Electron. J. Probab. **28** (2023), article no. 59, 1–40.
ISSN: 1083-6489 https://doi.org/10.1214/22-EJP883

Stein's density method for multivariate continuous distributions*

G. Mijoule† M. Raič‡ G. Reinert§ Y. Swan¶

Abstract

This paper provides a general framework for Stein's density method for multivariate continuous distributions. The approach associates to any probability density function a canonical operator and Stein class, as well as an infinite collection of operators and classes which we call standardizations. These in turn spawn an entire family of Stein identities and characterizations for any continuous distribution on $\mathbb{R}^d$, among which we highlight those based on the score function and the Stein kernel. A feature of these operators is that they do not depend on normalizing constants. A new definition of Stein kernel is introduced and examined; integral formulas are obtained through a connection with mass transport, as well as ready-to-use explicit formulas for elliptical distributions. The flexibility of the kernels is used to compare in Stein discrepancy (and therefore 2-Wasserstein distance) between two normal distributions, Student and normal distributions, as well as two normal-gamma distributions. Upper and lower bounds on the 1-Wasserstein distance between continuous distributions are provided, and computed for a variety of examples: comparison between different normal distributions (improving on existing bounds in some regimes), posterior distributions with different priors in a Bayesian setting (including logistic regression), centred Azzalini–Dalla Valle distributions. Finally the notion of weak Stein equation and weak Stein factors is introduced. Bounds for solutions of the weak Stein equation are obtained for Lipschitz test functions if the distribution admits a Poincaré constant. We use these bounds to compare different copulas on the unit square in 1-Wasserstein distance.

*GM gratefully acknowledges support by the Fonds de la Recherche Scientifique – FNRS under Grant MIS F.4539.16. YS acknowledges support by the Fonds de la Recherche Scientifique – FNRS under Grant CDR/OL J.0197.20, as well as the ULB ARC Consolidator grant. GR acknowledges support from EPSRC grants EP/T018445/1, EP/R018472/1, and EP/X0021951, and the Alan Turing Institute.

†INRIA Paris, Team MoKaPlan, 2 rue Simone Iff, 75012 Paris. E-mail: guillaume.mijoule@gmail.com

‡University of Ljubljana, Faculty of Mathematics and Physics, Jadranska 19, SI–1000 Ljubljana, Slovenia; University of Primorska, Faculty of Mathematics, Natural Sciences and Information Technologies, Glagoljaška 8, SI–6000 Koper, Slovenia; Institute of Mathematics, Physics and Mechanics, Jadranska 19, SI–1000 Ljubljana, Slovenia. E-mail: martin.raic@fmf.uni-lj.si

§University of Oxford, Department of Statistics, 24–29 St Giles, Oxford OX1 3LB, and The Alan Turing Institute, 96 Euston Rd, London NW1 2DB, UK. E-mail: reinert@stats.ox.ac.uk

¶Université libre de Bruxelles, Département de Mathématique, Campus Plaine, Boulevard du Triomphe CP210, B–1050 Brussels. E-mail: yvik.swan@ulb.be

Keywords: Stein operators; density method; elliptical distributions; Stein kernels.
MSC2020 subject classifications: 60B10; 60B12.
Submitted to EJP on November 13, 2019, final version accepted on November 10, 2022.
Supersedes arXiv:2101.05079.

1 Introduction

Stein's method of approximate computation of expectations is a collection of tools permitting to bound *integral probability metrics* or *Zolotarev metrics*

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

where $\mathcal{H}$ is a measure-determining class of functions and X, W are two random quantities of interest with X, say, distributed according to the *target distribution* over which it is assumed that we have some handles. In order to implement this method for X, one needs first to find a linear operator $\mathcal{A}$, called a *Stein operator*, and a class of functions $\mathcal{F}(\mathcal{A})$, called a *Stein class*, such that (i) the *Stein identity* $\mathbb{E}[\mathcal{A}f(X)] = 0$ holds for all functions $f \in \mathcal{F}(\mathcal{A})$ and (ii) for each $h \in \mathcal{H}$, there exists a solution $f = f_h \in \mathcal{F}(\mathcal{A})$ of the *Stein equation* $\mathcal{A}f(w) = h(w) - \mathbb{E}[h(X)]$. 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 the *Stein discrepancy* $\sup_{h\in\mathcal{H}} |\mathbb{E}[\mathcal{A}f_h(W)]|$. The popularity of Stein's method lies for a large part in the fact that, in many important cases, Stein discrepancies are amenable to a wide variety of approaches. Moreover the various equations, operators and discrepancies appearing in Stein's method provide mathematical objects which can be in themselves of intrinsic and independent interest.

There exist many frameworks in which Stein's method is well understood, particularly for univariate distributions. Comprehensive introductions to some of the most important aspects of the theory are available from the monographs [44] as well as [37, 14], with a particular focus on Gaussian approximation. We also refer to the survey [42] as well as the papers [33, 32].

Although the univariate case is the most studied, many references also tackle multivariate distributions. The starting point for a multivariate Stein's method is a *Stein characterization* for the Gaussian law which states that a random vector Z is a multivariate Gaussian d-random vector with mean ν and covariance Σ (short: $Z \sim \mathcal{N}(\nu,\Sigma)$) if and only if

$$\mathbb{E}\big[\,\langle \Sigma, \nabla^2 g(Z)\rangle_{\mathrm{HS}}\,\big] = \mathbb{E}\big[\,\langle Z-\nu, \nabla g(Z)\rangle\,\big], \tag{1.1}$$

for all absolutely continuous function $g\colon\ \mathbb{R}^d \to \mathbb{R}$ for which the expectations exist; here ∇ denotes the gradient operator for functions which act on $\mathbb{R}^d$ (a column vector), $\langle A, B\rangle_{\mathrm{HS}} = \sum_{i,j} A_{ij}B_{ij}$ is the Hilbert–Schmidt inner product between $d\times d$ matrices and $\langle u, v\rangle = \sum_i u_i v_i$ is the usual inner product between vectors in $\mathbb{R}^d$. From (1.1), one infers that the second order differential operator $\mathcal{A}g(x) = \langle \Sigma, \nabla^2 g(x)\rangle_{\mathrm{HS}} - \langle x-\nu, \nabla g(x)\rangle$ is a Stein operator for the Gaussian distribution $\mathcal{N}(\nu,\Sigma)$, with Stein class $\mathcal{F}(\mathcal{A})$ the collection of functions for which (1.1) holds. Classical Markov theory then provides a solution of the corresponding Stein equation $\mathcal{A}g(x) = h(x) - \mathbb{E}[h(Z)]$ via the so-called Mehler formula, leading to a variety of applications of Stein's method for multivariate normal approximation as treated e.g. in [9, 26, 23, 40, 36, 11]. Classical Markov theory also provides Stein operators, equations, and solutions outside the Gaussian case, e.g. for log-concave densities as well as for ergodic measures of stochastic differential equations satisfying regularity conditions, as studied e.g. by [35, 25, 19]. Other references extending the method beyond the multivariate Gaussian setting include [2, 3] for infinitely divisible distributions with finite first moment, and [7, 8]

for a framework which is applicable to many discrete multivariate distributions. In [41], stationary distributions of Glauber Markov chains are characterized. The Dirichlet distribution has been treated in [22], and the multinomial distribution is considered in [34]. This literature review is not intended to be exhaustive; more references can be found e.g. at the websites https://sites.google.com/site/steinsmethod and https://sites.google.com/site/malliavinstein.

Several general approaches to finding Stein operators exist in the univariate setting, among which we highlight Barbour's *generator approach* (which makes use of Markov generator theory, see [26, 9]) and Stein's *density approach* (which makes use of the properties of the underlying density, see [44, 45] as well as [16, 48, 32, 33]). The generator approach extends naturally to the multivariate setting. The density approach has, so far, not been extended to the multivariate setting. The aim of this paper is to fill this gap and to explore implications of Stein's density method in the multivariate setting. In particular this paper highlights the considerable flexibility in Stein operator choice which includes operators which are not gradient operators, thus complementing the generator approach in the multivariate setting by offering an alternative which does not require the analysis of a Markov operator.

