# Stein operators, kernels and discrepancies for multivariate continuous distributions

Guillaume Mijoule, Gesine Reinert, Yvik Swan

# Stein operators, kernels and discrepancies for multivariate continuous distributions

G. Mijoule[*], G. Reinert[†] and Y. Swan[‡]

**Abstract**

We present a general framework for setting up Stein's method for multivariate continuous distributions. The approach gives a collection of Stein characterizations, among which we highlight score-Stein operators and kernel-Stein operators. Applications include copulas and distance between posterior distributions. We give a general explicit construction for Stein kernels for elliptical distributions and discuss Stein kernels in generality, highlighting connections with Fisher information and mass transport. Finally, a goodness-of-fit test based on Stein discrepancies is given.

MSC 2010 classification: 60B10, 60B12

## 1 Introduction

Stein's method is a collection of tools permitting to bound quantities of the form

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

where $\mathcal{H}$ is a measure-determining class and $X, W$ are two random quantities of interest with $X$, say, following the target distribution $\mu$. The method can be summarised as follows. First find a so-called *Stein operator* $\mathcal{A}$ and a wide class of functions $\mathcal{F}(\mathcal{A})$ such that (i) $X \sim \mu$ if and only if $\mathbb{E}[\mathcal{A}f(X)] = 0$ for all functions $f \in \mathcal{F}(\mathcal{A})$ *and* (ii) for each $h \in \mathcal{H}$ there exists a well-defined and tractable solution $f = f_h \in \mathcal{F}(\mathcal{A})$ of the *Stein equation* $h(x) - \mathbb{E}[h(X)] = \mathcal{A}f(x)$. Then, upon noting that $\mathbb{E}[h(W)] - \mathbb{E}[h(X)] = \mathbb{E}[\mathcal{A}f_h(W)]$ for all $h$, the problem of bounding $d_{\mathcal{H}}(X,W)$ has been re-expressed in terms of that of bounding $\sup_{h\in\mathcal{H}} \mathbb{E}[\mathcal{A}f_h(W)]$. The success of the method lies in the fact that this last quantity is amenable to a wide variety of approaches.

There exist many frameworks in which Stein's method is well understood. We refer to the surveys [73, 13, 22] as well as the papers [56, 55] for the univariate setting. We also highlight [6] who provide an up-to-date overview in the context of infinitely divisible distributions. Comprehensive introductions to some of the most important aspects of

[*]INRIA Paris, Team MoKaPlan, 2 rue Simone Iff, 75012 Paris, guillaume.mijoule@inria.fr

[†]University of Oxford, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK, reinert@stats.ox.ac.uk

[‡]Université libre de Bruxelles, Département de Mathématique, Campus Plaine, Boulevard du Triomphe CP210, B-1050 Brussels, yvswan@ulb.ac.be

the theory are available from the monographs [77] as well as [62, 25], with a particular focus on Gaussian approximation. Although the univariate case is the most studied, many references also tackle multivariate distributions. For discrete multivariate distributions, [14] and [15] provide a framework which is applicable in many situations. In [72], stationary distributions of Glauber Markov chains are characterised. The multivariate Gaussian case has naturally received the most attention, starting with [12] and [39], see also [71, 34] and [36]. The Dirichlet distribution has been treated in [37] using a coupling approach. Log-concave densities are treated in [59], and more general settings have been dealt with in [43, 34]. Yet, a general approach for the multivariate case has been elusive. This is the task which our paper addresses.

The starting point for a multivariate Stein's method is a Stein characterization for the Gaussian law which states that a random vector $Y$ is a multivariate Gaussian $d$-random vector with mean $\nu$ and covariance $\Sigma$ (short: $Y \sim \mathcal{N}(\nu, \Sigma)$) if and only if

$$\mathbb{E}\big[(Y-\nu)^t \nabla f(Y)\big] = \mathbb{E}\big[\nabla^t \Sigma \nabla f(Y)\big], \tag{1}$$

for all absolutely continuous function $f : \mathbb{R}^d \to \mathbb{R}$ for which the expectations exist; here $\nabla$ denotes the gradient operator. Assume that $h : \mathbb{R}^d \to \mathbb{R}$ has three bounded derivatives. Then, if $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric and positive definite, and $Z \sim \mathcal{MVN}(0, \Sigma)$, there is a solution $f : \mathbb{R}^d \to \mathbb{R}$ to the Stein equation

$$\nabla^t \Sigma \nabla f(w) - w^t \nabla f(w) = h(w) - \mathbb{E}h(\Sigma^{1/2} Z), \tag{2}$$

given at every $w \in \mathbb{R}^d$ by the *Mehler formula*

$$f(w) = \int_0^1 \frac{1}{2t}\big(\mathbb{E}h(Z_{w,t}) - \mathbb{E}h(\Sigma^{1/2}Z)\big)dt \tag{3}$$

with $Z_{w,t} = \sqrt{t}w + \sqrt{1-t}\,\Sigma^{1/2}Z$. Generalizations to infinitely-dimensional functionals using Malliavin calculus are available in [62] and a generalization to Gaussian random measures is employed in [46].

Pursuing the analogy, classical Markov theory leads, under appropriate regularity assumptions, to Stein equations of the form $\big\langle a(x), \nabla^2 f(x)\big\rangle + \langle b(x), \nabla f(x)\rangle = h(x) - \mathbb{E}[h(Z)]$ for targets $Z \sim \mu$ the ergodic measure of SDE's of the form $\mathrm{d}Z_t = a(t)\mathrm{d}B_t + b(X_t)\mathrm{d}t$; see e.g. [10, 43, 34]. As in the Gaussian case, a solution $f_h$ to the Stein equation is also provided by semigroup approach to Markov generators.

In this paper we shall use first order directional derivatives as basic building blocks of Stein operators. In Definition 3.1 we define the *canonical Stein derivative for $p$ in the direction $e$* as $\mathcal{T}_{e,p}\phi = \partial_e(p\,\phi)/p$ acting on sufficiently smooth test functions $\phi$. Using this building block, we define the *Stein gradient operator* acting on real valued functions $f$, vector valued functions $\mathbf{f}$ or matrix valued functions $\mathbf{F}$ as $\bullet \mapsto \mathcal{T}_{\nabla,p}\bullet = \nabla\,(p\bullet)\; p = \sum_{i=1}^d e_i \mathcal{T}_{e_i,p}$ where $\{e_i, i = 1, \ldots, d\}$ is the canonical basis of $\mathbb{R}^d$. Similarly we define the *Stein divergence operator* acting on vector or matrix valued functions with compatible dimensions as $\bullet \mapsto \mathcal{T}_{\mathrm{div},p}\bullet = \mathrm{div}\,(p\bullet)/p$. As in the one-dimensional case, this operator satisfies a product rule for all smooth functions $f, g$, namely $\mathcal{T}_{e,p}(f\,g) = (\mathcal{T}_{e,p}f)g + f(\partial_e g)$ for all $e$ in the $d$-dimensional unit sphere $S^{d-1}$.

Our framework provides a mechanism for obtaining a broad family of Stein operators which we call *standardizations* of the canonical operator $\mathcal{T}_{\mathcal{D},p}$, where $\mathcal{D} = \nabla$ or div – see Section 3. We do not claim that all useful operators from the literature are of this form

– for instance [16] proposes an alternative construction. We do demonstrate, however, that the operators we construct provide crucial handles on a wide family of multivariable distributions.

For example, setting $\rho = \nabla \log p = \mathcal{T}_{\nabla,p} 1$ the multivariate score function, the analog to the one-dimensional score-Stein operator is the vector valued operator $\mathcal{A}_p g = \nabla g + \rho\, g$ acting on differentiable functions $g : \mathbb{R}^d \to \mathbb{R}$, see Definition 3.6. Many authors call this score-Stein operator *the* Stein operator for multivariate $p$; see e.g. [57, definition 2.1], [26, Section 2.1], [66, Section 2.2.2].

A second set of Stein operators are related to Stein kernels. In dimension $d = 1$, this is the unique bounded solution $x \mapsto \tau_p(x)$ in $\mathcal{F}(p)$ to the ODE $\mathcal{T}_p(\tau_p(x)) = \nu - x$, given by

$$\tau_p(x) = \frac{1}{p(x)} \int_x^\infty (y - \nu) p(y) dy, \quad x \in \mathcal{I} \tag{4}$$

(with $\nu = \mathbb{E}_p[X]$). Properties of the Stein kernel were first studied in [77] (although it had long been a known important handle on smooth densities $p$, see e.g. [18, 19], who refer to it as a covariance kernel). The Stein kernel has become a powerful tool in Stein's method, see e.g. [21, 63, 62] or the more recent works [35, 27]. In one dimension, Stein kernels are unique when they exist - the Stein kernel is the zero bias density from [41]. They have been studied in detail in [32, 74]. In higher dimensions even the definition of a Stein kernel is not obvious; see [64, 65, 27, 52]. The zero-bias coupling definition in [40] uses a collection of random elements in higher dimensions. In analogy, in Definition 3.8 we define a *directional Stein kernel* for each canonical direction $e_i \in \mathbb{R}^d$, as any function $x \mapsto \tau_{p,i}(x) \in \mathbb{R} \times \mathbb{R}^d$ belonging to $\mathcal{F}(p)$ such that $\mathcal{T}_{\mathrm{div},p}\big(\tau_{p,i}(x)\big) = \mathbb{E}_p[X_i] - x_i$ for all $x \in \mathrm{supp}(p)$. A *Stein kernel* is then *any* square matrix $\boldsymbol{\tau}$ such that each line $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{id})$ is a kernel in the direction $e_i$, and the *kernel-Stein operator* is the operator $\mathcal{A}_p g(x) = \boldsymbol{\tau}_p(x) \nabla g(x) - (x - \nu) g(x)$. Stein kernels need not exist and may not be unique when they exist. Aside from discussing this issue in Section 6, we also give an explicit – formula for Stein kernels of elliptical distributions in Section 5. In particular, we obtain new identities even for the standard Gaussian: aside from Stein's classical Gaussian covariance identity

$$X \sim \mathcal{N}(\nu, \Sigma) \Rightarrow \mathbb{E}[\Sigma \nabla g(X)] = \mathbb{E}[(X - \nu) g(X)] \tag{5}$$

we also prove that for all $\beta \neq 2$ and $d \neq 1$, if $X \sim \mathcal{N}(\nu, \Sigma)$ then

$$\frac{\beta}{(\beta - 2)(d - 1)} \mathbb{E}\left[\left(X^T \Sigma^{-1} X \Sigma - X X^T\right) \nabla g(X)\right] = \mathbb{E}\left[\frac{2}{\beta - 2} \Sigma \nabla g(X) + X g(X)\right];$$

both identities hold for a wide family of test functions $g$, see Section 5.2. Also, in the case of the multivariate Student $t$-distribution, we obtain the identity

$$X \sim t_k(\nu, \Sigma) \Rightarrow \mathbb{E}\left[(X X^T + k I_d) \nabla g(X)\right] = (k - 1) \mathbb{E}\left[X g(X)\right]$$

which generalizes the corresponding univariate result (see Section 5.4). We also provide, in Section 6, a connection between Stein kernels and transport maps. These results are closely linked to the seminal works [7, 11, 8]; we show that the key quantity they introduce to solve e.g. Shannon's conjecture actually is a Stein kernel. In particular, these results give new and explicit expressions for Stein kernels in $d = 2$ and $d = 3$ dimensions.

A natural question is which Stein operator to choose. Here practical issues may serve as guidance: solutions to multivariate Stein equations may not generally exist. In the

case of densities which possess a Poincaré constant, weak solutions of second-order Stein equations exist, and some regularity results for these solutions are available, see [70, 71, 23, 79, 43, 34, 36], see Section 3.3.

We illustrate our framework by assessing the (1-)Wasserstein distance between two absolutely continuous distributions. This result leads to three applications: assessing the difference between copulas, assessing the distance between prior and posterior in a Bayesian setting, and bounding the Wasserstein distance between skew-normal and standard normal distributions.

In practical applications and for obtaining information inequalities, the *Stein discrepancy* plays a key role. The Stein discrepancy from distribution $p$ to distribution $q$ using the Stein operator $\mathcal{A}_p$ for $p$ is given in (104); $\mathcal{S}_{\|\cdot\|}(q, \mathcal{A}_p, \mathcal{G}) = \sup_{g\in\mathcal{G}} \|\mathbb{E}\left[\mathcal{A}_p g(Y)\right]\|$. Here $Y \sim q$ and $\mathcal{G}$ is a class of smooth test functions. A particular case is that of kernelized Stein discrepancies, introduced in [57, 26]. We assess the performance of such a kernelized Stein discrepancy measure for simulations from a centred Student $t$-distribution, employing a Stein discrepancy for a goodness-of-fit test.

The paper is structured as follows. Section 2 introduces the notations and the product rule. Section 3.1 gives the general Stein operators, Stein identities, and Stein characterizations. Score-Stein operators and kernel-Stein operators are introduced, and properties of the solution of Stein equations are given for densities which admit a Poincaré constant. Applications of this framework are given in Section 4. Then the paper turns its attention to Stein kernels. In Section 5, Stein kernels for elliptical distributions are derived. Section 6 provides the general discussion on Stein kernels. Finally, Section 7 discusses information metrics, kernelized Stein discrepancies, and ends with a goodness-of-fit test based on Stein discrepancies.

# 2 Notations, gradients and product rules

In this paper we use small caps for elements of $\mathbb{R}^m, m \geq 1$ (by convention column vectors), capitalized letters for matrices in $\mathbb{R}^m \times \mathbb{R}^n$, small caps for real-valued functions, boldfaced small caps for vector valued functions and boldfaced capitalized letters for matrix valued functions.

Fix $d \in \mathbb{N}_0$ and let $e_1, \ldots, e_d$ be the canonical basis for Cartesian coordinates in $\mathbb{R}^d$ . Given $x, y \in \mathbb{R}^d$ and any symmetric positive-definite $d \times d$ matrix $A$ we set $\langle x, y\rangle_A = x^T A y$ (here $\cdot^T$ denotes the transpose) with associated norm $\|x\|_A = \sqrt{\langle x, x\rangle_A}$. With $\mathrm{Tr}(\cdot)$ the trace operator, the *Hilbert-Schmidt scalar product* between matrices $A, B$ of compatible dimension is $\langle A, B\rangle_{\mathrm{HS}} = \mathrm{Tr}(AB^T)$, with associated norm $\|A\|^2_{HS} := \mathrm{Tr}(A^T A) = \sum_{i,j=1}^d a^2_{i,j}$.

Let $S^{d-1}$ denote the unit sphere in $\mathbb{R}^d$ and let $e \in S^{d-1}$ be a unit vector in $\mathbb{R}^d$. The directional derivative of a smooth function $v : \mathbb{R}^d \to \mathbb{R}$ in the direction $e$ is the function $\partial_e v : \mathbb{R}^d \to \mathbb{R}$ given by

$$\partial_e v(x) = \lim_{h\to 0} \frac{v(x + he) - v(x)}{h}, \tag{6}$$

at every point where this limit exists. For $i = 1, \ldots, d$ we write $\partial_i v$ for the derivative in the direction of the unit vector $e_i$. The *gradient* of a smooth function $v : \mathbb{R}^d \to \mathbb{R}$ is

$$\nabla v = \mathrm{grad}(v) = (\partial_1 v, \ldots, \partial_d v)^T = \sum_{i=1}^d (\partial_i v)\, e_i \tag{7}$$

(by convention, a column vector). The *gradient* of a $1 \times m$ vector field $\mathbf{v} = (v_1, \ldots, v_m)$ (a line vector) is

$$\nabla \mathbf{v} = \begin{pmatrix} \nabla v_1 & \cdots & \nabla v_m \end{pmatrix} = \begin{pmatrix} \partial_1 v_1 & \cdots & \partial_1 v_m \\ \vdots & \ddots & \vdots \\ \partial_d v_1 & \cdots & \partial_d v_m \end{pmatrix}. \tag{8}$$

(by convention, a $d \times m$ matrix).

The *Jacobian matrix* of a $\mathbb{R}^m$-valued function $\mathbf{w} = (w_1, \ldots, w_m)^T$ (a column vector) is the $m \times d$ matrix $\mathrm{Jac}(\mathbf{w}) = \left(\partial_i w_j\right)_{1 \le i \le m, 1 \le j \le d}$. As $\mathrm{Jac}(\mathbf{v}^T) = \nabla \mathbf{v}$ we will simply use the generic notation $\nabla$ for both operations from here onwards. The *divergence* operator is defined for $d$ valued (line or column) vector fields $\mathbf{v}$ with components $v_j, j = 1, \ldots, d$ as

$$\mathrm{div}(\mathbf{v}) = \mathrm{Tr}\left(\nabla \mathbf{v}\right) = \sum_{i=1}^{d} \partial_i v_i.$$

More generally, given any $m, n \in \mathbb{N}$, the directional derivative of a matrix valued function

$$\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^m \times \mathbb{R}^d : \mathbf{x} \mapsto \mathbf{F}(\mathbf{x}) = \begin{pmatrix} \mathbf{f}_1(\mathbf{x}) \\ \vdots \\ \mathbf{f}_m(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} f_{11}(\mathbf{x}) & \ldots & f_{1d}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ f_{m1}(\mathbf{x}) & \ldots & f_{md}(\mathbf{x}) \end{pmatrix} \tag{9}$$

is defined componentwise; $(\partial_e \mathbf{F}(x))_{1 \le i \le m, 1 \le j \le n} = (\partial_e f_{ij}(x))_{1 \le i \le m, 1 \le j \le n}$ for all $e \in S^{d-1}$ and all $x \in \mathbb{R}^d$. The gradient of a $m \times r$ matrix valued function $\mathbf{F}$ of the form (9) is defined as the $d \times r \times m$ tensor with entry $(\nabla F)_{i,j,k} = \partial_i f_{jk}$, $1 \le i \le d, 1 \le j \le m, 1 \le k \le r$. The divergence of $\mathbf{F}$ as in (9) is defined as the $m \times 1$ column vector with components $\mathrm{div}(\mathbf{f}_j), j = 1, \ldots, m$; so that

$$\mathrm{div}(\mathbf{F}) = \begin{pmatrix} \mathrm{div}(\mathbf{f}_1) \\ \vdots \\ \mathrm{div}(\mathbf{f}_m) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{d} \frac{\partial f_{1i}}{\partial x_i} \\ \vdots \\ \sum_{i=1}^{d} \frac{\partial f_{mi}}{\partial x_i} \end{pmatrix}.$$

Similarly, the divergence of $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_m)$ with values in $\mathbb{R}^d \times \mathbb{R}^m$ is $\mathrm{div}(\mathbf{F}) = (\mathrm{div}(\mathbf{F}^T))^T$.

## The product rules

Given $v$ and $w$ two sufficiently smooth functions from $\mathbb{R}^d \to \mathbb{R}$ and $e \in S^{d-1}$, the directional derivative satisfies the *product rule*

$$\partial_e \left(v(x)w(x)\right) = (\partial_e v(x))w(x) + v(x)(\partial_e w(x), \quad \text{or short: } \partial_e(vw) = (\partial_e v)w + v\partial_e w, \tag{10}$$

for all $x \in \mathbb{R}^d$ at which all derivatives are defined. Gradients and divergences therefore also satisfy the corresponding product rules. We shall mainly consider three instances.

1. If $\mathbf{v}$ is a $m$-dimensional vector field and $\phi : \mathbb{R}^d \to \mathbb{R}$ then

$$\nabla(\phi \mathbf{v}) = \phi(\nabla \mathbf{v}) + (\nabla \phi)\mathbf{v} \tag{11}$$

(a $d \times m$ matrix if $\mathbf{v}$ is a line, a $m \times d$ matrix if $\mathbf{v}$ is a column). In particular if $m = d$ and $\mathbf{v}$ is a $d \times 1$ column vector, then

$$\operatorname{div}(\phi \mathbf{v}) = \phi \operatorname{div}(\mathbf{v}) + \langle \nabla \phi, \mathbf{v} \rangle \tag{12}$$

(a scalar). When $\mathbf{v} = \nabla \psi$ is a gradient of a sufficiently smooth function $\psi$, then

$$\operatorname{div}(\phi \nabla \psi) = \phi \, \Delta \psi + \langle \nabla \phi, \nabla \psi \rangle \, . \tag{13}$$

2. For $\mathbf{v} \in \mathbb{R}^m$ and $\mathbf{F}$ a $m \times d$ square matrix we have

$$\operatorname{div}(\mathbf{v}^T \mathbf{F}) = \langle \mathbf{v}, \operatorname{div} \mathbf{F} \rangle + \langle \nabla \mathbf{v}, \mathbf{F} \rangle_{\mathrm{HS}} \tag{14}$$

(a scalar).

3. If $\mathbf{v}, \mathbf{w} \in \mathbb{R}^m$ and $A$ is a $m \times m$ positive definite symmetric matrix then (with a slight abuse of notation)

$$\nabla \left( \langle \mathbf{v}, \mathbf{w} \rangle_A \right) = \langle \nabla \mathbf{v}, \mathbf{w} \rangle_A + \langle \mathbf{v}, \nabla \mathbf{w} \rangle_A \, . \tag{15}$$

# 3 Stein operators and their standardizations

In this paper, all random vectors are assumed to admit a probability density function (pdf) $p$ with respect to the Lebesgue measure. The support of $p$, denoted by $K_p$ , is the complement of the largest open set $\mathcal{O}$ such that $p \equiv 0$ on $\mathcal{O}$. We further make the following assumption:

**Assumption A**: There exists an open set $\Omega_p$ such that the Lebesgue measure of $K_p \backslash \Omega_p$ is zero, $p > 0$ on $\Omega_p$ and $p$ is $\mathcal{C}^1$ on $\Omega_p$.

If $p$ satisfies this assumption, then $p$ is $\mathcal{C}^1$ on $\Omega_p \cup K_p^C$, which by Assumption A is a set of full Lebesgue measure. For ease of notation the Lebesgue measure is left out in integrals, so that $\int_{\mathbb{R}^d} f(x) dx$ is written as $\int_{\mathbb{R}^d} f$. Similarly we abbreviate $\mathbb{E}_p f = \int_{K_p} fp = \int_{K_p} f(x) p(x) \, dx$.

