$ for $f, g \in dom(\mathcal{D})$ if and only if $f/g \equiv \alpha$ for some $\alpha \in \mathbb{R}$.*

Both assumptions are simultaneously satisfied in all examples discussed in Section 3.

Theorem 3.26. *Let Y be a random element with the same support as X and assume that the law of Y is absolutely continuous w.r.t. μ with Radon-Nikodym derivative q.*

1. *Suppose that $\mathcal{F}(X)$ is dense in $L^1_\mu(p)$ and that $\frac{q}{p} \in dom(\mathcal{D}^*)$. Take $g \in dom(\mathcal{D}, X)$ which is X-a.s. never 0 and assume that $g\frac{q}{p} \in dom(\mathcal{D}, X)$. Then*

$$Y \overset{\mathcal{D}}{=} X \text{ if and only if } \mathbb{E}\left[f(Y)\mathcal{D}^\star(g)(Y)\right] = -\mathbb{E}\left[g(Y)\mathcal{T}_X f(Y)\right] \tag{41}$$

   for all $f \in \mathcal{F}(X)$.
2. *Let $f \in \mathcal{F}(X)$ be X-a.s. never zero and assume that $dom(\mathcal{D}, X, f)$ is dense in $L^1_\mu(p)$. Then*

$$Y \overset{\mathcal{D}}{=} X \text{ if and only if } \mathbb{E}\left[f(Y)\mathcal{D}^\star(g)(Y)\right] = -\mathbb{E}\left[g(Y)\mathcal{T}_X f(Y)\right] \tag{42}$$

for all $g \in dom(\mathcal{D}, X, f)$.

Remark 3.27. *The assumptions leading to (41) and (42) can be relaxed by removing the assumption that* Y *and* X *share a support* $\mathcal{I}$ *but instead conditioning on the event that* $Y \in \mathcal{I}$ *and writing* $Y \,|\, Y \in \mathcal{I} \overset{\mathcal{D}}{=} X$ *to indicate that* $p = cq$ *on* $\mathcal{I}$*, for a constant* $c = P(Y \in \mathcal{I})$*, see [65].*

Proof. The sufficient conditions are immediate. Indeed, from (33), if Y has the same distribution as X then (41) and (42) hold true.

We now prove the necessity. We start with statement 1. Let g be such that $gq/p \in dom(\mathcal{D}, X)$. Then, $gq/p \in dom(\mathcal{D}^\star)$ and, for all $f \in \mathcal{F}(X)$, we have $\mathcal{D}^\star(gq/p)fp \in L^1(\mu)$ as well as

$$\int \mathcal{D}\left(g(\cdot + l)\frac{q(\cdot + l)}{p(\cdot + l)}f(\cdot)p(\cdot)\right)d\mu = 0$$

and we can apply (30) to get

$$\mathbb{E}\left[g(Y)\mathcal{T}_X f(Y)\right] = \int g\frac{q}{p}\mathcal{D}(fp)d\mu = -\int fp\mathcal{D}^\star\left(g\frac{q}{p}\right)d\mu.$$

Supposing (41) gives

$$\int \mathcal{D}^\star\left(g\frac{q}{p}\right)fpd\mu = \int f\mathcal{D}^\star\left(g\right)qd\mu = \int f\mathcal{D}^\star\left(g\right)\frac{q}{p}pd\mu$$

for all $f \in \mathcal{F}(X)$. On the one hand, as $\mathcal{F}(X)$ is assumed to be dense in $L^1_\mu(p)$, it follows that $\mathcal{D}^\star\left(g\frac{q}{p}\right) = \frac{q}{p}\mathcal{D}^\star\left(g\right)$ $p-$a.e. and, on the other hand, by Claim 2. in Lemma 3.22 we know that $\mathcal{D}^\star\left(g\frac{q}{p}\right) = \frac{q}{p}\mathcal{D}^\star g + g(\cdot + l)\mathcal{D}^\star\left(\frac{q}{p}\right)$. Equating these two expressions gives that $g(\cdot + l)\mathcal{D}^\star\left(\frac{q}{p}\right) = 0$ $p-$a.e. and, as g is p-a.e. never 0 we obtain that

$$\mathcal{D}^\star\left(\frac{q}{p}\right) = 0 \qquad p-\text{a.e.}.$$

Assumption 4 now gives that there is a constant c such that $p = cq$ except on a set of p-measure 0. As both p and q integrate to 1, it must be the case that $c = 1$, and so $p = q$ on $supp(p)$, which gives the first assertion.

We tackle statement 2. If $g\frac{q(\cdot - l)}{p(\cdot - l)} \in dom(\mathcal{D}, X, f)$ then

$$\int \mathcal{D}(f(\cdot)\frac{q(\cdot)}{p(\cdot)}g(\cdot + l))d\mu = \int \mathcal{D}(f(\cdot)g(\cdot + l)q(\cdot))d\mu = 0$$

so that

$$\mathbb{E}\left[f(Y)\mathcal{D}^\star\left(g\right)\left(Y\right)\right] = -\int g\mathcal{D}(fq)d\mu = -\int \frac{\mathcal{D}(fq)}{p}gpd\mu.$$

Supposing (42) gives

$$\int \frac{\mathcal{D}(fq)}{p} g p d\mu = \int \frac{\mathcal{D}(fp)}{p} g q d\mu = \int \frac{\mathcal{D}(fp)}{p} g \frac{q}{p} p d\mu$$

for all $g \in dom(\mathcal{D}, X, f)$. As $dom(\mathcal{D}, X, f)$ is assumed to be dense in $L^1(\mu)$ it follows that $\mathcal{D}(fq) = \mathcal{D}(fp)\frac{q}{p}$. On the one hand $\mathcal{D}(fp)\frac{q}{p} = f(\cdot - l)\frac{q}{p}\mathcal{D}p + q\mathcal{D}^\star f(\cdot - l)$ and, on the other hand, $\mathcal{D}(fq) = f(\cdot - l)\mathcal{D}(q) + q\mathcal{D}^\star f(\cdot - l)$. Simplifying and using the fact that f is never 0 we deduce the equivalent score-like condition

$$\frac{\mathcal{D}(q)}{q} = \frac{\mathcal{D}(p)}{p} \quad p - a.e.$$

Assumption 5 gives the conclusion. □

Theorem 3.26 generalizes the literature on this topic in a subtle, yet fundamental, fashion. To see this first take $g \equiv 1$ in (41) (recall that this is always permitted) to obtain the Stein characterization

$$Y \overset{\mathcal{D}}{=} X \text{ if and only if } \mathbb{E}\left[\mathcal{T}_X f(Y)\right] = 0 \text{ for all } f \in \mathcal{F}(X)$$

which is valid as soon as the densities of X and Y have same support and $q/p \in dom(\mathcal{D}, X, \cdot)$. This is the characterization given in [65, 64]. If $f \equiv 1$ is in $\mathcal{F}(X)$ then, for this choice of f in (42) we obtain the Stein characterization

$$Y \overset{\mathcal{D}}{=} X \iff \mathbb{E}[g'(Y)] = -\mathbb{E}\left[\frac{p'(Y)}{p(Y)} g(Y)\right] = 0 \text{ for all } g \in dom(\mathcal{D}, X, 1).$$

Here we assume that p and q share the same support. The condition $g \in dom(\mathcal{D}, X, 1)$ is equivalent to $g(\cdot + l) \in \mathcal{F}(X)$ and $\mathbb{E}\left|g(X)\mathcal{T}_X 1(X)\right| < \infty$. This is the general characterization investigated in [92].

Remark 3.28. *The hypothesis that the constant function 1 belongs to $\mathcal{F}(X)$ is not a small assumption. Indeed, we easily see that*

$$1 \in \mathcal{F}(X) \iff p'/p \in L^1_\mu(X) \text{ and } \int_{\mathcal{I}} p'(x)dx = 0.$$

This condition is not satisfied e.g. by the exponential distribution $p(x) = e^{-x}\mathbb{I}(x \geq 0)$ (because the integral of the derivative is not 0) nor by the arcsine distribution $p(x) = 1/\sqrt{x(1-x)}\mathbb{I}(0 < x < 1)$ (because the derivative is not integrable).

Remark 3.29. *Our approach is reminiscent of Stein characterizations of birth-death processes where one can choose the death rate and adjust the birth rate accordingly, see [52] and [34].*

3.6. Connection with biasing

In [41] the notion of a zero-bias random variable was introduced. Let X be a mean zero random variable with finite, nonzero variance σ^2. We say that X^*

has the X-zero biased distribution if for all differentiable f for which $\mathbb{E}[Xf(X)]$ exists,

$$\mathbb{E}[Xf(X)] = \sigma^2\mathbb{E}[f'(X^*)].$$

Furthermore the mean zero normal distribution with variance σ^2 is the unique fixed point of the zero-bias transformation.

More generally, if X is a random variable with density $p_X \in dom(\mathcal{D}^*)$ then for all $f \in dom(\mathcal{D})$, by (29) we have

$$p_X(x)\mathcal{T}_X(f)(x) = \mathcal{D}(f(x)p_X(x)) = p_X(x-l)\mathcal{D}f(x) + f(x)\mathcal{D}^\star p_X(x-l)$$

and so

$$\mathbb{E}\left[\frac{p_X(X-l)}{p_X(X)}\mathcal{D}f(X)\right] + \mathbb{E}\left[f(X)\frac{\mathcal{D}^\star p_X(X-l)}{p_X(X)}\right] = 0.$$

This equation leads to the definition of a transformation which maps a random variable Y to $Y^{(X)}$ such that, for all $f \in dom(\mathcal{D})$ for which the expressions exist,

$$\mathbb{E}\left[\frac{p_X(Y^{(X)}-l)}{p_X(Y^{(X)})}\mathcal{D}f(Y^{(X)})\right] = -\mathbb{E}\left[f(Y)\frac{\mathcal{D}^\star p_X(Y-l)}{p_X(Y)}\right].$$

For some conditions which give the existence of such Y^* see [42]. As an illustration, in the setting of Example 3.1, if the density p is log-concave (so that $-p'/p$ is increasing) then the existence of the coupling $Y^{(X)}$ is straightforward via the Riesz representation theorem, as in [41].

Finally assume that $\mathcal{F}(X) \cap dom(\mathcal{D})$ is dense in $L^1_\mu(X)$. To see that $Y =_d X$ if and only if $Y^{(X)} =_d Y$, first note that by construction if $Y =_d X$ then $Y^{(X)} =_d Y$. On the other hand, if $Y^{(X)} =_d Y$, then $\mathbb{E}[\mathcal{T}_X(f)(Y)] = 0$ for all $f \in \mathcal{F}(X) \cap dom(\mathcal{D})$, and the assertion follows from the density assumption and (41). Hence (41) can be used to establish distributional characterizations based on biasing equations.

4. Stein operators

Let X be a random variable with support $\mathcal{X}$, let $\mathcal{D}$ be a linear operator acting on $\mathcal{X}^\star$ and satisfying Assumptions 1 and 2. There are now two seemingly antagonistic points of view:

- In the Introduction we mention the fact that Stein's method for X relies on a pair $(\mathcal{A}_X, \mathcal{F}(\mathcal{A}_X))$ with $\mathcal{A}_X$ a differential operator acting on $\mathcal{F}(\mathcal{A}_X)$ a class of functions. For any given X, the literature on Stein's method contains *many* different such (not necessarily first order!) operators and classes.
- In Section 3, we claim to obtain "the" canonical operator associated to X, denoted $\mathcal{T}_X$, acting on "the" canonical class $\mathcal{F}(X)$ (uniqueness up to the choice of $\mathcal{D}$) with unique inverse $\mathcal{T}_X^{-1}$.

In this section we merge these two points of view. Our general point of view is that a Stein operator for a random variable X is any operator that can be

written in the form

$$\mathcal{A}_X : \mathcal{F}(X) \times dom(\mathcal{D}, X, \cdot) \to \mathcal{X}^\star : (f, g) \mapsto \mathcal{T}_X(fg), \tag{43}$$

and, given $h \in L^1_\mu(X)$, the corresponding Stein equation is

$$h - \mathbb{E}[h(X)] = \mathcal{A}_X(f, g)$$

whose solutions are any functions $f \in \mathcal{F}(X)$ and $g \in dom(\mathcal{D}, X, f)$ such that $fg = \mathcal{T}_X^{-1}(h - \mathbb{E}[h(X)])$. There are many ways to particularise (43), such as

1. fix $f \in \mathcal{F}(X)$ and let g vary in $dom(\mathcal{D}, X, f)$,
2. fix $g \in dom(\mathcal{D}, X)$ and let f vary in $\mathcal{F}(X)$,
3. let f and g vary simultaneously.