1.1 Stein's multivariate density method

The probability distributions in this paper are assumed to admit a probability density function (pdf) p with respect to Lebesgue measure on $\mathbb{R}^d$. The expectation under p is denoted by $\mathbb{E}_p$; for a function h, we write $\mathbb{E}_p h = \mathbb{E}\big[h(X)\big]$ with X having pdf p. The building blocks of our construction are the *canonical directional Stein derivatives*

$$f \mapsto \mathcal{T}_{e,p} f = \frac{\partial_e (p\, f)}{p};$$

the class of functions which satisfy the "elementary" Stein identity $\mathbb{E}_p[\mathcal{T}_{e,p} f] = 0$ is then denoted $\mathcal{F}_1(p)$ (Section 3.4 provides many more Stein identities which are perhaps not as elementary). Such pairs $(\mathcal{T}_{e,p}, \mathcal{F}_1(p))$ can be defined in any direction e, and we collect in the *canonical Stein class* $\mathcal{F}(p)$ all functions for which the elementary Stein identity holds for every component in (almost) every direction. More precise definitions, along with clarifications of the various functional spaces we consider, will be provided below. The resulting objects can be combined in virtually endless possibilities. In this spirit, we introduce the *Stein gradient operator* $\bullet \mapsto \mathcal{T}_{\nabla,p}\bullet = \nabla(p\bullet)/p$ which many authors call *the* Stein operator for multivariate p. Similarly, we define the *Stein divergence operator* acting on vector- or matrix-valued functions with compatible dimensions as $\bullet \mapsto \mathcal{T}_{\mathrm{div},p}\bullet = \mathrm{div}\,(p\bullet)/p$. More generally, a *standardization* of the canonical operator is any operator of the form $\mathcal{A}\bullet = \sum_{i=1}^d \mathbf{A}_i \mathcal{T}_{e_i,p}(\mathbf{T}_i \bullet)$ where $e_i, i = 1, \dots, d$ are the unit vectors in $\mathbb{R}^d$ and $\mathbf{A}_i, \mathbf{T}_i, i = 1, \dots, d$ are some linear operators.

The freedom of choice in the standardizations is only as useful as one can find operators $\mathbf{A}_i, \mathbf{T}_i$ leading to tractable expressions for $\mathcal{A}$. Many distributions p have a tractable score function $\nabla \log p$ (including, of course, the Gaussian distribution but also *elliptical distributions*, convolutions of independent components, etc.). In such cases a direct application of the Stein gradient or divergence operators leads to tractable first order vector-valued operators of the form $\mathcal{A}_1 g = \mathcal{T}_{\nabla,p}(\Sigma g) = \Sigma\, \nabla g + (\Sigma\, \nabla \log p) g$ (where $\Sigma \in \mathbb{R}^{d\times d}$) with class $\mathcal{F}(\mathcal{A}_1) = \mathcal{F}_1(p)$, as well as second order scalar-valued operators $\tilde{\mathcal{A}}_1 g = \mathcal{T}_{\mathrm{div},p}(\Sigma\, \nabla g) = \big\langle \Sigma, \nabla^2 g \big\rangle_{\mathrm{HS}} + \langle \Sigma\, \nabla \log p, \nabla g \rangle$ acting on $\mathcal{F}(\tilde{\mathcal{A}}_1)$ the collection of functions such that $\nabla g \in \mathcal{F}(\mathcal{A}_1)$. The Gaussian Stein operator derived from (1.1) is of this second order form. Much of the more applied literature on multivariate Stein's method focuses on such second order operators, because $\nabla \log p$ (and hence the resulting operator) does not depend on the normalizing constants of p, which is very useful e.g.

in the study of posterior densities in a Bayesian context. This independence of the normalizing constants is inherited from the nature of the canonical directional Stein derivatives and is a feature of all operators provided by Stein's density method.

Another family of standardizations with which much of the paper is concerned relates to the *Stein kernel*. First studied in [44] (although already earlier used to tackle a smooth pdf p, see e.g. [12] where it is referred to as a covariance kernel), this quantity has become an important component of univariate Stein's method, see e.g. [37, 17, 43] for an overview. In dimension $d = 1$, the Stein kernel for a pdf p is the unique solution $x \mapsto \tau_p(x)$ in $\mathcal{F}(p)$ of the ordinary differential equation

$$\mathcal{T}_p \tau_p(x) := \frac{\big(\tau_p(x)\, p(x)\big)'}{p(x)} = \nu - x$$

with ν the mean of p. It exists and is unique when the distribution has finite variance – the Stein kernel is the zero bias density from [24]. Moreover, for $X \sim p$, $\mathbb{E}[\tau_p(X) g'(X)] = \mathbb{E}[(X-\nu) g(X)]$ for all g such that $\mathbb{E}|(X-\nu)g(X)| < \infty$. Uniqueness of the kernel is lost for multivariate distributions and several concurrent definitions exist. In [38, 15], a Stein kernel $\boldsymbol{\tau}$ for a pdf p is defined as any matrix-valued function satisfying

$$\mathbb{E}\left[\langle \boldsymbol{\tau}(X), \nabla g(X) \rangle_{\mathrm{HS}}\right] = \mathbb{E}\left[\langle X - \nu, g(X) \rangle\right] \tag{1.2}$$

at least for all smooth $g\colon \mathbb{R}^d \to \mathbb{R}^d$ with compact support. In [30], a Stein kernel $\boldsymbol{\tau}$ for a pdf p is required to satisfy (1.2) only for smooth functions which are gradients, to read

$$\mathbb{E}\left[\left\langle \boldsymbol{\tau}(X), \nabla^2 g(X) \right\rangle_{\mathrm{HS}}\right] = \mathbb{E}\left[\langle X - \nu, \nabla g(X) \rangle\right]. \tag{1.3}$$

Associated with this definition is a second order scalar-valued operator $\mathcal{A}_2 g(x) = \left\langle \boldsymbol{\tau}(x), \nabla^2 g \right\rangle_{\mathrm{HS}} - \langle x - \nu, \nabla g(x) \rangle$. Comparing (1.3) with (1.1) shows that the covariance matrix Σ is a Stein kernel for the $\mathcal{N}(\nu, \Sigma)$ distribution; hence the Stein-kernel operator $\mathcal{A}_2$ also generalizes the Gaussian operator. Moreover, the difference $\boldsymbol{\tau} - \Sigma$ has been used as a Gaussian discrepancy metric, see e.g. [38, 15, 30], where it is shown that it captures some essential features of the underlying distribution, relating Stein kernels with log-Sobolev inequalities, Poincaré constants and moment maps.

In this paper we adopt a direct approach and define a *directional Stein kernel* for each canonical direction $e_i \in \mathbb{R}^d$, as any differentiable function $x \mapsto \tau_{p,i}(x) \in \mathbb{R}^d$ such that $\mathcal{T}_{\mathrm{div},p}\big(\tau_{p,i}\big)(x) = \mathbb{E}[\langle X, e_i \rangle] - x_i$ for Lebesgue almost all x in the support of p. A *Stein kernel* is then *any* matrix-valued function $\boldsymbol{\tau}$ such that each row $\boldsymbol{\tau}_i = (\tau_{i1}, \ldots, \tau_{id})$ is a kernel in the direction e_i, different kernels leading to different associated operators. This definition opens the way for identifying several Stein kernels (and therefore Stein operators) for any given pdf, even for the Gaussian $\mathcal{N}(\nu, \Sigma)$: Example 4.18 will show that, aside from the Stein kernel $\boldsymbol{\tau} = \Sigma$, if $d \neq 1$ then the matrix-valued functions

$$\boldsymbol{\tau}_\delta(x) = \frac{1}{(2-\delta)(d-1)} \Big[\big(d - 1 + (1-\delta) x^T x\big)\, \Sigma - (1-\delta)(x-\nu)(x-\nu)^T \Big]$$

are Gaussian Stein kernels for all $\delta \neq 2$. More generally, we find Stein kernels for elliptical distributions and formulas allowing to obtain Stein kernels for smooth multivariate distributions with densities which are available up to a normalizing constant.