## 3.1 Definition and some general comments

**Definition 3.1** (Multivariate Stein class and operator)**.** *Let $X$ be a $d$-dimensional random vector with pdf $p : \mathbb{R}^d \to \mathbb{R}$ satisfying Assumption A.*

1. *The* canonical directional Stein class for $p$ *in direction $e$ is the vector field $\mathcal{F}_{1,e}(p)$ of all functions $\phi : \mathbb{R}^d \to \mathbb{R}$ such that the restriction of $\phi$ to $\Omega_p$ is $\mathcal{C}^1$, $p \, \phi$ has integrable gradient and satisfies $\int_{\mathbb{R}^d} \partial_e (p \, \phi) = 0$.*
2. *The* canonical Stein derivative for $p$ in the direction $e$ *is the operator $\mathcal{T}_{e,p}$ on $\mathcal{F}_{1,e}(p)$ defined by $\mathcal{T}_{e,p} \phi = \frac{\partial_e (p \, \phi)}{p}$.*

*We set $\mathcal{F}_1(p)$ the collection of all scalar valued functions $f$ which belong to $\mathcal{F}_{1,e_i}(p)$ for all unit vectors $e_i, i = 1, \ldots, d$, and define the Stein class of $p$, denoted $\mathcal{F}(p)$, as the collection of all scalar, vector and matrix valued functions whose components belong to $\mathcal{F}_1(p)$.*

Note in particular that if $\phi \in \mathcal{F}_1(p)$, then $\int_{\mathbb{R}^d} \partial_e(p\,\phi) = 0$ for all unit vectors $e$. Furthermore, any function $\phi \in \mathcal{C}^1(\mathbb{R}^d)$ with compact support $K \subset \Omega_p \cup K_p^C$ lies in $\mathcal{F}_{1,e}(p)$. Indeed, if $e = (a_1, \ldots, a_d)$, since $p\phi \in \mathcal{C}^1(\mathbb{R}^d)$, we have $\partial_e(p\phi) = \sum_{i=1}^d a_i \partial_i(p\phi)$, then

$$\int_{\mathbb{R}^d} \partial_i(p\phi) = \int_{\mathbb{R}^{d-1}} \left( \int_{\mathbb{R}} \partial_i(p\phi) dx_i \right) \prod_{j \neq i} dx_j,$$

and the inside integral is zero since $p(x)\phi(x)$ cancels out for $x$ away enough from 0.

On $\mathcal{F}(p)$ we define the *Stein gradient operator* acting on real valued functions $f$, vector valued functions $\mathbf{f}$ or matrix valued functions $\mathbf{F}$, as

$$\bullet \mapsto \mathcal{T}_{\nabla,p}\bullet = \frac{\nabla\,(p\bullet)}{p}$$

(expressed in terms of $\mathcal{T}_{e,p}$'s as $\mathcal{T}_{\nabla,p} = \sum_{i=1}^d e_i \mathcal{T}_{e_i,p}$); we also define the *Stein divergence operator* acting on vector or matrix valued functions with compatible dimensions as

$$\bullet \mapsto \mathcal{T}_{\mathrm{div},p}\bullet = \frac{\mathrm{div}\,(p\bullet)}{p}$$

(this operator also can be expressed in terms of $\mathcal{T}_{e,p}$'s). Although in the literature, it is not $\mathcal{T}_{e,p}$ but $\mathcal{T}_{\nabla,p}$ and $\mathcal{T}_{\mathrm{div},p}$ that are the operators used, their properties follow from those of $(\mathcal{T}_{e,p}), e \in \mathcal{B}$ with $\mathcal{B}$ a unit basis of $\mathbb{R}^d$.

**Remark 3.2.** *Let $\mathcal{F}^{(0)}(p)$ denote the collection of functions (in any dimension) with mean 0 under $p$. The definition of $\mathcal{F}(p)$ is tailored to ensure $\mathcal{T}_{\mathcal{D},p}(\mathbf{F}) \in \mathcal{F}^{(0)}(p)$ for all $\mathbf{F} \in \mathcal{F}(p)$ (here and below $\mathcal{D}$ denotes $\partial_e, \nabla$ or* div*). Hence, the Stein operators are also a machinery for producing mean 0 functions under $p$. This can be useful in applications, particularly when $p$ is intractable, since by construction the operators are invariant under scalar multiplication, and therefore do not depend on normalizing constants.*

### Stein identities

Product rule (10) directly leads to Stein-type product rules of the type

$$\mathcal{T}_{e,p}(f\,g) = (\mathcal{T}_{e,p}f)g + f(\partial_e g) \tag{16}$$

on a suitable set of functions. The next definition introduces such sets.

**Definition 3.3** (Stein adjoint class)**.** *To every (scalar/vector/matrix valued function) $\mathbf{F} \in \mathcal{F}(p)$ we associate* $\mathrm{dom}(p, \mathbf{F})$ *the collection of all matrix-valued functions $\mathbf{G}$ with compatible dimensions such that $\mathbf{G}$ is $\mathcal{C}^1$ on $\Omega_p$, $\mathbf{FG} \in \mathcal{F}(p)$ and $\mathbf{F}(\partial_e \mathbf{G}) \in L^1(p)$ for all $e \in S^{d-1}$.*

Using the product rules from Section 2, we deduce similar identities for Stein gradients and divergences. Here we give two instances.

1. For all scalar functions $f \in \mathcal{F}_1(p)$ and $g \in \mathrm{dom}(p, f)$, from (11)

$$\mathcal{T}_{\nabla,p}\,(fg) = (\mathcal{T}_{\nabla,p}f)g + f\nabla g. \tag{17}$$

2. For all $\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^m \times \mathbb{R}^d \in \mathcal{F}_1(p)$ and $\mathbf{g} : \mathbb{R}^d \to \mathbb{R} \times \mathbb{R}^m \in \mathrm{dom}(p, \mathbf{F})$, using (14),

$$\mathcal{T}_{\mathrm{div},p}(\mathbf{gF}) = \langle \mathcal{T}_{\mathrm{div},p}\mathbf{F}, \mathbf{g}\rangle + \langle \mathbf{F}, \nabla \mathbf{g}\rangle_{\mathrm{HS}}\,. \tag{18}$$

We deduce a family of Stein operators for $p$ by fixing $\mathbf{F} = (F_{ij})_{1\le i\le m, 1\le j\le d}$ and considering $\mathcal{A}_p := \mathcal{A}_{\mathbf{F},p}$ defined by

$$\mathbf{g} \mapsto \mathcal{A}_p\mathbf{g} = \langle \mathcal{T}_{\mathrm{div},p}\mathbf{F}, \mathbf{g}\rangle + \langle \mathbf{F}, \nabla \mathbf{g}\rangle_{\mathrm{HS}} \tag{19}$$

with domain $\mathcal{F}(\mathcal{A}_p) := \mathrm{dom}(p, \mathbf{F})$. Note that $\mathcal{A}_p(\mathbf{g})$ is a scalar function. We also deduce the Stein identities (one for each $\mathbf{F}$)

$$\mathbb{E}_p\left[\langle \mathcal{T}_{\mathrm{div},p}\mathbf{F}, \mathbf{g}\rangle\right] = -\mathbb{E}_p\left[\langle \mathbf{F}, \nabla \mathbf{g}\rangle_{\mathrm{HS}}\right] \text{ for all } \mathbf{g} \in \mathcal{F}(\mathcal{A}_p). \tag{20}$$

More product rules are provided later, see e.g. (25), (29).

It follows from the definition of the classes of functions $\mathrm{dom}(p, \mathbf{F})$ and the Stein operators and the Stein class that the left hand side in each of (16), (17), and their variations such as (18), (19), (23), (25), integrates to 0. These probabilistic integration by parts formulas lead to identities known as *Stein (covariance) identities*; they are inspired by Stein's original Gaussian identity (1).

**Proposition 3.4** (Stein identities)**.** *For all $f$: $\mathbb{R}^d \to \mathbb{R} \in \mathcal{F}_1(p)$ and all $e \in S^{d-1}$ we have*

$$\mathbb{E}_p\left[f(\mathcal{T}_{e,p}g)\right] = -\mathbb{E}_p\left[g(\partial_e f)\right] \tag{21}$$

*for all $g$: $\mathbb{R}^d \to \mathbb{R} \in \mathrm{dom}(p, f)$*

*Proof.* Let $f \in \mathcal{F}_1(p)$ and $g \in \mathrm{dom}(p, f)$ and $e \in S^{d-1}$. Then, from (16), $f(\mathcal{T}_{e,p}g) = \mathcal{T}_{e,p}(f\,g) - g\partial_e f$. Since $\mathbb{E}_p\mathcal{T}_{e,p}(f\,g) = 0$ for all $f \in \mathcal{F}_1(p)$ and $g \in \mathrm{dom}(p, f)$, identity (21) ensues. □

Identity (21) shows that the Stein operators are *the* skew adjoint operators to the classical directional derivatives, with respect to integration in $p$. In this sense the operator from Definition 3.1 is "canonical".

The Stein identities (21) for $(e_1, \ldots, e_d)$ the standard unit basis in $\mathbb{R}^d$ yield corresponding Stein identities for the gradient Stein operator $\mathcal{T}_{\nabla,p}$ and for the divergence Stein operator $\mathcal{T}_{\mathrm{div},p}$. Here we give three instances.

1. For all $f \in \mathcal{F}_1(p)$ and $g \in \mathrm{dom}(p, f)$,

$$\mathbb{E}_p\left[(\mathcal{T}_{\nabla,p}f)g\right] = -\mathbb{E}_p\left[f(\nabla g)\right]\text{‘} \tag{22}$$

2. For all $f \in \mathcal{F}(p)$ and $g$ such that $\nabla g \in \mathrm{dom}(p, f)$, from (13) we obtain

$$\mathcal{T}_{\mathrm{div},p}(f\nabla g) = f\Delta g + \langle \mathcal{T}_{\nabla,p}f, \nabla g\rangle\,. \tag{23}$$

Consequently,

$$\mathbb{E}_p\left[f\Delta g\right] = -\mathbb{E}_p\left[\langle \mathcal{T}_{\nabla,p}f, \nabla g\rangle\right] \tag{24}$$

for all such $g$.

3. For all $m \times d$ square matrices $\mathbf{F} \in \mathcal{F}(p)$ and $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^m \in \mathrm{dom}(p, \mathbf{F})$, using (14) gives

$$\mathcal{T}_{\mathrm{div},p}(\mathbf{g}^T\mathbf{F}) = \langle \mathbf{g}, \mathcal{T}_{\mathrm{div},p}(\mathbf{F})\rangle + \langle \nabla \mathbf{g}, \mathbf{F}\rangle_{\mathrm{HS}}\,. \tag{25}$$

Consequently,

$$\mathbb{E}_p\left[\langle \mathbf{g}, \mathcal{T}_{\mathrm{div},p}(\mathbf{F})\rangle\right] = -\mathbb{E}_p\left[\langle \nabla \mathbf{g}, \mathbf{F}\rangle_{\mathrm{HS}}\right]. \tag{26}$$

### Stein characterizations

Recall that $\mathbb{E}_p\left[\mathcal{T}_{e,p}\mathbf{F}\right] = 0$ for every $\mathbf{F} \in \mathcal{F}(p)$. Under well chosen assumptions, the collection of operators $\mathcal{T}_{e_i,p}$ for $(e_1, \ldots, e_d)$ the standard unit basis of $\mathbb{R}^d$ characterises $p$ in the sense that if $\mathbb{E}_q\left[\mathcal{T}_{e_i,p}\mathbf{F}\right] = 0, i = 1, \ldots, d$ for a sufficiently wide class of $\mathbf{F}$, then necessarily $q = p$.

**Theorem 3.5** (Stein characterizations)**.** *Let $X \sim p$ and $Y \sim q$; assume that $p$ and $q$ both satisfy Assumption A. Assume moreover that $\Omega_p = \Omega_q$, and that this set is connected. Let $f : \mathbb{R}^d \to \mathbb{R} \in \mathcal{F}_1(p)$, and assume $f > 0$ on $\Omega_p$.*

1. *It holds that $Y \overset{\mathcal{L}}{=} X$ if and only if for all $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d \in \mathrm{dom}(p, f)$*

$$\mathbb{E}_q\left[(\mathcal{T}_{\mathrm{div},p}\mathbf{g})f\right] = -\mathbb{E}_q\left[\langle \mathbf{g}, \nabla f\rangle\right]. \tag{27}$$

2. *It holds that $Y \overset{\mathcal{L}}{=} X$ if and only if for all $g : \mathbb{R}^d \to \mathbb{R} \in \mathrm{dom}(p, f)$*

$$\mathbb{E}_q\left[(\mathcal{T}_{\nabla,p}g)f\right] = -\mathbb{E}_q\left[g(\nabla f)\right]. \tag{28}$$

*Proof.* First let $f : \mathbb{R}^d \to \mathbb{R} \in \mathcal{F}_1(p)$ and $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d \in \mathrm{dom}(p, f)$. By the product rule (12), $f\mathrm{div}(p\mathbf{g}) = \mathrm{div}(fp\mathbf{g}) - p\langle \nabla f, \mathbf{g}\rangle$ and hence $f\mathcal{T}_{\mathrm{div},p}\mathbf{g} = \mathcal{T}_{\mathrm{div},p}(f\mathbf{g}) - \langle \nabla f, \mathbf{g}\rangle$. Taking expectations gives

$$\mathbb{E}_p[f\mathcal{T}_{\mathrm{div},p}\mathbf{g}] = \mathbb{E}_p[\mathcal{T}_{\mathrm{div},p}(f\mathbf{g})] - \mathbb{E}_p[\langle \nabla f, \mathbf{g}\rangle].$$

By construction of $\mathrm{dom}(p, f)$, $\mathbb{E}_p[\mathcal{T}_{\mathrm{div},p}(f\mathbf{g})] = 0$ and the only-if direction follows.

For the converse direction, let $\mathbf{g}$ be an infinitely differentiable vector field with compact support $K \subset \Omega_p$; recall that in this case $g \in \mathrm{dom}(p, f)$. Suppose that (27) holds. Rewriting the expectations,

$$\int_K \mathrm{div}(p\mathbf{g}) f \frac{q}{p} = -\int_K q\langle \nabla f, \mathbf{g}\rangle.$$

Now, by the divergence (or Ostrogradski) theorem, and since $\mathbf{g} \equiv 0$ on $\partial K$,

$$\int_K \mathrm{div}(p\mathbf{g}) f \frac{q}{p} = -\int_K p \left\langle \mathbf{g}, \nabla\left(f\frac{q}{p}\right)\right\rangle.$$

Thus $\int_K p \left\langle \mathbf{g}, \nabla\left(f\frac{q}{p}\right)\right\rangle = \int_K \langle \mathbf{g}, \nabla g\rangle q$. Hence $p\nabla\left(f\frac{q}{p}\right) = q\nabla g$ in the weak sense on $\Omega_p$; but it is also true in the strong sense, because all functions are continuous (recall that $\Omega_p = \Omega_q$). We deduce that $f\nabla(q/p) = 0, \quad$ on $\Omega_p$, but since $f > 0$ on this connected domain, $q/p$ is constant on $\Omega_p$. Finally, using $\int_{\Omega_p} p = \int_{\Omega_p} q$ gives the first result.

For Item 2, proceeding as for the first part, we arrive at the differential equation $p\nabla\left(f\frac{q}{p}\right) = q\nabla f$ from which we deduce that $p/q = 1$. □

The proof of the claims in Theorem 3.5 are greatly facilitated by the tailored assumptions which hide some difficulties. It still remains to write such characterizations out in detail for specific targets; this can turn out to be quite complicated and will not be the focus of the paper. Also, there are many more general ways to formulate similar characterizations as in Theorem 3.5 be it by changing the starting operator, by relaxing the assumptions on the test functions or indeed by relaxing the assumptions on $Y$.

## 3.2 Standardizations

Similarly as in the univariate case (see [56, Section 4]), we introduce a broad family of Stein operators $\mathcal{A}_p : \mathcal{F}(\mathcal{A}_p) \to \mathcal{F}^{(0)}(p)$ which we call *standardizations* of the canonical operator $\mathcal{T}_{\mathcal{D},p}$, where $\mathcal{D} = \nabla$ or div. These are operators such that there exists a transformation $\mathbf{T} : \mathcal{F}(\mathcal{A}_p) \to \mathcal{F}(p) : \mathbf{u} \mapsto \mathbf{T}(\mathbf{u})$ such that $\mathcal{A}_p(\mathbf{u}) = \mathcal{T}_p(\mathbf{T}(\mathbf{u}))$ for all $\mathbf{u} \in \mathcal{F}(\mathcal{A}_p)$. Such standardizations were studied in [56]; [67, Appendix A.2] suggests some standardizations for densities on a manifold. Here we concentrate on standardizations which are connected to score functions and Stein kernels.

### 3.2.1 Gradient based operators and the score function

In this section we focus on the Stein identity (22), which arises from the product rule (17) extended as

$$\mathcal{T}_{\nabla,p}(f\mathbf{g}) = (\mathcal{T}_{\nabla,p}f)\mathbf{g} + f\nabla\mathbf{g} \tag{29}$$

for $f \in \mathcal{F}_1(p)$ and $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^m \in \mathrm{dom}(p, f)$. From this, we deduce a family of Stein operators for $p$ obtained by fixing some $f \in \mathcal{F}(p)$ and considering the operator $\mathcal{A}_p := \mathcal{A}_{f,p}$ given in (29) acting through $\mathbf{g} \mapsto \mathcal{A}_p\mathbf{g} := \mathcal{T}_{\nabla,p}(f\mathbf{g})$ on $\mathbf{g} : \mathbb{R}^d \to \times\mathbb{R}^m \in \mathcal{F}(\mathcal{A}_p) = \mathrm{dom}(p, f)$. Note that $\mathcal{A}_p(\mathbf{g})$ is a $d \times m$ matrix, and $\mathbb{E}_p[\mathcal{A}_p\mathbf{g}] = \mathbf{0}$ for all $\mathbf{g} \in \mathrm{dom}(p, f)$. Each particular $f \in \mathcal{F}_1(p)$ thus gives rise to a Stein identity

$$\mathbb{E}_p\left[(\mathcal{T}_{\nabla,p}f)\mathbf{g}\right] = -\mathbb{E}_p\left[f\nabla\mathbf{g}\right] \text{ for all } \mathbf{g} \in \mathcal{F}(\mathcal{A}_p). \tag{30}$$

One particular choice for $f$ stands out: $f = 1$. The following definition is classical.

**Definition 3.6** (Score function)**.** *Let $p$ be differentiable. The* score function *of $p$ is the function defined on $\Omega_p$ by*

$$\rho_p(x) = \nabla \log p(x) = \frac{1}{p(x)}\nabla p(x). \tag{31}$$

*The* score-Stein operator *is the vector valued operator*

$$\mathcal{A}_p g = \nabla g + \rho_p\, g \tag{32}$$

*acting on differentiable functions $g : \mathbb{R}^d \to \mathbb{R}$.*

**Remark 3.7.** *One can easily extend Definition 3.6 to introduce a score (and score-Stein operator) in any direction $e \in S^{d-1}$, by considering the gradient $\nabla = (\partial_e, \partial_{e^\perp})$ along $e$.*

Equation (32) is often called *the* Stein operator for multivariate $p$; see e.g. [57, definition 2.1], [26, Section 2.1], [66, Section 2.2.2]. This operator is particularly interesting when constant functions $1 \in \mathcal{F}_1(p)$, an assumption which holds if $p$ is a differentiable density such that $\partial_i p$ is integrable for all $i = 1, \ldots, d$ and $\int_{\mathbb{R}^d} \partial_i p = 0$. It is easy to see that $\mathcal{C}_0^\infty(\mathbb{R}^d) \subset \mathrm{dom}(p, 1)$. The resulting (characterizing) Stein identity is

$$\mathbb{E}_p\left[\rho_p\, g\right] = -\mathbb{E}_p\left[\nabla g\right] \text{ for all } g \in \mathcal{F}(\mathcal{A}_p). \tag{33}$$

Equation (33) is reminiscent of the classical Stein identity (1); one can easily check that $\rho_\phi(x) = \nabla \log \phi(x) = -\Sigma^{-1}(x - \nu)$ when $X \sim \phi$ the multivariate normal density with mean $\nu$ and covariance $\Sigma$. We shall see that for the Student-t distribution the choice of a non-constant $f$ in (30) leads to more tractable operators and identities. We will discuss operator (32) and Stein identity (33) – and variations thereon – in the more general context of score functions and Fisher information in Section 7.2.

### 3.2.2 Divergence based first order operators and Stein kernels

In this section, we focus on the product rule (12) rewritten as

$$\mathcal{T}_{\mathrm{div},p}(\mathbf{F}g) = (\mathcal{T}_{\mathrm{div},p}\mathbf{F})g + \mathbf{F}\nabla g \tag{34}$$

for all properly chosen $\mathbf{F} : \mathbb{R}^d \to \mathbb{R}^m \times \mathbb{R}^d$ and all $g : \mathbb{R}^d \to \mathbb{R}$. By construction, $\mathcal{T}_{\mathrm{div},p}(\mathbf{F}g) \in \mathbb{R}^m$ for all $\mathbf{F}, g$. From this we deduce a family of Stein operators for $p$ obtained by fixing $\mathbf{F}$ in $\mathcal{F}(p)$ and considering $\mathcal{A}_p := \mathcal{A}_{\mathbf{F},p}$ defined by $g \mapsto \mathcal{A}_p g = (\mathcal{T}_{\mathrm{div},p}\mathbf{F})g + \mathbf{F}\nabla g$ with domain $\mathcal{F}(\mathcal{A}_p) := \mathrm{dom}(p, \mathbf{F})$. We also deduce the Stein identities (one for each $\mathbf{F}$)

$$\mathbb{E}_p\left[(\mathcal{T}_{\mathrm{div},p}\mathbf{F})g\right] = -\mathbb{E}_p\left[\mathbf{F}\nabla g\right] \text{ for all } g \in \mathcal{F}(\mathcal{A}_p). \tag{35}$$

One particularly important choice of $\mathbf{F}$ in (35) is the *Stein kernel* first defined in dimension $d = 1$ in [77]. Here we consider general dimension $d \geq 1$ and propose the following definition.