We refer to these mechanisms as *standardizations*.

For the first approach pick a function $f \in \mathcal{F}(X)$ and define the operator

$$\mathcal{A}_X g = \mathcal{T}_X\left(f(\cdot)g(\cdot + l)\right) = f\mathcal{D}^\star(g) + g\mathcal{T}_X f \tag{44}$$

acting on functions $g \in \mathcal{F}(\mathcal{A}_X) = dom(\mathcal{D}, X, f)$. The corresponding Stein equation is

$$h - \mathbb{E}[h(X)] = \mathcal{A}_X g$$

whose solutions are $g \in dom(\mathcal{D}, X, f)$ given by $g = \mathcal{T}_X^{-1}(h - \mathbb{E}[h(X)])/f$.

The second option is to fix a function $g \in dom(\mathcal{D}, X)$ and define the operator

$$\mathcal{A}_X f = \mathcal{T}_X\left(f(\cdot)g(\cdot + l)\right) = f\mathcal{D}^\star(g) + g\mathcal{T}_X f \tag{45}$$

acting on functions $f \in \mathcal{F}(X)$. In this case solutions of the Stein equation are $f \in \mathcal{F}(X)$ given by $f = \mathcal{T}_X^{-1}(h - \mathbb{E}[h(X)])/g$.

The third option is to consider operators of the form

$$\mathcal{A}_X(f, g) = \mathcal{T}_X\left(f(\cdot)g(\cdot + l)\right) = f\mathcal{D}^\star(g) + g\mathcal{T}_X f \tag{46}$$

acting on functions $(f, g) \in \mathcal{G}_1 \times \mathcal{G}_2$ where $\mathcal{G}_1, \mathcal{G}_2 \subseteq \mathcal{X}^\star$ are such that $f(\cdot)g(\cdot + l) \in \mathcal{F}(X)$. For example we could consider $\mathcal{G}_i$ polynomial functions or exponentials and pick $\mathcal{G}_j$ with $j \neq i$ so as to satisfy the assumptions. Solutions of the Stein equation are pairs of functions such that $f(\cdot)g(\cdot + l) = \mathcal{T}_X^{-1}(h - \mathbb{E}[h(X)])$.

Remark 4.1. *The use of the notation c in (44) relates to the notation in [43], where the idea of using a c-function to generate a family of Stein operators (44) was first proposed (in a less general setting).*

Remark 4.2. *Although appearances might suggest otherwise, operators (44) and (45) are not necessarily first order differential/difference operators. One readily obtains higher order operators by considering, for example, classes $\mathcal{F}_A(X)$ of functions of the form $f = \mathcal{D}^k\tilde{f}$ for $\tilde{f}$ appropriately chosen; see Section 4.6.*

The difference between (44), (45) and (46) is subtle (the first two being particular cases of the third). The guiding principle is to find a form of Stein equation for which the solutions are smooth. The remainder of the Section is dedicated to illustrating standardizations under several general assumptions on the target density, hereby providing interesting and important families of Stein operators.

4.1. Stein operators via score functions

Suppose that X is such that the constant function $1 \in \mathcal{F}(X)$ and define

$$u(x) = \mathcal{T}_X 1(x) = \frac{\mathcal{D}p(x)}{p(x)} \tag{47}$$

the so-called score function of X. Then taking $f = 1$ in (44) we introduce the operator

$$\mathcal{A}_X g(x) = \mathcal{D}^\star g(x-l) + u(x)g(x-l) \tag{48}$$

acting on $\mathcal{F}(\mathcal{A}_X) = dom(\mathcal{D}, X, 1)$. The corresponding Stein equation is

$$\bar{h}(u) = \mathcal{D}^\star g(x-l) + g(x-l)u(x)$$

for $\bar{h}$ any function with X-mean zero; solutions of this equation are the functions

$$g_h = \mathcal{T}_X^{-1}\left(\bar{h}\right)$$

and bounds on these functions (as well as on their derivatives) are crucial to the applicability of Part B of Stein's method through operator (48).

In the continuous setting of Example 3.1 we recover operator (7). In this case $\mathcal{F}(\mathcal{A}_X)$ is the set of all differentiable functions g such that

$$\mathbb{E}\,|g'(X)| < \infty \text{ and } \mathbb{E}\,|g(X)u(X)| < \infty.$$

These are the conditions (27) and (28) from [92, Proposition 4].

Remark 4.3. *The terminology "score function" for the function $\mathcal{D}p(x)/p(x)$ is standard (at least in the continuous case); it is inherited from the statistical literature.*

4.2. Stein operators via the Stein kernel

Suppose that X has finite mean ν and define

$$\tau(x) = \mathcal{T}_X^{-1}(\nu - Id) \tag{49}$$

a function which we call the *Stein kernel* of X (see Equation (10) in Section 2.2, forthcoming Remark 4.6 as well as Sections 5.2 and 5.3). Next take $f = \tau$ in (44) (this is always permitted) and introduce the operator

$$\mathcal{A}_X g(x) = \tau(x)\mathcal{D}^\star g(x-l) + (\nu - x)g(x-l) \tag{50}$$

acting on $\mathcal{F}(\mathcal{A}_X) = dom(\mathcal{D}, X, \tau)$. The corresponding Stein equation is

$$\bar{h}(x) = \tau(x)\mathcal{D}^\star g(x-l) + (\nu - x)g(x-l)$$

for $\bar{h}$ any function with X-mean 0; solutions of this equation are the functions

$$g_h = \frac{1}{\tau}\mathcal{T}_X^{-1}(\bar{h})$$

and bounds on these functions (as well as on their derivatives) are crucial to the applicability of Part B of Stein's method via operator (50).

In the continuous setting of Example 3.1, $\mathcal{F}(\mathcal{A}_X)$ is the set of all differentiable functions such that

$$\mathbb{E}\,|g(X)(X-\nu)| < \infty \text{ and } \mathbb{E}\,|g'(X)\tau(X)| < \infty.$$

These integrability conditions are the same as in [77, Lemma 2.1]; see also [14].

The Stein kernel (49) has a number of remarkable properties. In particular, it plays a pivotal role in the connection between information inequalities and Stein's method, see [60, 74, 73].

Proposition 4.4. *Let Assumptions 1-5 hold. Suppose furthermore that there exists $\delta > 0$ such that $\mathcal{D}^\star(a\,Id + b) = a\,\delta$ for all $a, b \in \mathbb{R}$ and $Id(x) = x$ the identity. Then*

$$\mathbb{E}\,[\tau(X)\mathcal{D}^\star g(X-l)] = \mathbb{E}\,[(X-\nu)g(X)]] \tag{51}$$

for all $g \in dom(\mathcal{D}, X, \tau)$ and

$$\mathbb{E}\,[\tau(X)] = \delta^{-1}\,Var(X). \tag{52}$$

Proof. Identity (51) is obvious and (52) follows by taking $g(x-l) = x-\nu$ (which is allowed) in (51). □

Remark 4.5. *It is easy to show that, moreover, $\tau(x) \geq 0$ if $\mathcal{D}$ is either the strong derivative or the discrete forward/backward difference.*

Remark 4.6. *Although the function $\tau = \mathcal{T}_X^{-1}(\nu - Id)$ has already been much used in the literature, it has been given various names all marked with some ambiguity. Indeed [70, 71, 73] (among others) refer to τ as the "Stein factor" despite the fact that this term also refers to the bounds on the solutions of the Stein equations, see [86, 25, 7]. Other authors, including [16, 15, 13], rather refer to this function as the "ω-function" or the "covariance kernel" of X. We prefer to unify the terminology by calling τ a* Stein kernel.

Two particular instances of (50) have already been perused in the literature in the following case.

Definition 4.7 (Pearson's class of distributions)**.** *A continuous distribution p with support $supp(p)$ is a member of Pearson's family of distributions if it is solution to the differential equation*

$$\frac{p'(x)}{p(x)} = \frac{\alpha - x}{\beta_2(x-\lambda)^2 + \beta_1(x-\lambda) + \beta_0} \tag{53}$$

for some constants λ, α, β_j, $j = 0, 1, 2$.

Properties of the differential operator $\mathcal{T}_X f = (fp)'/p$ have been studied in quite some detail for distributions p which belong to Pearson's class of distributions, see e.g. [26, 57, 55, 78, 61, 67, 2]. If $X \sim p$, a Pearson distribution, then by definition its derivative p' exists and, using (34), its canonical Stein operator is

$$\mathcal{T}_X f(x) = f'(x) + \frac{\alpha - x}{\beta_2(x-\lambda)^2 + \beta_1(x-\lambda) + \beta_0} f(x)$$

for $x \in \text{supp}(p)$. In general this operator is not easy to handle.

It is shown in [57] that, in the setting of Example 3.1, a density p satisfies (53) if and only if its Stein kernel $\tau(x)$ is quadratic. This function can be calculated (using e.g. [26, Equation (3.5)]), and is given by

$$\tau(x) = \frac{\beta_0 + \beta_1 x + \beta_2 x^2}{1 - 2\beta_2},$$

see also [78]. This observation leads us to considering distributions, discrete or continuous, which have a Stein kernel of the form

$$\mathcal{T}_X^{-1}(\nu - Id)(x) = a + bx + cx^2 \tag{54}$$

for some constants a, b and c. For distributions satisfying (54) we deduce a natural family of Stein operators

$$\mathcal{A}_X g(x) = \left(a + bx + cx^2\right) \mathcal{D}^\star g(x) + (\nu - x) g(x)$$

acting on the class $\mathcal{F}(\mathcal{A}_X)$ of functions such that $g\tau \in \mathcal{F}(X)$ as well as

$$\mathbb{E}\left|g(X)(\nu - X)\right| < \infty \text{ and } \mathbb{E}\left|\mathcal{D}^\star g(X)\left(a + bX + cX^2\right)\right| < \infty.$$

Remark 4.8. *[88, 2] call the class of densities satisfying (54) the Pearson class (in the continuous case) and the Ord class (in the discrete case); Ord's class as originally defined in [76] is, in fact, larger. In the case of integer valued random variables, [57, Theorem 4.6] shows that, under conditions on the coefficients, condition (54) is equivalent to requiring that* $p(x) = \binom{a}{x}\binom{b}{n-x}/\binom{a+b}{n}$ *for some constants* a, b *and* n*, so that* X *has a generalized hypergeometric distribution. See also [1] where distributions satisfying (54) are referred to as Cumulative Ord distributions; see in particular their Proposition 2.1 for a characterization.*

Example 4.9. *Many "useful" densities satisfy (54) in which case the operator (50) has a nice form as well. The following examples are easy to compute and will be useful in the sequel; for future reference we also provide the log-derivative of the density.*

1. *Continuous setting, strong derivative:*
   (a) *Gaussian* $\mathcal{N}(0, \sigma^2)$ *with* $p(x) = (2\pi)^{-1} e^{-x^2/2}$ *on* $\mathcal{I} = \mathbb{R}$*:*

$$\frac{p'(x)}{p(x)} = -\frac{x}{\sigma^2} \text{ and } \tau(x) = \sigma^2;$$

   (b) *Gamma* $\Gamma(\alpha, \beta)$ *with* $p(x) = \beta^{-\alpha}\Gamma(\alpha)^{-1} e^{-x/\beta} x^{\alpha-1}$ *on* $\mathcal{I} = \mathbb{R}^+$*:*

$$\frac{p'(x)}{p(x)} = \frac{-1 + \alpha}{x} - \frac{1}{\beta} \text{ and } \tau(x) = \frac{x}{\beta};$$

   (c) *Beta* $\mathcal{B}(\alpha, \beta)$ *with* $p(x) = B(\alpha, \beta)^{-1} x^{\alpha-1}(1 - x)^{\beta-1}$ *on* $\mathcal{I} = [0, 1]$*:*

$$\frac{p'(x)}{p(x)} = \frac{\alpha - 1}{x} - \frac{\beta - 1}{x - 1} \text{ and } \tau(x) = \frac{x(1 - x)}{\alpha + \beta};$$