As our formalism provides an infinite family of Stein operators and classes $(\mathcal{A}, \mathcal{F}(\mathcal{A}))$ for any distribution P having a pdf (even intractable distributions), a variety of versions of Stein's method of distributional approximation can be deployed by considering quantities of the form $\mathcal{S}(q, \mathcal{A}, \mathcal{G}) = \sup_{g \in \mathcal{G}} \|\mathbb{E}_q \mathcal{A} g\|$ for q some distribution of interest, $\mathcal{A}$ any Stein operator for P and $\mathcal{G}$ a well-chosen class of functions. These quantities are called *Stein discrepancies* and have proven to be particularly useful when studying questions of

convergence to equilibrium, see e.g. [25]. The freedom of choice and ease of use of our operator approach $\mathcal{A}$ now allows for optimization over possible Stein operators $\mathcal{A}$ for P in $\mathcal{S}(q, \mathcal{A}, \mathcal{G})$.

In a first application, in Section 4.3, we concentrate on differences of Stein kernels, which provides a general and easy to use discrepancy metric; our examples include comparison of Gaussians (Example 4.20), comparison of Student and Gaussian (Example 4.21), and comparison of normal-gamma distributions (Example 4.22). In the second example we also exploit the freedom of choice in the Stein kernels to optimize the resulting discrepancy, hereby demonstrating the use of disposing of several kernels for a fixed comparison problem.

In a second application, in Section 5 we estimate discrepancies between distributions P_1, P_2 with supports Ω_1 and Ω_2 which are open subsets of $\mathbb{R}^d$ and nested, so that $\Omega_2 \subseteq \Omega_1$. We express the discrepancies in terms of the (1-)Wasserstein distance

$$\mathcal{W}_1(P_1, P_2) = \sup_{h \in \mathrm{Lip}(\Omega_1, 1)} \left| \int h \, \mathrm{d}P_1 - \int h \, \mathrm{d}P_2 \right|,$$

where $\mathrm{Lip}(\Omega, 1)$ is the collection of Lipschitz functions $\Omega \to \mathbb{R}$ with Lipschitz constant 1. For P_1 and P_2 with pdfs p_1, p_2 and chosen Stein operator and class $(\mathcal{A}_{p_1}, \mathcal{F}(\mathcal{A}_{p_1}))$ for p_1 and $(\mathcal{A}_{p_2}, \mathcal{F}(\mathcal{A}_{p_2}))$ for p_2, fixing P_1 as the target, we consider the *Stein equations*

$$\mathcal{A}_{p_1} g = h - \mathbb{E}_{p_1} h; \quad h \in \mathrm{Lip}(\Omega_1, 1). \tag{1.4}$$

Taking expectations with respect to P_2 on either side of (1.4) shows that the Wasserstein distance satisfies

$$\mathcal{W}_1(P_1, P_2) = \sup_{g \in \mathcal{G}(\mathrm{Lip}(\Omega_1, 1))} |\mathbb{E}_{p_2} [\mathcal{A}_{p_1} g]| \tag{1.5}$$

with $\mathcal{G}(\mathrm{Lip}(\Omega_1, 1))$ the collection of all solutions of (1.4) which belong to $\mathcal{F}(\mathcal{A}_{p_1})$. Moreover, if $\mathcal{A}_{p_2}$ for p_2 is chosen such that $\mathcal{G}(\mathrm{Lip}(\Omega_1, 1)) \subseteq \mathcal{F}(\mathcal{A}_{p_2})$ then $\mathbb{E}_{p_2} [\mathcal{A}_{p_2} g] = 0$ and $\mathcal{W}_1(P_1, P_2) = \sup_{g \in \mathcal{G}(\mathrm{Lip}(\Omega_1, 1))} |\mathbb{E}_{p_2} [(\mathcal{A}_{p_1} - \mathcal{A}_{p_2}) g]|$. The freedom of choice in $\mathcal{A}_{p_i}$, $i = 1, 2$, makes this last expression a good starting point for comparison in Wasserstein distance: Theorem 5.1 provides a general bound on the Wasserstein distance between different pdfs on $\mathbb{R}^d$ which we particularize in Proposition 5.10 to even obtain lower bounds (depending on the Stein kernel). We apply these results to a number of concrete applications; in Section 5 we compare in Wasserstein distance different normal distributions (Example 5.9), posterior distributions with different priors in a Bayesian setting (Examples 5.13 and 5.14); we also study the effect of the skewness parameter on centred Azzalini–Dalla Valle distributions where our upper and lower lead to an explicit expression for the Wasserstein distance (Example 5.12).

The above approach to distributional comparisons with Stein's method relies on a good understanding of Stein equations (1.4) and their solutions. Much is already known about the regularity properties (often referred to as *Stein factors*) of these solutions in several important settings; see Example 5.5 for Gaussians, Example 5.6 for log-concave pdfs, and Example 5.7 for more general results. Such ready-to-use regularity properties are not always available and, finally, as a further contribution of the paper, Proposition 5.15 shows that (1.5) also holds under a weakened form of the Stein equation (1.4), namely (5.20), which states

$$\mathbb{E}_{p_2} [\mathcal{A}_{p_1} g] = \mathbb{E}_{p_2} [h - \mathbb{E}_{p_1} h].$$

Solutions of this equation exist under the relatively weak assumption of existence of a Poincaré constant for P_1. This observation allows to provide bounds even when regularity properties of solutions of Stein equations are hard to establish, e.g. for Wasserstein distance between copulas on the unit square (Example 5.17).

1.2 Overview of the paper

The paper is structured as follows. In Section 2 we provide notations that are used throughout the paper. The multivariate Stein's density method is presented and studied in Section 3, first by defining the canonical directional Stein operators and classes (Section 3.1), then by providing sufficient conditions for obtaining Stein identities and characterizations (Section 3.2), and finally by making the connection with current literature through the concept of standardizations of the canonical operators (Section 3.4). Section 4 is devoted to the study of Stein kernels, first in general (Section 4.1) then under an assumption of ellipticity of the distribution (Section 4.2); first applications towards distributional approximation are also outlined (Section 4.3). Section 5 contains the main applications, namely a flexible set of bounds on Wasserstein distance between densities admitting a Stein operator under an additional assumption of nested supports (Section 5.1). The idea of *weak Stein equations* and *weak Stein factors* is introduced in Section 5.2, leading to bounds on Wasserstein distance under the assumption of existence of a Poincaré constant; these bounds are illustrated by comparing copulas.

This paper is a complete overhaul of a paper that previously appeared on the arXiv (https://arxiv.org/abs/1806.03478) which includes more material, such as an excursion into kernelized Stein discrepancies and goodness-of-fit tests.

2 Notations, gradients and product rules

We first settle the notation that will be used throughout the paper. Fix $d \in \mathbb{N}_0$ and let $e_1, \dots, e_d$ be the canonical basis for Cartesian coordinates in $\mathbb{R}^d$. Vectors of $\mathbb{R}^d$ are understood as column vectors. Given $x, y \in \mathbb{R}^d$ the Euclidean inner product is $\langle x, y\rangle = x^T y = \sum_{i=1}^d x_i y_i$ (here $\cdot^T$ denotes the usual transpose) with associated norm $\|x\| = \sqrt{\langle x, x\rangle}$. With $\mathrm{Tr}(\cdot)$ the trace operator, the *Hilbert–Schmidt inner product* between matrices A, B of compatible dimensions is $\langle A, B\rangle_{\mathrm{HS}} = \mathrm{Tr}(A^T B)$, with associated norm $\|A\|_{\mathrm{HS}}^2 := \mathrm{Tr}(A^T A)$. For $W \in \mathbb{R}^{d\times d}$ a matrix, $\|W\|_{\mathrm{op}} = \sum_{v\in\mathbb{R}^d, \|v\|=1} \|Wv\|$. The identity function on $\mathbb{R}^d$ is denoted by Id, so that $\mathrm{Id}(x) = x$; for a unit vector e, $\mathrm{Id}_e = \langle \mathrm{Id}, e\rangle$ denotes the marginal projection in direction e; we abbreviate $\mathrm{Id}_{e_i}(x) = \mathrm{Id}_i(x) (= x_i)$. The identity on function spaces is denote $\mathbf{I}$; the $d \times d$ identity matrix is denoted by I_d. Thus, for a scalar-valued function f:

- $\mathrm{Id} f$ is a vector-valued function mapping x to $f(x)\, x$;
- $\mathbf{I} f$ is a scalar-valued function mapping x to $f(x)$;
- $\mathrm{I}_d f$ is a matrix-valued function mapping x to $f(x)\, \mathrm{I}_d$.