**Definition 3.8** (Stein kernel)**.** *Consider a density $p : \mathbb{R}^d \to \mathbb{R}^+$ satisfying Assumption A, and suppose $p$ has finite mean $\nu \in \mathbb{R}^d$. For each canonical direction $e_i \in \mathbb{R}^d$, $i = 1, \ldots, d$,* a Stein kernel *for $p$ in the direction $e_i$ is any vector field $x \mapsto \tau_{p,i}(x) \in \mathbb{R}^d$ belonging to $\mathcal{F}(p)$ such that*

$$\mathcal{T}_{\mathrm{div},p}\big(\tau_{p,i}(x)\big) = \nu_i - x_i \tag{36}$$

*for all $x \in \Omega_p$. A* Stein kernel *is any square matrix $\boldsymbol{\tau}_p = (\tau_{i,j})_{1 \leq i,j \leq d}$ such that each line $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{id})$ is a kernel in the direction $e_i$. The* $\boldsymbol{\tau}_p$-kernel-Stein operator *is the $\mathbb{R}^d$-valued operator*

$$\mathcal{A}_p g(x) = \boldsymbol{\tau}_p(x)\nabla g(x) - (x - \nu)g(x) \tag{37}$$

*with domain $\mathcal{F}(\mathcal{A}_p) = \mathrm{dom}(p, \boldsymbol{\tau}_p)$.*

When the kernel is clear, we leave out the $\boldsymbol{\tau}_p$-prefix in the kernel-Stein operator. To illustrate the concept we give two examples; Section 6 contains more constructions for some cases.

**Example 3.9.** *A Stein kernel in the sense of Definition 3.8 is immediately obtained in the following way. For each $i = 1, \ldots, d$ let $\tau_{p,i} = (0, \ldots, \tau_i, \ldots, 0)$ with*

$$\tau_i(x) = \frac{1}{p(x)} \int_{x_i}^{\infty} (\nu_i - u_i) p(x_{\setminus i}, u_i) du_i$$

*where $(x_{\setminus i}, u_i) = (x_1, \ldots, x_{i-1}, u_i, x_{i+1}, \ldots, x_d)$. The diagonal matrix $\boldsymbol{\tau}$ obtained by stacking up the above vectors is a Stein kernel. If $X$ has finite first moment, this Stein kernel is defined for almost every $x$ by Fubini's theorem. Such diagonal kernels are often not interesting (mainly because the Stein identity* (40) *below will not hold for a large enough class of functions).*

**Example 3.10.** *The score function and the divergence operator are linked through*

$$\mathcal{T}_{\mathrm{div,p}}\mathbf{F} = \mathbf{F}\rho_p + \mathrm{div}(\mathbf{F}),$$

*where $\mathbf{F} \in \mathcal{F}(p)$. Hence for a density $p$ with mean $\nu$, a Stein kernel is any matrix $\boldsymbol{\tau}$ satisfying*

$$\boldsymbol{\tau}(x)\rho_p(x) + \mathrm{div}(\boldsymbol{\tau}(x)) = -x$$

*for all $x \in \Omega_p$. A general construction of Stein kernels is thus given by finding an $\mathbf{F} \in \mathcal{F}(p)$ such that for some constants $\alpha, \beta \in \mathbb{R}$, $\alpha + \beta \neq 0$,*

$$\mathbf{F}(x)\rho_p(x) = \alpha(x-\nu) + r(x) \tag{38}$$
$$\operatorname{div}(\mathbf{F}(x)) = \beta(x-\nu) - r(x) \tag{39}$$

*($r(x)$ is* any *function) and setting*

$$\boldsymbol{\tau} = \frac{1}{\alpha+\beta}\mathbf{F}.$$

**Remark 3.11.** *1. With Definition 3.8, existence of the Stein kernel is readily checked, and the resulting Stein identity is*

$$\mathbb{E}\left[\boldsymbol{\tau}(X)\nabla g(X)\right] = \mathbb{E}\left[(X-\nu)g(X)\right] \text{ for all } g \in \mathcal{F}(\mathcal{A}_p) = \operatorname{dom}(p, \boldsymbol{\tau}). \tag{40}$$

*Equation* (40) *is reminiscent of the classical Stein identity* (1)*; one can easily check that $\boldsymbol{\tau}_p(x) = \Sigma$ is a Stein kernel for the multivariate standard Gaussian distribution.*

2. *The question of whether a density $p$ admits a Stein kernel is non trivial, if one wants the Stein identity to hold for a larger class of functions than the regular ones with compact support in $\Omega_p$ (see Section 6). Definition* (3.8) *agrees with the one of [27], except for the class of test functions we impose.*

3. *The Stein kernel $\tau_{p,i} \in \mathcal{F}(p)$ in the direction $e_i$ in Definition* (3.8) *could be defined equivalently by requiring*

$$\mathbb{E}\left[\langle \tau_{p,i}(X), \nabla g(X)\rangle\right] = \mathbb{E}\left[(X_i - \nu_i)g(X)\right], \tag{41}$$

*for all $g \in \mathcal{C}_c^\infty(\Omega_p)$, the set of functions infinitely differentiable with compact support in $\Omega_p$. Thus, showing some matrix is a Stein kernel can be done equivalently by checking pointwisely* (36)*, or by checking the previous identity holds at least for infinitely differentiable functions with compact support included in $\Omega_p$ (and that it is in $\mathcal{F}(p)$). However this class of functions is narrow, and in most applications, one would want* (41) *to hold for more test functions $g$ – e.g. for bounded $g$. In this work, we use the above definition to find Stein kernels, and we extend the class of functions for which the identity holds on a case-by-case basis. This will be done, for instance, for the Stein kernels we build for elliptical distributions. From the divergence theorem, when the boundary of $\Omega_p$ is smooth enough, any $g \in \mathcal{C}^1(\Omega_p)$ such that $\tau g p \to 0$ on the boundary of $\Omega_p$ will verify the identity* (41)*.*

### 3.2.3 Divergence based second order operators

The starting point here is the divergence product rule (14) and the corresponding Stein operator equation (25), extended as follows. Choose $\mathbf{A}, \mathbf{B}$ two $d \times d$ matrix valued functions in $\mathcal{F}(p)$. Then

$$\mathcal{T}_{\operatorname{div},p}\left((\nabla g)^T \mathbf{A}\mathbf{B}^T\right) = \langle \nabla g, \mathcal{T}_{\operatorname{div},\mathrm{p}}(\mathbf{B})\rangle_{\mathbf{A}} + \left\langle \nabla\left((\nabla g)^T\mathbf{A}\right), \mathbf{B}\right\rangle_{\mathrm{HS}} \tag{42}$$

for all $g \in \mathcal{C}^1(\Omega_p)$. This leads to a family of second order scalar valued operators obtained by fixing $\mathbf{A}$ and/or $\mathbf{B}$ and considering $\mathcal{A}_p g := \mathcal{A}_{\mathbf{A},\mathbf{B},p} g$ defined in (42) with domain the collection of $g$ such that $(\nabla g^T \mathbf{A} \in \operatorname{dom}(p, \mathbf{B})$. We single out three particular choices for $\mathbf{A}$, and $\mathbf{B}$.

1. $\mathbf{A} = \mathbf{B} = I_d$ yields
$$\mathcal{A}_p g = \langle \nabla \log p, \nabla g \rangle + \Delta g, \tag{43}$$
($\Delta$ being the Laplacian on $\mathbb{R}^d$) with domain $\mathrm{dom}(\mathcal{A}_p)$;
2. $\mathbf{A} = I_d$ and $\mathbf{B} = \boldsymbol{\tau}_p^T$ the (transpose of a) Stein kernel of $X$ yields
$$\mathcal{B}_p g = \langle \nabla g, \nu - \bullet \rangle + \left\langle \nabla^2 g, \boldsymbol{\tau}_p \right\rangle_{\mathrm{HS}} \tag{44}$$
(where $\nabla^2(f) = \nabla(\nabla f^T)$ is the Hessian of $f$ and $\nu$ is the mean of $X$) with domain $\mathrm{dom}(\mathcal{B}_p)$.
3. $\mathbf{A}$ symmetric definite positive and $\mathbf{B}$ such that $\mathcal{T}_{\mathrm{div},p}(\mathbf{B}) = \mathbf{b}$ yields
$$\mathcal{C}_p g = \langle \nabla g, \mathbf{b} \rangle_{\mathbf{A}} + \left\langle \nabla^2_{\mathbf{A}} g, \mathbf{B} \right\rangle_{\mathrm{HS}} \tag{45}$$
($\nabla^2_{\mathbf{A}} g = \nabla\left((\nabla g)^T \mathbf{A}\right)$) with domain $\mathrm{dom}(\mathcal{C}_p)$.

One recognizes in operators such as (44) and (45) the infinitesimal generators of multivariate diffusions, see e.g. [43].

## 3.3 Stein equations and Stein factors

Let $X \sim p$ with Stein canonical class and operator $(\mathcal{F}(p), \mathcal{T}_{\mathcal{D},p})$, where $\mathcal{D}$ denotes $\partial_e, \nabla$, or div, and consider some standardization of the form
$$\mathcal{A}_p : \mathcal{F}(\mathcal{A}_p) \to \mathrm{im}(\mathcal{A}_p) : g \mapsto \mathcal{A}_p g \tag{46}$$
as detailed in Subsection 3.2.

**Definition 3.12** (Stein's equations and (magic) factors)**.** *Instate all previous notations. Let $\mathcal{H} \subset L^1(p)$ be a family of test functions, and suppose that $h - \mathbb{E}_p h \in \mathrm{im}(\mathcal{A}_p)$. The* $(\mathcal{A}_p - \mathcal{H})$ Stein equation *for $X$ is the family of differential equations*
$$\mathcal{A}_p g = h - \mathbb{E}_p h; \quad h \in \mathcal{H}. \tag{47}$$
*For a given $h$, a solution to* (47) *is an absolutely continuous function $g_h$ such that there exists a version of the derivatives for which* (47) *is satisfied at all $x \in \mathbb{R}^d$. An* $(\mathcal{A}_p - \mathcal{H})$ Stein factor *for $X$ is* any *uniform bound on some moment of (derivatives of) $g_h = \mathcal{A}_p^{-1}(h - \mathbb{E}_p h)$ over all $h \in \mathcal{H}$ of solutions to Stein equations* (47)*.*

Solving, and bounding the solution of, the Stein equation (47) is well studied in the univariate case under quite general conditions on $p$, see for example [77, 25, 62, 28, 29] and [31]. Matters are much more complicated in the multivariate setting. When $p$ is a multivariate Gaussian, then [12] identified a solution of the Stein equation (2) to be given by the Mehler formula (3). Such explicit dependence of solution $f_h$ on the function $h$ permits to study regularity properties of $f_h$ in terms of those of $h$, see for example [71, 70, 23].

For solutions of the Stein equation here we focus on the general score equation based on (43), namely
$$\Delta u + \langle \nabla \log p, \nabla u \rangle = h - \mathbb{E}_p h. \tag{48}$$

Bounds on the solution of this Stein equation are available (a) when $p$ is strongly log-concave, or (b) when $p$ admits a Poincaré constant. For the first condition, recall that a smooth density $p$ is $k$-strongly log-concave for some $k > 0$ if

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d,\ \langle y - x, \nabla \log p(y) - \nabla \log p(x) \rangle\ < -k\|y - x\|^2.$$

[59] proved that if $p$ is $k$-strongly log-concave and $h$ is 1-Lipschitz, then (48) admits a solution $f = f_h$ such that for $p$-almost every $x$,

$$|\nabla f_h(x)| \leqslant \frac{1}{k}. \tag{49}$$

Similar bounds based on Malliavin calculus are obtained in [34]. [43] computed uniform Stein factors for diffusions satisfying "distant dissipativity" and, more generally, for diffusions that couple rapidly. This class admits not only strong log-concave densities but also a large class of non-log concave, multimodal, and heavy-tailed distributions. We also point out that the authors manage to obtain bounds on higher-order derivatives of $f_h$, under suitable assumptions. Related results can be found in [34].

For the second set of results, if $X \sim p$, recall that $C_p$ is a Poincaré constant associated to $p$ if for every differentiable function $\varphi \in L^2(p)$ such that $\mathbb{E}\varphi(X) = 0$, $\mathbb{E}\big[\varphi^2(X)\big] \leqslant C_p \mathbb{E}\big[|\nabla \varphi(X)|^2\big]$. Not all densities $p$ admit a finite $C_p$ and probability distributions which possess a Poincaré constant are also referred to as having a spectral gap. Let $\mathcal{C}^\infty_{c,0}(\mathbb{R}^d) = \{f \in \mathcal{C}^\infty_c(\mathbb{R}^d)\ :\ \int f\, p = 0\} \subset L^2(p)$. Define on $\mathcal{C}^\infty_{c,0}(\mathbb{R}^d)^2$ the Dirichlet form $\mathcal{E}(f, g) = \int \langle \nabla f, \nabla g \rangle\, p$ and the scalar product $\mathcal{E}_1(f, g) = \int (f\, g + \langle \nabla f, \nabla g \rangle) p$. Assume that $\mathcal{E}$ is *closable* (i.e. for all $u_n \in \mathcal{C}^\infty_{c,0}(\mathbb{R}^d)$; $n \geqslant 0$, if $u_n \to 0$ in $L^2(p)$ and $(u_n)$ is $\mathcal{E}$-Cauchy, then $\mathcal{E}(u_n, u_n) \to 0$). This condition ensures that there exists a set of functions $W^0_{1,2}(p) \supset \mathcal{C}^\infty_{c,0}(\mathbb{R}^d)$, which admit a gradient, and such that $W^0_{1,2}(p)$ is a Hilbert space for the scalar product $\mathcal{E}_1$. In particular, by passing to the limit, $\int \phi\, p = 0$ for all $\phi \in W^0_{1,2}(p)$, and the Poincaré inequality holds for such $\phi$ if it holds for any differentiable $f \in L^2(p)$ such that $\int f\, p = 0$. For background and more material on closable Dirichlet forms as well as sufficient conditions on $p$ for a Dirichlet form to be closable, we refer to [58].

**Proposition 3.13.** *Let $h$ be a 1-Lipschitz function. Let $X$ be a random vector with density $p$ with Poincaré constant $C_p$ and satisfying the closability property. Let $W^0_{1,2}(p)$ be defined as above. Then there exists a weak solution $u \in W^0_{1,2}(p)$ to* (48) *such that*

$$\sqrt{\int |\nabla u|^2 p} \leqslant C_p.$$

The proof of the result follows exactly the lines of the proof in [27] and is hence omitted. We point out that here $p$ does not need to satisfy Assumption A.

**Remark 3.14.** *For strongly $k$-log concave $p$,* (49) *gives that there exists a (strong) solution $u$ such that $|\nabla u(x)| \leqslant 1/k$, for all $x$ in the domain. Now it is known that when $X$ has $k$-log-concave density, then the law of $X$ possesses a Poincaré constant $C_p = 1/k$ see [17] (it is not necessary to assume that the density is $k$-strongly log-concave). Hence we can also apply Proposition 3.13, which gives gives $\sqrt{\int |\nabla u|^2 p} \leqslant 1/k$. Thus, the constants in* (49) *and in Proposition* (3.13) *are the same; the bound in Proposition* (3.13) *is weaker only because the norm is weaker.*

**Remark 3.15.** *In [27] it is shown that for any mean zero multivariate distribution $p$ which is absolutely continuous with respect to the Lebesgue measure with finite second moment, and which satisfies a Poincaré inequality with constant $C_p$, there exists a unique function $g \in W_p^{1,2}$ such that $\boldsymbol{\tau}_p = \nabla g$ is a Stein kernel for $p$. Moreover,*

$$\int \|\boldsymbol{\tau}_p\|_{HS}^2 \, p \leq C_p \int |x|^2 \, p. \tag{50}$$

*In the one-dimensional case, [55] (inspired by a similar result from [28]) show that there exists a solution $u$ such that*

$$|u'(x)| \leqslant \tau_p(x), \tag{51}$$

*$\tau_p$ being the univariate Stein kernel associated to $X$. Under log concavity of the density, in the univariate case a stronger bound is available in [74]: Assume $X$ is centered and has a smooth, non-vanishing $k$-log concave density $p$ on $\mathbb{R}$. Then $\tau_p(x) \leqslant 1/k$, for all $x \in \mathbb{R}$. Thus, in the univariate case where $p$ is $k$-log concave, the bound* (51) *implies* (49)*, which in turn implies the $L^2$ bound from Proposition 3.13.*

# 4 Wasserstein distance between nested distributions

In this section we illustrate Stein's method in the multivariate setting in three examples which are based on a novel bound on the Wasserstein distance between nested distributions. The (1-)Wasserstein distance between two distributions $F$ and $G$ is

$$d_{\mathcal{W}}(F, G) = \sup_{h \in \mathcal{W}} |F(h) - G(h)|$$

with $\mathcal{W} = \mathrm{Lip}(1)$ the collection of Lipschitz functions $h : \mathbb{R}^d \to \mathbb{R}$ with Lipschitz constant smaller than 1. Here $F(h) = \int h \, dF$ is the expectation of $h$ under $F$. Abusing notation, we also write $d_{\mathcal{W}}(X, Y) = d_{\mathcal{W}}(F, G)$ when $X \sim F$ and $Y \sim G$.

Let $P_1$ and $P_2$ be two probability measures on $\mathbb{R}^d$, with respective pdfs $p_1$ and $p_2 = \pi_0 p_1$, so that the two densities are nested. Assume $p_1$ is $k$-log concave. Consider as in (43) a Stein operator for $P_i, i = 1, 2$ defined by

$$\mathcal{A}_i u = \langle \nabla \log p_i, \nabla u \rangle + \Delta u.$$

Then $\mathcal{A}_2 u = \mathcal{A}_1 u + \langle \nabla \log \pi_0, \nabla u \rangle$. This relationship between generators makes it straightforward to bound the Wasserstein distance between the distributions, as follows.

**Proposition 4.1.** *Assume $p_2 = \pi_0 p_1$ and that $\mathbb{E}[|\nabla \pi_0(X_1)|] < \infty$.*

1. *Assume that $p_1$ is $k$-strongly log concave. Then with $X_1 \sim p_1$ and $X_2 \sim p_2$,*

$$d_{\mathcal{W}}(X_1, X_2) \leqslant \frac{1}{k} \mathbb{E}[|\nabla \pi_0(X_1)|]. \tag{52}$$

2. *Assume that $p_1$ admits a Poincaré constant $C_p$ and $\pi_0 \in W_{1,2}(p_1)$. Then with $X_1 \sim p_1$ and $X_2 \sim p_2$,*

$$d_{\mathcal{W}}(X_1, X_2) \leqslant C_p \sqrt{\mathbb{E}[|\nabla \pi_0(X_1)|^2]}. \tag{53}$$

*Proof.* Let $h : \mathbb{R}^d \mapsto \mathbb{R}$ be a 1-Lipschitz function. To show (52), by (49), there exists a solution $u_h$ to $\mathcal{A}_1 u_h = h - \int h p_1$ such that $\|\nabla u_h\|_\infty \leqslant 1/k$. Let $X_1$ ($X_2$) have distribution $P_1$ ($P_2$). Then

$$\begin{aligned}\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)] &= \mathbb{E}[\mathcal{A}_1 u_h(X_2)] \\ &= \mathbb{E}\left[\mathcal{A}_2 u_h(X_2) - \langle \nabla \log \pi_0(X_2), \nabla u_h(X_2)\rangle\right] \\ &= -\mathbb{E}\left[\langle \nabla \log \pi_0(X_2), \nabla u_h(X_2)\rangle\right].\end{aligned}$$

By (49) we have $|\langle \nabla \log \pi_0(X_2), \nabla u_h(X_2)\rangle| \leqslant |\nabla \log \pi_0(X_2)|/k$, and thus

$$|\mathbb{E}[h(X_2)] - \mathbb{E}[h(X_1)]| \leqslant \frac{1}{k}\mathbb{E}[|\nabla \log \pi_0(X_2)|], \tag{54}$$

and the same bound holds for the Wasserstein distance.