(d) *Student* t_ν *(for* $\nu > 1$*) with* $p(x) = \nu^{-1/2}B(\nu/2, 1/2)^{-1}(\nu/(\nu + x^2))^{(1+\nu)/2}$ *on* $\mathbb{R}$*:*

$$\frac{p'(x)}{p(x)} = -\frac{x(1+\nu)}{\nu + x^2} \text{ and } \tau(x) = \frac{x^2+\nu}{\nu - 1}.$$

2. *Discrete setting, forward derivative:*

(a) *Poisson* $Po(\lambda)$ *with* $p(x) = e^{-\lambda}\lambda^x/x!$ *on* $\mathcal{I} = \mathbb{Z}$*:*

$$\frac{\Delta^+ p(x)}{p(x)} = \frac{\lambda}{x+1} - 1 \text{ and } \tau(x) = x;$$

(b) *Binomial* $Bin(n, p)$ *with* $p(x) = \binom{n}{x}p^x(1-p)^{n-x}$ *on* $\mathcal{I} = [0, n] \cap \mathbb{Z}$*:*

$$\frac{\Delta^+ p(x)}{p(x)} = \frac{(n-x)}{x+1}\frac{p}{1-p} - 1 \text{ and } \tau(x) = (1-p)x.$$

4.3. Invariant measures of diffusions

Recent papers [30, 58, 59] provide a general framework for performing Stein's method with respect to densities p which are supposed to admit a variance and be continuous (with respect to the Lebesgue measure), bounded with open interval support. Specifically, [58] suggest studying operators of the form

$$\mathcal{A}_X g(x) = \frac{1}{2}\beta(x)g'(x) + \gamma(x)g(x) \tag{55}$$

with $\gamma \in L^1(\mu)$ a continuous function with strictly one sign change on the support of X, negative on the right-most interval and such that γp is bounded and $\mathbb{E}[\gamma(X)] = 0$,

$$\beta(x) = \frac{2}{p(x)}\int_a^x \gamma(y)p(y)dy,$$

for $g \in \mathcal{F}(\mathcal{A}_X)$ the class of functions such that $g \in C^1$ and

$$\mathbb{E}|\gamma(X)g(X)| < +\infty \text{ and } \mathbb{E}|\beta(X)g'(X)| < +\infty.$$

Then [58] (see as well [59] for an extension) use diffusion theory to prove that such $\mathcal{A}_X$ as defined in (55) are indeed Stein operators in the sense of the Introduction (their approach falls within the generator approach). In our framework, (55) is a particular case of (44), with $f = \beta/2 = \mathcal{T}_X^{-1}\gamma \in \mathcal{F}(X)$ and $\gamma = \mathcal{T}_X f$ (which necessarily satisfies $\mathbb{E}[\gamma(X)] = 0$) and $\mathcal{F}(\mathcal{A}_X) = dom((\cdot)', X, f)$.

4.4. Gibbs measures on non-negative integers

We can treat any discrete univariate distribution on non-negative integers by writing it as a Gibbs measure

$$\mu(x) = \frac{1}{\kappa}\exp(V(x))\frac{\omega^x}{x!}, \quad x = 0, 1, \ldots, N,$$

where $N \in \{0, 1, 2, \ldots\} \cup \{\infty\}$ and κ is a normalizing constant. Here the choice of V and ω is not unique. In [34], Stein's method for discrete univariate Gibbs measures on non-negative integers is developed, with operator

$$\mathcal{A}_\mu(f)(x) = f(x+1)\omega \exp\left(V(x+1) - V(x)\right) - xf(x) \tag{56}$$

acting on the class of functions such that $f(0) = 0$ and, if N is infinite, $\lim_{x\to\infty} f(x)\exp(V(x))\frac{\omega^x}{x!} = 0$. The canonical operator (31) is (with $\mathcal{D} = \Delta^+$)

$$\mathcal{T}_\mu f(x) = f(x+1)\frac{\omega}{x+1}\exp\left(V(x+1) - V(x)\right) - f(x)$$

which yields (56) via (46) using the pair $(f(x), g(x)) = (f(x), x+1)$. In [34], other choices of birth and death rates were discussed; here the birth rate b_x is the pre-factor of $g(x+1)$, and the death rate d_x is the pre-factor of $g(x)$. Indeed any choice of birth and death rates which satisfy the detailed balance conditions

$$\mu(x)b_x = \mu(x+1)d_{x+1}$$

for all x are viable. Our canonical Stein operator can be written as

$$\mathcal{T}_\mu g(x) = \frac{b_x}{d_{x+1}}g(x+1) - g(x).$$

Choosing $f(x) = d_x$ and applying (44) gives the general Stein operator $b_x g(x+1) - d_x g(x)$. The Stein kernel here is

$$\tau(x) = \sum_{y=0}^{x} e^{V(y)-V(x)}\frac{x!}{y!w^{x-y}}(\nu - y)$$

with ν the mean of the distribution. This expression can be simplified in special cases; for example in the Poisson case V is constant and we obtain $\tau(x) = w$, as before. Similar developments are also considered by [52].

4.5. Higher order operators

So far, in all examples provided we only consider first-order difference or differential operators. One way of constructing higher order operators is to consider

$$\mathcal{A}_X f = \mathcal{T}_X(c\mathcal{D}^k f)$$

for c well chosen and $\mathcal{D}^k$ the kth iteration of $\mathcal{D}$. This approach is strongly connected with Sturm-Liouville theory and will be the subject of a future publication. Here we merely give examples illustrating that our results are not restricted to first-order operators. The first example is the Kummer-U distribution in Example 2.5.

Similar considerations as in Example 2.5 provide tractable operators for other distributions involving special functions.

Example 4.10 (Variance Gamma distribution)**.** *Let K_ν be the modified Bessel function of the second kind, of index ν. A random variable has the Variance Gamma distribution $VG(\nu, \alpha, \beta, \eta)$ if its density is given on $\mathbb{R}$ by*

$$p(x) = \frac{(\alpha^2 - \beta^2)^{\nu+\frac{1}{2}}}{\sqrt{\pi}\Gamma\left(\nu + \frac{1}{2}\right)} \left(\frac{|x-\eta|}{2\alpha}\right)^\nu e^{\beta x} K_\nu(\alpha|x-\eta|),$$

where $\alpha > |\beta| > 0, \nu > -\frac{1}{2}, \eta \in \mathbb{R}$. For simplicity we take $\eta = 0, \alpha = 1$, and $\nu > 0$. A generator for this distribution is

$$\mathcal{A}f(x) = xf''(x) + (2\nu + 1 + 2\beta x)f'(x) + \{(2\nu+1)\beta - (1-\beta^2)x\}f(x), \quad (57)$$

see [38]. The canonical operator is (with $\mathcal{D}$ the usual strong derivative)

$$\mathcal{T}(f)(x) = f'(x) + f(x)\left(\frac{2\nu}{x} + \beta\right) - \frac{K_{\nu+1}(x)}{K_\nu(x)}.$$

Applying (46) via the pair $(f, g) = (f, g(f))$ with

$$g(f)(x) = x\frac{f'(x)}{f(x)} + x\left(\beta + \frac{K_{\nu+1}(x)}{K_\nu(x)}\right)$$

we retrieve (57).

4.6. Densities satisfying a differential equation

Lastly we consider the case where the density of interest p with interval support $I = \{a, b\}$ is defined as the solution of some differential equation, say

$$\mathcal{L}(p) = 0$$

along with some boundary conditions. Suppose that $\mathcal{L}$ admits an adjoint (w.r.t. Lebesgue integration) which we denote $\mathcal{L}^\star$ so that, for $X \sim p$, we can apply integration by parts to get

$$\begin{aligned} 0 &= \int_a^b g(x)\mathcal{L}(p)(x)dx = C_a^b(g,p) + \int_a^b \mathcal{L}^\star(g)(x)p(x)dx \\ &= C_a^b(g,p) + \mathbb{E}\left[\mathcal{L}^\star(g)(X)\right] \end{aligned}$$

with $C_a^b(g,p)$ the constant arising through the integration by parts. We define $\mathcal{A}_X(g) = \mathcal{L}^\star(g)$ acting on the class $\mathcal{F}(\mathcal{A}_X)$ of sufficiently smooth functions such that $C_a^b(g,p) = 0$. To qualify $\mathcal{A}_X$ as a Stein operator in the sense of (1), it still remains to identify conditions on g which ensure that this operator characterises the density.

This point of view blends smoothly into our canonical approach to Stein operators; we can moreover provide conditions on g in a generic way. To see this fix a function g of interest and choose f such that

$$\frac{(fp)'}{p} = \mathcal{L}^\star(g)$$

if such an f exists. Then, reversing the integration by parts argument provided above, we get

$$\begin{aligned} f(x) &= \frac{1}{p(x)} \int_a^x \mathcal{L}^\star(g)(u)p(u)du \\ &= \frac{1}{p(x)} C_a^x(g,p) + \frac{1}{p(x)} \int_a^x g(u)\mathcal{L}(p)(u)du \\ &= \frac{1}{p(x)} C_a^x(g,p) =: F(g,p)(x) \end{aligned}$$

with $\frac{1}{p(x)} C_a^x(g,p)$ the quantities resulting from the integration by parts (and using the fact that now $\mathcal{L}(p) = 0$, by assumption). This leads to the standardization

$$\mathcal{A}_X(g) = \mathcal{T}_X\left(F(g,p)\right)$$

acting on the class of functions $\mathcal{F}(\mathcal{A}_X) = \{g \text{ such that } F(g,p) \in \mathcal{F}(X)\}$. Note how, in particular, the assumption $F(g,p) \in \mathcal{F}(X)$ implies that $C_a^b(g,p) = 0$, as demanded in the beginning of the Section.

Example 4.11. *We illustrate this point of view in the case of the spectral density h_n on $[-2,2]$ of a $GUE(n,1/n)$ random matrix studied in [47, 49]. This density is defined through the third order differential equation*

$$\mathcal{L}(h_n)(x) = \frac{1}{n^2} h_n'''(x) + (4-x^2)h_n'(x) + xh_n(x) = 0, x \in \mathbb{R},$$

along with a boundary condition. Letting $X \sim h_n$ it is straightforward to show that

$$\mathcal{L}^\star(g)(x) = -\frac{1}{n^2} g'''(x) - ((4-x^2)g(x))' + xg(x)$$

acting on the collection

$$\left\{ g \in C^3 \text{ such that } \left[\frac{h_n''(x)g(x) - h_n'(x)g'(x)}{n^2} + h_n(x)g(x)(4-x^2) \right]_{-2}^{2} = 0 \right\}.$$

Integrating by parts we then get

$$F(g,h_n)(x) = \frac{1}{n^2} \left(g''(x) - g'(x)\frac{h_n'(x)}{h_n(x)} + \frac{h_n''(x)}{h_n(x)} g(x) \right) + (4-x^2)g(x) - \mathrm{c}$$

with $\mathrm{c} = g''(-2) + g'(-2)\frac{h_n'(-2)}{h_n(-2)} - \frac{h_n''(-2)}{h_n(-2)} g(-2)$. Considering only functions g such that $F(g,h_n) \in \mathcal{F}(X)$ leads to a Stein operator for X.

5. Distributional comparisons

Resulting from our framework, in this Section we provide a general "comparison of generators approach" (Theorem 5.1) which provides bounds on the probability distance between univariate distributions in terms of their Stein operators. This result is formal and abstract; it is our take on a general version of Part B of Stein's method. Specific applications to concrete distributions will be deferred to Section 6.

5.1. Comparing Stein operators

Let $(\mathcal{X}_1, \mathcal{B}_1, \mu_1)$ and $(\mathcal{X}_2, \mathcal{B}_2, \mu_2)$ be two measure spaces as in Section 3.1. Let X_1 and X_2 be two random variables on $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively, and suppose that their respective densities p_1 and p_2 have interval support. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two linear operators acting on $\mathcal{X}_1$ and $\mathcal{X}_2$ and satisfying Assumption 1 (with l_1 and l_2, respectively) and Assumption 2. Denote by $\mathcal{T}_1$ and $\mathcal{T}_2$ the Stein operators associated with $(X_1, \mathcal{D}_1)$ and $(X_2, \mathcal{D}_2)$, acting on Stein classes $\mathcal{F}_1 = \mathcal{F}(X_1)$ and $\mathcal{F}_2 = \mathcal{F}(X_2)$, respectively. Finally let $\mathbb{E}_i h = \mathbb{E}[h(X_i)]$ denote the expectation of a function h under the measure $p_i d\mu$, $i = 1, 2$.