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 function $v\colon\ \mathbb{R}^d \to \mathbb{R}$ in the direction e is denoted by the real-valued function $\partial_e v$. For $i = 1, \dots, d$ we write $\partial_i v$ for the derivative in the direction of the unit vector e_i. Higher order derivatives are denoted accordingly. The directional derivative of a matrix-valued function $\mathbf{F}\colon\ \mathbb{R}^d \to \mathbb{R}^{m+r}\colon\ x \mapsto \mathbf{F}(x) = \big(f_{ij}(x)\big)_{1\leqslant i\leqslant m, 1\leqslant j\leqslant r}$ is defined component-wise: $(\partial_e \mathbf{F})_{1\leqslant i\leqslant m, 1\leqslant j\leqslant r} = (\partial_e f_{ij})_{1\leqslant i\leqslant m, 1\leqslant j\leqslant r}$ (an $m \times r$ matrix).

The gradient of a smooth function $v\colon\ \mathbb{R}^d \to \mathbb{R}$ is the $d \times 1$ column vector valued function ∇v with entries $(\nabla v)_i = \partial_i v$, $i = 1, \dots, d$, and the Hessian is the symmetric $d \times d$ matrix-valued function $\nabla^2 v$ with entries $(\nabla^2 v)_{i,j} = \partial_{ij} v$, $i, j = 1, \dots, d$;

$$\nabla v = \begin{pmatrix} \partial_1 v \\ \vdots \\ \partial_d v \end{pmatrix} \text{ and } \nabla^2 v = \begin{pmatrix} \partial_{11} v & \partial_{12} v & \cdots & \partial_{1d} v \\ \partial_{21} v & \partial_{22} v & \cdots & \partial_{2d} v \\ \vdots & \vdots & \ddots & \vdots \\ \partial_{d1} v & \partial_{d2} v & \cdots & \partial_{dd} v \end{pmatrix}.$$

The divergence of a d-vector-valued function $\mathbf{v}\colon\ \mathbb{R}^d \to \mathbb{R}^d$ with components $v_j, j = 1, \ldots, d$ is the scalar-valued function

$$\operatorname{div} \mathbf{v}\colon\ \mathbb{R}^d \to \mathbb{R} : x \mapsto \operatorname{div} \mathbf{v}(x) = \sum_{i=1}^{d} \partial_i v_i(x).$$

The divergence of a $m \times d$ matrix-valued function $\mathbf{F} = \big(f_{ij}(x)\big)_{1 \leqslant i \leqslant m, 1 \leqslant j \leqslant d}$ is

$$\operatorname{div} \mathbf{F} = \begin{pmatrix} \sum_{j=1}^{d} \partial_j f_{1j} \\ \sum_{j=1}^{d} \partial_j f_{2j} \\ \vdots \\ \sum_{j=1}^{d} \partial_j f_{mj} \end{pmatrix}$$

a m-column vector; the divergence acts on the rows. For $v\colon\ \mathbb{R}^d \to \mathbb{R}$ we also use the Laplacian

$$\Delta v = \operatorname{div}(\nabla v) = \sum_{i=1}^{d} \partial_i^2 v.$$

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*

$$\partial_e(vw) = (\partial_e v)w + v\partial_e w, \tag{2.1}$$

at all points in $\mathbb{R}^d$ at which all derivatives are defined, implying product rules for gradients and divergences. We shall mainly consider three instances.

1. For $f\colon\ \mathbb{R}^d \to \mathbb{R}$ and $g\colon\ \mathbb{R}^d \to \mathbb{R}$

$$\nabla(fg) = (\nabla f)g + f(\nabla g). \tag{2.2}$$

2. For a matrix-valued function $\mathbf{F}\colon\ \mathbb{R}^d \to \mathbb{R}^{m \times d}$ and $g\colon\ \mathbb{R}^d \to \mathbb{R}$, we have

$$\operatorname{div}(\mathbf{F}g) = (\operatorname{div} \mathbf{F})\, g + \mathbf{F}\, \nabla g \tag{2.3}$$

   (an m-vector). In particular if $\mathbf{F} = \mathrm{I}_d$, then $\operatorname{div}(\mathrm{I}_d g) = \nabla g$.

3. For $\mathbf{F}\colon\ \mathbb{R}^d \to \mathbb{R}^{d \times d}$ and $g\colon\ \mathbb{R}^d \to \mathbb{R}$ we have

$$\operatorname{div}(\mathbf{F}^T \nabla g) = \langle \operatorname{div} \mathbf{F}, \nabla g \rangle + \big\langle \mathbf{F}, \nabla^2 g \big\rangle_{\mathrm{HS}} \tag{2.4}$$

   (a scalar).

Throughout this paper, all random vectors are assumed to live on the same probability space and, unless explicitly mentioned otherwise, are distributed according to a measure P which is (i) absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$, and (ii) with pdf p whose support $\Omega_p = \{x\ ;\ p(x) > 0\}$ is an open subset of $\mathbb{R}^d$; unless otherwise stated the pdf is assumed to be continuous on Ω_p.

The collection of all functions whose components are integrable with respect to p is denoted by $L^1(p)$. For ease of notation, the Lebesgue measure is often left out in integrals; thus, for $f \in L^1(p)$, we use the notations $\mathbb{E}_p f = \int_{\Omega_p} fp = \int_{\Omega_p} f(x)\, p(x)\, \mathrm{d}x$. We shall consider several function spaces on open sets $\Omega \subseteq \mathbb{R}^d$:

- $\mathrm{Lip}_{\mathrm{loc}}(\Omega)$ denotes the set of all locally Lipschitz functions $g\colon\ \Omega \to \mathbb{R}$.

- $\mathcal{C}^k(\Omega)$ (with possibly $k=\infty$) denotes the set of real-valued functions on Ω with k continuous derivatives.
- $\mathcal{C}^k_{\mathrm{c}}(\Omega)$ (with possibly $k=\infty$) denotes the set of functions belonging to $\mathcal{C}^k(\Omega)$ with a compact support $K \subseteq \Omega$.
- $W^{1,1}(\Omega)$ denotes the Sobolev space of functions $g\colon\ \Omega \to \mathbb{R}$ such that the Sobolev norm
$$\|g\|_{1,1} = \int_\Omega |g| + \sum_{i=1}^d \int_\Omega |\partial_i g|$$
is finite. Here, ∂_i denotes the *weak* derivative. Following [31], p.320, we denote by ∂_i also the usual derivative, the interpretation of ∂_i depending on the context. For Sobolev spaces it is understood as the weak derivative.
- $\dot{W}^{1,1}(\Omega)$ denotes the homogeneous Sobolev space of functions $g \in L^1_{\mathrm{loc}}(\Omega)$ with $\sum_{i=1}^d \int_\Omega |\partial_i g| < \infty$.
- $W^{1,1}_0(\Omega)$ denotes the space of all functions $g \in W^{1,1}(\Omega)$, such that $\int_\Omega \partial_i g = 0$ for all $i = 1, 2, \ldots, d$ or, equivalently, $\int_\Omega \partial_e g = 0$ for all $e \in S^{d-1}$.
- $\dot{W}^{1,1}_0(\Omega)$ denotes the space of all functions $g \in \dot{W}^{1,1}(\Omega)$, such that $\int_\Omega \partial_e g = 0$ for all $e \in S^{d-1}$.
- $\mathrm{ACL}(\Omega)$ denotes the set of all Borel measurable functions $g\colon\ \Omega \to \mathbb{R}$, such that for each $e \in S^{d-1}$, it is true that for almost all lines L parallel to e with respect to the $(d-1)$-dimensional Lebesgue measure, the restriction of g to any compact interval I contained in $\Omega \cap L$ is absolutely continuous. Notice that $\mathrm{ACL}(\Omega)$ is closed under multiplication.
- $\mathrm{ACL}^1(\Omega)$ denotes the set of functions $g \in \mathrm{ACL}(\Omega)$ with $\partial_e g \in L^1(\Omega)$ for all $e \in S^{d-1}$.
- $\mathrm{ACL}^1_{\mathrm{loc}}(\Omega)$ denotes the set of all functions $g\colon\ \Omega \to \mathbb{R}$ with the property that each $x \in \Omega$ has an open neighbourhood $U \subseteq \Omega$, such that the restriction of g to U belongs to $\mathrm{ACL}^1(U)$.
- $\mathrm{ACL}^1_0(\Omega)$ denotes the set of functions $g \in \mathrm{ACL}^1(\Omega)$ with $\int_\Omega \partial_e g = 0$ for all $e \in S^{d-1}$.