To show (53), by Proposition 3.13, there exists a weak solution $u_h \in W^0_{1,2}(p)$ to (48) such that $\sqrt{\int |\nabla u|^2 p} \leqslant C_p$. It is straightforward to see that

$$\mathbb{E}\left[\langle \nabla u_h(X), \nabla v(X)\rangle\right] = \mathbb{E}\left[\bar{h}(X) v(X)\right] \tag{55}$$

for any $v \in W_{1,2}(p_i)$. Indeed, by definition, $\mathcal{A} f = \mathcal{T}_{\nabla,p}(\nabla f)$ so that, by (30),

$$\mathbb{E}\left[\mathcal{A} f(X) g(X)\right] = \mathbb{E}\left[\mathcal{T}_{\nabla,p}(\nabla f(X)) g(X)\right] = -\mathbb{E}\left[\langle \nabla f(X), \nabla g(X)\rangle\right];$$

and (55)follows. Let $\bar{h}(x) = h(x) - \mathbb{E}[h(X_1)]$. By 55, we have for any $v \in W_{1,2}(p)$,

$$\int \langle \nabla u_h, \nabla v\rangle \, p_1 = -\int \bar{h}\, v \, p_1.$$

Applying this equation to $v = -\pi_0 \in W_{1,2}(p_1)$,

$$\mathbb{E}[h(X_2) - h(X_1)] = \int \langle \nabla u_h, \nabla \pi_0\rangle p_1 \leqslant \left(\int |\nabla u_h|^2 p_1 \int |\nabla \pi_0|^2 p_1\right)^{1/2} \leqslant C_p \sqrt{\mathbb{E}[|\nabla \pi_0(X_1)|^2]}.$$

□

**Remark 4.2.** *In dimension 1, in [55], Equation (4.2), gives*

$$d_{\mathcal{W}}(X_1, X_2) \leqslant \mathbb{E}[\tau_1(X_1)|\pi_0'(X_1)|], \tag{56}$$

*where $\tau_1$ is the Stein kernel associated to $X_1$. If $p_1$ admits a Poincaré constant $C_p$, then from Proposition 3.13 (applied to $h(x) = -x$), we have that $\int (\tau(x))^2 p_1 \leqslant C_p^2$. Thus by the Cauchy-Schwarz inequality,* (56) *is a stronger bound than* (53)*.*

## 4.1 Example 1: Copulas

Let $(V_1, V_2)$ be a 2-dimensional random vector, such that the marginals $V_1$ and $V_2$ have a uniform distribution on $[0, 1]$. We want to bound the Wasserstein distance between $(V_1, V_2)$ and its independent version $(U_1, U_2)$ ($U_1$ and $U_2$ are uniform and independent), in terms of the copula of $(V_1, V_2)$ defined as

$$C(x_1, x_2) = \mathbb{P}[V_1 \leqslant x_1, V_2 \leqslant x_2], \quad (x_1, x_2) \in [0, 1]^2.$$

(Note that the copula for $(U_1, U_2)$ is $(x_1, x_2) \mapsto x_1 x_2$.) Assume that $V_1, V_2$ has a pdf $c = \partial^2_{x_1 x_2} C$. An optimal Poincaré constant for the uniform distribution on $[0, 1]^2$ is $C_p = 2/\pi^2$, see [68]. Then, a simple application of (52) yields

**Corollary 4.3.** *Let $(V_1, V_2)$ have uniform marginals on $[0,1]$, and pdf $c$. Let $(U_1, U_2)$ also have uniform marginals, and let $U_1$ be independent of $U_2$. Then*

$$d_{\mathcal{W}}[(V_1, V_2)\ ,\ (U_1, U_2)] \leqslant \frac{2}{\pi^2}\sqrt{\int_{[0,1]^2} |\nabla c(x_1, x_2)|^2 dx_1\, dx_2}.$$

In some cases, one can compute the gradient of $c$ in a closed form.

**Example 4.4.** *The Ali-Mikhail-Haq copula [5] has pdf*

$$c(x_1, x_2) = \frac{(1-\theta)\{1-\theta(1-x_1)(1-x_2)\} + 2\theta x_1 x_2}{\{1-\theta(1-x_1)(1-x_2)\}^3}.$$

*Here $\theta \in [-1, 1]$ is a measure of association between the two components $V_1$ and $V_2$ of the vector $(V_1, V_2)$ with uniform marginals each. If $\theta = 0$ then the uniform copula $(x_1, x_2) \mapsto x_1 x_2$ is recovered. Using Corollary 4.3 we can assess the Wasserstein distance between the Ali-Mikhail-Haq copula and the uniform copula in terms of $\theta$. for $-1 < \theta < 1$*

$$\int_{[0,1]^2} |\nabla c(x_1, x_2)|^2 dx_1\, dx_2 \quad \leq \quad 128\theta^2 \frac{1}{\{1-|\theta|\}^8}.$$

*Hence*

$$d_{\mathcal{W}}([(V_1, V_2)\ ,\ (U_1, U_2)]) \leq 2.3\,|\theta|\{1-|\theta|\}^{-4}.$$

*This bound shows the expected behaviour – it tends to 0 for $\theta \to 0$, whereas it diverges for $|\theta| \to 1$.*

## 4.2 Example 2: Normal model with normal prior

Consider a normal $\mathcal{N}(\theta, \Sigma_2)$ model with mean $\theta \in \mathbb{R}^d$ and positive definite covariance matrix $\Sigma$. The likelihood of a sample $(x_1, \ldots, x_n)$ (where $x_i \in \mathbb{R}^d$ for all $i$) is given by $\mathcal{L}(x_1, \ldots, x_n|\theta) = \prod_i f(x_i|\theta)$ where

$$f(x|\theta) = (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(x-\theta)^T \Sigma^{-1}(x-\theta)\right).$$

To compare the posterior distribution $P_1$ of $\theta$ with uniform prior with the posterior $P_2$ with normal prior with parameters $(\mu, \Sigma_2)$ we employ the operator norm $|\!|\!|A|\!|\!| = \sup_{\|x\|=1} \|Ax\|$.

**Corollary 4.5.** *Let $P_1$ denote the posterior distribution of $\theta$ with uniform prior and $P_2$ the posterior of $\theta$ with prior $\mathcal{N}(\mu, \Sigma_2)$; $\Sigma_2$ is assumed positive definite. Then*

$$\begin{aligned} d_{\mathcal{W}}(P_1, P_2) \leqslant\ & |\!|\!|\Sigma|\!|\!|\ |\!|\!|(\Sigma + n\Sigma_2)^{-1}|\!|\!|\ \|\bar{x} - \mu\| \\ & + \frac{\sqrt{2}\Gamma(d/2 + 1/2)}{\Gamma(d/2)} \frac{|\!|\!|\Sigma|\!|\!|}{n} |\!|\!|(\Sigma_2 + n\Sigma_2\Sigma^{-1}\Sigma_2)^{-1/2}|\!|\!|. \end{aligned}$$

*Proof.* It is a standard calculation that $P_1 \sim \mathcal{N}(\bar{x}, n^{-1}\Sigma)$, with $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ Thus (see Example 2.10 in [75]) $p_1$ is strongly $1/\lambda$-log concave,where $\lambda$ is the greatest eigenvalue of $\Sigma$; $\lambda$ is also the operator norm of $\Sigma$,. and (52) can be applied. From (54), we deduce

$$d_{\mathcal{W}}(P_1, P_2) \leqslant \frac{|\!|\!|\Sigma|\!|\!|}{n} \mathbb{E}[\|\nabla \log \pi_0(X_2)\|]$$

where $X_2$ is a r.v. with law $P_2$. It remains to calculate this last expectation. It is again a standard calculation that $P_2 \sim \mathcal{N}(\tilde{\mu}, \tilde{\Sigma}_n)$ with

$$\begin{aligned} \tilde{\mu} &= (\Sigma_2^{-1} + (n^{-1}\Sigma)^{-1})^{-1}((n^{-1}\Sigma)^{-1}\bar{x} + \Sigma_2^{-1}\mu) = \mu + n\tilde{\Sigma}_n\Sigma^{-1}(\bar{x} - \mu) \\ \tilde{\Sigma}_n &= (\Sigma_2^{-1} + n\Sigma^{-1})^{-1}. \end{aligned} \tag{57}$$

Since $p_2(\theta) \propto p_1(\theta)\exp[(\theta-\mu)^T\Sigma_2^{-1}(\theta-\mu)]$, it holds that

$$\nabla \log \pi_0(\theta) = -\Sigma_2^{-1}(\theta - \mu).$$

Together with (57), it follows that $\nabla \log \pi_0(X_2)$ has a normal distribution with mean $-n\Sigma_2^{-1}\tilde{\Sigma}_n\Sigma^{-1}(\bar{x}-\mu)$ and covariance matrix $\Sigma_2^{-1}\tilde{\Sigma}_n\Sigma_2^{-1}$. Note that

$$\Sigma_2^{-1}\tilde{\Sigma}_n\Sigma^{-1} = \Sigma_2^{-1}(\Sigma_2^{-1} + n\Sigma^{-1})^{-1}\Sigma^{-1} = (\Sigma + n\Sigma_2)^{-1}$$

and, in the same way, $\Sigma_2^{-1}\tilde{\Sigma}_n\Sigma_2^{-1} = (\Sigma_2 + n\Sigma_2\Sigma^{-1}\Sigma_2)^{-1}$. Let $N$ stand for a $d$-dimensional standard normal vector. From the above, we deduce that $\nabla \log \pi_0(X_2) \sim -n(\Sigma+n\Sigma_2)^{-1}(\bar{x}-\mu) + (\Sigma_2 + n\Sigma_2\Sigma^{-1}\Sigma_2)^{-1/2}N$. Thus,

$$\begin{aligned} \frac{|||\Sigma|||}{n}\mathbb{E}[\|\nabla \log \pi_0(X_2)\|] \leqslant & |||\Sigma||| \, |||(\Sigma + n\Sigma_2)^{-1}||| \, \|\bar{x} - \mu\| \\ & + \frac{|||\Sigma|||}{n}|||(\Sigma_2 + n\Sigma_2\Sigma^{-1}\Sigma_2)^{-1/2}|||\mathbb{E}[\|N\|]. \end{aligned}$$

Since $\|N\|^2 \sim \chi^2(d)$, $\|N\|$ follows a chi distribution with $d$ degrees of freedom and $\mathbb{E}[\|N\|] = \frac{\sqrt{2}\Gamma(d/2+1/2)}{\Gamma(d/2)}$ . The assertion follows. □

**Remark 4.6.** *The bound in Corollary* (4.5) *has a contribution of order* $n^{-1}$ *from the different covariance matrices, but also a contribution which depends on* $\|\bar{x} - \mu\|$. *By the law of large numbers, this term is order* $n^{-1}$ *in probability, but crucially depends on the actual sample which was drawn.*

**Remark 4.7.** *When* $d = 1$, *we retrieve the bound of [55].*

As in [55], our Stein framework lends itself naturally for assessing distributional distances between nested densities more generally than in the previous example.

**Proposition 4.8.** *Assume* $p_2 = p_1\pi_0$, *where* $p_1$ *and* $p_2$ *are densities which satisfy Assumption A. Let* $X_1 \sim p_1$ *and* $X_2 \sim p_2$ *denote two random variables with distributions having densities* $p_1$ *and* $p_2$, *respectively. Then*

$$\|\mathbb{E}[\boldsymbol{\tau}_1(X_1)\nabla\pi_0(X_1)]\| \leqslant d_{\mathcal{W}}(X_1, X_2) \leqslant \sup_{f_h} |\mathbb{E}[\langle \nabla \log \pi_0(X_2), \nabla f_h(X_2)\rangle)]| \tag{58}$$

*where the supremum is taken over all* $f_h$ *solving the* $p_1$*-Stein equation* (48) *for* $h$ *a 1-Lipschitz function.*

*Proof.* Consider the lower bound first. Let $e$ be a unit vector. Since $x \mapsto \langle x, e\rangle$ is 1-Lipschitz, using the nested structure we have

$$\begin{aligned} d_{\mathcal{W}}(X_1, X_2) &\geqslant \mathbb{E}[\langle X_1, e\rangle - \langle X_2, e\rangle] \\ &= \mathbb{E}[\langle X_1, e\rangle(1 - \pi_0(X_1))] \\ &= -\langle \mathbb{E}[\boldsymbol{\tau}_1(X_1)\nabla\pi_0(X_1)], e\rangle. \end{aligned}$$

Taking $e = -\mathbb{E}[\boldsymbol{\tau}_1(X_1)\nabla(p_2/p_1)(X_1)]/\|\mathbb{E}[\boldsymbol{\tau}_1(X_1)\nabla\pi_0(X_1)]\|$ gives

$$d_{\mathcal{W}}(X_1, X_2) \geqslant \|\mathbb{E}[\boldsymbol{\tau}_1(X_1)\nabla\pi_0(X_1)]\|.$$

For the upper bound, if $h$ is 1-Lipschitz, using the score Stein operator for $p_1$,

$$\begin{aligned}\mathbb{E}[h(X_2) - h(X_1)] &= \mathbb{E}[(\Delta f_h)(X_2) - \langle \nabla \log p_1(X_2), \nabla f_h(X_2)\rangle)] \\ &= \mathbb{E}[\langle \nabla \log \pi_0(X_2), \nabla f_h(X_2)\rangle)].\end{aligned}$$

The last equality follows from the nested structure $p_2 = p_1\theta_0$. □

**Remark 4.9.** *The following argument shows that the gradient of the likelihood ratio between two densities arises naturally in the Stein framework. Suppose that $q$ is another density on $\mathbb{R}^d$ and that $K_p$, the support of $p$, is a subset of $K_q$, the support of $q$. Then the likelihood ratio $r = p/q$ is well defined over $\mathbb{R}^d$ (with the convention that $r = 0$ outside of $K_q$) and, for every $\mathbf{f} \in \mathcal{F}(p) \cap \mathcal{F}(q)$, the product rule* (12) *gives*

$$\begin{aligned}\mathcal{T}_{\mathrm{div},p}\mathbf{f} &= \frac{\mathrm{div}(\mathbf{f}\, p)}{p} = \frac{\mathrm{div}(\mathbf{f}\, q\, r)}{q}\frac{1}{\ell} \\ &= \frac{\mathrm{div}(\mathbf{f}\, q)}{q} + \Big\langle \mathbf{f}, \frac{\nabla r}{r}\Big\rangle = \mathcal{T}_{\mathrm{div},q}(\mathbf{f}) + \Big\langle \mathbf{f}, \frac{\nabla r}{r}\Big\rangle. \end{aligned} \tag{59}$$

*In combination with the ideas of Proposition 4.8, this leads to a general bound on the difference between densities in terms of the likelihood ratio.*

**Remark 4.10.** *Based on Remark 4.9, the comparisons of the posteriors arising from different priors can be carried out more generally. Here is a sketch. In the Bayesian framework, we aim to compare $\Theta_1$ and $\Theta_2$ obtained through different priors, where for $\ell_\theta(x)$ the likelihood function,*

$$\begin{aligned}\Theta_1 &\sim \pi_1(\theta) = \kappa_1(x)\ell_\theta(x) \\ \Theta_2 &\sim \pi_2(\theta) = \kappa_2(x)\pi_0(\theta)\ell_\theta(x)\end{aligned}$$

*for $\pi_0$ a nonnegative function with support a subset of $\mathbb{R}^d$, and $\kappa_i(x)$, $i = 1, 2$ the normalizing constants – here $x = (x_1, \ldots, x_J) \in \mathbb{R}^J$ is a fixed sample of size $J$ and $\theta \in \mathbb{R}^d$ is the variable. Then, in the notations of Remark 4.9, we have $r(\theta) = \kappa_2(x)/\kappa_1(x)\pi_0(\theta)$ so that, introducing $\rho_0 = \nabla\pi_0/\pi_0$ – which does not depend on the normalizing constants – we obtain, from* (59),

$$\mathcal{T}_{\mathrm{div},\pi_2}\mathbf{f} = \mathcal{T}_{\mathrm{div},\pi_1}\mathbf{f} + \langle \mathbf{f}, \rho_0\rangle \tag{60}$$

*for any function $\mathbf{f} \in \mathbb{R}^n \times \mathbb{R}^d \in \mathcal{F}(\pi_1) \cap \mathcal{F}(\pi_2)$, where $n \geq 1$ is arbitrary. In particular,*

$$\mathbb{E}\left[\mathcal{T}_{\mathrm{div},\pi_1}\mathbf{f}(\Theta_2)\right] = \mathbb{E}\left[\langle \mathbf{f}(\Theta_2), \rho(\Theta_2)\rangle\right] \tag{61}$$

*for all sufficiently regular vector-valued functions $\mathbf{f}$. Next suppose that there exists some well chosen matrix valued function $\mathbf{B}$ for which $\pi_1$ is characterized by a second-order divergence Stein operator acting on real valued functions $u : \mathbb{R}^d \to \mathbb{R}$ via*

$$u \mapsto \mathcal{A}_1 u = \mathcal{T}_{\mathrm{div},\pi_1}(\mathbf{B}\nabla u) = \langle \mathbf{b}, \nabla u\rangle + \big\langle \mathbf{B}, \nabla^2 u\big\rangle_{\mathrm{HS}} \tag{62}$$

*with $\mathbf{b} = \mathcal{T}_{\mathrm{div},\pi_1}\mathbf{B}$ (recall* (45)*); then in light of* (60)*, we know that*

$$\mathcal{A}_2 u = \mathcal{T}_{\mathrm{div},\pi_2}(\mathbf{B}\nabla u) = \mathcal{A}_1 u + \langle \mathbf{B}\nabla u, \rho_0\rangle \tag{63}$$

*satisfies* $\mathbb{E}\big[\mathcal{A}_2 u(\Theta_2)\big] = 0$ *for all admissible functions* $u$*. In particular, if* $u_h$ *is a solution to the (second order) Stein equation*

$$\mathcal{A}_1 u_h(\theta) = h(\theta) - \mathbb{E}\left[h(\Theta_1)\right]$$

*we get, with the help of* (63) *and under suitable conditions,*

$$\begin{aligned}\mathbb{E}\left[h(\Theta_2)\right] - \mathbb{E}\left[h(\Theta_1)\right] &= \mathbb{E}\left[\mathcal{A}_1 u_h(\Theta_2)\right] \\ &= \mathbb{E}\left[\mathcal{A}_2 u_h(\Theta_2)\right] - \mathbb{E}\left[\langle \mathbf{B}(\Theta_2)\nabla u_h(\Theta_2), \rho_0(\Theta_2)\rangle\right] \\ &= -\mathbb{E}\left[\langle \mathbf{B}(\Theta_2)\nabla u_h(\Theta_2), \rho_0(\Theta_2)\rangle\right].\end{aligned}$$

*Investigating this approach in more detail will be part of future research.*

## 4.3 Example 3: Comparing Azzalini-Dalla Valle skew-normal distributions to multivariate normal

The density of the Azzalini-Dalla Valle type r.v. $X$ is given by

$$p_\alpha(x) = 2\omega_d(x-\mu;\Sigma)\Phi(\alpha^T x),$$

where $\omega_d(x-\mu:\Sigma)$ is the density of $\mathcal{N}(\mu,\Sigma)$, $\Phi$ the c.d.f. of the standard normal on $\mathbb{R}$, and $\alpha \in \mathbb{R}^d$ is a skew parameter, see [9]. Now we assume for simplicity $\Sigma = I_d$ and $\mu = 0$. In [56], an exact expression for the Wasserstein distance is given for $d = 1$; here we extend this result to general $d$.

**Proposition 4.11.** *Let* $X \in \mathbb{R}^d$ *have pdf* $p_\alpha(x) = 2\omega_d(x; Id)\Phi(\alpha^T x)$,, *and let* $Z$ *be a* $d$*-dimensional standard normal element. Then*

$$d_{\mathcal{W}}(X,Z) = \sqrt{\frac{2}{\pi}}\frac{\|\alpha\|}{\sqrt{1+\|\alpha\|^2}}. \tag{64}$$

*Proof.* First we show the upper bound on the Wasserstein distance. As the multivariate standard normal distribution is 1-strongly log concave,(52) can be applied with $\pi_0(x) = 2\Phi(\alpha^T x)$. As $\nabla\pi_0(x) = 2\alpha\phi(\alpha^T x)$ with $\phi$ the one-dimensional standard normal pdf, it suffices to calculate that

$$\begin{aligned}\mathbb{E}[\phi(\alpha^T Z)] &= (2\pi)^{-(d+1)/2}\int_{\mathbb{R}^d}\exp\left[-\frac{1}{2}\left((\alpha^T x)^2 + \|x\|^2\right)\right]dx \\ &= (2\pi)^{-(d+1)/2}\,\frac{(2\pi)^{d/2}}{\sqrt{\det(M)}} \\ &= \frac{1}{\sqrt{2\pi(1+\|\alpha\|^2)}}.\end{aligned}$$

Thus,

$$d_{\mathcal{W}}(X,Z) \leqslant \sqrt{\frac{2}{\pi}}\frac{\|\alpha\|}{\sqrt{1+\|\alpha\|^2}}. \tag{65}$$

Now we prove we have actually equality. Consider the 1-Lipschitz test function $h(x) = \langle \frac{\alpha}{\|\alpha\|}, x\rangle$; note that $\mathbb{E}[h(Z)] = 0$. Then $u_h(x) = -h(x)$ is a solution to the Stein equation

$\Delta u - \langle x, \nabla u(x)\rangle = h(x)$ and $\Delta u = 0$, and

$$\begin{aligned}\mathbb{E}[h(X) - h(Z)] &= 2\mathbb{E}[\Delta u_h(X) - \langle X, \nabla u_h(X)\rangle] \\ &= 2\mathbb{E}[\langle \nabla u_h(Z), \alpha\phi(\alpha^T Z)\rangle] \\ &= 2\mathbb{E}[\langle \nabla h(Z), \alpha\phi(\alpha^T Z)\rangle] \\ &= 2\mathbb{E}\left[\frac{1}{\|\alpha\|}\langle \alpha, \alpha\phi(\alpha^T Z)\rangle\right] \\ &= \sqrt{\frac{2}{\pi}}\frac{\|\alpha\|}{\sqrt{1+\|\alpha\|^2}},\end{aligned}$$

so that $d_{\mathcal{W}}(X, Z) \geqslant \sqrt{\frac{2}{\pi}}\frac{\|\alpha\|}{\sqrt{1+\|\alpha\|^2}}$. Thus we obtained (64). □

# 5 Stein operators for elliptical distributions

In this section we detail the constructions explicitly for the entire family of elliptical distributions, defined as follows, see [49].