The framework outlined in Section 3 (specifically Section 3.4) is tailored for the following result to hold.

Theorem 5.1. *Let h be a function such that $\mathbb{E}_i|h| < \infty$ for $i = 1, 2$.*

1. *Let (f, g) with $f \in \mathcal{F}_1$ and $g \in dom(\mathcal{D}_1, X_1, f)$ solve the X_1-Stein equation (37) for h. Then*
$$\mathbb{E}_2 h - \mathbb{E}_1 h = \mathbb{E}_2 \left[f(X_2)\mathcal{D}_1^\star g(X_2) - g(X_2)\mathcal{T}_1 f(X_2) \right]. \tag{58}$$
2. *Fix $f_1 \in \mathcal{F}_1$ and define the function $g_h := \frac{1}{f_1}\mathcal{T}_1^{-1}(h - \mathbb{E}_1 h)$. Then*
$$\begin{aligned}\mathbb{E}_2 h - \mathbb{E}_1 h = \mathbb{E}_2 \big[& f_1(\cdot)\mathcal{D}_1^\star g_h(\cdot) - f_2(\cdot)\mathcal{D}_2^\star g_h(\cdot) \\ & + g_h(\cdot)\mathcal{T}_1 f_1(\cdot) - g_h(\cdot)\mathcal{T}_2 f_2(\cdot) \big]\end{aligned} \tag{59}$$
for all $f_2 \in \mathcal{F}_2$ such that $g_h \in dom(\mathcal{D}_2, X_2, f_2)$.
3. *Fix $g_1 \in dom(\mathcal{D}_1, X_1)$ and define the function $f_h := \frac{1}{g_1}\mathcal{T}_1^{-1}(h - \mathbb{E}_1 h)$. If $f_h \in \mathcal{F}_1 \cap \mathcal{F}_2$ then*
$$\begin{aligned}\mathbb{E}_2 h - \mathbb{E}_1 h = \mathbb{E}_2 \big[& f_h(\cdot)\mathcal{D}_1^\star g_1(\cdot) - f_h(\cdot)\mathcal{D}_2^\star g_2(\cdot) \\ & + g_1(\cdot)\mathcal{T}_1 f_h(\cdot) - g_2(\cdot)\mathcal{T}_2 f_h(\cdot) \big].\end{aligned} \tag{60}$$
for all $g_2 \in dom(\mathcal{D}_2, X_2)$.

Remark 5.2. *Our approach contains the classical "direct" approach described in the Introduction (see (4)). Indeed, if allowed, one can take $f_1 = 1$ and $f_2 = 0$ in (59) to get*
$$\mathbb{E}_2 h - \mathbb{E}_1 h = \mathbb{E}_2 \left[\mathcal{D}_1^\star g_h(\cdot) + u_1(\cdot) g_h(\cdot) \right]$$
with u_1 the score of X_1 (defined in (47)) and g_h now the usual solution of the Stein equation. This yields the bound
$$d_{\mathcal{H}}(X_1, X_2) \leq \sup_{\mathcal{H}} \left| \mathbb{E}_2 \left[\mathcal{A}(g_h)(X_2) \right] \right|$$
with $\mathcal{A}(g_h) = \mathcal{D}_1^\star g_h + u_1 g_h$. In this case one does not need to calculate $\mathcal{T}_2$.

Proof. The starting point is the Stein equation (37) which, in the current context, becomes
$$h(x) - \mathbb{E}[h(X_\bullet)] = f(x)\mathcal{D}_\bullet^\star g(x) + g(x)\mathcal{T}_\bullet f(x) = \frac{\mathcal{D}_\bullet \left(f g p_\bullet \right)}{p_\bullet}(x) \tag{61}$$

with $\bullet \in \{1,2\}$. Solutions of this equation are pairs of functions (f,g) with $f \in \mathcal{F}(X_\bullet)$ and $g \in dom(\mathcal{D}_\bullet, X_\bullet, f)$. Using $\bullet = 1$, replacing x by X_2 and taking expectations gives (58).

For (59), first fix $f_1 \in \mathcal{F}_1$ and choose $g = g_h$ the corresponding solution of (61) with $\bullet = 1$. By construction we can then take expectations and write

$$\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)] = \mathbb{E}\left[f_1(X_2)\mathcal{D}_1^\star g_h(X_2) + g_h(X_2)\mathcal{T}_1 f_1(X_2)\right]$$

because $h \in L^1(X_1) \cap L^1(X_2)$. Finally we know that for all $f_2 \in \mathcal{F}_2$ such that $g_h \in dom(\mathcal{D}_2, X_2, f_2)$ we can use (61) with $\bullet = 2$ to get

$$\mathbb{E}\left[f_2(X_2)\mathcal{D}_2^\star g_h(X_2) + g_h(X_2)\mathcal{T}_2 f_2(X_2)\right] = 0.$$

Taking differences we get (59). Equation (60) follows in a similar fashion, fixing this time $f = f_h$ and letting g_1 and g_2 vary. □

The power of Theorem 5.1 and of Stein's method in general lies in the freedom of choice on the r.h.s. of the identities : all functions $f_\bullet, g_\bullet$ (where now $\bullet$ needs to be replaced by $h, 1$ or 2 according to which of (59) or (60) is used) can be chosen so as to optimise resulting bounds. We can even optimise the bounds over all suitable pairs (f,g). We will discuss two particular choices of functions in Section 5.2 which lead to well-known Stein bounds. We will also provide illustrations (discrete vs discrete, continuous vs continuous and discrete vs continuous) in Section 6.

In particular (59) and (60) provide tractable (and still very general) versions of (5). Indeed taking suprema over all $h \in \mathcal{H}$ some suitably chosen class of functions we get, in the notations of the Introduction,

$$d_{\mathcal{H}}(X_1, X_2) = \sup_{h\in\mathcal{H}} |\mathbb{E}_2 h - \mathbb{E}_1 h| \leq A_1 + A_2$$

with

$$A_1 = A_1(\mathcal{H}) = \sup_{h\in\mathcal{H}} \left|\mathbb{E}_2\left[f_\bullet(\cdot)\mathcal{D}_1^\star g_\bullet(\cdot) - f_\bullet(\cdot)\mathcal{D}_2^\star g_\bullet(\cdot)\right]\right|$$

and

$$A_2 = A_2(\mathcal{H}) = \sup_{h\in\mathcal{H}} \left|\mathbb{E}_2\left[g_\bullet(\cdot)\mathcal{T}_1 f_\bullet(\cdot) - g_\bullet(\cdot)\mathcal{T}_2 f_\bullet(\cdot)\right]\right|.$$

Different choices of functions f_1 and f_2 (resp. g_1 and g_2) will lead to different expressions bounding all distances $d_{\mathcal{H}}(X_1, X_2)$ in terms of properties of $\mathcal{T}_1$ and $\mathcal{T}_2$.

Remark 5.3. *If there exist no functions f_1, f_2 (resp. g_1, g_2) such that the assumptions are satisfied, then the claims of Theorem 5.1 are void. Such is not the case whenever p_1 and p_2 are "reasonable".*

Remark 5.4 (About the Stein factors)**.** *In view of* (59) *and* (60)*, good bounds on $\mathbb{E}_1 h - \mathbb{E}_2 h$ will depend on the control we have on functions*

$$g_h = \frac{\mathcal{T}_1^{-1}(h - \mathbb{E}_1 h)}{f_1} \text{ and/or } f_h = \frac{\mathcal{T}_1^{-1}(h - \mathbb{E}_1 h)}{g_1}. \tag{62}$$

Bounds on these functions and on their derivatives are called, in the dedicated literature, Stein (magic) factors *(see for example [25, 86]). There is an important connection between such constants and Poincaré / variance bounds or spectral gaps, as already noted for example in [23, 56, 54, 61, 72]. This connection is quite transparent in our framework and will be explored in future publications.*

In the sequel we will not use the full freedom of choice provided by Theorem 5.1, but rather focus on applications of identity (59) only. Indeed in this case much is known about $\|g_h\|$ and $\|\mathcal{D}g_h\|$ in case $f_1 = 1$ and X_1 is Gaussian (see [24]), Binomial (see [31]), Poisson (see [8]), Gamma (see [19, 71]), etc. See also [30, 58, 28, 65] for computations under quite general assumptions on the density of X_1. We will make use of these results in Section 6. It is hopeless to wish for useful bounds on (62) in all generality (see also the discussion in [3]). Of course one could proceed as in [28, 21] or [65] by imposing *ad hoc* assumptions on the target density which ensure that the functions in (62) have good properties. Such approaches are not pursued in this paper. Specific bounds will therefore only be discussed in particular examples.

5.2. Comparing Stein kernels and score functions

There are two obvious ways to exploit (59), namely either by trying to make the first summand equal zero, or by trying to make the second summand equal zero. In the rest of this section we do just that, in the case $\mathcal{X}_1 = \mathcal{X}_2$ and $\mathcal{D}_1 = \mathcal{D}_2 = \mathcal{D}$ (and hence $l_1 = l_2 = l$); extension of this result to mixtures is straightforward.

Cancelling the first term in (59) and ensuring that all resulting assumptions are satisfied immediately leads to the following result.

Corollary 5.5. *Let $\mathcal{H} \subset L^1(X_1) \cap L^1(X_2)$. Take $f \in \mathcal{F}_1 \cap \mathcal{F}_2$ and suppose that $(1/f)\mathcal{T}_1^{-1}(h - \mathbb{E}_1 h) \in dom(\mathcal{D}, X_1, f) \cap dom(\mathcal{D}, X_2, f)$ for all $h \in \mathcal{H}$. Then*

$$\sup_{h\in\mathcal{H}} |\mathbb{E}_1 h - \mathbb{E}_2 h| \leq \kappa_{\mathcal{H},1}(f)\mathbb{E}_2|\mathcal{T}_1 f - \mathcal{T}_2 f| \tag{63}$$

with $\kappa_{\mathcal{H},1}(f) = \sup_{h\in\mathcal{H}} \|(1/f)\,\mathcal{T}_1^{-1}(h - \mathbb{E}_1 h)\|_\infty$.

Remark 5.6. *1. If the constant function $1 \in \mathcal{F}_1 \cap \mathcal{F}_2$, then we can take $f = 1$ in (63) to deduce that*

$$d_{\mathcal{H},1}(X_1, X_2) \leq \kappa_{\mathcal{H},1}(1)\mathbb{E}_2\,|u_1 - u_2| \leq \kappa_{\mathcal{H},1}\sqrt{\mathbb{E}_2\left[(u_1 - u_2)^2\right]},$$

with $u_i = \mathcal{T}_i(1)$ the score function of X_i (defined in (47)) and $\kappa_{\mathcal{H},1}$ an explicit constant that can be computed in several important cases, see e.g. [65, Section 4] and [89, 54] for applications in the Gaussian case. Note that $\mathcal{J}(X_1, X_2) = \mathbb{E}_2\left[(u_1 - u_2)^2\right]$ is the so-called generalized Fisher information distance (see e.g. [53, 65]).

2. The assumption that $f \in \mathcal{F}_1 \cap \mathcal{F}_2$ can be relaxed; if $\int_I D_2(fp_2)d\mu \neq 0$ then this just adds terms which relate to the boundaries of I.

Cancelling the second term in (59) and ensuring that all resulting assumptions are satisfied immediately leads to the following result.

Corollary 5.7. *Let $\mathcal{H} \subset L^1(X_1) \cap L^1(X_2)$. Take $\omega \in Im(\mathcal{T}_1) \cap Im(\mathcal{T}_2)$ such that $\mathcal{T}_1^{-1}(h-\mathbb{E}_1 h)/\mathcal{T}_1^{-1}(\omega) \in dom(\mathcal{D}, X_1, \mathcal{T}_1^{-1}(\omega)) \cap dom(\mathcal{D}, X_2, \mathcal{T}_2^{-1}(\omega))$. Then*

$$|\mathbb{E}_1 h - \mathbb{E}_2 h| \leq \kappa_{\mathcal{H},2}(\omega)\mathbb{E}_2|\mathcal{T}_1^{-1}(\omega) - \mathcal{T}_2^{-1}(\omega)| \tag{64}$$

with $\kappa_{\mathcal{H},2}(\omega) = \sup_{h\in\mathcal{H}} \|\mathcal{D}\left(\mathcal{T}_1^{-1}(h-\mathbb{E}_1 h)/\mathcal{T}_1^{-1}(\omega)\right)\|_\infty$.