Here ACL stands for 'absolutely continuous on lines'. This concept is closely related with the concept of *almost differentiability* used by Stein [46, Definition 1]. Indeed, each function in $\mathrm{ACL}(\mathbb{R}^d)$ is almost differentiable, but the converse is not true: take a straight line in $\mathbb{R}^2$ and let $g\colon\ \mathbb{R}^2 \to \mathbb{R}$ be defined as $g \equiv 1$ on that line and $g \equiv 0$ elsewhere. Then g is almost differentiable, but it is not in $\mathrm{ACL}(\mathbb{R}^2)$.

3 Stein's multivariate density method

3.1 The canonical operator

All Stein operators to be constructed in this paper have as building blocks the *canonical* Stein operator and the *canonical* Stein class which we now define.

Definition 3.1 (Canonical directional Stein operator)**.** *Let $e \in S^{d-1}$. The* canonical Stein derivative for p in the direction e *is the differential operator $\phi \mapsto \mathcal{T}_{e,p}\phi := \partial_e(p\phi)/p$ where $\mathcal{T}_{e,p}\phi \equiv 0$ outside of Ω_p. The domain of $\mathcal{T}_{e,p}$ is $\mathrm{dom}(\mathcal{T}_{e,p})$ the collection of all functions that are differentiable in direction e.*

Definition 3.2 (Canonical Stein class)**.** *The* canonical scalar Stein class for p *is the collection $\mathcal{F}_1(p)$ of all functions $f\colon\ \Omega_p \to \mathbb{R}$ with $fp \in \mathrm{ACL}^1_0(\Omega_p)$. The* canonical Stein class for p *is the collection $\mathcal{F}(p)$ of all scalar-, vector-, and matrix-valued functions whose components belong to $\mathcal{F}_1(p)$.*

To illustrate the Stein class we first give the next result.

Proposition 3.3. *Let $u \in \mathrm{ACL}(\Omega_p)$ If, in addition, $\partial_e u \in L^1_{\mathrm{loc}}(\Omega)$ and u is compactly supported, then $\partial_e u \in L^1(\Omega)$ and $\int_\Omega \partial_e u = 0$. In particular, any scalar-valued function $\phi \in \mathcal{C}^1_{\mathrm{c}}(\Omega_p)$ lies in $\mathcal{F}_1(p)$ as long as the pdf p is continuously differentiable.*

Proof. Since u has a compact support, so does $\partial_e u$. Therefore, $\partial_e u \in L^1(\Omega)$. Without loss of generality, we may assume that $e = e_1$. For Lebesgue-almost all $x' \in \mathbb{R}^{d-1}$, the function $u_{x'}(x_1) := u(x_1, x')$ is absolutely continuous on all compact intervals contained in $\Omega_{x'} := \{x_1 \in \mathbb{R} \,;\, (x_1, x') \in \Omega\}$ and $u'_{x'} \in L^1(\Omega_{x'})$. Now observe that $\Omega_{x'}$ is a union of countably many open intervals. If I is such an interval, the restriction of $u_{x'}$ to I has a compact support. By the fundamental theorem of calculus, we have $\int_I u'_{x'} = 0$. Since $u'_{x'} \in L^1(\Omega_{x'})$, we can sum over the intervals to obtain $\int_I u'_{x'} = 0$. The main assertion now follows by Fubini's theorem. The last assertion is immediate by taking $u = p\phi$. □

Remark 3.4.

1. The directional Stein operator requires the pdf p to be available only up to a normalizing constant.
2. For a Borel function $f\colon \Omega_p \to \mathbb{R}$, a version of f belongs to $\mathcal{F}_1(p) \cap L^1(p)$ if and only if $fp \in W^{1,1}_0(\Omega_p)$.

Example 3.5 (Gaussian directional Stein operators)**.** For the multivariate Gaussian distribution with location $\nu \in \mathbb{R}^d$, positive definite covariance matrix $\Sigma \in \mathbb{R}^d \times \mathbb{R}^d$ and pdf γ, with e_i a unit vector,

$$\mathcal{T}_{e_i,\gamma} f(x) = \frac{\partial_{e_i}(f(x)\gamma(x))}{\gamma(x)} = \partial_i f(x) - (\Sigma^{-1}(x-\nu))_i f(x)$$

for $i = 1, \dots, d$ and the scalar Stein class is the class of all almost differentiable functions $f\colon \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}_\gamma[\|\nabla f\|] < \infty$, see [46, Lemma 2]. Indeed the integrability condition $\int \partial_i(f\gamma) = 0$ is automatically satisfied.

Example 3.6 (Student-t operators)**.** For 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

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

with normalizing constant $c_{k,d}$ and full support $\mathbb{R}^d$, the directional derivatives are

$$\mathcal{T}_{e_i,t_k} f(x) = \partial_{e_i} f(x) - \frac{k+d}{2} \left(1 + \frac{(x-\nu)^T \Sigma^{-1}(x-\nu)}{k}\right)^{-1} (\Sigma^{-1}(x-\nu))_i f(x)$$

for $i = 1, \dots, d$. Using a similar argument as in Example 3.5, the scalar Stein class contains all functions $f\colon \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{E}_{t_k}[\|\nabla f\|] < \infty$.

We stress the difference between the domain of the canonical directional operator, which simply consists of differentiable functions, and its Stein class $\mathcal{F}(p)$ which contains all functions $\mathbf{F}$ such that $\mathcal{T}_{e,p}$ is integrable with respect to p in all directions and, moreover, satisfy that for all $e \in S^{d-1}$,

$$\mathbb{E}_p[\mathcal{T}_{e,p}\mathbf{F}] = 0. \tag{3.1}$$

Identity (3.1) is our *canonical Stein identity* in the spirit of (1.1). It is the identity from which all other Stein identities in this paper will follow. In particular, given any differentiable $\mathbf{F}, g$, the product rule (2.1) yields the Stein-type product rule:

$$\mathcal{T}_{e,p}(\mathbf{F}\, g) = (\mathcal{T}_{e,p}\mathbf{F})g + \mathbf{F}\, \partial_e g. \tag{3.2}$$

Plugging (3.2), for well-chosen functions $\mathbf{F}$ or g (or both), into (3.1), then leads to Stein-type integration by parts identities which are the basis of many of the forthcoming results. We first introduce the appropriate sets of functions.

Definition 3.7 (Stein adjoint class)**.** *Let $\mathcal{F}_{1,\mathrm{loc}}(p)$ be the class of all functions $f\colon \Omega_p \to \mathbb{R}$ with $fp \in \mathrm{ACL}(\Omega_p)$, and denote by $\mathcal{F}_{\mathrm{loc}}(p)$ the class of all scalar-, vector- and matrix-valued functions $\mathbf{F}$ with all components belonging to $\mathcal{F}_{1,\mathrm{loc}}(p)$. To every (scalar-, vector-, or matrix-valued) function $\mathbf{F} \in \mathcal{F}_{\mathrm{loc}}(p)$ we denote by $\mathrm{dom}(p,\mathbf{F})$ the collection of all functions $g \in \mathrm{ACL}(\Omega_p)$ which satisfy $(\mathbf{F}\, g) \in \mathcal{F}(p)$ and $\mathbf{F}(\partial_e g) \in L^1(p)$ for all $e \in S^{d-1}$.*

By definition of $\mathcal{F}_{\mathrm{loc}}(p)$, for all $g\colon \mathbb{R}^d \to \mathbb{R} \in \mathrm{dom}(p,\mathbf{F})$, the left-hand side in (3.2) integrates to 0 under p and the integrals of the summands on the right-hand side may be taken separately so that

$$\mathbb{E}_p\left[(\mathcal{T}_{e,p}\mathbf{F})g + \mathbf{F}\,\partial_e g\right] = 0 \tag{3.3}$$

for all $e \in S^{d-1}$, generalizing (3.1) to a Stein identity which is valid for all $g \in \mathrm{dom}(p,\mathbf{F})$. Our approach to the Stein operator machinery for p is to fix some $\mathbf{F}$ and study the operator (in g) that can be obtained from (3.3), viewed through its action on the class $\mathrm{dom}(p,\mathbf{F})$. We call this process a *standardization* of the canonical operator, and will dive more deeply into this in the next section. The quality of the operators obtained in this manner depend on the choice of $\mathbf{F}$. If $\mathbf{F}$ has all components in $\mathcal{F}(p)$ then $\mathrm{dom}(p,\mathbf{F})$ contains at least the constant functions $g \equiv 1$. Even if the components of $\mathbf{F}$ do not belong to $\mathcal{F}(p)$, $\mathrm{dom}(p,\bullet)$ remains actually quite large, as we now show.