**Definition 5.1.** *An absolutely continuous d-random vector has multivariate elliptical distribution $\mathbb{E}_d(\nu, \Sigma, \phi)$ if its density is of the form*

$$p(x) = \kappa|\Sigma|^{-1/2}\phi\left(\frac{1}{2}(x-\nu)^T\Sigma^{-1}(x-\nu)\right), x \in \mathbb{R}^d, \tag{66}$$

*for $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ a measurable function called* density generator*, $\nu \in \mathbb{R}^d$ the location parameter, $\kappa$ the normalising constant and $\Sigma = (\sigma_{ij})$ a symmetric positive definite $d \times d$ dispersion matrix.*

A particular important case is $\mathbb{E}_d(0, I_d, \phi)$ called *spherical distribution*. Note that the matrix $\Sigma$ in definition (66) is not necessarily the covariance matrix; also not all choices of $\phi$ lead to well-defined densities, see [49] or [69] for a discussion and references. Prominent members of the elliptical family are

1. Gaussian distribution $\mathcal{N}_d(\nu, \Sigma)$, with $\phi(t) = e^{-t}$ and $\kappa = (2\pi)^{-d/2}$.
2. Power exponential distribution $\phi(t) = \exp(-b_{p,\zeta}t^\zeta)$ with $\zeta > 0$, $b_{p,\zeta}$ a scale factor and $\kappa$ defined accordingly, see [42, 4] for details.
3. Multivariate Student-$t$ distribution, with $\phi(t) = (1 + 2t/k)^{-(k+d)/2}$ and $\kappa = c_{k,d,\Sigma}$.
4. Symmetric generalized hyperbolic distribution with density

$$p(x) = \frac{\left(\sqrt{\chi\psi}\right)^{-\lambda}\psi^{d/2}}{(2\pi)^{d/2}\,|\Sigma|\,K_\lambda\left(\sqrt{\chi\psi}\right)}\frac{K_{\lambda-d/2}\left(\sqrt{(\chi + (x-\nu)^T\Sigma^{-1}(x-\nu))\,\psi}\right)}{\left(\sqrt{(\chi + (x-\nu)^T\Sigma^{-1}(x-\nu))\,\psi}\right)^{(d/2)-\lambda}} \tag{67}$$

   where $K_\lambda$ denotes a modified Bessel function of the third kind and $\lambda, \chi, \psi$ are real parameters, see [61, Example 3.8] for details. To put in the parameterization (66) we take $\phi(t) = K_{\lambda-d/2}(\sqrt{2t})/(\sqrt{2t})^{(d/2)-\lambda}$ and $\kappa = \frac{\left(\sqrt{\chi\psi}\right)^{-\lambda}\psi^{d/2}}{(2\pi)^{d/2}\left(\sqrt{\chi\psi}\right)}$.

Throughout this section we let $\Omega_\phi$ be the image of $\Omega_p$ through $x \mapsto \frac{1}{2}(x-\nu)^T\Sigma^{-1}(x-\nu)$. We make the following assumption.

**Assumption B**: There is a non-empty open set $\Omega_\phi$ such that $\phi > 0$ and $\phi$ is $\mathcal{C}^1$ on this open set, and such that the Lebesgue measure of $K_\phi \backslash \Omega_\phi$ is zero.

Under Assumption B, it is readily checked that $p$ satisfies Assumption A with $\Omega_p$ defined by

$$\begin{array}{rcl} \Omega_p & = & \left\{x \in \mathbb{R}^d \ : \ \frac{1}{2}(x-\nu)^T\Sigma^{-1}(x-\nu) \in \Omega_\phi\right\} \\ & = & \left\{x \in \mathbb{R}^d \ : \ \phi\left(\frac{1}{2}(x-\nu)^T\Sigma^{-1}(x-\nu)\right) > 0\right\}. \end{array} \tag{68}$$

It is straightforward to calculate that if $p$ is of the form (66) and satisfies Assumption B then

$$\rho_p(x) = \Sigma^{-1}(x-\nu)\frac{\phi'((x-\nu)^T\Sigma^{-1}(x-\nu)/2)}{\phi((x-\nu)^T\Sigma^{-1}(x-\nu)/2)} \tag{69}$$

is the score function of $p$ (defined on $\Omega_p$). Hence the score-Stein operator (32) is easily obtained for this family of distributions.

## 5.1 Stein kernels for elliptical distributions

The following proposition shows that in order to find Stein kernels for elliptical distributions, it suffices to consider the case $\Sigma = I_d$ and $\nu = 0$.

**Proposition 5.2.** *The application*

$$\boldsymbol{\tau} \mapsto (x \mapsto \Sigma^{1/2}\boldsymbol{\tau}(\Sigma^{-1/2}(x-\nu))\Sigma^{1/2}),$$

*is a bijection between the set of Stein kernels of $\mathbb{E}_d(0, I_d, \phi)$ and that of $\mathbb{E}_d(\nu, \Sigma, \phi)$.*

*Proof.* Let $p$ (resp. $q$) be the density of $\mathbb{E}_d(0, I_d, \phi)$ (resp. $\mathbb{E}_d(\nu, \Sigma, \phi)$). If $\boldsymbol{\tau}_X$ is a Stein kernel for $X \sim \mathbb{E}_d(0, Id, \phi)$, then for all $f \in \mathcal{C}_c^\infty(\Omega_p)$, $\mathbb{E}[\boldsymbol{\tau}_X(X)\nabla f(X)] = \mathbb{E}[Xf(X)]$. Setting $f(x) = g(\Sigma^{1/2}x+\nu)$, we deduce that for all $g \in \mathcal{C}_c^\infty(\Omega_q)$, we have $\mathbb{E}[\boldsymbol{\tau}_X(X)\Sigma^{1/2}\nabla g(\Sigma^{1/2}X+\nu)] = \mathbb{E}[Xg(\Sigma^{1/2}X+\nu)]$. Thus

$$\mathbb{E}[\Sigma^{1/2}\boldsymbol{\tau}_X(\Sigma^{-1/2}((\Sigma^{1/2}X+\nu)-\nu))\Sigma^{1/2}\nabla g(\Sigma^{1/2}X+\nu)] = \mathbb{E}[(\Sigma^{1/2}X+\nu-\nu)g(\Sigma^{1/2}X+\nu)].$$

It follows $x \mapsto \Sigma^{1/2}\boldsymbol{\tau}_X(\Sigma^{-1/2}(x-\nu))\Sigma^{1/2}$ is a Stein kernel for $\Sigma^{1/2}X+\nu$. The converse is shown in the same way. □

Now we state the following result, due to [49, 51] but for which we give a new proof.

**Lemma 5.3** (Proposition 2, [51])**.** *If $X \sim E_d(\nu, \Sigma, \phi)$ then the matrix*

$$\boldsymbol{\tau}_1(x) = \left(\frac{1}{\phi((x-\nu)^T\Sigma^{-1}(x-\nu)/2)}\int_{(x-\nu)^T\Sigma^{-1}(x-\nu)/2}^{+\infty}\phi(u)du\right)\Sigma, \tag{70}$$

*is a Stein kernel for $X$ if $\boldsymbol{\tau}_1 \in \mathcal{F}(p)$.*

*Proof.* For transparency here we consider general $\nu$ and $\Sigma$. We will show the Stein identity holds for functions in $\mathcal{C}_c^\infty(\Omega_p)$. Consider any functions $f : \mathbb{R} \to \Omega_\phi \in \mathcal{F}(p)$ and $g : \Omega_p \to \mathbb{R} \in \mathrm{dom}(p, f)$. We start by inverting (69) to get

$$x - \nu = \Sigma\rho_p(x)\frac{\phi((x-\nu)^T\Sigma^{-1}(x-\nu)/2)}{\phi'((x-\nu)^T\Sigma^{-1}(x-\nu)/2)}.$$

Fixing $\nu = 0$ (see Proposition 5.2) and introducing the temporary notations $\psi(t) = \phi(t)/\phi'(t)$ and $t = x^T\Sigma^{-1}x/2$, $T = X^T\Sigma^{-1}X/2$, by (33),

$$\begin{aligned}
&\mathbb{E}_p\left[Xf(T)g(X)\right] = \mathbb{E}_p\left[\Sigma\rho_p(X)\psi(T)f(T)g(X)\right]\\
&= -\Sigma\mathbb{E}_p\left[\nabla\left\{\psi(T)f(T)g(X)\right\}\right]\\
&= -\Sigma\mathbb{E}_p\left[\nabla\left\{\psi(T)f(T)\right\}g(X)\right] - \Sigma\mathbb{E}_p\left[\psi(T)f(T)\nabla g(X)\right]\\
&= -\Sigma\mathbb{E}_p\left[\left(\psi'(T)f(T) + \psi(T)f'(T)\right)\Sigma^{-1}Xg(X)\right] - \Sigma\mathbb{E}_p\left[\psi(T)f(T)\nabla g(X)\right]
\end{aligned}$$

and thus

$$\mathbb{E}_p\left[\left(f(T)(1+\psi'(T)) + \psi(T)f'(T)\right)Xg(X)\right] = -\mathbb{E}_p\left[\psi(T)f(T)\Sigma\nabla g(X)\right]. \tag{71}$$

In order to obtain a Stein kernel, it suffices to choose $f$ solution to the ODE

$$f(t)(1+\psi'(t)) + \psi(t)f'(t) = -1$$

to ensure that the function $x \mapsto \psi(t)f(t)\Sigma$ satisfies (40), and is a Stein kernel in the sense of Definition 3.8. Now note that the function $u(t) := \frac{1}{\phi(t)}\int_t^{+\infty}\phi(u)du$ satisfies $u' = -\frac{1}{\psi}u - 1$; hence the choice $f = u/\psi$ satisfies $(f\psi)' = (u)' = -\frac{1}{\psi}u - 1$ and thus is exactly what we need. Setting $t = x^T\Sigma^{-1}x/2$ the claim follows. □

**Remark 5.4.** *In dimension $d = 1$, the Stein kernel (4) of $p$ is the function $\tau_p(x) = \frac{1}{p(x)}\int_x^{+\infty}(u-\nu)p(u)du$. Changing variables in (70) for $x \geq 0$,*

$$\begin{aligned}
\boldsymbol{\tau}_p(x) &= \frac{1}{\phi((x-\nu)^2/2)}\int_{(x-\nu)^2/2}^{+\infty}\phi(u)du\\
&= \frac{1}{p(x)}\int_{(x-\nu)^2/2}^{+\infty}p(\sqrt{2u})du = \frac{1}{p(x)}\int_x^{\infty}(y-\nu)p(y)dy.
\end{aligned}$$

*The case $x < 0$ is treated similarly. Hence (70) indeed recovers the Stein kernel.*

The identity (70) resulting from Lemma 5.3 has found many applications, [1, 2, 50, 3, 51, 80] and the references therein.

The following proposition gives a way of finding Stein kernels of a particular form which generalizes Lemma 5.3.

**Proposition 5.5.** *Let $a, b : \Omega_\phi \to \mathbb{R}$ two $\mathcal{C}^1$ functions such that, for all all $u \in \Omega_\phi$,*

$$\frac{(a(t)\phi(t))'}{\phi(t)} + 2t\frac{(b(t)\phi(t))'}{\phi(t)} + (d+1)b(t) + 1 = 0 \tag{72}$$

*and set*

$$\boldsymbol{\tau}_{a,b}(t) = a(t)\Sigma + b(t)\,(x-\nu)(x-\nu)^T$$

*with $t = \frac{1}{2}(x-\nu)^T\Sigma^{-1}(x-\nu)$. If $\boldsymbol{\tau}_{a,b} \in \mathcal{F}(p)$ then this function is a Stein kernel for $X \sim \mathbb{E}_d(\nu, \Sigma, \phi)$. Moreover, if $\phi$ is continuous and positive on $[0, +\infty)$, then the Stein identity (41) holds for every test function $g \in \mathcal{C}^1(\mathbb{R}^d)$ such that $g(x)a(t)\phi(t)$ and $g(x)b(t)t$ go to zero when $|x| \to +\infty$.*

*Proof.* From Proposition 5.2, we can assume $\nu = 0$ and $\Sigma = I_d$. It is readily checked that $\mathrm{div}(I_d) = 0$, $\mathrm{div}(xx^T) = (d+1)x$, $xx^T x = 2tx$. Thus, by the chain rule and noting that $\nabla t = x$,

$$\begin{aligned}
&\mathrm{div}(\phi(t)(a(t)I_d + b(t)xx^T)) \\
&= (a(t)\phi(t))' I_d\, x + a(t)\phi(t)\mathrm{div}(I_d) + (b(t)\phi(t))' xx^T\, x + b(t)\phi(t)\mathrm{div}(xx^T) \\
&= (a(t)\phi(t))' x + 2t(b(t)\phi(t))' x + (d+1)b(t)\phi(t)x.
\end{aligned}$$

Hence $a(t)I_d + b(t)xx^T$ is a Stein kernel if the last quantity is equal to $-\phi(t)x$, so that the result follows.

To see that a test function satisfying the stated conditions verifies (41), simply note that since $(x-\nu)(x-\nu)^T$ is of order $t$ when $|x|$ is large, the conditions imply that $\boldsymbol{\tau}(x)p(x)f(x) \to 0$ when $|x|$ is large, and the result follows from the divergence theorem (see discussion below Definition 3.8). □

**Remark 5.6.** *1. A particular instance of Proposition 5.5 is given by the following expression:*

$$\boldsymbol{\tau}_{2,\beta}(x) = \frac{(\beta+2) - 2\frac{\phi''(t)/\phi'(t)}{\phi'(t)/\phi(t)}}{(\beta-2)(d-1)} \left( 2\left( \frac{d-1}{(\beta+2)\frac{\phi'(t)}{\phi(t)} - 2\frac{\phi''(t)}{\phi'(t)}} + t \right) \Sigma - (x-\nu)(x-\nu)^T \right) \tag{73}$$

*is a Stein kernel for $X$ for all $\beta \neq 2$ as long as $\boldsymbol{\tau}_{2,\beta} \in \mathcal{F}(p)$.*

*2. It is straightforward to generalize the previous proposition in the following way. Here without loss of generality we take $\Sigma = I_d$ and $\nu = 0$. Assume we are given matrices $\mathbf{U}_1, \ldots, \mathbf{U}_m$ such that for every $i = 1, \ldots, m$ and some functions $\alpha_i, \beta_i : \mathbb{R} \to \mathbb{R}$,*

$$\mathrm{div}\mathbf{U}_i = \alpha_i(t)x, \quad \mathbf{U}_i x = \beta_i(t)x.$$

*If*

$$1 + \sum_{i=1}^m a_i\alpha_i + \frac{(a_i\phi)'}{\phi}\beta_i = 0,$$

*then $a_1(t)\mathbf{U}_1(t) + \ldots + a_m(t)\mathbf{U}_m(t)$ is a Stein kernel for $X \sim \mathbb{E}_d(0, I_d, \phi)$ if this function is in the class $\mathcal{F}(p)$.*

By setting $b \equiv 0$, we obtain $a(t) = \frac{1}{\phi(t)}\int_t^{+\infty} \phi(u)du$, and we retrieve Lemma 5.3. Setting $a \equiv 0$ leads to the following

**Corollary 5.7.** *Set $t = \frac{1}{2}(x-\nu)^T\Sigma^{-1}(x-\nu)$. If $\int_0^{+\infty} u^{\frac{d-1}{2}}\phi(u)du < \infty$, then*

$$\left( \frac{t^{-\frac{d+1}{2}}}{2\phi(t)} \int_t^{+\infty} u^{\frac{d-1}{2}}\phi(u)du \right) (x-\nu)(x-\nu)^T, \tag{74}$$

*is a Stein kernel for $X \sim \mathbb{E}_d(\nu, \Sigma, \phi)$. Moreover, if $\Omega_p = \mathbb{R}^d$, the Stein identity (41) holds for every function $f \in \mathcal{C}^1(\mathbb{R}^d)$ such that $f(x)t^{\frac{d-1}{2}}$ is bounded.*

*Proof.* With $a \equiv 0$, (72) becomes

$$(b\phi)' + \frac{d+1}{2t}b\phi = -\frac{\phi}{2t},$$

which integrates to

$$b(t) = \frac{t^{-\frac{d+1}{2}}}{2\phi(t)} \int_t^{+\infty} u^{\frac{d-1}{2}} \phi(u) du.$$

Now if $f$ is as stated in the corollary, since $(x-\nu)(x-\nu)^T$ is of order $t$ for large $|x|$, then $\tau(x)f(x)p(x) = \mathcal{O}(\int_t^{+\infty} u^{\frac{d-1}{2}} \phi(u)du)$ for large $|x|$, so that $\boldsymbol{\tau}(x)f(x)p(x)$ goes to zero at infinity, and the Stein identity follows again from the divergence theorem. □

**Remark 5.8.** *For $d = 1$, (74) leads to the classical Stein kernel (4). Indeed, assuming $\nu = 0$, for $x > 0$,*

$$\begin{aligned} \frac{t^{-1}}{2\phi(t)} \int_t^{+\infty} \phi(u)du &= \frac{1}{x^2 p(x)} \int_{x^2/2}^{+\infty} p(2\sqrt{u})du \\ &= \frac{1}{x^2 p(x)} \int_x^{+\infty} s\, p(s)ds, \end{aligned}$$

*and multiplying by $x^2$ yields the claim. The case $x < 0$ is treated similarly.*

In the next three subsections we develop Stein kernels for three distributional families: the multivariate Gaussian, the power exponential, and the multivariate Student $t$-distribution. Similar computations are possible for the symmetric generalized hyperbolic distribution but are not pursued here. We refer to [51, 80] and the references therein.

## 5.2 The multivariate Gaussian distribution

Consider a Gaussian $d$-dimensional random vector $Z \sim \mathcal{N}_d(\nu, \Sigma)$ with pdf $\varphi$ on $\mathbb{R}^d$ and let $\mu(dx) = \varphi(x)dx$ be the corresponding probability measure. As $Z \sim E_d(\nu, \Sigma, \phi)$ with $\phi(t) = e^{-t}$ and $\phi'(t)/\phi(t) = -1$, from (69), we recover that $\rho_\varphi(x) = -\Sigma^{-1}(x-\nu)$ is the score function of $\varphi$. Since $\frac{1}{\phi(t)} \int_t^\infty \phi(u)du = 1$ for all $t$, Lemma 5.3 shows that $\boldsymbol{\tau}_1 = \Sigma$ is, as is well-known, a Stein kernel for $\varphi$. Moreover, (73) gives, after some simplifications, the following family of Stein kernels which are indexed by $\beta \neq 2$ (we set $\nu = 0$ to save space):

$$\boldsymbol{\tau}_{2,\beta}(x) = \frac{\beta}{(\beta-2)(d-1)} \big(x^T\Sigma^{-1}x\Sigma - xx^T\big) - \frac{2}{\beta-2}\Sigma.$$

It is easy to check that these functions are in the class $\mathcal{F}(p)$. Several interesting cases stand out. Sending $\beta$ to 0, on the one hand, and to $+\infty$ on the other hand, we obtain

$$\boldsymbol{\tau}_{2,0}(x) = \Sigma \text{ and } \boldsymbol{\tau}_{2,\infty}(x) = \frac{1}{d-1} \big(x^T\Sigma^{-1}x\Sigma - xx^T\big).$$

In dimension $d \geq 3$, setting $\beta = 2(d-1)$ we get

$$\boldsymbol{\tau}_{2,2(d-1)}(x) = \frac{1}{d-2} \big(x^T\Sigma^{-1}x\Sigma - xx^T - \Sigma\big)$$

Moreover, we find that the Gaussian multivariate normal satisfies

$$\frac{\beta}{(\beta-2)(d-1)} \mathbb{E}\left[\big(X^T\Sigma^{-1}X\Sigma - XX^T\big)\nabla g(X)\right] = \mathbb{E}\left[\frac{2}{\beta-2}\Sigma\nabla g(X) + Xg(X)\right] \quad (75)$$

for all $g \in \mathrm{dom}(\varphi, \boldsymbol{\tau}_{2,\beta})$ and all $\beta \neq 2$. In particular, every $g \in \mathcal{C}^1(\mathbb{R}^d)$ with at most polynomial growth at infinity lies in this domain.

## 5.3 Power exponential distribution

Consider a $d$ random vector $Z \sim \mathrm{PE}_{d,\zeta}(\nu, \Sigma)$ distributed according to the multivariate power exponential distribution with power $\zeta > 0$, location $\mu$, scale $b$, shape $\Sigma \in \mathbb{R}^d \times \mathbb{R}^d$ and pdf

$$\varphi_\zeta(x) = a_{d,\zeta} \left|\Sigma\right|^{-1/2} \exp\left(-b\left((x-\nu)^T\Sigma^{-1}(x-\nu)\right)^\zeta\right) \tag{76}$$

on $\mathbb{R}^d$ ($a_{d,\zeta}$ is the normalizing constant), $\zeta \in (0, \infty)$, and let $\mu(dx) = \varphi_\zeta(x)dx$ be the corresponding probability measure. Clearly $Z \sim \mathrm{E}_d(\nu, \Sigma, \phi)$ with $\phi(t) = e^{-bt^\zeta}$ so that $\phi'(t)/\phi(t) = -b\zeta t^{\zeta-1}$ and $\phi''(t)/\phi'(t) = -b\zeta t^{\zeta-1} + \frac{\zeta-1}{t}$. From (69), the score function of $\varphi_\zeta$ is

$$\rho_\zeta(x) = -2b\zeta((x-\nu)^T\Sigma^{-1}(x-\nu))^{\zeta-1}\Sigma^{-1}(x-\nu). \tag{77}$$

While, except when $\zeta = 1$,the kernel from (5.3) does not lead to palatable expressions, applying (73) we obtain (for $\zeta \neq 1$)

$$\boldsymbol{\tau}_{2,\zeta}(x) = \frac{\beta + 2\frac{\zeta-1}{b\zeta t^\zeta}}{(\beta-2)(d-1)}\left(\left(1 - \frac{d-1}{\beta b\zeta t^\zeta + 2(\zeta-1)}\right)(x-\nu)^T\Sigma^{-1}(x-\nu)\Sigma - (x-\nu)(x-\nu)^T\right). \tag{78}$$

These functions are Stein kernels : they are in the class $\mathcal{F}(p)$, since $\boldsymbol{\tau}_{2,\zeta}\,\varphi_\zeta \to 0$ when $x \to +\infty$. %eqrefeq:26. We do not provide details here.

Note again that the Stein identity (41) holds for every $g \in \mathcal{C}^1(\mathbb{R}^d)$ with at most polynomial growth at infinity, since in this case we have $\boldsymbol{\tau}_{2,\zeta}\, g\, \varphi_\zeta \to 0$ at infinity.