If, moreover, X_1 and X_2 have common finite mean ν then one can choose $\omega(x) = \nu - x$ in (64) to get

$$|\mathbb{E}_1 h - \mathbb{E}_2 h| \leq \kappa_{\mathcal{H},2}\mathbb{E}_2|\tau_1 - \tau_2| \tag{65}$$

with τ_j, $j = 1, 2$, the Stein kernel of X_j (defined in (49)) and $\kappa_{\mathcal{H},2}$ an explicit constant that can be computed in several cases. In [13], and references therein, consequences of (65) are explored in quite some detail. In particular in the Gaussian and central Gamma cases, (65) has been exploited fruitfully in conjunction with Malliavin calculus, leading to the important new stream of research known as Nourdin-Peccati analysis, see [71, 70]. See also aforementioned references [59, 58, 30] where several extensions of the Nourdin-Peccati analysis are discussed. Note that, in the Gaussian case $X_1 \sim \mathcal{N}(0,1)$ we readily obtain $\tau_1 = 1$. The quantity

$$S(X) = \sqrt{\mathbb{E}\left[(1-\tau_2)^2\right]} \tag{66}$$

is the *Stein discrepancy* from [73, 60].

5.3. Sums of independent random variables and the Stein kernel

We begin by relaxing the definition of Stein kernel. This approach is similar to that advocated in [74].

Definition 5.8. *Let $\mathcal{X}$ be a set and $\mathcal{D}$ a linear operator acting on $\mathcal{X}^\star$ satisfying the Assumptions of Section 3.1. Let $X \sim p$ have mean ν and $\mathcal{D}$-Stein pair $(\mathcal{T}_X, \mathcal{F}(X))$. A random variable $\tau_X(X)$ is a $\mathcal{D}$-Stein kernel for X if it is measurable in X and if*

$$\mathbb{E}\left[\tau(X)\mathcal{D}^\star g(X-l)\right] = \mathbb{E}\left[(X-\nu)g(X)\right]] \tag{67}$$

for all $g \in dom(\mathcal{D}, X, \tau)$. If, moreover, $dom(\mathcal{D}, X, \tau)$ is dense in $L^1(\mu)$ then the Stein kernel is unique.

Applying (35) one immediately sees that $\mathcal{T}_p^{-1}(Id-\nu)$ is a Stein kernel for X.

Proposition 5.9. *If $\mathcal{D}^\star$ satisfies a chain rule $\mathcal{D}^\star f(ax) = a\mathcal{D}_a^\star f(x)$ for some operator $\mathcal{D}_a^\star$ satisfying the same assumptions as $\mathcal{D}$ but now on $a\mathcal{X}$ then*

$$\tau_{aX}(aX) = a^2\tau_X(X) \tag{68}$$

is a Stein kernel for aX.

Proof. The claim follows immediately from the definition. □

Let $X_i, i = 1, \ldots, n$, be independent random variables with respective means ν_i, and put $W = \sum_{i=1}^n X_i$. Following [91, Lecture VI] and [74, 73] we obtain an almost sure representation formula for the Stein kernel of sums of independent random variables.

Lemma 5.10. *Suppose that (i) $Id - \nu_i \in Im(\mathcal{T}_i)$ for $i = 1, \ldots, n$ and (ii) $Id - \sum_{i=1}^n \nu_i \in Im(\mathcal{T}_W)$ and (iii) the collection of functions of the form $\mathcal{D}^\star g$ with $g \in dom(\mathcal{D}, W, \tau_W) \cap (\bigcap_{i=1}^n dom(\mathcal{D}, X_i, \tau_{X_i}))$ is dense in $L^1(\mu)$. Then*

$$\tau_W(W) = \mathbb{E}\left[\sum_{i=1}^n \tau_{X_i}(X_i) \,|\, W\right] \quad a.s.$$

Proof. For every $g \in dom(\mathcal{D}, W, \tau_W)$ we have with (35) that

$$\begin{aligned} -\mathbb{E}[\tau_W(W)\mathcal{D}^\star g(W)] &= \mathbb{E}\left[\left(W - \sum_{i=1}^n \nu_i\right) g(W)\right] \\ &= \sum_{i=1}^n \mathbb{E}\left\{\mathbb{E}[(X_i - \nu_i) g(W) | W_i]\right\} \end{aligned}$$

where $W_i = W - X_i = \sum_{j \neq i} X_j$ is independent of X_i. Therefore, conditionally on W_i we can use (an appropriate version of) (35) for each X_i, turning the previous expression into

$$\begin{aligned} -\sum_{i=1}^n \mathbb{E}\left\{\mathbb{E}\left[\tau_{X_i}(X_i)\mathcal{D}^\star g(W) | W_i\right]\right\} &= -\sum_{i=1}^n \mathbb{E}\left\{\mathbb{E}\left[\tau_{X_i}(X_i)\mathcal{D}^\star g(W) | W\right]\right\} \\ &= -\mathbb{E}\left\{\mathbb{E}\left[\sum_{i=1}^n \tau_{X_i}(X_i) | W\right] \mathcal{D}^\star g(W)\right\} \end{aligned}$$

where the first equality follows de-conditioning w.r.t. W_i and then conditioning w.r.t. W. The assertion follows by denseness. □

Combining this representation lemma with Corollary 5.7 leads to the following general result, which in particular implies inequality (28) from Section 2.7.

Proposition 5.11. *Suppose that the assumptions in Lemma 5.10 are satisfied. Let X be a random variable with finite mean $\nu = \sum_{i=1}^n \nu_i$. If $g_h = \mathcal{T}_X^{-1}(h - \mathbb{E}[h(X)])/\tau_X \in dom(\mathcal{D}, W, \tau_W) \cap dom(\mathcal{D}, X, \tau_X)$ then*

$$|\mathbb{E}[h(X)] - \mathbb{E}[h(W)]| \leq ||\mathcal{D}g_h||_\infty \mathbb{E}\left|\tau_X(W) - \sum_{i=1}^n \tau_{X_i}(X_i)\right|$$

for all $h \in \mathcal{H}$ a class of functions as in Corollary 5.7.

Proof. Lemma 5.10 with Corollary 5.7 (whose conditions are satisfied) gives that

$$|\mathbb{E}[h(X)] - \mathbb{E}[h(W)]| \leq ||\mathcal{D}g_h||_\infty \, |\mathbb{E}[\tau_X(W) - \tau_W(W)]|$$

$$\leq ||\mathcal{D}g_h||_\infty \mathbb{E}\left|\tau_X(W) - \mathbb{E}\left[\sum_{i=1}^n \tau_{X_i}(X_i)|W\right]\right|.$$

The assertion now follows by Jensen's inequality for conditional expectations. □

Proposition 5.12. *Let $W = \frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i$ with $\xi_i, i = 1,\dots,n$, centered independent random variables with $\mathcal{D}$-Stein kernels $\tau_i, i = 1,\dots,n$. Then*

$$\tau_W(W) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\tau_i(\xi_i)\,|\,W\right] \tag{69}$$

is a Stein kernel for W. Furthermore the Stein discrepancy of W satisfies

$$S(W) := \sqrt{\mathbb{E}\left[(1-\tau_W(W))^2\right]} \leq \frac{1}{n}\sqrt{\sum_{i=1}^n Var(\tau_i(\xi_i))}. \tag{70}$$

Proof. Identity (69) follows from a straightforward conditioning argument. To see (70) note how under the assumptions of the proposition we have

$$\begin{aligned}\mathbb{E}\left[(1-\tau_W(W))^2\right] &= \mathbb{E}\left[\left(\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n(1-\tau_i(\xi_i))\,|\,W\right]\right)^2\right]\\ &\leq \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n(1-\tau_i(\xi_i))\right)^2\right]\\ &\leq \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(\tau_i(\xi_i)).\end{aligned}$$

□

Our general setup also caters for comparison of distributions with Stein pair based on different linear operators $\mathcal{D}$; this has already been explored in [43] for Beta approximation of the Pólya-Eggenberger distribution. Here we illustrate the technique for Gaussian comparison in terms of Stein discrepancies.

Proposition 5.13. *Let $\mathcal{D}$ be a linear operator satisfying the Assumptions from Section 3.1; let l be as in Assumption 1. Let W be centered with variance σ^2, and $\mathcal{D}$-Stein pair $(\mathcal{T}_W, \mathcal{F}(W))$; let τ_W be the corresponding Stein kernel. Let $Z \sim \mathcal{N}(0,1)$ and $S(W)$ be as above the Stein discrepancy between W and Z. Then for all $g \in dom((\cdot)', Z) \cap dom(\mathcal{D}, W, \tau_W)$ we have*

$$|\mathbb{E}\left[g'(W) - Wg(W)\right]| \leq S(W)\|g'\| + \sigma^2\|g'(\cdot) - \mathcal{D}^\star g(\cdot - l)\|_\infty + \|g(\cdot - l) - g(\cdot)\|_\infty. \tag{71}$$

Proof. Applying Proposition 4.4 to $\nu = 0$ we get

$$\mathbb{E}\left[Wg(W-l)\right] = \mathbb{E}\left[\tau_W(W)\mathcal{D}^\star g(W-l)\right] \tag{72}$$

for all $g \in dom(\mathcal{D}, W, \tau_W)$. If furthermore $g \in dom((\cdot)', Z)$ then

$$\begin{aligned}
&\mathbb{E}\left[g'(W) - Wg(W)\right] \\
&= \mathbb{E}\left[g'(W) - Wg(W-l)\right] + \mathbb{E}\left[W\left(g(W-l) - g(W)\right)\right] \\
&= \mathbb{E}\left[g'(W) - \tau_W(W)\mathcal{D}^\star g(W-l)\right] + \mathbb{E}\left[W\left(g(W-l) - g(W)\right)\right] \\
&= \mathbb{E}\left[g'(W)(1-\tau_W(W))\right] + \mathbb{E}\left[\tau_W(W)\left(g'(W) - \mathcal{D}^\star g(W-l)\right)\right] \\
&\quad + \mathbb{E}\left[W\left(g(W-l) - g(W)\right)\right].
\end{aligned}$$

Applying Cauchy-Schwarz to the first summand in the last equality yields the first summand of (71). To get the second summand of (71) note that $\tau_W(W) \geq 0$ almost surely (recall Remark 4.5) so that

$$\left|\mathbb{E}\left[\tau_W(W)\left(g'(W) - \mathcal{D}^\star g(W-l)\right)\right]\right| \leq \mathbb{E}\left[\tau_W(W)\right] \left\| \left(g'(\cdot) - \mathcal{D}^\star g(\cdot - l)\right) \right\|_\infty$$

and now we use $\mathbb{E}\left[\tau_W(W)\right] = \mathrm{Var}(W) = \sigma^2$. The last term in (71) follows by a similar reasoning. □

As an illustration we now provide a Gaussian approximation bound in the Wasserstein distance under a Stein kernel assumption.

Proposition 5.14. *Let W be centered with variance σ^2 and support in $\delta\mathbb{Z}$ for some $\delta > 0$. Consider $\mathcal{D} = \delta^{-1}\Delta_\delta^+$ as in Example 3.3. Suppose that the assumptions in Lemma 5.10 are satisfied. Then*

$$d_{\mathrm{Wass}}(W, Z) \leq S(W) + (1+\sigma^2)\delta \tag{73}$$

with $d_{\mathrm{Wass}}(W, Z)$ the Wasserstein distance between the laws of W and Z.

Proof. We aim to apply (71), with $g = g_h$ the classical solution to the Gaussian Stein equation

$$g'(x) - xg(x) = h(x) - \mathbb{E}[h(Z)]$$

where h is a Lipschitz function with constant 1. The properties of such g are well understood, see e.g. [6, Lemma 2.3]. In particular these functions are differentiable and bounded with $\|g'\|_\infty \leq 1$ so that

$$|g(x-\delta) - g(x)| = \int_{-\delta}^{0} g'(x+u)du \leq \delta$$

for all $x \in \mathbb{R}$. Also, $\|g''\|_\infty \leq 2$ and hence

$$\begin{aligned}
|g'(x) - \mathcal{D}^\star g(x-l)| &= |g'(x) - \delta^{-1}(g(x) - g(x-\delta))| \\
&= \left|\frac{1}{\delta}\int_{-\delta}^{0}\int_{0}^{u} g''(x+v)dvdu\right| \\
&\leq \delta,
\end{aligned}$$

again for all $x \in \mathbb{R}$. The claim follows. □

Finally, following up on the results presented in Section 2.7, we conclude with a central limit theorem for sums of centered Rademacher random variables.