Lemma 3.8. *Suppose that all components f of $\mathbf{F}$ are such that $fp \in \mathrm{ACL}^1_{\mathrm{loc}}(\Omega_p)$. Then any function $g \in \mathrm{Lip}_{\mathrm{loc}}(\Omega)$ such that $\mathbf{F}gp$ is compactly supported belongs to $\mathrm{dom}(p,\mathbf{F})$.*

Proof. By Proposition A.5, we have $fp \in L^1_{\mathrm{loc}}(\Omega_p)$. Since $g \in \mathrm{Lip}_{\mathrm{loc}}(\Omega)$, it belongs to $\mathrm{ACL}(\Omega_p)$. Moreover, as g and $\partial_e g$ both bounded for any $e \in S^{d-1}$, $fp(\partial_e g) \in L^1_{\mathrm{loc}}(\Omega_p)$. Since $fp(\partial_e g)$ has compact support, it belongs to $L^1(\Omega_p)$. From the product rule for the classical derivatives and local boundedness of g and $\partial_e g$, it follows that $fgp \in \mathrm{ACL}^1_{\mathrm{loc}}(\Omega_p)$. As fgp has compact support, Proposition 3.3 yields that $fgp \in \mathrm{ACL}^1_0(\Omega_p)$. □

An in-depth discussion of $\mathrm{dom}(p,\mathbf{F})$ in the one-dimensional case, along with simple sufficient conditions, can be found in [17, Section 2.3].

3.2 Standardizations

The canonical operators are our building blocks for a large family of Stein operators, as follows.

Definition 3.9. *A differential operator $\mathbf{G} \mapsto \mathcal{A}\mathbf{G}$ acting on scalar-, or vector-, or matrix-valued valued functions is a* standardization *of the canonical operators $\mathcal{T}_{e_i,p}, i = 1,\dots,d$ if there exist linear operators $\{\mathbf{A}_1,\dots,\mathbf{A}_d\}$ and $\{\mathbf{T}_1,\dots,\mathbf{T}_d\}$ such that*

$$\mathcal{A}\mathbf{G} = \sum_{i=1}^{d} \mathbf{A}_i \mathcal{T}_{e_i,p}(\mathbf{T}_i\mathbf{G}).$$

The associated Stein class is the collection $\mathcal{F}(p,\mathcal{A})$ of all functions $\mathbf{G}$ of appropriate dimension such that $\mathcal{A}\mathbf{G}$ is integrable with mean 0 under p.

Remark 3.10. We have kept the definition vague in our assumptions on the operators $\mathbf{A}_1,\dots,\mathbf{A}_d$ and $\mathbf{T}_1,\dots,\mathbf{T}_d$ which are allowed to be any linear operators on the corresponding function spaces. In particular, they are allowed to be multiplications by fixed functions, shift or differential operators which typically shrinks the associated Stein class $\mathcal{F}(p,\mathcal{A})$ to a very small set: if $\mathbb{E}_p h = 0$, we typically do not have $\mathbb{E}_p \mathbf{A}_i h = 0$. In practice, however, the operators $\mathbf{A}_1,\dots,\mathbf{A}_d$ we consider are in fact left compositions

by linear maps acting between scalars, vectors or matrices. This preserves the size of the associated Stein class $\mathcal{F}(p, \mathcal{A})$. In particular if $\mathbf{A}_i = \mathbf{T}_i = \mathbf{I}$ (the identity functional operator) for $i = 1, \dots, d$, then the canonical Stein class $\mathcal{F}(p)$ is a subset of $\mathcal{F}(p, \mathcal{A})$.

In the sequel, we will say that a linear operator is *a* Stein operator for p if it is a standardization of the canonical operators obtained through Definition 3.9. Such standardizations were studied in [33] in dimension $d = 1$. Among the many possible options, two stand out most naturally:

- $\mathbf{A}_i = e_i$ and $\mathbf{T}_i = 1$ for $i = 1, \dots, d$ leading to what we call the gradient operator,
- $\mathbf{A}_i = \mathbf{T}_i = 1$ for $i = 1, \dots, d$ leading to what we call the divergence operator.

Definition 3.11 (Gradient and divergence operator)**.** *The* Stein gradient operator *acting on real-valued functions* $g\colon \mathbb{R}^d \to \mathbb{R}$ *is*

$$\mathcal{T}_{\nabla,p} g = \sum_{i=1}^{d} e_i \mathcal{T}_{e_i,p} g = \frac{\nabla(pg)}{p}.$$

The Stein divergence operator *acting on vector-valued functions* $\mathbf{g}\colon \mathbb{R}^d \to \mathbb{R}^d$ *is*

$$\mathcal{T}_{\mathrm{div},p} \mathbf{g} = \sum_{i=1}^{d} \mathcal{T}_{e_i,p} \mathbf{g}_i = \frac{\mathrm{div}(p\mathbf{g})}{p}.$$

The Stein divergence operator acting on matrix-valued functions $\mathbf{G}\colon \mathbb{R}^d \to \mathbb{R}^{m\times d}$ *is*

$$(\mathcal{T}_{\mathrm{div},p} \mathbf{G})_i = \sum_{j=1}^{d} \mathcal{T}_{e_j,p} \mathbf{G}_i = \frac{\mathrm{div}(p\mathbf{G}_i)}{p} \qquad \text{for } i = 1, \dots, m.$$

Example 3.12 (Gaussian operators)**.** Take $\nu \in \mathbb{R}^d$ and Σ an invertible matrix in $\mathbb{R}^{d\times d}$. For $\mathcal{N}(\nu, \Sigma)$ with pdf γ,

$$\mathcal{T}_{\nabla,\gamma} g(x) = \nabla g(x) - \Sigma^{-1}(x - \nu) g(x) \tag{3.4}$$

for all $g\colon \mathbb{R}^d \to \mathbb{R}$ ($\mathcal{T}_{\nabla,\gamma} g(x)$ is a vector) and

$$\mathcal{T}_{\mathrm{div},p} \mathbf{g}(x) = \mathrm{div}\, \mathbf{g}(x) - \big\langle \Sigma^{-1}(x - \nu), \mathbf{g}(x) \big\rangle \tag{3.5}$$

for all $\mathbf{g}\colon \mathbb{R}^d \to \mathbb{R}^d$ ($\mathcal{T}_{\mathrm{div},\gamma} \mathbf{g}(x)$ is a scalar).

From (3.3) we thus inherit an entire collection of Stein identities for functions defined on an open set Ω and with components belonging to $\mathrm{ACL}(\Omega)$. In the sequel we will make use of the following instances:

1. For all scalar-valued functions $f, g\colon \Omega \to \mathbb{R}$ we have from (2.2)

$$\mathcal{T}_{\nabla,p}(fg) = (\mathcal{T}_{\nabla,p} f)\, g + f\, \nabla g\,. \tag{3.6}$$

   For each $f \in \mathcal{F}_{1,\mathrm{loc}}(p)$ we obtain the Stein identity

$$\mathbb{E}_p\big[(\mathcal{T}_{\nabla,p} f)\, g\big] = -\mathbb{E}_p[f\, \nabla g]\,, \tag{3.7}$$

   which holds for all $g \in \mathrm{dom}(p, f)$.