## 5.4 The multivariate Student $t$-distribution with $k > 1$

Consider a $d$ random vector $X \sim t_k(\nu, \Sigma)$ distributed according to the multivariate Student-$t$ distribution with $k > 1$ degrees of freedom, location $\nu \in \mathbb{R}^d$, shape $\Sigma \in \mathbb{R}^d \times \mathbb{R}^d$ and pdf

$$t_k(x) = c_{k,d}|\Sigma|^{-1/2}\left[1 + \frac{(x-\nu)^T\Sigma^{-1}(x-\nu)}{k}\right]^{-(k+d)/2} \tag{79}$$

with normalizing constant $c_{k,d,\Sigma} = \Gamma((k+d)/2)/(\Gamma(k/2)k^{d/2}\pi^{d/2})$. Let $\mu(dx) = t_k(x)dx$ be the corresponding probability measure. The assumption that $k > 2$ ensures that this distribution has finite mean and finite variance.

This distribution is an elliptical distribution with $\phi(t) = (1 + 2t/k)^{-(k+d)/2}$ and hence $\phi'(t)/\phi(t) = -(k+d)/(k+2t)$ leading to the score function

$$\rho_{t_k}(x) = -(k+d)\frac{\Sigma^{-1}(x-\nu)}{k + (x-\nu)^T\Sigma^{-1}(x-\nu)}. \tag{80}$$

Moreover,

$$\frac{1}{\phi(t)}\int_t^{+\infty}\phi(u)du = \frac{k+2t}{d+k-2}$$

(from $k > 1$ it follows that $d + k > 2$) and hence Lemma 5.3 gives that

$$\boldsymbol{\tau}_1(x) = \frac{(x-\nu)^T\Sigma^{-1}(x-\nu) + k}{d+k-2}\Sigma \tag{81}$$

is a Stein kernel for the multivariate Student distribution for $k > 1$, as then $\boldsymbol{\tau}_1 \in \mathcal{F}(t_k)$.

Similarly, using that $\phi''(t)/\phi'(t) = -(d+k+2)/(k+2t)$, Lemma 73 gives a family of Stein kernels which are indexed by $\beta \in \mathbb{R}$:

$$\begin{aligned}
&\boldsymbol{\tau}_{2,\beta}(x) \\
&= \frac{\beta(d+k)-4}{(d+k)(\beta-2)(d-1)}\left(2\left(\frac{(d-1)(k+2t)}{4-\beta(d+k)}+t\right)\Sigma-(x-\nu)(x-\nu)^T\right) \\
&= \frac{\beta(d+k)-4}{(d+k)(\beta-2)(d-1)}\left(2\left(\frac{(d-1)k+t\left(2(d+1)-\beta(d+k)\right))}{4-\beta(d+k)}\right)\Sigma-(x-\nu)(x-\nu)^T\right).
\end{aligned}$$

It is easy to verify that $\boldsymbol{\tau}_{2,\beta} \in \mathcal{F}(t_k)$. If we choose $\beta$ so that $(d+k)\beta = 2(d+1)$, i.e. $\beta = 2(d+1)/(d+k)$ then, after simplifications, we obtain for $k > 2$

$$\boldsymbol{\tau}_2(x) = \frac{1}{k-1}\left((x-\nu)(x-\nu)^T + k\Sigma\right). \tag{82}$$

Since $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ are of order $t$ for large $|x|$, and since $\phi(t) \underset{t\to+\infty}{\sim} t^{-(k+d)/2}$, the Stein identity (41) holds for every $g \in \mathcal{C}^1(\mathbb{R}^d)$ such that $t^{-(k+d-2)/2}g(x) \to 0$ at infinity. In particular constant functions verify (41) and the Stein kernels are in $\mathcal{F}(p)$. Note that both $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ simplify to $\tau(x) = (x^2 + k\sigma^2)/(k-1)$ when $d = 1$; this last quantity is well-known to be the univariate kernel for the Student-$t$ distribution with $k$ degrees of freedom and centrality parameter $\nu$, see e.g. [56, page 30].

There are several types of operators and identities that can be obtained; below are some examples.

1. **Vector valued operators.** Applying the product rule (29) directly with $f(x) = k+(x-\nu)^T\Sigma^{-1}(x-\nu)$ we obtain for $g : \mathbb{R}^d \to \mathbb{R} \in \mathrm{dom}(f, t_k)$ the operator $\mathcal{A}_{t_k}g(x) = (k+(x-\nu)^T\Sigma^{-1}(x-\nu))\nabla g(x) + (2-k-d)\Sigma^{-1}(x-\nu)g(x)$. Taking expectations for $X \sim t_k(\nu, \Sigma)$ we obtain the vector-Stein identity

$$\mathbb{E}\left[(k+(X-\nu)^T\Sigma^{-1}(X-\nu))\nabla g(X)\right] = (k+d-2)\mathbb{E}\left[\Sigma^{-1}(X-\nu)g(X)\right] \tag{83}$$

   By definition of the Stein kernel we also get new Stein operators and identities. Using $\boldsymbol{\tau}_1$ recovers (83), whereas using $\boldsymbol{\tau}_2$ we obtain

$$\mathbb{E}\left[((X-\nu)(X-\nu)^T + k\Sigma)\nabla g(X)\right] = (k-1)\mathbb{E}\left[(X-\nu)g(X)\right] \tag{84}$$

   (still with $X \sim t_k(\nu, \Sigma)$).

2. **Scalar valued operators.** Suppose for simplicity that $\Sigma = I_d$ and $\nu = 0$. Taking $\mathbf{B}$ successively equal to $\boldsymbol{\tau}_1$ then $\boldsymbol{\tau}_2$ in (44) leads to

$$\begin{aligned}
\mathcal{B}_1 g(x) &= -\langle \nabla g(x), x\rangle + \frac{1}{d+k-2}\left\langle x^T x + 2k, \nabla^2 g(x)\right\rangle_{\mathrm{HS}} \\
\mathcal{B}_2 g(x) &= -\langle \nabla g(x), x\rangle + \frac{1}{k-1}\left\langle xx^T + kI_d, \nabla^2 g(x)\right\rangle_{\mathrm{HS}}
\end{aligned}$$

   acting on functions $g$ such that $\nabla g \in \mathcal{F}_1(t_k)$.

3. **A covariance identity.** Starting from (84) with $X = (X_1, X_2)^T \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2}$ multivariate student with location $(\nu_1, \nu_2)$ and shape $\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$, taking $g(x) = g(x_2)$ and considering only the first $d_1$ components of the resulting identity we obtain

$$\mathbb{E}\left[((X_1-\nu_1)(X_2-\nu_2)^T + k\Sigma_{12})\nabla g(X_2)\right] = (k-1)\mathbb{E}\left[(X_1-\nu_1)g(X_2).\right] \tag{85}$$

   Many more such covariance identities can be obtained by this approach, thus complementing those obtained in [1].

# 6 Generalities on Stein kernels

Let $\mathcal{T}_{\mathrm{div},p}$ be the canonical Stein operator (34) acting on $\mathcal{F}(p)$, the corresponding Stein class. In this section we explore properties of the Stein kernels from Definition 3.8.

**Proposition 6.1** (Properties)**.** *Let $\tau_{p,i}$ be a Stein kernel for $p$ in the direction $e_i$ and $\boldsymbol{\tau}_p$ the matrix with $i^{th}$ raw being $\tau_{p,i}$. Then*

1. *For all $j = 1, \dots, d$ we have*

$$\frac{\partial}{\partial x_j}\left(\int_{\mathbb{R}^d} \tau_{p,i}(x)p(x)dx\right) = \int_{\mathbb{R}^d} \frac{\partial}{\partial x_j}(\tau_{p,i}(x)p(x))dx = 0.$$

2. *If $p$ admits a second moment, and if $x - \nu \in \mathrm{dom}(p, \boldsymbol{\tau}_p)$, then*

$$\mathbb{E}\left[\boldsymbol{\tau}_p(X)\right] = \mathrm{Var}(X).$$

*Proof.* The first statement follows by the requirement that the kernel belongs to $\mathcal{F}(p)$, which in particular imposes that all components of $\tau_{p,i}$ belong to $\mathcal{F}_1(p)$. To see the second claim, taking expectations in (37) yields that the Stein kernel necessarily satisfies

$$\mathbb{E}\left[\boldsymbol{\tau}_p(X)\nabla g(X)\right] = \mathbb{E}\left[(X-\nu)g(X)\right] \tag{86}$$

for all $g : \mathbb{R}^d \to \mathbb{R}$ belonging to $\mathrm{dom}(p, \boldsymbol{\tau}_p)$. By assumption, $g_i(x) = x_i - \nu_i$ belongs to $\mathrm{dom}(p, \boldsymbol{\tau}_p)$ for all $i = 1, \dots, d$. Plugging these functions in (86) leads to

$$\mathbb{E}\left[(\boldsymbol{\tau}_{p,i}(X))_j\right] = \mathbb{E}\left[(X_i - \nu_i)(X_j - \nu_j)\right]$$

for all $i, j = 1, \dots, d$. The claim follows. □

**Proposition 6.2.** *Given $k \le d$ and $\{i_1, \dots, i_k\} \subset \{1, \dots, n\}$ denote by $\mathcal{V} = \langle e_{i_1}, \dots, e_{i_k}\rangle$ the space generated by $e_{i_1}, \dots, e_{i_k}$. Also, write any $x \in \mathbb{R}^d$ as $x = (x^{\mathcal{V}}, x^{\mathcal{V}^\perp})$ and let $p_{\mathcal{V}} = \int_{\mathcal{V}^\perp} p$ be the marginal of $p$ on $\mathcal{V}$. Suppose that $p$ admits a $p$-integrable Stein kernel $\tau_{i_j} = \left(\tau_{i_j,1}, \dots, \tau_{i_j,d}\right)$ in each direction $e_{i_1}, \dots, e_{i_k}$. Then the vector $\tau^{\mathcal{V}}_{i_j} = \left(\tau^{\mathcal{V}}_{i_j,1}, \dots, \tau^{\mathcal{V}}_{i_j,k}\right)$ with components*

$$\tau^{\mathcal{V}}_{i_j,\ell}(x^{\mathcal{V}}) = \mathbb{E}\left[\tau_{i_j,i_\ell}(X) \,|\, X^{\mathcal{V}} = x^{\mathcal{V}}\right], \quad \ell = 1, \dots, k \tag{87}$$

*is a Stein kernel for $p_{\mathcal{V}}$ in the direction $e_{i_j}$.*

*Proof.* Without loss of generality we suppose that $p$ is centered. Fix $x^{\mathcal{V}} \in \mathcal{V}$. Then

$$\begin{aligned}
&\sum_{\ell=1}^k \frac{\partial}{\partial_{i_\ell}}\left(\tau^{\mathcal{V}}_{i_j,\ell}(x^{\mathcal{V}})p_{\mathcal{V}}(x^{\mathcal{V}})\right)\\
&= \sum_{\ell=1}^k \frac{\partial}{\partial_{i_\ell}}\left(\int_{\mathcal{V}^\perp} \tau_{i_j,i_\ell}(x^{\mathcal{V}}, x^{\mathcal{V}^\perp})p(x^{\mathcal{V}}, x^{\mathcal{V}^\perp})dx^{\mathcal{V}^\perp}\right)\\
&= \sum_{\ell=1}^k \int_{\mathcal{V}^\perp} \frac{\partial}{\partial_{i_\ell}}\left(\tau_{i_j,i_\ell}(x^{\mathcal{V}}, x^{\mathcal{V}^\perp})p(x^{\mathcal{V}}, x^{\mathcal{V}^\perp})\right)dx^{\mathcal{V}^\perp}\\
&= -x_{i_j}\int_{\mathcal{V}^\perp} p(x^{\mathcal{V}}, x^{\mathcal{V}^\perp})dx^{\mathcal{V}^\perp} + \sum_{m\in\{1,\dots,n\}\setminus\{i_1,\dots,i_k\}} \int_{\mathcal{V}^\perp} \frac{\partial}{\partial_m}\left(\tau_{i_j,m}(x^{\mathcal{V}}, x^{\mathcal{V}^\perp})p(x^{\mathcal{V}}, x^{\mathcal{V}^\perp})\right)dx^{\mathcal{V}^\perp}
\end{aligned}$$

where the second-last line is allowed thanks to integrability of the Stein kernel and the last follows from (36), giving the identity

$$\sum_{m=1}^{d} \frac{\partial}{\partial_m} \left( \tau_{i_j,m}(x^{\mathcal{V}}, x^{\mathcal{V}^\perp}) p(x^{\mathcal{V}}, x^{\mathcal{V}^\perp}) \right) = -x_{i_j} p(x^{\mathcal{V}}, x^{\mathcal{V}^\perp})$$

which is valid for any $x^{\mathcal{V}^\perp}$ such that $(x^{\mathcal{V}}, x^{\mathcal{V}^\perp})$ lies in the support of $\Omega_p$. Now using the requirement $\tau_{i_j} \in \mathcal{F}(p)$ ensures that $\int_{\mathcal{V}^\perp} \frac{\partial}{\partial_m} \left( \tau_{i_j,m}(x^{\mathcal{V}}, x^{\mathcal{V}^\perp}) p(x^{\mathcal{V}}, x^{\mathcal{V}^\perp}) \right) dx^{\mathcal{V}^\perp} = 0$ for all $m \notin \{i_1, \ldots, i_k\}$ so that

$$\sum_{\ell=1}^{k} \frac{\partial}{\partial_{i_\ell}} \left( \tau^{\mathcal{V}}_{i_j,\ell}(x^{\mathcal{V}}) p_{\mathcal{V}}(x^{\mathcal{V}}) \right) = -x_{i_j} p_{\mathcal{V}}(x),$$

as required. It remains to check that $\tau^{\mathcal{V}}_{i_j,\ell}(x^{\mathcal{V}}) \in \mathcal{F}(p_{\mathcal{V}})$, but this is a direct consequence of the definitions. □

Next, we provide formulas for computing Stein kernels explicitly based on univariate Stein kernels.

**Proposition 6.3** (Bivariate Stein kernels)**.** *Let $X = (X_1, X_2)^T \sim p$ with $p$ a continuous pdf on $\mathbb{R}^2$ satisfying Assumption A. Let $p_1$ be the marginal of $p$ in direction $e_1$, $\rho_1$ the corresponding univariate score and $\tau_1$ the corresponding univariate kernel (which we suppose to exist). Set $\tau_{11}(x_1, x_2) = \tau_1(x_1)$, and*

$$\tau_{12}(x_1, x_2) = \tau_1(x_1) \frac{p_1(x_1)}{p(x_1, x_2)} \partial_1 \left( \int_{x_2}^{\infty} p(x_1, v) dv / p_1(x_1) \right). \tag{88}$$

*Then the vector $(x_1, x_2) \mapsto (\tau_1(x_1, x_2), \tau_{12}(x_1, x_2))_{1 \le i,j \le 2}$ is a Stein kernel for $p$ in the direction $e_1$.*

Note that

$$\begin{aligned}
&\tau_1(x_1) \frac{p_1(x_1)}{p(x_1, x_2)} \partial_1 \left( \int_{x_2}^{\infty} p(x_1, v) dv / p_1(x_1) \right) \\
&= \tau_1(x_1) \frac{p_1(x_1)}{p(x_1, x_2)} \left( -\frac{p_1'(x_1)}{(p_1(x_1))^2} \int_{x_2}^{\infty} p(x_1, v) dv + \frac{1}{p_1(x_1)} \int_{x_2}^{\infty} \partial_1 p(x_1, v) dv \right) \\
&= \frac{1}{p(x_1, x_2)} \tau_1(x_1) \int_{x_2}^{\infty} \big( -\rho_1(x_1) p(x_1, v) + \partial_1 p(x_1, v) \big) dv.
\end{aligned} \tag{89}$$

This alternative form of (88) is often more convenient than (88).

*Proof.* We need to prove that

$$\sum_{j=1}^{2} \partial_j (\tau_{1j}(x) p(x)) = -(x_1 - \nu_1) p(x) \tag{90}$$

for all $x = (x_1, x_2)^T \in \mathbb{R}^2$. Applying (89) we have

$$\begin{aligned}
\tau_{12}(x) p(x_1, x_2) &= p_1(x_1) \tau_1(x_1) \partial_1 \left( \int_{x_2}^{\infty} p(x_1, v) dv / p_1(x_1) \right) \\
&= -\tau_1(x_1) \frac{p_1'(x_1)}{p_1(x_1)} \int_{x_2}^{\infty} p(x_1, v) dv + \tau_1(x_1) \int_{x_2}^{\infty} \partial_1 p(x_1, v) dv
\end{aligned} \tag{91}$$

so that

$$\partial_2 \left(\tau_{12}(x)p(x_1,x_2)\right)) = \tau_1(x_1)\frac{p_1'(x_1)}{p_1(x_1)}p(x_1,x_2) - \tau_1(x_1)\partial_1 p(x_1,x_2). \tag{92}$$

Also, as $\partial_1(p_1(x_1)\tau_1(x_1)) = (\nu_1 - x_1)p_1(x_1)$,

$$\begin{aligned}
\partial_1(\tau_1(x)p(x_1,x_2)) &= \partial_1\left(p_1(x_1)\tau_1(x_1)\frac{p(x_1,x_2)}{p_1(x_1)}\right) \\
&= \partial_1(p_1(x_1)\tau_1(x_1))\frac{p(x_1,x_2)}{p_1(x_1)} + p_1(x_1)\tau_1(x_1)\partial_1\left(\frac{p(x_1,x_2)}{p_1(x_1)}\right) \\
&= -(x_1-\nu_1)p(x_1,x_2) + p_1(x_1)\tau_1(x_1)\left(-\frac{p_1'(x_1)}{(p_1(x_1))^2}p(x_1,x_2) + \frac{\partial_1 p(x_1,x_2)}{p_1(x_1)}\right) \\
&= -(x_1-\nu_1)p(x_1,x_2) - \tau_1(x_1)\frac{p_1'(x_1)}{p_1(x_1)}p(x_1,x_2) + \tau_1(x_1)\partial_1 p(x_1,x_2).
\end{aligned} \tag{93}$$

Adding up (92) and (93) we get (90). □

**Remark 6.4.** *The proof of Proposition 6.3 is of a purely computational nature. The inspiration for formula* (88) *is [11, equation (9)], where a similar quantity is introduced via a transport argument. To see the connection, note that* (89) *gives*

$$\tau_{12}(x_1,x_2)p(x_1,x_2) = \tau_1(x_1)\int_{-\infty}^{x_2}\left(\rho_1(x_1)p(x_1,v) - \partial_1 p(x_1,v)\right)dv. \tag{94}$$

*We introduce* $p^{X\,|\,X_i=x_i}(x_1,x_2) = p(x_1,x_2)/p_i(x_i)$ *the conditional density of* $X$ *at* $X_i = x_i$*. Fix* $i=1$ *and, for each* $t,t',x_2$ *let* $x_2 \mapsto T_{t,t'}(x_2)$ *be the mapping transporting the conditional density at* $x_1 = t$ *to that at* $x_1 = t'$*, implicitly defined via*

$$p^{X\,|\,X_1=t}(x_2) = p^{X\,|\,X_1=t'}\big(T_{t,t'}(x_2)\big)\partial_{x_2}T_{t,t'}(x_2). \tag{95}$$

*Taking derivatives in* (95) *with respect to* $t'$ *and setting* $t' = t = x_1$ *we deduce (using the fact that* $T_{t,t}(x_2) = x_2$*) that*

$$\begin{aligned}
0 = \frac{\partial_1 p(x_1,x_2)}{p_1(x_1)} &+ \frac{\partial_2 p(x_1,x_2)}{p_1(x_1)}\partial_{t'}T_{t,t'}(x_2)\big|_{t'=t=x_1} - \frac{p(x_1,x_2)}{p_1(x_1)}\frac{p_1'(x_1)}{p_1(x_1)} \\
&+ \frac{p(x_1,x_2)}{p_1(x_1)}\partial_2(\partial_{t'}T_{t,t'}(x_2))\big|_{t'=t=x_1},
\end{aligned}$$

*that is,*

$$\partial_2\left(p(x_1,x_2)\partial_{t'}T_{t,t'}(x_2)\big|_{t'=t=x_1}\right) = p(x_1,x_2)\frac{p_1'(x_1)}{p(x_1)} - \partial_1 p(x_1,x_2)$$

*and we recognize from* (94) *that, up to a function which depends only on* $x_1$*,*

$$\frac{\tau_{12}(x_1,x_2)}{\tau_1(x_1)} = \partial_{t'}T_{t,t'}(x_2)\big|_{t'=t=x_1}. \tag{96}$$

*This is not the only Stein kernel in connection with transport maps, see [35] where yet another construction is introduced.*

Further, inspired by [7, 8], we can directly postulate our next result which guarantees existence of Stein kernels under smoothness conditions.