Corollary 5.15. *Let* $W = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i$ *with* $\xi_i, i = 1, \dots, n$, *independent centered with support in* $\{-1, 1\}$. *Fix* $\mathcal{D}f = f(x+1) - f(x-1)$ *and let* $\tau_i(\xi_i) = \mathbb{I}(\xi_i = 1)$. *The* $\tau_i(\xi_i)_{i=1,\dots,n}$ *are* $\mathcal{D}$*-Stein kernels for* $(\xi_i)_{i=1,\dots,n}$ *and*

$$d_{\text{Wass}}(W, Z) \leq \frac{3}{\sqrt{n}}. \tag{74}$$

Proof. The first claim is immediate. Next we use (70) to deduce that

$$S(W) \leq \frac{1}{n} \sqrt{\sum_{i=1}^{n} \text{Var}(\tau_i(\xi_i))} = \frac{1/2}{\sqrt{n}}.$$

Finally we apply (73) with $\sigma^2 = 1$ and $\delta = \frac{1}{\sqrt{n}}$. □

Remark 5.16. *It is straightforward to extend the results of this Section to random sums of independent random variables and therefore deduce central limit theorems for randomly centered random variables. A much more challenging task is to deal with non-randomly centered random sums, as e.g. in [27].*

6. Stein bounds

As anticipated, in this Section we discuss non-asymptotic approximation via Stein differentiation in several concrete examples. The main purpose of this Section is illustrative and most of the examples we discuss lead to well-known situations. Relevant references are given in the text.

6.1. Binomial approximation to the Poisson-binomial distribution

An immediate application of Proposition 5.11 can be found in binomial approximation for a sum of independent Bernoulli random variables. Writing X for a $\text{Bin}(n, p)$ and $W = \sum_{i=1}^{n} X_i$ with $X_i \sim \text{Bin}(1, p_i)$, $i = 1, \dots, n$, and $np = \sum_{i=1}^{n} p_i$ (the distribution of W is called a Poisson-binomial distribution, see e.g. [31]), we readily compute

$$\tau_X(x) = (1-p)x \text{ and } \tau_{X_i}(x) = (1-p_i)x.$$

Here we use $\mathcal{D} = \Delta^+$, the forward difference. Thus for any measurable function h such that $\mathbb{E}|h(X)| < \infty$ and $\mathbb{E}|h(W)| < \infty$,

$$\begin{aligned} |\mathbb{E}[h(X)] - \mathbb{E}[h(W)]| &\leq ||\mathcal{D}g_h||_\infty \mathbb{E} \left| (1-p)W - \sum_{i=1}^{n} (1-p_i) X_i \right| \\ &\leq ||\mathcal{D}g_h||_\infty \sum_{i=1}^{n} |p_i - p| p_i. \end{aligned} \tag{75}$$

An alternative angle on this problem is to use the score function approach, although here with $\mathcal{T}(Id)$ instead of $\mathcal{T}(1)$. It is easy to show (see e.g. Example 4.9.2.(b)) that

$$\mathcal{T}_{Bin(n,p)}(f)(x) = \frac{p(n-x)}{(1-p)(x+1)} f(x+1) - f(x)$$

so that for $f = Id$, the identity function,

$$\mathcal{T}_{Bin(n,p)}(Id)(x) = \frac{np-x}{1-p}.$$

By Example 3.13 we find that $f = Id \in \mathcal{F}(X) \cap \mathcal{F}(W)$ because $Id(0) = 0$. Now let h be such that $\mathbb{E}|h(X)| < \infty$ and $\mathbb{E}|h(W)| < \infty$, and let $g_h = \mathcal{T}_X^{-1}(h - \mathbb{E}[h(X)])/Id$; then $g \in dom(\Delta^+, W, Id) \cap dom(\Delta^+, X, Id)$. From (59) we obtain that

$$\mathbb{E}[h(W)] - \mathbb{E}[h(X)] = \mathbb{E}\left[g_h(W+1)\left\{\mathcal{T}_{Bin(n,p)}(Id)(W) - \mathcal{T}_{\mathcal{L}(W)}(Id)(W)\right\}\right].$$

By (33), using the notation $g_a(x) = g(x+a)$ for a function in x,

$$\begin{aligned}
&\mathbb{E}[g_h(W+1)\mathcal{T}_{\mathcal{L}(W)}(Id)(W)] \\
&= -\mathbb{E}[W\Delta^- g(W+1)] \\
&= -\sum_{i=1}^n \mathbb{E}\left[\mathbb{E}\left\{X_i \Delta^- g_{\sum_{j\neq i} X_j + 1}(X_i) | X_j, j \neq i\right\}\right] \\
&= \sum_{i=1}^n \mathbb{E}\left[\mathbb{E}\left\{\mathcal{T}_{Bin(1,p_i)}(Id)(X_i) g_{\sum_{j\neq i} X_j + 1}(X_i) | X_j, j \neq i\right\}\right] \\
&= \sum_{i=1}^n \mathbb{E}\left\{g(W+1)\mathcal{T}_{Bin(1,p_i)}(Id)(X_i)\right\}.
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathbb{E}[h(W)] - \mathbb{E}[h(X)] \\
&= \mathbb{E}\left[g_h(W+1)\left\{\mathcal{T}_{Bin(n,p)}(Id)(W) - \sum_{i=1}^n \mathcal{T}_{Bin(1,p_i)}(Id)(X_i)\right\}\right] \\
&= \mathbb{E}\left[g_h(W+1)\left\{\frac{np-W}{1-p} - \sum_{i=1}^n \frac{p_i - X_i}{1-p_i}\right\}\right] \\
&= \mathbb{E}\left[g_h(W+1)\sum_{i=1}^n (p_i - X_i)\left\{\frac{1}{1-p} - \frac{1}{1-p_i}\right\}\right]
\end{aligned}$$

and so

$$|\mathbb{E}[h(X)] - \mathbb{E}[h(W)]| \leq \frac{\|g_h\|_\infty}{1-p} \sum_{i=1}^n |p - p_i| \mathbb{E}\left|\frac{p_i - X_i}{1-p_i}\right| = \frac{2\|g_h\|_\infty}{1-p} \sum_{i=1}^n |p - p_i| p_i. \tag{76}$$

The fact that we obtain two different bounds, (75) and (76), for the same problem illustrates the freedom of choice in specifying f and g in the Stein equa-

tion. In [31], bounds for $\sup_x \left|\mathcal{D}\frac{g_h(x)}{x+1}\right|$ are calculated, and in [34] a bound for $\sup_x \left|\frac{g_h(x)}{x+1}\right|$ is given.

6.2. Distance between Gaussians

Consider two centered Gaussian random variables X_1 and X_2 with respective variances $\sigma_1^2 \leq \sigma_2^2$, say. Denote ϕ the density of Z, a standard normal random variable. The canonical Stein operators are then of the form

$$\mathcal{T}_i f(x) = f'(x) - \frac{x}{\sigma_i^2} f(x)$$

acting on the classes $\mathcal{F}_1(X_1) = \mathcal{F}_2(X_2) = \mathcal{F}(Z)$ of Z-integrable differentiable functions such that $(f\phi)' \in L^1(dx)$. In this simple toy-setting it is possible to write out (59) in full generality. Indeed we have

$$\begin{aligned} f_1 g_h &= \mathcal{T}_1^{-1}(h - \mathbb{E}_1 h) \\ &= e^{x^2/(2\sigma_1^2)} \int_{-\infty}^{x} (h(y) - \mathbb{E}[h(X_1)]) e^{-y^2/(2\sigma_1^2)} dy \\ &= e^{(x/\sigma_1)^2/2} \sigma_1 \int_{-\infty}^{x/\sigma_1} (h(\sigma_1 u) - \mathbb{E}[h(\sigma_1 Z)])\, e^{-u^2/2} du \quad =: \sigma_1 g_{\tilde{h},0}(x/\sigma_1) \end{aligned}$$

with $\tilde{h}(u) = h(\sigma_1 u)$ and $g_{h,0}$ the solution of the classical Stein equation given by

$$g_{h,0}(x) = e^{x^2/2} \int_{-\infty}^{x} (h(y) - \mathbb{E}[h(Z)]) e^{-y^2/2} dy.$$

In the particular case where one is interested in the total variation distance, one only considers $h : \mathbb{R} \to [0,1]$ Borel functions for which $\|g_{h,0}\| \leq \sqrt{\frac{\pi}{2}}$ and $\|g'_{h,0}\| \leq 2$ (see e.g. [71, Theorem 3.3.1]). In the rest of this Section we focus on such h, although similar results are available for $h = \mathbb{I}_{(-\infty,z]}$ (leading to bounds on the Kolmogorov distance, see [24, Lemma 2.3]) and for $h \in Lip(1)$ (leading to bounds on the Wasserstein distance, see [71, Proposition 3.5.1]). Identity (59) becomes

$$\begin{aligned} \mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)] = \mathbb{E}\Bigg[&(f_1(X_2) - f_2(X_2)) \left(\frac{\sigma_1 g_{\tilde{h},0}(X_2/\sigma_1)}{f_1(X_2)} \right)' \\ &+ (\mathcal{T}_1 f_1(X_2) - \mathcal{T}_2 f_2(X_2)) \left(\frac{\sigma_1 g_{\tilde{h},0}(X_2/\sigma_1)}{f_1(X_2)} \right) \Bigg]. \end{aligned}$$

for any $f_1, f_2 \in \mathcal{F}(Z)$. There are many directions that can be taken from here, of which we illustrate three (to simplify notation we write g_h for $g_{\tilde{h},0}$).

- Taking $f_1 = 1$ and $f_2 = 0$ (see Remark 5.2) leads to the identity

$$\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)] = \mathbb{E}\left[g'_h\left(\frac{X_2}{\sigma_1}\right) - \frac{X_2}{\sigma_1} g_h\left(\frac{X_2}{\sigma_1}\right) \right]$$

because $\mathcal{T}_1(1)(x) = -x/\sigma_1^2$. Recalling that $\mathbb{E}[X_2\zeta(X_2)] = \sigma_2^2\mathbb{E}[\zeta'(X_2)]$ for any differentiable function ζ, and also noting that one can interchange the roles of X_1 and X_2, we deduce the bound

$$d_{\mathrm{TV}}(X_1, X_2) \leq \frac{2}{\sigma_2^2}\left|\sigma_1^2 - \sigma_2^2\right|, \tag{77}$$

already obtained e.g. in [71, Proposition 3.6.1].

- Taking $f_1 = \sigma_1^2$ and $f_2 = \sigma_2^2$ (thus a particular case of the comparison of kernels from Corollary 5.7) also yields (77).
- Taking $f_1 = f_2 = 1$ (thus a particular case of the comparison of scores from Corollary 5.5) yields the identity

$$\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)] = \mathbb{E}\left[X_2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)\left(\sigma_1 g_{h,0}\left(\frac{X_2}{\sigma_1}\right)\right)\right]$$

because $\mathcal{T}_i(1)(x) = -x/\sigma_i^2$. Using $\mathbb{E}|X_2| = \sqrt{\frac{2}{\pi}}\sigma_2$ and $\|\sigma_1 g_{h,0}(\cdot/\sigma_1)\|_\infty \leq \sigma_1\sqrt{\frac{\pi}{2}}$ leads to

$$d_{\mathrm{TV}}(X_1, X_2) \leq \frac{\left|\sigma_1^2 - \sigma_2^2\right|}{\sigma_1\sigma_2},$$

which is better than (77) whenever $\sigma_2/\sigma_1 < 2$.