2. For all matrix-valued functions $\mathbf{F}\colon \Omega \to \mathbb{R}^{m\times d}$ and all $g\colon \mathbb{R}^d \to \mathbb{R}$, using (2.3),

$$\mathcal{T}_{\mathrm{div},p}(\mathbf{F}\, g) = (\mathcal{T}_{\mathrm{div},p} \mathbf{F})\, g + \mathbf{F}\, \nabla g\,. \tag{3.8}$$

   For each $\mathbf{F} \in \mathcal{F}_{\mathrm{loc}}(p)$ we obtain the Stein identity

$$\mathbb{E}_p\big[(\mathcal{T}_{\mathrm{div},p} \mathbf{F})\, g\big] = -\mathbb{E}_p[\mathbf{F}\, \nabla g]\,, \tag{3.9}$$

   which holds for all $g \in \mathrm{dom}(p, \mathbf{F})$.

3. For all vector-valued function $\mathbf{f}\colon\ \Omega \to \mathbb{R}^d$ and all $g\colon\ \mathbb{R}^d \to \mathbb{R}$, using (3.8),

$$\mathcal{T}_{\mathrm{div},p}(\mathbf{f}g) = (\mathcal{T}_{\mathrm{div},p}\mathbf{f})\, g + \langle \mathbf{f}, \nabla g\rangle\,. \tag{3.10}$$

   For each $\mathbf{f} \in \mathcal{F}_{\mathrm{loc}}(p)$ we obtain the Stein identity

$$\mathbb{E}_p\big[(\mathcal{T}_{\mathrm{div},p}\mathbf{f})\, g\big] = -\mathbb{E}_p\big[\langle \mathbf{f}, \nabla g\rangle\big]\,, \tag{3.11}$$

   which holds for all $g \in \mathrm{dom}(p, \mathbf{f})$.

4. For all matrix-valued functions $\mathbf{F}\colon\ \Omega \to \mathbb{R}^{d\times d}$ and all $g\colon\ \mathbb{R}^d \to \mathbb{R}$, using (2.4),

$$\mathcal{T}_{\mathrm{div},p}\big(\mathbf{F}^T\nabla g\big) = \langle \mathcal{T}_{\mathrm{div},p}\mathbf{F}, \nabla g\rangle + \big\langle \mathbf{F}, \nabla^2 g\big\rangle_{\mathrm{HS}}\,. \tag{3.12}$$

   For each $\mathbf{F} \in \mathcal{F}_{\mathrm{loc}}(p)$ we obtain the Stein identity

$$\mathbb{E}_p\big[\langle \mathcal{T}_{\mathrm{div},p}\mathbf{F}, \nabla g\rangle\big] = -\mathbb{E}_p\big[\big\langle \mathbf{F}, \nabla^2 g\big\rangle_{\mathrm{HS}}\big], \tag{3.13}$$

   which holds for all g with $\nabla g \in \mathrm{dom}(p, \mathbf{F})$.

Example 3.13 (Gaussian Stein identities)**.** Following up on Examples 3.5 and 3.12, take $\nu \in \mathbb{R}^d$ and an invertible matrix $\Sigma \in \mathbb{R}^{d\times d}$ and let γ denote the pdf of $\mathcal{N}(\nu, \Sigma)$. Taking the constant function $f = 1$ in (3.7), which is in $\mathcal{F}_1(\gamma)$, we reap

$$\mathbb{E}_\gamma[\Sigma^{-1}(\mathrm{Id} - \nu)g] = \mathbb{E}_\gamma[\nabla g] \tag{3.14}$$

for all $g\colon\ \mathbb{R}^d \to \mathbb{R}$ belonging to $\mathrm{dom}(\gamma, 1)$. We could have also obtained this directly from operator (3.4). Similarly, now taking the constant function $\mathbf{F} = \Sigma$ in (3.13), which is also in $\mathcal{F}(\gamma)$, we recover the classical second order Stein identity (1.1) for the Gaussian. This could have been obtained directly from operator (3.5) applied to functions $\mathbf{g} = \Sigma\nabla g$. Example 4.18 provides other choices of $\mathbf{F}$.

3.3 Stein characterizations

Identity (3.3) (and the corresponding identities from Section 3.2) holds when integrating under p. Our purpose now is to deduce "reverse" implications when changing the measure p to some other measure q. The starting point is the following "directional" observation.

Proposition 3.14 (Stein characterizations)**.** *Let p and q be pdfs with $\Omega_q \subseteq \Omega_p$. Suppose that Ω_p is connected and $q/p \in \mathrm{Lip}_{\mathrm{loc}}(\Omega_p)$. Fix $f \in \mathcal{F}_{1,\mathrm{loc}}(p)$ with $f \neq 0$ over Ω_p. Then $p = q$ if and only if $\mathbb{E}_q[g\mathcal{T}_{e,p}f] = -\mathbb{E}_q[f\partial_e g]$ for all $e \in S^{d-1}$ and all $g \in \mathrm{dom}(p, f)$ with $g(\mathcal{T}_{e,p}f), f(\partial_e g) \in L^1(q)$.*

Proof. If $p = q$, then the identity follows from (3.3). For the opposite direction, assume that $\mathbb{E}_q[g\mathcal{T}_{e,p}f] = -\mathbb{E}_q[f\,\partial_e g]$. Then with (3.2), we have

$$0 = \mathbb{E}_q\big[\mathcal{T}_{e,p}(fg)\big] = \mathbb{E}_p\left[\big(\mathcal{T}_{e,p}(fg)\big)\,\frac{q}{p}\right] = -\mathbb{E}_p\left[fg\,\partial_e\left(\frac{q}{p}\right)\right],$$

where the last identity follows by (3.3) applied with q/p in place of g and $\mathbf{F} = fg$, provided that $fg \in \mathcal{F}_{1,\mathrm{loc}}(p)$ and $q/p \in \mathrm{dom}(p, fg)$. In particular, by Lemma 3.8, the latter condition is satisfied by all $g \in \mathcal{C}_c^\infty(\Omega_p)$. Therefore, $fp\,\partial_e(q/p)$ vanishes Lebesgue-almost everywhere over Ω_p. The same is true for $\partial_e(q/p)$ because fp vanishes nowhere over Ω_p. Since this is true for all $e \in S^{p-1}$ and Ω_p is connected, the ratio q/p is Lebesgue-almost everywhere constant on Ω_p by Proposition A.2. However, as q/p is continuous, it must be constant on the entire Ω_p. In particular, $\Omega_p = \Omega_q$ automatically. Since both p and q integrate to 1 on their support, $p = q$ follows. □

Remark 3.15. Proposition A.2, which was applied in the proof, can be regarded as the extreme case of the *Poincaré inequality* – see Definition 4.5 – in which the inequality turns to an equality. Similarly, Stein characterizations can be regarded as the extreme case of inequalities for Stein disrepancies in which the Stein discrepancy takes on the value 0. In Section 5.2, we derive bounds which are based precisely on the Poincaré inequality.

The directional characterizations from Proposition 3.14 lead to a wide variety of characterizations, among which we highlight the following two.

Proposition 3.16 (Some Stein characterizations)**.** *Let p and q be as in Proposition 3.14.*

1. *Let $f\colon \mathbb{R}^d \to \mathbb{R} \in \mathcal{F}_{1,\mathrm{loc}}(p)$. Then $p = q$ if and only if*

$$\mathbb{E}_q\big[g\mathcal{T}_{\nabla,p}f\big] = -\mathbb{E}_q\big[f\,\nabla g\big]$$

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

2. *Let $\mathbf{F}\colon \mathbb{R}^d \to \mathbb{R}^{m\times d} \in \mathcal{F}_{\mathrm{loc}}(p)$ be such that the matrix $\mathbf{F}(x)$ has zero nullity for Lebesgue-almost all $x \in \Omega_p$. Then $p = q$ if and only if*

$$\mathbb{E}_q\left[g\mathcal{T}_{\mathrm{div},p}\mathbf{F}\right] = -\mathbb{E}_q\big[\mathbf{F}\,\nabla g\big]$$

   for all $g\colon \mathbb{R}^d \to \mathbb{R} \in \mathrm{dom}(p, \mathbf{F})$.