**Theorem 6.5.** *Let $p : \mathbb{R}^d \to (0, \infty)$ be a continuously twice differentiable density on $\mathbb{R}^d$ with*

$$\int \frac{\|\nabla p\|^2}{p}, \quad \int \|\nabla^2(p)\| < \infty.$$

*Let $\tau_i^{(1)}, i = 1, \dots, d$ be the marginal Stein kernels. Then, for any direction $e_i, i = 1, \dots, d$ there exists a Stein kernel for $p$ in direction $e_i$*

$$\tau_{p,i}^{(d)}(x) = \tau_i^{(1)}(x_i) \left( \tau_{i,1}^{(d)}(x \,|\, x_i) \quad \cdots \quad \tau_{i,i-1}^{(d)}(x \,|\, x_i) \quad 1 \quad \tau_{i,i+1}^{(d)}(x \,|\, x_i) \quad \cdots \quad \tau_{i,d}^{(d)}(x \,|\, x_i) \right)$$

*with coefficients $\tau_{i\bullet}, i = 1, \dots, d$ solving the equations*

$$\mathcal{T}_{\mathrm{div},p}(\tau_{i\bullet}(x \,|\, x_i)) = \rho_i(x_i). \tag{97}$$

*Here $\rho_i(x_i) = p_i'(x_i)/p_i(x_i)$ is the score function of the $i$th marginal and $x = (x_1, \dots, x_d)$.*

*Proof.* The result is almost immediate from [7, Theorem 4] where it is proved (see middle of page 978) that, under the stated conditions, there exist continuously differentiable vector fields $R_h$ such that

$$\frac{\mathrm{div}(R_h(x)p(x))}{p(x)} = \frac{h'(u)}{h(u)}$$

for any marginal $u \mapsto h(u)$ of $p$, in any direction. Collecting these into a single vector and adapting the notations leads to (97). To see the connection with Stein kernels, write

$$\tau_{ij}(x) = \tau_i(x_i)\tau_{ij}(x \,|\, x_i). \tag{98}$$

Then

$$\begin{aligned}
\sum_{j=1}^d \partial_j(\tau_{ij}(x)p(x)) &= \sum_{j=1}^d \partial_j(\tau_{ij}(x \,|\, x_i)p(x)\tau_i(x_i)) \\
&= \sum_{j=1}^d \partial_j(\tau_{ij}(x \,|\, x_i)p(x))\tau_i(x_i) + \sum_{j=1}^d \tau_{ij}(x \,|\, x_i)p(x)\partial_j(\tau_i(x_i)) \\
&= \rho_i(x_i)p(x)\tau_i(x_i) + p(x)\partial_i(\tau_i(x_i))
\end{aligned}$$

where in the last line we use (97) in the first sum and $\partial_j(\tau_i(x_i)) = 0$ for all $j \neq i$ in the second sum. Clearly by the definition of the univariate Stein kernel

$$\partial_i(\tau_i(x_i)) = -\rho_i(x_i)\tau_i(x_i) + \mathbb{E}[X_i] - x_i$$

and the claim follows. □

**Remark 6.6.** *Theorem 6.5 gives a mechanism allowing to generalize the bivariate construction from Proposition 6.3. Under the conditions of Theorem 6.5, for $d = 3$, we*

*set* $p_{ij}(x_i,x_j)=\int_{-\infty}^{\infty}p(x)dx_k$ *and* $P_i(x)=\int_{-\infty}^{x}p_i(v)dv, i=1,2,3$. *Then we can choose* $\tau_{i,i}=\tau_i$ *and*

$$\begin{aligned}\tau_{1,2}^{(3)}(x\,|\,x_1)p(x_1,x_2,x_3)&=\int_{-\infty}^{x_2}\left(\rho_1(x_1)p(x_1,v,x_3)-\partial_1 p(x_1,v,x_3)\right)dv\\&\quad-P_2(x_2)\big\{\rho_1(x_1)p_{13}(x_1,x_3)-\partial_1 p_{13}(x_1,x_3)\big\}\\\tau_{1,3}^{(3)}(x\,|\,x_1)p(x_1,x_2,x_3)&=p_2(x_2)\int_{-\infty}^{x_3}\left(\rho_1(x_1)p_{13}(x_1,w)-\partial_1 p_{13}(x_1,w)\right)dw\end{aligned}$$

*and similarly for* $\tau_{i,j}^{(3)}(x\,|\,x_i)$ *for all* $i,j$. *Direct computations suffice for this claim. Also, setting*

$$\tau_{12}^{(2)}(x_1,x_2)=\tau_1^{(1)}(x_i)\mathbb{E}[\tau_{12}^{(3)}(x\,|\,x_1)\,|\,X_1=x_1,X_2=x_2],\tag{99}$$

*the vector* $(\tau_1^{(1)}(x_1),\tau_{12}^{(2)}(x_1,x_2))$ *forms a bivariate Stein kernel for* $(X_1,X_2)$. *Moreover,*

$$\begin{aligned}\int_{-\infty}^{+\infty}\tau_{1,2}^{(3)}(x\,|\,x_1)\frac{p(x_1,x_2,x_3)}{p(x_1,x_2)}dx_3&=\frac{1}{p(x_1,x_2)}\int_{-\infty}^{x_2}\left(\rho_1(x_1)p(x_1,v)-\partial_1 p(x_1,v)\right)dv\\&\quad-P_2(x_2)\big\{\rho_1(x_1)p_1(x_1)-\partial_1 p_1(x_1)\big\}\\&=\frac{1}{p(x_1,x_2)}\int_{-\infty}^{x_2}\left(\rho_1(x_1)p(x_1,v)-\partial_1 p(x_1,v)\right)dv\end{aligned}$$

*which is equivalent to the expression* (89).

*The* $k$*-variate extension can also be constructed, as follows. For all* $k\geq 1$, *and under the same conditions, for all* $1\leq j\leq d-1$ *we can define*

$$\begin{aligned}\tau_{1,j}^{(k)}(x\,|\,x_1)p(x)&=\int_{-\infty}^{x_j}\left(\rho_1(x_1)p(x_1,x_j=v,\ldots,x_d)-\partial_1 p(x_1,x_j=v,\ldots,x_d)\right)dv\\&\quad-P_j(x_j)\left(\rho_1(x_1)p(x_1,x_{j+1}\ldots,x_d)-\partial_1 p(x_1,x_{j+1},\ldots,x_d)\right)\end{aligned}\tag{100}$$

*and for* $j=d$

$$\tau_{1,d}^{(k)}(x\,|\,x_1)p(x)=P_{d-1}(x_{d_1})\int_{-\infty}^{x_d}\left(\rho_1(x_1)p(x_1,x_d=v)-\partial_1 p(x_1,x_d=v)\right)dv\tag{101}$$

**Example 6.7** (Multivariate Gaussian)**.** *If* $X\sim\mathcal{N}_2(\nu,\Sigma)$ *is multivariate* $d$*-dimensional Gaussian then direct computations of the kernel as provided by Proposition 6.3 leads to* $\boldsymbol{\tau}_2$ *given in* (82). *The expression is more complicated in dimension* $d\geq 3$, *and so far we have not been able to give a probabilistic interpretation of it.*

**Example 6.8.** *If* $X=(X_1,X_2)^T\sim t_k(\nu,\Sigma)$ *follows the bivariate Student distribution then direct computations of the kernel as provided by Proposition 6.3 leads to* $\boldsymbol{\tau}_2$ *given in* (82). *Again, we have not been able to give a probabilistic interpretation of the expression in dimension* $d\geq 3$.

# 7 Stein discrepancies

Instead of using the Wasserstein metric which uses Lipschitz functions, more general classes functions $\mathcal{G}$ in $\sup_{g\in\mathcal{G}}|E[\mathcal{A}_p g(Y)]|$ can be useful to assess distributional distances, leading to the notion of *Stein discrepancies*.

## 7.1 Integral probability metrics and Stein discrepancies

Differences between distributions can be measured using probability metrics. For applying Stein's method, so-called *integral probability metrics* are well suited.

**Definition 7.1** (Integral Probability Metrics)**.** *Let $\mathbb{F}(\mathbb{R}^d)$ be a collection of cumulative distribution functions on $\mathbb{R}^d$ and denote $L^1(\mathbb{F}(\mathbb{R}^d))$ the class of Borel measurable functions $h : \mathbb{R}^d \to \mathbb{R}$ (for some $q \geq 1$) such that $F(|h|) = \int |h| dF < \infty$ for all $F \in \mathbb{F}(\mathbb{R}^d)$. A metric on $\mathbb{F}(\mathbb{R}^d)$ is an* integral probability metric *(IPM) if it can be written in the form*

$$d_{\mathcal{H}}(F, G) := \sup_{h \in \mathcal{H}} |F(h) - G(h)| \tag{102}$$

*for some class of real-valued bounded measurable test functions $\mathcal{H} \subset L^1(\mathbb{F}(\mathbb{R}^d))$ ($|\cdot|$ is the Euclidean norm). The expression on the right-hand side of* (102) *is called an* IPM-discrepancy.

Many important probability metrics can be represented as integral probability metrics; classical references are [81, 38]. The Wasserstein distance between $X$ and $Y$, which we have already used in this paper, takes $\mathcal{H} = \mathcal{W}$ the collection of Lipschitz functions $h : \mathbb{R}^d \to \mathbb{R}$ with Lipschitz constant smaller than 1. The Kolmogorov distance between two random vectors $X \sim F$ and $Y \sim G$ is $\mathrm{Kol}(X, Y) = d_{\mathcal{H}_{\mathrm{Kol}}}(F, G)$ with $\mathcal{H}_{\mathrm{Kol}} = \left\{\mathbb{I}_{(-\infty, z]}, z \in \mathbb{R}^d\right\}$. The Wasserstein distance between $X$ and $Y$ takes $\mathcal{H} = \mathcal{W}$ the collection of Lipschitz functions $h : \mathbb{R}^d \to \mathbb{R}$ with Lipschitz constant smaller than 1. The total variation distance takes $\mathcal{H}_{\mathrm{TV}}$ the collection of Borel measurable functions $h : \mathbb{R}^d \to [0, 1]$. For other examples and references see [62, Appendix E].

In Definition 7.1 the Euclidean norm $|\cdot|$ is used, but the definition generalises easily to other norms $\|\cdot\|$ as long as

$$d_{\mathcal{H}}(F, G) := \sup_{h \in \mathcal{H}} \|F(h) - G(h)\| \tag{103}$$

defines a distance between probability distributions. This intuition leads to the following general definition.

**Definition 7.2** (Stein discrepancy)**.** *Let $p$ be a density on $\mathbb{R}^d$ and $\mathcal{A}_p$ a Stein operator acting on some class* $\mathrm{dom}(\mathcal{A}_p)$*. Then for any random $Y \sim q$, any $\mathcal{G} \subset \mathrm{dom}(\mathcal{A}_p)$ and any norm $\|\cdot\|$ , the quantity*

$$\mathcal{S}_{\|\cdot\|}(q, \mathcal{A}_p, \mathcal{G}) = \sup_{g \in \mathcal{G}} \|\mathbb{E}\left[\mathcal{A}_p g(Y)\right]\| \tag{104}$$

*is the ($\|\cdot\| - \mathcal{G} - \mathcal{A}_p$)* Stein discrepancy *from $Y$ to $X$.*

Definition 7.2 is motivated by the reference [44] where, to the best of our knowledge, such a unified notation for general Stein-based discrepancies (with freedom of choice both in the operator and the class of functions) is first introduced.

The choice of norm $\|\cdot\|$ is generally fixed by context such as dimensionality, basic properties of the operator and the random variables $X, Y$ under study. In the sequel we will generally drop the indexation in the norm and simply write $\mathcal{S}(Y, \mathcal{A}_p, \mathcal{G})$ instead. The next subsection links Stein discrepancies to information metrics.

## 7.2 Information metrics and kernelized Stein discrepancies

In principle, the Stein discrepancy (7.2) can be used as a basis for goodness-of-fit tests, and could be estimated numerically. In high-dimensional problems, the class $\mathcal{G}$ is often too large to allow numerical integration. In high-dimensional goodness of fit tests, restricting the class of functions to a ball in a reproducing kernel Hilbert space associated with a positive definite kernel $k(x, x')$ has been shown in [26, 57] to be an efficient way of estimating Stein discrepancies. In this context these discrepancies are called *kernelized Stein discrepancies*, with the kernel $k(x, x')$ in mind. The framework of the present paper shows how to generalise their approach, as follows.

Let $X \sim p$ and $Y \sim q$ be two random variables on $\mathbb{R}^d$ with differentiable densities and respective Stein classes $\mathcal{F}(p)$ and $\mathcal{F}(q)$. Suppose, for simplicity, that both share the same mean and the same support $\mathcal{S}$, satisfying Assumption A. Fix $d \times d$ matrix valued functions $\mathbf{A}_p \in \mathcal{F}(p)$ and $\mathbf{A}_q \in \mathcal{F}(q)$, set $\mathbf{a}_p = \mathcal{T}_{\mathrm{div},p}(\mathbf{A}_p)$, $\mathbf{a}_q = \mathcal{T}_{\mathrm{div},q}(\mathbf{A}_q)$ and introduce the divergence based vector valued standardizations

$$\mathcal{A}_p g(x) = \mathbf{A}_p(x)\nabla g(x) + \mathbf{a}_p(x)g(x), \quad g : \mathbb{R}^d \to \mathbb{R} \in \mathcal{F}(\mathcal{A}_p) \tag{105}$$

$$\mathcal{A}_q g(x) = \mathbf{A}_q(x)\nabla g(x) + \mathbf{a}_q(x)g(x), \quad g : \mathbb{R}^d \to \mathbb{R} \in \mathcal{F}(\mathcal{A}_q) \tag{106}$$

as in (35).

The Stein heuristic that if $p$ and $q$ are close, then $\mathbb{E}_{q\otimes q}[\mathcal{A}_{p\otimes p}k(Y, Y')]$ should be small still holds, where $\mathcal{A}_{p\otimes p}$ is the concatenated operator $\mathcal{A}_p$ operating on functions $g : \mathbb{R}^{2d} \to \mathbb{R}$ by treating the first $d$ and the last $d$ components independently. It turns out that iterating the operator is a more elegant way to obtain a kernelized discrepancy. Write $\mathcal{A}_p^T$ for the transpose of the operator $\mathcal{A}_p$ in the matrix transpose sense. Then for any positive definite symmetric kernel $k$ with marginals in $\mathcal{F}(p)$, using (105),

$$\begin{aligned}\mathcal{A}_p^T\mathcal{A}_p k(x, x') = {} & \mathbf{A}_p(x)^T\nabla_x^T\mathbf{A}_p(x')\nabla_{x'}k(x, x') + \mathbf{A}_p(x)^T\nabla_x^T\mathbf{a}_p(x')k(x, x') \\ & + \mathbf{a}_p(x)^T\mathbf{A}_p(x')\nabla_{x'}k(x, x') + \mathbf{a}_p(x)^T\mathbf{a}_p(x')k(x, x').\end{aligned} \tag{107}$$

By conditioning on $X$ it is easy to see that

$$\mathbb{E}_{p\otimes p}[\mathcal{A}_p^T\mathcal{A}_p k(X, X')] = 0.$$

The operator (107) has been used in [26] for the particular choice $\mathbf{A}_p(x) = I_d$ for which $\mathbf{a}_p = \rho_p$, the score operator. In this case, evaluating (107) does not require knowledge of the normalising constant for the density $p$ and is hence particularly attractive for applications in Bayesian inference. Equation (107) motivates our general definition of kernelized Stein discrepancies. We use the convention that $\mathcal{L}_i, i = 1, 2$ denotes the operator $\mathcal{L}$ applied with respect to the $i$th variable of the function $k(\cdot, \cdot)$.

**Definition 7.3** (Kernelized Stein discrepancies)**.** *Let $\mathcal{A}_p$ (resp., $\mathcal{A}_p$) be a Stein operator for $p$ (resp., for $q$) with class $\mathcal{F}(p)$ (resp., $\mathcal{F}(q)$). Let $k$ be some kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that $y \mapsto k(y, \cdot)$ and $y' \mapsto k(\cdot, y')$ belong to $\mathcal{F}(\mathcal{A}_p) \cap \mathcal{F}(\mathcal{A}_p)$. The $k$-kernelized Stein discrepancy from $p$ to $q$ is*

$$\mathcal{S}(p, q, k) = \mathbb{E}\left[\mathcal{A}_{p,1}^T\mathcal{A}_{p,2}k(Y, Y')\right]. \tag{108}$$

*When $X \sim p$ and $Y \sim q$, in abuse of notation we also write $\mathcal{S}(X, Y, k) = \mathcal{S}(p, q, k)$.*

Our set-up allows for a combination of the operators from (105) and (106) which is sometimes more suitable for the problem at hand. All classes of functions are designed to ensure that, for all $g \in \mathcal{G} = \mathcal{F}(\mathcal{A}_p) \cap \mathcal{F}(\mathcal{A}_q)$, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{A}_p g(Y)\right] = \mathbb{E}\left[\mathcal{A}_p g(Y)\right] - \mathbb{E}\left[\mathcal{A}_q g(Y)\right] \\
&= \mathbb{E}\left[\left(\mathbf{A}_p(Y) - \mathbf{A}_q(Y)\right)\nabla g(Y)\right] + \mathbb{E}\left[\left(\mathbf{a}_p(Y) - \mathbf{a}_q(Y)\right)g(Y)\right] \\
&=: \mathbb{E}\left[\mathbf{A}_{p/q}(Y)\nabla g(Y)\right] + \mathbb{E}\left[\mathbf{a}_{p/q}(Y)g(Y)\right]. \qquad (109)
\end{aligned}$$

Thus in particular we can take

$$\mathbf{A}_{p/q} \text{ and } \mathbf{a}_{p/q} \text{ in (105)} \qquad (110)$$

$$-\mathbf{A}_{p/q} \text{ and } -\mathbf{a}_{p/q} \text{ in (106).} \qquad (111)$$

Two particular choices of input matrices $\mathbf{A}_p$ and $\mathbf{A}_q$ stand out:

- $\mathbf{A}_p(x) = \boldsymbol{\tau}_p(x)$ and $\mathbf{A}_q(x) = \boldsymbol{\tau}_q(x)$ for which $\mathbf{a}_{p/q} = 0$ and (109) becomes

$$\mathbb{E}\left[\mathcal{A}_p g(Y)\right] = \mathbb{E}\left[\left(\boldsymbol{\tau}_p(Y) - \boldsymbol{\tau}_q(Y)\right)\nabla g(Y)\right] =: E\left[\boldsymbol{\tau}_{p/q}(Y)\nabla g(Y)\right] \qquad (112)$$

- $\mathbf{A}_p(x) = \mathbf{A}_q(x) = I_d$ for which $\mathbf{A}_{p/q} = 0$ and $\mathbf{a}_{p/q} = \rho_p - \rho_q$ and (109) becomes

$$\mathbb{E}\left[\mathcal{A}_p g(Y)\right] = \mathbb{E}\left[\left(\rho_p(Y) - \rho_q(Y)\right)g(Y)\right] =: \mathbb{E}\left[\rho_{p/q}(Y)g(Y)\right]. \qquad (113)$$

The Stein operators (105) and (106) can also be applied jointly to functions $k(x, x')$ with marginals in $\mathcal{G}$.

$$\begin{aligned}
\mathcal{A}_{p,1}^T \mathcal{A}_{p,2} k(x, x') = {} & \mathbf{A}_p^T(x)\mathbf{A}_q(x')\nabla_x^T \nabla_{x'} k(x, x') + A_p^T(x)\mathbf{a}_q(x')\nabla_x^T k(x, x') \\
& + \mathbf{a}_p(x)^T \nabla_{x'} \mathbf{A}_q(x') k(x, x') + \mathbf{a}_p(x)^T \mathbf{a}_q(x') k(x, x').
\end{aligned}$$

In [26], such kernelized expressions are studied for the score function choice (113). Inspired by [26, 57] we give the following result which follows immediately from (109) with the choices (110) and (111). This result shows that kernelized methods have a larger range of applicability than usually assumed. At the same time, it illustrates the power of our general set-up.

**Theorem 7.4.** *Let $Y, Y'$ be* independently *drawn from $q$ on the same space and consider functions $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that both $y \mapsto k(y, \cdot)$ and $y' \mapsto k(\cdot, y')$ belong to $\mathcal{F}(\mathcal{A}_p) \cap \mathcal{F}(\mathcal{A}_q)$. Then*

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{A}_{p,1}^T \mathcal{A}_{p,2} k(Y, Y')\right] \\
&= \mathbb{E}\left[\nabla_y^T \left(\mathbf{A}_{p/q}(Y)\mathbf{A}_{p/q}(Y')\right)\nabla_{y'} k(y, y')|_{y=Y, y'=Y'}\right] \\
&\quad + \mathbb{E}\left[\nabla_y^T \left(\mathbf{A}_{p/q}(Y)\mathbf{a}_{p/q}(Y')k(y, Y')\right)|_{y=Y}\right] + \mathbb{E}\left[\mathbf{a}_{p/q}(Y)^T \nabla_{y'}\left(\mathbf{A}_{p/q}(Y)k(y, Y')\right)|_{y=Y}\right] \\
&\quad + \mathbb{E}\left[\mathbf{a}_{p/q}(Y)^T k(Y, Y')\mathbf{a}_{p/q}(Y')\right] \qquad (114)
\end{aligned}$$

*with $\mathcal{L}_i, i = 1, 2$ the operator $\mathcal{L}$ applied with respect to the $i$th variable of the function $k(\cdot, \cdot)$.*

Theorem 7.4, with the particular choice that $\mathbf{A}_p(x) = \mathbf{A}_q(x) = I_d$, so that $\mathbf{A}_{p/q} = 0$ and $\mathbf{a}_{p/q} = \rho_p - \rho_q$, was proposed in [57]. If both $p$ and $q$ are known up to a normalising constant, then (114) does not depend on the normalising constant and hence, again, is particularly attractive in high-dimensional Bayesian inference.

From (112) and (113) we introduce the following two special cases of particular interest.