6.3. From student to Gauss

Set $X_1 = Z$ standard Gaussian and $X_2 = W_\nu$ a Student t random variable with $\nu > 2$ degrees of freedom. In this case the Stein kernels for both distributions are well defined and given, respectively, by $\tau_1 = 1$ and $\tau_2(x) = \frac{x^2+\nu}{\nu-1}$, see Example 4.9. All assumptions in Corollary 5.7 are satisfied so that we can plug these functions with $\mathcal{H}$ the class of Borel functions in $[0,1]$ to get

$$d_{\mathrm{TV}}(Z, W_\nu) \leq 2\mathbb{E}\left|\frac{W_\nu^2 + \nu}{\nu - 1} - 1\right| \tag{78}$$

where, as in the previous example, we make use of our knowledge on the solutions of the Gaussian Stein equation. It is straightforward to compute (78) explicitly (under the assumption $\nu > 2$, otherwise the expectation does not exist) to get

$$d_{\mathrm{TV}}(Z, W_\nu) \leq \frac{4}{\nu - 2}. \tag{79}$$

A similar result is obtained with Corollary 5.5, namely

$$d_{\mathrm{TV}}(Z, W_\nu) \leq \sqrt{\frac{\pi}{2}}\frac{-2 + 8\left(\frac{\nu}{1+\nu}\right)^{(1+\nu)/2}}{(\nu-1)\sqrt{\nu}B(\nu/2, 1/2)},$$

which is of the same order as (79), with a better constant, but arguably much less elegant.

Remark 6.1. *It is of course possible to exchange the roles of the Student and the Gaussian in the above computations.*

6.4. Exponential approximation

Let $X_{(n)}$ be the maximum of n i.i.d. uniform random variables on $[0,1]$. It is known that $M_n = n(1-X_{(n)})$ converges in distribution to X_1 a rate-1 exponential random variable. Note that $\mathbb{E}[M_n] = \frac{n}{n+1} \neq 1$. In order to apply Corollary 5.7 most easily we are led to consider the slightly transformed random variable $X_2 = \frac{n+1}{n} M_n = (n+1)(1-X_{(n)})$.

The canonical operator for X_1 is $\mathcal{T}_1 f = f' - f$ acting on the class of differentiable f such that $f(0) = 0$. The Stein equation (37) becomes

$$h(x) - \mathbb{E}[h(X_1)] = f(x)g'(x) + f'(x)g(x) - g(x)f(x) = (fg)'(x) - (fg)(x).$$

Then the solution pairs $(f,g) = (f_h, g_h)$ are such that $(fg)(x) = \mathcal{T}_{\exp}^{-1}(h)$ so that

$$(fg)(x) = e^x \int_0^x (h(u) - \mathbb{E}[h(X_1)]) e^{-u} du \tag{80}$$

for $x > 0$. If $h(x) = \mathbb{I}(x \leq t)$ we need to understand the properties of

$$(fg)(x) = e^{-(t-x)^+} - e^{-t}.$$

This function is bounded and differentiable on $\mathbb{R}$, with limit 0 at the left boundary and constant with value $1 - e^{-t}$ for all $x \geq t$ (see also [19, Lemma 3.2]). Taking $g(x) = x^\epsilon$ in (80) the corresponding function f from (80) is

$$f_{t,\epsilon}(x) = x^{-\epsilon} \left(e^{-(t-x)^+} - e^{-t} \right)$$

with $a^+ = \max(a, 0)$. For all choices $0 < \epsilon < 1$ we have

$$\lim_{x \to 0} f_{t,\epsilon}(x) = 0 \text{ and } \lim_{x \to \infty} f_{t,\epsilon}(x) = 0 \text{ and } \|f_{t,\epsilon}\|_\infty = t^{-\epsilon}(1 - e^{-t}),$$

as well as $\lim_{x \to 0} f_{t,1}(x) = e^{-t}$ (see [19] for details on the cases $\epsilon = 0$ and $\epsilon = 1$).

We now turn our attention to the problem of approximating the law of X_2, whose density is $p(x) = \frac{n}{n+1}(1 - \frac{x}{n+1})^{n-1}$ with support $[0, n+1]$. Taking derivatives we get

$$\mathcal{T}_2 f(x) = f'(x) - \frac{n-1}{n+1-x} f(x)$$

acting, as above, on the class of differentiable functions such that $f(0) = 0$. Clearly $f_{t,\epsilon}(0) g_\epsilon(0) = 0$ for all $0 < \epsilon < 1$ and therefore

$$\begin{aligned} P(X_2 \leq t) - P(X_1 \leq t) &= \mathbb{E}[(f_{t,\epsilon} g_\epsilon)'(X_2) - (f_{t,\epsilon} g_\epsilon)(X_2)] \\ &= \mathbb{E}\left[(f_{t,\epsilon} g_\epsilon)(X_2) \left\{ \frac{n-1}{n+1-X_2} - 1 \right\} \right] \end{aligned}$$

which yields the non-uniform bound

$$|P(X_1 \leq t) - P(X_2 \leq t)| \leq t^{-\epsilon}(1 - e^{-t}) \mathbb{E}\left[X_2^\epsilon \left| \frac{n-1}{n+1-X_2} - 1 \right| \right]. \tag{81}$$

The quantity on the rhs of (81) can be optimised numerically in (ϵ, t). For example for $n = 100$ and $t = 1/2$, we can compute the upper bound at $\epsilon = 0$ to get 0.00497143 and 0.00852033 at $\epsilon = 1$. The optimal choice of ϵ in this case is $\epsilon \approx 0.138$ for which the bound is 0.00488718. Obviously, in this simple situation, it is also easy to evaluate the expressions $\Delta(t) = \sup_t |P(X_2 \leq t) - P(X_1 \leq t)|$ numerically; explorations show that there is some interesting optimization (depending on the magnitude of t) to be performed in order to obtain good bounds.

6.5. Gumbel approximation

Let $X_{(n)}$ be the maximum of n i.i.d. exponential random variables. It is known that $M_n = X_{(n)} - \log n$ converges in distribution to X_1 a Gumbel random variable with density $p(x) = e^{-x}e^{-e^{-x}}$ on $\mathbb{R}$. The Stein kernel of the Gumbel does not take on a tractable form, hence we shall here rather use Corollary 5.7 with another choice of function ω.

A natural choice for ω is the score function, here $u_{\text{Gumbel}}(x) = e^{-x} - 1$, since in this case $\mathcal{T}^{-1}_{\text{Gumbel}}(u_{\text{Gumbel}}) = 1$. As for the exponential example, we here also run into the difficulty that $\mathbb{E}[e^{-M_n} - 1] = \frac{n}{n+1} - 1 \neq 0$, leading us to consider the transformed random variable $X_2 = M_n + \log \frac{n}{n+1}$. Simple calculations give $\mathcal{T}_2^{-1}(e^{-x} - 1) = 1 - \frac{e^{-x}}{n+1}$ and we can use Corollary 5.7 to obtain

$$\begin{aligned} |\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)]| &\leq ||g_h'||_\infty \mathbb{E}\left|1 - \left(1 - \frac{e^{-X_2}}{n+1}\right)\right| \\ &= ||g_h'||_\infty \frac{1}{n+1}\mathbb{E}\left[e^{-X_2}\right] \end{aligned}$$

with $g_h(x) = \mathcal{T}^{-1}_{\text{Gumbel}}(h)$. Since, furthermore, $\mathbb{E}\left[e^{-X_2}\right] = 1$ we deduce

$$|\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)]| \leq \frac{1}{n+1}||g_h'||_\infty.$$

Again it is easy to express g_h explicitly in most cases. For example, taking $h(x) = \mathbb{I}(x \leq t)$ we readily compute $g_h(x) = e^x\left(e^{-(e^{-t}-e^{-x})^+} - e^{-e^{-t}}\right)$ which can be shown to satisfy $\|g_h\| \leq e^t(1 - e^{-e^{-t}}) \leq 1$ and $\|g_h'\| \leq 1$. This provides the uniform bound

$$|P(X_2 \leq t) - P(X_1 \leq t)| \leq \frac{1}{n+1},$$

which is of comparable order (though with a worse constant) with, e.g., [50].

Acknowledgements

This work has been initiated when Christophe Ley and Yvik Swan were visiting Keble College, Oxford. Substantial progress was also made during a stay at the

CIRM in Luminy. Christophe Ley thanks the Fonds National de la Recherche Scientifique, Communauté française de Belgique, for support via a Mandat de Chargé de Recherche FNRS. Gesine Reinert was supported in part by EPSRC grant EP/K032402/1. Yvik Swan gratefully acknowledges support from the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy). We thank Carine Bartholmé for discussions which led to the application given in Section 2.6. The authors would further like to thank Oliver Johnson, Larry Goldstein, Giovanni Peccati and Christian Döbler for the many discussions about Stein's method which have helped shape part of this work. In particular, we thank Larry for his input on Section 3.6, Christian for the idea behind Section 4.6 and Oliver for the impetus behind the computations shown in Section 6.1. Finally, we thank an anonymous referee for their suggestions.

References

[1] Afendras, G., Balakrishnan, N. and Papadatos, N. (2014). Orthogonal polynomials in the cumulative Ord family and its application to variance bounds, preprint arXiv:1408.1849. MR3160581

[2] Afendras, G., Papadatos, N. and Papathanasiou, V. (2011). An extended Stein-type covariance identity for the Pearson family with applications to lower variance bounds, *Bernoulli* **17**, 507–529. MR2787602

[3] Arras, B., Azmoodeh, E., Poly, G. and Swan, Y. (2016). Stein's method on the second Wiener chaos: 2-Wasserstein distance, preprint arXiv:1601.03301.

[4] Baldi, P., Rinott, Y. and Stein, C. (1989). A normal approximation for the number of local maxima of a random function on a graph, *Probability, Statistics, and Mathematics, Academic Press, Boston, MA*, 59–81. MR1031278

[5] Barbour, A. D. (1990). Stein's method for diffusion approximations, *Probability Theory and Related Fields* **84**, 297–322. MR1035659

[6] Barbour, A. D. and Chen L. H. Y. (2005). *An introduction to Stein's method*, *Lecture Notes Series Inst. Math. Sci. Natl. Univ. Singap.* **4**, Singapore University Press, Singapore. MR2205339

[7] Barbour, A. D., Gan, H. L. and Xia, A. (2015). Stein factors for negative binomial approximation in Wasserstein distance, *Bernoulli* **21**, 1002–1013. MR3338654

[8] Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson approximation*, Oxford Studies in Probability, **2**, The Clarendon Press Oxford University Press, New York, Oxford Science Publications. MR1163825

[9] Barbour, A. D. and Eagleson, G. K. (1985). Multiple comparisons and sums of dissociated random variables, *Advances in Applied Probability* **17**, 147–162. MR0778597

[10] Barbour, A. D., Luczak M. J. and Xia A. (2015). Multivariate approximation in total variation, preprint arXiv:1512.07400.

[11] Bonis T. (2015). Stable measures and Stein's method: rates in the Central Limit Theorem and diffusion approximation, preprint arXiv:1506.06966.

[12] Brown, T. C. and Xia, A. (1995). On Stein-Chen factors for Poisson approximation, *Statistics & Probability Letters* **23**, 327–332. MR1342742

[13] Cacoullos, T., Papadatos, N. and Papathanasiou, V. (2001). An application of a density transform and the local limit theorem, *Teor. Veroyatnost. i Primenen.* **46**, 803–810. MR1971837

[14] Cacoullos, T. and Papathanasiou, V. (1989). Characterizations of distributions by variance bounds, *Statistics & Probability Letters* **7**, 351–356. MR1001133

[15] Cacoullos, T., Papadatos, N. and Papathanasiou, V. (1998). Variance inequalities for covariance kernels and applications to central limit theorems, *Theory of Probability & Its Applications* **42**, 149–155. MR1453340

[16] Cacoullos, T. and Papathanasiou, V. (1995). A generalization of covariance identity and related characterizations, *Mathematical Methods of Statistics* **4**, 106–113. MR1324694

[17] Cacoullos, T., Papathanasiou, V. and Utev, S. A. (1994). Variational inequalities with examples and an application to the central limit theorem, *Annals of Probability* **22**, 1607–1618. MR1303658

[18] Chatterjee, S. (2014). A short survey of Stein's method, *Proceedings of ICM 2014*, to appear.

[19] Chatterjee, S., Fulman, J. and Röllin, A. (2011). Exponential approximation by exchangeable pairs and spectral graph theory, *ALEA Latin American Journal of Probability and Mathematical Statistics* **8**, 1–27. MR2802856

[20] Chatterjee, S. and Meckes, E. (2008). Multivariate normal approximation using exchangeable pairs, *ALEA Latin American Journal of Probability and Mathematical Statistics* **4**, 257–283. MR2453473

[21] Chatterjee, S. and Shao, Q.-M. (2011). Nonnormal approximation by Stein's method of exchangeable pairs with application to the Curie-Weiss model, *Annals of Applied Probability* **21**, 464–483. MR2807964

[22] Chen, L. H. Y. (1975). Poisson approximation for dependent trials, *Annals of Probability* **3**, 534–545. MR0428387

[23] Chen, L. H. Y. (1980). An inequality for multivariate normal distribution, *Technical report, MIT*.