1. If $\mathbf{A}_p(x) = \boldsymbol{\tau}_p(x)$ and $\mathbf{A}_q(x) = \boldsymbol{\tau}_q(x)$ for which $\mathbf{a}_{p/q} = 0$ then

$$\mathbb{E}\left[\nabla_y^T \left(\boldsymbol{\tau}_{p/q}(Y)\boldsymbol{\tau}_{p/q}(Y')\right) \nabla_y k(t,t')|_{y=Y,y'=Y'}\right] = \mathbb{E}\left[\mathcal{A}_{p,1}^T \mathcal{A}_{p,2} k(Y,Y')\right]. \tag{115}$$

2. If $\mathbf{A}_p(x) = \mathbf{A}_q(x) = I_d$ for which $\mathbf{A}_{p/q} = 0$ and $\mathbf{a}_{p/q} = \rho_p - \rho_q$ then

$$\mathbb{E}\left[\rho_{p/q}(Y)^T k(Y,Y')\rho_{p/q}(Y')\right] = \mathbb{E}\left[\mathcal{A}_{p,1}^T \mathcal{A}_{p,2} k(Y,Y')\right] \tag{116}$$

**Example 7.5** (Fisher information distance)**.** *Pick $k(y,y') = \delta_{y=y'}$ the Dirac delta on the diagonal. Then* (108) *with the choice* (116) *becomes*

$$\mathcal{S}(Y,X,\delta) = \mathbb{E}\left[\rho_{X/Y}(Y)^T \rho_{X/Y}(Y)\right] =: J(Y/X) \tag{117}$$

*with $J(Y/X)$ the classical* Fisher Information Distance *between $X$ and $Y$, see [48].*

**Example 7.6** (Independent kernels)**.** *Let $(e_i)_{i=1,\dots,n}$ be a sequence of functions in $\mathcal{F}(X)$ and $k(x,y) = \sum_{i=1}^d \alpha_i e_i \otimes e_i$ (which belongs to $\mathcal{G}$ for any $(\alpha_i)_{i=1,\dots,n}$). Then*

$$\mathcal{S}(Y,X,(\alpha,e)_n) = \sum_{i=1}^n \alpha_i (\mathbb{E}\left[\mathcal{A}_p e_i(Y)\right])^T (\mathbb{E}\left[\mathcal{A}_p e_i(Y)\right]). \tag{118}$$

**Example 7.7.** *(Kernelized Stein discrepancies for comparing Gaussian random vectors) Let $X$ and $Y$ be independent centered multivariate normal random variables in $\mathbb{R}^d$ with variances $\Sigma_1$ and $\Sigma_2$, respectively. Take $\mathcal{A}_p g(x) = \Sigma \nabla g(x) - x g(x)$. Then, for any sufficiently regular function $e : \mathbb{R}^d \to \mathbb{R}$ we have*

$$\mathbb{E}\left[\mathcal{A}_p e(Y)\right] = E\left[(\Sigma_1 - \Sigma_2)\nabla e(Y)\right]$$

*so that the kernelized discrepancy* (118) *becomes*

$$\mathcal{S}(Y,X,(\alpha,e)_n) = \sum_{i=1}^n \alpha_i \mathbb{E}\left[\nabla e_i(Y)\right]^T (\Sigma_1 - \Sigma_2)^2 \mathbb{E}\left[\nabla e_i(Y)\right].$$

*Taking $n = d$, $\alpha_i = 1$ and $e_i(y) = y_i$ and supposing that all marginals have unit variance leads to the natural measure of discrepancy*

$$\mathcal{S}_n(Y,X) = 2\sum_{i<j} (\sigma_{ij}^X - \sigma_{ij}^Y)^2$$

*with $\sigma_{ij}^X$ (resp., $\sigma_{ij}^Y$) the covariance between the marginals $i$ and $j$ of $X$ (resp., of $Y$).*

In terms of potential applications, one of the most interesting aspects of identity (114) is the fact that the right-hand side justifies the use of the left-hand side as a discrepancy metric. Applications of (116) have begun to be explored [26, 57], and more general versions have been touched upon in [45]. The freedom of choice in the input matrices $\mathbf{A}_p, \mathbf{A}_q$ encourages us to be hopeful that these quantities will have numerous applications.

Inspired by [26, 57], we conclude the section with an illustration of how to apply Stein discrepancies to obtain a goodness-of-fit test for a Student-$t$ distribution.

**Example 7.8** (Dimension 1)**.** *Let $p$ be the centered Student-t distribution with $\ell$ degrees of freedom, with score function given by*

$$\rho_{t_\ell}(y) = -\frac{y(\ell+1)}{\ell+y^2}. \tag{119}$$

*The operator obtained from* (119) *is*

$$\mathcal{A}^1_{t_\ell} f(y) = f'(y) - \frac{(\ell+1)y}{\ell+y^2} f(y) \tag{120}$$

*The preceding developments lead to postulating the sample-based discrepancy*

$$\mathcal{S}^\rho(t_\ell, q, k) = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} u^\nu_q(y_i, y'_j),$$

*with $y_1, \ldots, y_{n_1}$ and $y'_1, \ldots, y'_{n_2}$ two i.i.d. samples independently drawn from $q$ and*

$$u^\rho_q(y, y') = \mathcal{A}^\rho_{t_\ell, y} \mathcal{A}^\rho_{t_\ell, y'} k(y, y'), \tag{121}$$

*and $k(y, y')$ a well chosen kernel. Particularizing* (121) *we get*

$$\begin{aligned} u^1_q(y, y') = \; & \partial_y \partial_{y'} k(y, y') - (\ell+1) \frac{y'}{((y')^2+\ell)} \partial_y k(y, y') \\ & - (\ell+1) \frac{y}{(y^2+\ell)} \partial_{y'} k(y, y') + (\ell+1)^2 \frac{yy'}{(y^2+\ell)((y')^2+\ell)} k(y, y'). \end{aligned}$$

*Let $Y, Y' \sim q$ be two independent copies. Since under natural conditions, $\mathbb{E}_q\left[u^\rho_q(Y, Y')\right] = 0$ if and only if $q = p$, a natural goodness-of-fit test in this context is to reject the null assumption $\mathcal{H}_0 : q = p$ whenever $\mathcal{S}^\nu(t_\ell, q, k)$ is too large.*

*For the sake of proof of concept rather than anything else, here are the result of simulations comparing $X \sim p$ a Student with $\ell = 5$ degrees of freedom with $Y \sim q$ a Student with $\ell$ degrees of freedom, via the kernelized discrepancies based on the RBF kernel $k(x, y) = e^{-(x-y)^2/2}$. The quantiles for $\mathcal{S}^\rho(t_\ell, q, k)$ under the null hypothesis were estimated by simulation, with $J = 10^5$ experiments; we obtained*

| *2.5%* | *97.5%* |
|---|---|
| *-0.03837828* | *0.03970307* |

*The results for $10^4$ simulations with $n_1 = n_2 = 100$ run for each value of degrees of freedom $\ell \in \{1, 4, 5, 6, 8, 10, 12, 100, 1000\}$ (with $\ell = 5$ corresponding to the null hypothesis) are reported below (first line) as well as the corresponding results for the classical Kolmogorov Smirnov test (R implementation* `ks.test`*), each time on the same data:*

| $\ell$ | *1* | *4* | *5* | *6* | *8* | *10* | *12* | *100* | *1000* |
|---|---|---|---|---|---|---|---|---|---|
| *kernel* | *0.9049* | *0.0576* | *0.0507* | *0.0445* | *0.0433* | *0.0550* | *0.0587* | *0.1330* | *0.1510* |
| *ks* | *0.9384* | *0.0497* | *0.0459* | *0.0479* | *0.0443* | *0.0461* | *0.0480* | *0.0593* | *0.0618* |

*It appears that the test based on $u^\rho_q(\cdot, \cdot)$ is not as powerful as the Kolmogorov Smirnov test, at least in our implementation. The numerical values are not reported. A more detailed study of such Stein-based discrepancy tests is under way ([30]).*

**Example 7.9** (Dimension 2)**.** *Fix $d = 2$ and let $p$ be the centered Student-t distribution with $\ell$ degrees of freedom and $\Sigma = Id$ the identity matrix. We only consider the operator*

$$\mathcal{A}_1 f(y) = \boldsymbol{\tau}(y)\nabla f(y) - yf(y)$$

*with $\boldsymbol{\tau}$ the Stein kernel matrix given in* (82)*, with entries $(\tau_{ij})_{1\leq i,j\leq 2}$; the resulting kernelized discrepancy is*

$$\mathcal{S}^{\nu}(t_\ell, q, \ell) = \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} u_q(y_i, y'_j).$$

*with discrepancy generating function*

$$\begin{aligned}
u_q(y,y') =& \mathcal{A}_{p,1}^T A_{p,2} k(y,y') \\
=& [\tau_{11}(y)\tau_{11}(y') + \tau_{12}(y)\tau_{12}(y')]\partial_{y_1}\partial_{y'_1}k(y,y') \\
&+ [\tau_{11}(y)\tau_{12}(y') + \tau_{12}(y)\tau_{22}(y')]\partial_{y_1}\partial_{y'_2}k(y,y') \\
&+ [\tau_{12}(y)\tau_{11}(y') + \tau_{22}(y)\tau_{12}(y')]\partial_{y_2}\partial_{y'_1}k(y,y') \\
&+ [\tau_{12}(y)\tau_{12}(y') + \tau_{22}(y)\tau_{22}(y')]\partial_{y_2}\partial_{y'_2}k(y,y') \\
&- (\tau_{11}(y)y'_1 + \tau_{12}(y)y'_2)\partial_{y_1}k(y,y') - (\tau_{12}(y)y'_1 + \tau_{22}(y)y'_2)\partial_{y_2}k(y,y') \\
&- (y_1\tau_{11}(y') + y_2\tau_{12}(y'))\partial_{y'_1}k(y,y') - (y_1\tau_{12}(y') + y_2\tau_{22}(y'))\partial_{y'_2}k(y,y') \\
&+ (y_1 y'_1 + y_2 y'_2)k(y,y')
\end{aligned}$$

*As in the previous example, we present simulation results on a rather modest simulation study. We compare $X \sim p$ a bivariate (centered scaled) Student with $\ell = 5$ degrees of freedom with $Y \sim q$ a bivariate (centered scaled) Student with $\ell$ degrees of freedom, via the kernelized discrepancies based on the RBF kernel $k(x,y) = e^{-(x-y)^2/2}$ with $n_1 = n_2 = 100$. The quantiles were estimated by simulation, with $J = 10^3$ experiments; we obtained*

| *2.5%* | *97.5%* |
|---|---|
| *-0.07256331* | *0.08441458* |

*(which indicates some asymmetry in the sample distribution). The results for $10^3$ simulations run for each value of degrees of freedom $\ell \in \{0.1, 1, 5, 10, 100, 1000\}$ (with $\ell = 5$ corresponding to the null hypothesis) are reported below:*

| $\ell$ | *0.1* | *1* | *4* | *5* | *6* | *10* | *100* | *1000* |
|---|---|---|---|---|---|---|---|---|
| *kernel* | *0.339* | *0.690* | *0.056* | *0.043* | *0.046* | *0.038* | *0.048* | *0.052* |

*Our naive implementation of the bivariate test appears to have difficulties in distinguishing the bivariate Student from the bivariate Gaussian (obtained at $\ell = 1000$). Such an observation is perhaps not so surprising, see e.g. [60] where a similar problem is tested (by different means) with low power for the case of Gaussian vs Student, see page 1126. The problem of devising tractable powerful goodness-of-fit tests for multivariate distributions seems to be difficult; we will concentrate on this in future publications.*

**Acknowledgements.** GM and YS gratefully acknowledges support by the Fonds de la Recherche Scientifique - FNRS under Grant MIS F.4539.16. GR acknowledges partial support from EPSRC grant EP/K032402/1 and the Alan Turing Institute. We also thank Christophe Ley and Guillaume Poly for interesting discussions, as well as Lester Mackey and Steven Vanduffel for suggesting some references which we had overlooked.

# References


[1] C. J. Adcock. Extensions of Stein's lemma for the skew-normal distribution. *Communications in Statistics – Theory and Methods*, 36(9), 1661-1671, 2007.

[2] C. J. Adcock. Asset pricing and portfolio selection based on the multivariate extended skew-Student-t distribution. *Annals of Operations Research*, 176(1), 221-234, 2010.

[3] C. J. Adcock. Mean-variance-skewness efficient surfaces, Stein's lemma and the multivariate extended skew-Student distribution. *European Journal of Operational Research*, 234(2), 392–401, 2014.

[4] S. Aerts and G. Haesbroeck. Robust asymptotic tests for the equality of multivariate coefficients of variation. *Test*, 26(1):163–187, 2017.

[5] M. M. Ali, N.N. Mikhail, and M. S. Haq. A class of bivariate distributions including the bivariate logistic. *Journal of Multivariate Analysis*, 8(3):405–412, 1978.

[6] B. Arras and C. Houdré. *On Stein's Method for Infinitely Divisible Laws with Finite First Moment.* Springer International Publishing, 2019.

[7] S. Artstein, K. M. Ball, F. Barthe, and A. Naor. Solution of Shannon's problem on the monotonicity of entropy. *Journal of the American Mathematical Society*, 17(4): 975–982, 2004a.

[8] S. Artstein, K. M. Ball, F. Barthe, and A. Naor. On the rate of convergence in the entropic central limit theorem. *Probability Theory and Related Fields*, 129(3):381–390, 2004b.

[9] A. Azzalini and A. Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715-26, 1996.

[10] D. Bakry, I. Gentil, and M. Ledoux. *Analysis and geometry of Markov diffusion operators.* Springer Science & Business Media. 2013.

[11] K. Ball, F. Barthe, and A. Naor. Entropy jumps in the presence of a spectral gap. *Duke Mathematical Journal*, 119(1):41–63, 2003.

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

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

[14] A. D. Barbour, M. J. Luczak, and A. Xia. Multivariate approximation in total variation, i: equilibrium distributions of markov jump processes. *The Annals of Probability*, 46(3):1351–1404, 2018.

[15] A. D. Barbour, M. J. Luczak, A. Xia. Multivariate approximation in total variation, ii: Discrete normal approximation. *The Annals of Probability*, 46(3):1405–1440, 2018.

[16] T. Bonis. Rates in the central limit theorem and diffusion approximation via Stein's method. *arXiv preprint arxiv:1506.06966*, 2015.

[17] H. J. Brascamp and E. H. Lieb. On extensions of the Brunn-Minkowski and Prékopa-Leindler theorems, including inequalities for log concave functions, and with an application to the diffusion equation. *Journal of Functional Analysis*, 22(4):366–389, 1976.

[18] T. Cacoullos and V. Papathanasiou. Characterizations of distributions by variance bounds. *Statistics & Probability Letters*, 7(5):351–356, 1989.

[19] T. Cacoullos. On upper and lower bounds for the variance of a function of a random variable. *The Annals of Probability*, 10(3):799–809, 1982.

[20] T. Cacoullos, V. Papathanasiou, and S.Utev. Another characterization of the normal distribution and a related proof of the central limit theorem. *Teoriya Veroyatnostei i ee Primeneniya*, 37(4):648–657, 1992.

[21] T. Cacoullos, V. Papathanasiou, and S. Utev. Variational inequalities with examples and an application to the central limit theorem. *The Annals of Probability*, 22(3): 1607–1618, 1994.

[22] S. Chatterjee. A short survey of Stein's method. In *Proceedings of the International Congress of MathematiciansâĂŤSeoul* (S. Y. Jang, Y. R. Kim, D.-W. Lee and I. Yie, eds.) IV:1âĂŞ24, 2014.

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

[24] S. Chatterjee and Q.-M. Shao. Nonnormal approximation by Stein's method of exchangeable pairs with application to the Curie-Weiss model. *The Annals of Applied Probability*, 21(2):464–483, 2011.

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

[26] K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, 2606–2615, 2016.

[27] T. A. Courtade, M. Fathi, and A. Pananjady. Existence of Stein kernels under a spectral gap, and discrepancy bound. *Annales de l'IHP: Probabilites et Statistiques*, 55(2): 777-790, 2019.

[28] C. Döbler. Stein's method of exchangeable pairs for the beta distribution and generalizations. *Electronic Journal of Probability*, 20(109):1–34, 2015.

[29] C. Döbler, R. E Gaunt, and S. J. Vollmer. An iterative technique for bounding derivatives of solutions of Stein equations. *Electronic Journal of Probability*, 22(96): 1–39, 2017.

[30] M. Ernst and Y. Swan. Stein based goodness-of-fit tests. In preparation, 2019.

[31] M. Ernst and Y. Swan. Distances between distributions via Stein's method. *arXiv preprint arXiv:1909.11518*, 2019.

[32] M. Ernst, G. Reinert, and Y. Swan. First order covariance inequalities via Stein's method. *arXiv preprint*, 2019. Submitted for publication.

[33] M. Ernst, G. Reinert, and Y. Swan. On infinite covariance expansions. *arXiv preprint arXiv:1906.08376*, 2019.

[34] X. Fang, Q.M. Shao, and L. Xu. Multivariate approximations in Wasserstein distance by Stein's method and Bismut's formula. *Probability Theory and Related Fields*, 1–35, 2018.

[35] M. Fathi. Stein kernels and moment maps. *The Annals of Probability* 47(4), 2172–2185, 2019.

[36] T. Gallouët, G. Mijoule, and Y. Swan. Regularity of solutions of the Stein equation and rates in the multivariate central limit theorem. *arXiv preprint arXiv:1805.01720*, 2018.

[37] H. L. Gan, A. Röllin, and N. Ross. Dirichlet approximation of equilibrium distributions in Cannings models with mutation. *Advances in Applied Probability*, 49(3): 927–959, 2017.

[38] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, 70(3):419–435, 2002.

[39] F. Götze. On the rate of convergence in the multivariate clt. *The Annals of Probability*, 19(2):724–739, 1991.

[40] L. Goldstein and G. Reinert. Zero biasing in one and higher dimensions, and applications. In: A. D. Barbour and L. H. Y. Chen eds, *Stein's Method and Applications* (Vol. 5), World Scientific:1–18, 2005.

[41] L. Goldstein and G. Reinert. Stein's method and the zero bias transformation with application to simple random sampling. *The Annals of Applied Probability*, 7(4): 935–952, 1997.

[42] E. Gómez-Sánchez-Manzano, M.A. Gómez-Villegas, and J.M. Marín. A multivariate generalization of the power exponential family of distributions. *Communications in Statistics-Theory and Methods*, 27(3), 589–600, 1998.

[43] J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *arXiv preprint arXiv:1611.06972*, 2016.

[44] J. Gorham and L. Mackey. Measuring sample quality with Stein's method. In *Advances in Neural Information Processing Systems*, 226–234, 2015.

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

[46] S. Holmes and G. Reinert. Stein's method for the bootstrap. In Persi Diaconis and Susan Holmes, editors, *Stein's method: expository lectures and applications*, volume 46 of *IMS Lecture Notes Monogr. Ser*, pages 93–132. Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2004.

[47] J. H Huggins and J. Zou. Quantifying the accuracy of approximate diffusions and Markov chains. *Artificial Intelligence and Statistics*, 382–391. 2017.

[48] O. Johnson. *Information theory and the central limit theorem*. Imperial College Press, London, 2004. ISBN 1-86094-473-6.

[49] Z. Landsman and J. Nešlehová. Stein's lemma for elliptical random vectors. *Journal of Multivariate Analysis*, 99(5):912–927, 2008.

[50] Z. Landsman, S. Vanduffel, and J. Yao. A note on Stein's lemma for multivariate elliptical distributions. *Journal of Statistical Planning and Inference*, 143(11):2016–2022, 2013.

[51] Z. Landsman, S. Vanduffel, and J. Yao. Some Stein-type inequalities for multivariate elliptical distributions and applications. *Statistics & Probability Letters*, 97:54–62, 2015.

[52] M. Ledoux, I. Nourdin, and G. Peccati. Stein's method, logarithmic Sobolev and transport inequalities. *Geometric and Functional Analysis*, 25(1):256–306, 2015.

[53] C. Ley and Y. Swan. Stein's density approach and information inequalities. *Electronic Communications in Probability*, 18(7) 1–14, 2013.

[54] C. Ley and Y. Swan. A general parametric Stein characterization. *Statistics & Probability Letters*, 111, 67–71, 2016.

[55] C. Ley, G. Reinert, and Y. Swan. Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. *Annals of Applied Probability*, 27(1): 216–241, 2017.

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

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

[58] Z. Ma and M. Röckner. Introduction to the theory of (non-symmetric) Dirichlet forms. *Springer Science & Business Media*, 2012.

[59] L. Mackey and J. Gorham. Multivariate Stein factors for a class of strongly log-concave distributions. *Electronic Communications in Probability*, 21, 2016.

[60] M. P. McAssey. An empirical goodness-of-fit test for multivariate distributions. *Journal of Applied Statistics*, 40(5):1120–1131, 2013.

[61] A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative risk management: Concepts, techniques and tools*. Princeton University Press, 2015.

[62] I. Nourdin and G. Peccati. *Normal approximations with Malliavin calculus : from Stein's method to universality*. Cambridge Tracts in Mathematics. Cambridge University Press, 2012.

[63] I. Nourdin and Frederi G. Viens. Density formula and concentration inequalities with Malliavin calculus. *Electronic Journal of Probability*, 14:2287–2309, 2009.

[64] I. Nourdin, G. Peccati, and Y. Swan. Entropy and the fourth moment phenomenon. *Journal of Functional Analysis*, 266:3170–3207, 2014.

[65] I. Nourdin, G. Peccati, and Y. Swan. Integration by parts and representation of information functionals. *IEEE International Symposium on Information Theory (ISIT)*, pages 2217–2221, 2014.

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

[67] C. J. Oates, A. Barp and M. Girolami. Posterior Integration on an Embedded Riemannian Manifold. *arXiv preprint arXiv:1712.01793* (2017).

[68] L. E. Payne and H. F. Weinberger. An optimal Poincaré inequality for convex domains. *Archive for Rational Mechanics and Analysis*, 5(1):286–292, 1960.

[69] D. Paindaveine. Elliptical symmetry. *Wiley StatsRef: Statistics Reference Online*, 2014.

[70] M. Raič. A multivariate CLT for decomposable random vectors with finite second moments. *Journal of Theoretical Probability*, 17(3):573–603, 2004.

[71] G. Reinert and A. Röllin. Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *The Annals of Probability*, 37 (6):2150–2173, 2009.

[72] G. Reinert and N. Ross. Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs. *The Annals of Applied Probability*, 29(5):3201–29, 2019.

[73] N. Ross. Fundamentals of Stein's method. *Probability Surveys*, 8:210–293, 2011.

[74] A. Saumard. Weighted Poincaré inequalities, concentration inequalities and tail bounds related to the behavior of the Stein kernel in dimension one. *arXiv preprint arXiv:1804.03926*, 2018.

[75] A. Saumard and J.A. Wellner. Log-concavity and strong log-concavity: a review. *Statistics Surveys, 8*, 45, 2014.

[76] C. Stein. *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables*. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory, 1972.

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

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

[79] G. Valiant and P. Valiant. A CLT and tight lower bounds for estimating entropy. In *Electronic Colloquium on Computational Complexity (ECCC)*, volume 17, 2010.

[80] S. Vanduffel and J. Yao. A Stein type lemma for the multivariate generalized hyperbolic distribution. *European Journal of Operational Research*. 261(2):606–612, 2017.

[81] V. M. Zolotarev. Probability metrics. *Teoriya Veroyatnostei i ee Primeneniya*, 28(2): 264–287, 1983.