[24] Chen, L. H. Y., Goldstein, L. and Shao, Q.-M. (2011). *Normal approximation by Stein's method*, Probability and its Applications (New York), Springer, Heidelberg. MR2732624

[25] Daly, F. (2008). Upper bounds for Stein-type operators, *Electronic Journal of Probability* **13**, 566–587. MR2399291

[26] Diaconis, P. and Zabell, S. (1991). Closed form summation for classical distributions: variations on a theme of de Moivre, *Statistical Science* **6**, 284–302. MR1144242

[27] Döbler, C. (2012). On rates of convergence and Berry-Esseen bounds for random sums of centered random variables with finite third moments, preprint arXiv:1212.5401.

[28] Döbler, C. (2015). Stein's method of exchangeable pairs for the beta distribution and generalizations, *Electronic Journal of Probability* **20**, 1–34. MR3418541

[29] Döbler, C., Gaunt, R. E. and Vollmer, S. J. (2015). An iterative technique for bounding derivatives of solutions of Stein equations, preprint arXiv:1510.02623.

[30] Eden, R. and Viquez, J. (2015). Nourdin-Peccati analysis on Wiener and Wiener-Poisson space for general distributions, *Stochastic Processes and their Applications* **125**, 182–216. MR3274696

[31] Ehm, W. (1991). Binomial approximation to the Poisson binomial distribution, *Statistics & Probability Letters* **11**, 7–16. MR1093412

[32] Eichelsbacher, P. and Löwe, M. (2010). Stein's method for dependent random variables occurring in statistical mechanics, *Electronic Journal of Probability* **15**, 962–988. MR2659754

[33] Eichelsbacher, P. and Martschink, B. (2014). Rates of convergence in the Blume–Emery–Griffiths model, *Journal of Statistical Physics* **154**, 1483–1507. MR3176416

[34] Eichelsbacher, P. and Reinert, G. (2008). Stein's method for discrete Gibbs measures, *Annals of Applied Probability* **18**, 1588–1618. MR2434182

[35] Fulman, J. and Goldstein, L. (2015). Stein's method and the rank distribution of random matrices over finite fields, *Annals of Probability* **43**, 1274–1314. MR3342663

[36] Fulman, J. and Goldstein, L. (2014). Stein's method, semicircle distribution, and reduced decompositions of the longest element in the symmetric group, preprint arXiv:1405.1088.

[37] Gaunt, R. E. (2016). On Stein's method for products of normal random variables and zero bias couplings, *Bernoulli*, to appear.

[38] Gaunt, R. E. (2014). Variance-Gamma approximation via Stein's method, *Electronic Journal of Probability* **19**, 1–33. MR3194737

[39] Gaunt, R. E., Mijoule, G. and Swan, Y. (2016). Stein operators for product distributions, with applications. preprint arXiv:1604.06819.

[40] Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics, *International Statistical Review / Revue Internationale de Statistique* **70**, 419–435.

[41] Goldstein, L. and Reinert, G. (1997). Stein's method and the zero bias transformation with application to simple random sampling, *Annals of Applied Probability* **7**, 935–952. MR1484792

[42] Goldstein, L. and Reinert, G. (2005). Distributional transformations, orthogonal polynomials, and Stein characterizations, *Journal of Theoretical Probability* **18**, 237–260. MR2132278

[43] Goldstein, L. and Reinert, G. (2013). Stein's method for the Beta distribution and the Pólya-Eggenberger urn, *Journal of Applied Probability* **50**, 1187–1205. MR3161381

[44] Goldstein, L. and Rinott, Y. (1996). Multivariate normal approximations by Stein's method and size bias couplings, *Journal of Applied Probability* **33**, 1–17. MR1371949

[45] Gorham, J. and Mackey, L. (2015). Measuring sample quality with Stein's method. In: *Advances in Neural Information Processing Systems (NIPS)*, 226–234.
[46] Götze, F. and Tikhomirov, A. N. (2003). Rate of convergence to the semi-circular law, *Probability Theory and Related Fields* **127**, 228–276. MR2013983
[47] Götze, F. and Tikhomirov, A. N. (2006). Limit theorems for spectra of random matrices with martingale structure, *Teor. Veroyatnost. i Primenen.* **51**, 171–192. MR2324173
[48] Götze, F. (1991). On the rate of convergence in the multivariate CLT, *Annals of Probability* **19**, 724–739. MR1106283
[49] Haagerup, U. and Thorbjørnsen, S. (2012). Asymptotic expansions for the Gaussian unitary ensemble, *Infinite Dimensional Analysis, Quantum Probability and Related Topics* **15**, no. 01. MR2922846
[50] Hall, W. J. and Wellner, J. A. (1979). The rate of convergence in law of the maximum of an exponential sample, *Statistica Neerlandica* **33**, 151–154. MR0552259
[51] Hillion, E., Johnson, O. and Yu, Y. (2014). A natural derivative on [0, n] and a binomial Poincaré inequality. *ESAIM: Probability and Statistics* **18**, 703–712. MR3334010
[52] Holmes, S. (2004). Stein's method for birth and death chains, *Stein's Method: Expository Lectures and Applications, IMS Lecture Notes Monogr. Ser.,* **46**, 45–67. MR2118602
[53] Johnson, O. (2004). *Information Theory and the Central Limit Theorem*, Imperial College Press, London. MR2109042
[54] Johnson, O. and Barron, A. (2004). Fisher information inequalities and the central limit theorem, *Probability Theory and Related Fields* **129**, 391–409. MR2128239
[55] Johnson, R. W. (1993). A note on variance bounds for a function of a Pearson variate, *Statistics & Risk Modeling* **11**, 273–278. MR1257861
[56] Klaassen, C. A. J. (1985). On an inequality of Chernoff, *Annals of Probability* **13**, 966–974. MR0799431
[57] Korwar, R. M. (1991). On characterizations of distributions by mean absolute deviation and variance bounds, *Annals of the Institute of Statistical Mathematics* **43**, 287–295. MR1128869
[58] Kusuoka, S. and Tudor, C. A. (2012). Stein's method for invariant measures of diffusions via Malliavin calculus, *Stochastic Processes and Their Applications* **122**, 1627–1651. MR2914766
[59] Kusuoka, S. and Tudor, C. A. (2013). Extension of the fourth moment theorem to invariant measures of diffusions, preprint arXiv:1310.3785.
[60] Ledoux, M., Nourdin, I. and Peccati, G. (2015). Stein's method, logarithmic Sobolev and transport inequalities, *Geometric and Functional Analysis* **25**, 256–306. MR3320893
[61] Lefèvre, C., Papathanasiou, V. and Utev, S. (2002). Generalized Pearson distributions and related characterization problems, *Annals of the Institute of Statistical Mathematics* **54**, 731–742. MR1954042

[62] Ley, C., Reinert, G. and Swan, Y. (2016). Distances between nested densities and a measure of the impact of the prior in Bayesian statistics, *Annals of Applied Probability*, to appear.
[63] Ley, C. and Swan, Y. (2016). A general parametric Stein characterization, *Statistics & Probability Letters* **111**, 67–71. MR3474784
[64] Ley, C. and Swan, Y. (2013). Local Pinsker inequalities via Stein's discrete density approach, *IEEE Transactions on Information Theory* **59**, 5584–4491. MR3096942
[65] Ley, C. and Swan, Y. (2013). Stein's density approach and information inequalities, *Electronic Communications in Probability* **18**, 1–14. MR3019670
[66] Ley, C. and Swan, Y. (2016). Parametric Stein operators and variance bounds, *Brazilian Journal of Probability and Statistics* **30**, 171–195. MR3481100
[67] Loh, W.-L. (2004). On the characteristic function of Pearson type IV distributions, *A Festschrift for Herman Rubin, Institute of Mathematical Statistics*, 171–179. MR2126896
[68] Luk, H. M. (1994). *Stein's method for the Gamma distribution and related statistical applications*, Ph.D. thesis, University of Southern California. MR2693204
[69] Mackey, L. and Gorham, J. (2016). Multivariate Stein factors for a class of strongly log-concave distributions. Electronic Communications in Probability **21** paper no. 56, 14 pp.
[70] Nourdin, I. and Peccati, G. (2009). Stein's method on Wiener chaos, *Probability Theory and Related Fields* **145**, 75–118. MR2520122
[71] Nourdin, I. and Peccati, G. (2012). *Normal approximations with Malliavin calculus: from Stein's method to universality*, Cambridge Tracts in Mathematics, Cambridge University Press. MR2962301
[72] Nourdin, I., Peccati, G. and Reinert, G. (2009). Second order Poincaré inequalities and CLTs on Wiener space, *Journal of Functional Analysis* **257**, 593–609. MR2527030
[73] Nourdin, I., Peccati, G. and Swan, Y. (2014). Entropy and the fourth moment phenomenon, *Journal of Functional Analysis* **266**, 3170–3207. MR3158721
[74] Nourdin, I., Peccati, G. and Swan, Y. (2014). Integration by parts and representation of information functionals, *IEEE International Symposium on Information Theory (ISIT)*, 2217–2221.
[75] Novak, S. Y. (2011). *Extreme Value Methods with Applications to Finance*, Chapman & Hall/CRC Press, Boca Raton. MR2933280
[76] Ord, J. K. (1967). On a system of discrete distributions, *Biometrika* **54**, 649–656. MR0224192
[77] Papadatos, N. and Papathanasiou, V. (1995). Distance in variation between two arbitrary distributions via the associated w-functions, *Theory of Probability & Its Applications* **40**, 567–575. MR1402000
[78] Papathanasiou, V. (1995). A characterization of the Pearson system of distributions and the associated orthogonal polynomials, *Annals of the Institute of Statistical Mathematics* **47**, 171–176. MR1341214

[79] Peköz, E. and Röllin, A. (2011). New rates for exponential approximation and the theorems of Rényi and Yaglom, *Annals of Probability* **39**, 587–608. MR2789507

[80] Peköz, E. Röllin, A. and Ross, N. (2013). Degree asymptotics with rates for preferential attachment random graphs, *Annals of Applied Probability* **23**, 1188–1218. MR3076682

[81] Pickett, A. (2004). *Rates of convergence of χ^2 approximations via Stein's method*, Ph.D. thesis, Lincoln College, University of Oxford.

[82] Pike, J. and Ren, H. (2014). Stein's method and the Laplace distribution, *ALEA Latin American Journal of Probability and Mathematical Statistics* **11**, 571–587. MR3283586

[83] Rachev, S. T. (1991). *Probability Metrics and the Stability of Stochastic Models*, vol. 334, Wiley New York. MR1105086

[84] Reinert, G. (1998). Couplings for normal approximations with Stein's method, *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, **41**, 193–207. MR1630415

[85] Reinert, G. and Röllin, A. (2009). Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition, *Annals of Probability* **37**, 2150–2173. MR2573554

[86] Röllin, A. (2012). On magic factors and the construction of examples with sharp rates in Stein's method. *Probability Approximations and Beyond, Lecture Notes in Statistics* **205**, Springer. MR3289376

[87] Ross, N. (2011). Fundamentals of Stein's method, *Probability Surveys* **8**, 210–293. MR2861132

[88] Schoutens, W. (2001). Orthogonal polynomials in Stein's method, *Journal of Mathematical Analysis and Applications* **253**, 515–531. MR1808151

[89] Shimizu, R. (1975). On Fisher's amount of information for location family, *A Modern Course on Statistical Distributions in Scientific Work*, Springer, 305–312.

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

[91] Stein, C. (1986). *Approximate computation of expectations*, Institute of Mathematical Statistics Lecture Notes—Monograph Series, 7, Institute of Mathematical Statistics, Hayward, CA. MR0882007

[92] Stein, C., Diaconis, P., Holmes, S. and Reinert, G. (2004). Use of exchangeable pairs in the analysis of simulations, *Stein's method: expository lectures and applications (Persi Diaconis and Susan Holmes, eds.), IMS Lecture Notes Monogr. Ser, vol. 46, Beachwood, Ohio, USA: Institute of Mathematical Statistics*, 1–26. MR2118600

[93] Upadhye, N. S., Cekanavicius, V. and Vellaisamy, P. (2016). On Stein operators for discrete approximations, *Bernoulli*, to appear.