Proof. The first part follows directly from Proposition 3.14. The second one can be proved similarly: if $p = q$, the desired equality follows from (3.9). For the opposite direction, (3.8) and (3.9) give $\mathbb{E}_p\big[\mathbf{F}g\,\nabla(q/p)\big] = 0$ for all $g \in \mathcal{C}_c^\infty(\Omega_p)$. Therefore, $\mathbf{F}\,\nabla(q/p)$ vanishes Lebesgue-almost everywhere over Ω_p. Since $\mathbf{F}(x)$ has zero nullity for Lebesgue-almost all $x \in \Omega_p$, $\nabla(q/p)$ also vanishes Lebesgue-almost everywhere. As a result, q/p is constant on Ω_p and the result follows. □

3.4 The score function and the Stein kernel

We conclude the section with three key examples of Stein operators.

3.4.1 Gradient based first order operators and the score function

We first focus on operator (3.6) and its companion Stein identity (3.7). We deduce a family of Stein operators for p obtained by fixing some differentiable f and considering the first order operator $\mathcal{A}_p := \mathcal{A}_{f,p}$ given by

$$g \mapsto \mathcal{A}_p g := \mathcal{T}_{\nabla,p}(fg)$$

with associated Stein class $g\colon \mathbb{R}^d \to \mathbb{R} \in \mathcal{F}(\mathcal{A}_p) = \mathrm{dom}(p, f)$. Each particular a.s. differentiable f thus gives rise to a particular operator, acting on a particular class of functions, entailing a particular Stein identity. One choice for f stands out: $f = 1$.

Definition 3.17 (Score function and operator)**.** *Let p be differentiable. The* score function *of p is the function*

$$\rho_p = \mathcal{T}_{\nabla,p}1 = \nabla \log p = \frac{\nabla p}{p}$$

(still with the convention that $\rho_p \equiv 0$ outside of Ω_p). The score-Stein operator *is the vector-valued operator*

$$\mathcal{A}_p = \nabla + \rho_p \mathbf{I} \tag{3.15}$$

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

Operator (3.15) is the most classical Stein operator for multivariate p. It is particularly useful in the context of Stein's method when $\mathcal{F}_1(p)$ contains all constant functions. The latter assumption holds if p is a differentiable pdf such that $\partial_i p$ is integrable for all $i = 1, \dots, d$ and $\int \partial_i p = 0$ (this is not guaranteed if we only assume that p is continuously differentiable: consider for example $p(x) \propto x^2 \sin(x^{-2})$ on $(0,1)$). Then we can take $\mathcal{F}(\mathcal{A}_p) = \mathrm{dom}(p, 1)$; the resulting (characterizing) Stein identity is

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

If $1 \notin \mathcal{F}_1(p)$ then a version of this identity still holds, but with integration constants which must be taken into account in the various identities, see Theorem 5.1.

Example 3.18 (Gaussian score operator)**.** For $\mathcal{N}(\nu, \Sigma)$ with pdf γ, we have $\rho_\gamma(x) = \nabla \log \gamma(x) = -\Sigma^{-1}(x - \nu)$; the corresponding operator is

$$\mathcal{A}_\gamma g(x) = \nabla g(x) - \Sigma^{-1}(x - \nu) g(x).$$

The flexibility of choice of functions f clarifies a relationship between Stein operators, as follows.

Example 3.19 (Gradient based operators and change of measure)**.** Suppose that f is such that fp is a pdf which is continuously differentiable on its support Ω_p. Then, for any g which is continuously differentiable on Ω_p,

$$\mathcal{T}_{\nabla,p}\left(fg\right) = \frac{\nabla(gfp)}{p} = f\,\frac{\nabla(gfp)}{fp} = f\,\mathcal{T}_{\nabla,fp}(g)\,.$$

Such a change of measure operation can be useful for finding solutions of Stein equations. The Stein equation

$$\mathcal{T}_{\nabla,p}(fg) = h - \mathbb{E}_p h$$

then translates into

$$\mathcal{T}_{\nabla,fp}(g) = \frac{1}{f}\left(h - \mathbb{E}_p h\right).$$

For instance, if p is log-concave then, under additional regularity conditions, solutions of the Stein equation with bounded derivatives are available (see [35]). In Proposition 5.15 we shall show a similar result under the assumption of a Poincaré constant. Such bounds may also be applied even when p is not log-concave or does not possess a Poincaré constant, but there exists f such that fp is has this desired property and the change of measure illustrated above can be applied. As an illustration, the one-dimensional Beta distribution with pdf $p(x) = x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha,\beta)$ on $(0,1)$ is not log-concave if $\min(\alpha,\beta) < 1$. Choosing $f(x) = B(\alpha,\beta)x(1-x)/B(\alpha+1,\beta+1)$ results in fp being log-concave. However, as a different Stein equation is being solved, Stein's method need not bound the error in the same metric. For log-concave densities, such as fp, one can for instance obtain bounds in the Wasserstein distance, while for p, the bound is expressed in a different metric. This change of measure leads to p and fp to be nested in the sense of Section 5.1, where the focus is on Wasserstein distance.

3.4.2 Divergence based first order operators and Stein kernels

We now start from the product rule (3.8) and the corresponding Stein identity (3.9), with $\mathbf{F}\colon\ \mathbb{R}^d \to \mathbb{R}^{m\times d}$ and $g\colon\ \mathbb{R}^d \to \mathbb{R}$. We deduce a family of Stein operators for p obtained by fixing $\mathbf{F}$ and considering

$$g \mapsto \mathcal{A}_p g = \mathcal{T}_{\mathrm{div},p}\big(\mathbf{F} g\big) = (\mathcal{T}_{\mathrm{div},p}\mathbf{F})g + \mathbf{F}\,\nabla g$$

with associated Stein class $\mathcal{F}(\mathcal{A}_p) := \mathrm{dom}(p, \mathbf{F})$.

Taking $\mathbf{F} = I_d$ the identity matrix results in (3.15). If p has finite covariance Σ, a more natural choice may be $\mathbf{F} \equiv \Sigma$, but the resulting operator is not intrinsically different from (3.15). A different group of choices for $\mathbf{F}$ stands out: *Stein kernels*, which we define as follows.

Definition 3.20 (Stein kernel and operator)**.** *Suppose p has finite mean $\nu \in \mathbb{R}^d$. For each unit vector $e_i \in \mathbb{R}^d$, $i = 1, \dots, d$,* a Stein kernel *for p* in direction e_i *is any vector field $x \mapsto \tau_i(x) \in \mathbb{R}^d$ such that $\tau_i p \in \mathrm{ACL}^1_{\mathrm{loc}}(\Omega_p)$ and, for Lebesgue-almost all $x \in \Omega_p$,*

$$\mathcal{T}_{\mathrm{div},p}\big(\tau_i\big)(x) = \nu_i - x_i.$$

A Stein kernel *is any square matrix-valued function $\boldsymbol{\tau} = (\tau_{i,j})_{1 \leqslant i,j \leqslant d}$ such that each row $\boldsymbol{\tau}_i = (\tau_{i1}, \dots, \tau_{id})$ is a Stein kernel for p in direction e_i;*

$$\mathcal{T}_{\mathrm{div},p}\boldsymbol{\tau} = \nu - \mathrm{Id} \quad \textit{a.e. on } \Omega_p. \tag{3.16}$$

For a given Stein kernel $\boldsymbol{\tau}$ the $\boldsymbol{\tau}$-kernel-Stein operator *acting on scalar-valued functions g is the $\mathbb{R}^d$-valued operator*

$$\mathcal{A}_p g(x) = \mathcal{T}_{\mathrm{div},p}\big(\boldsymbol{\tau} g\big)(x) = \boldsymbol{\tau}(x)\, \nabla g(x) - (x - \nu) g(x)$$

with domain $\mathcal{F}(\mathcal{A}_p)$.

Example 3.21 (Gaussian Stein kernel)**.** For $\mathcal{N}(\nu, \Sigma)$ with pdf γ, as we shall see in Section 4, the constant function with value Σ is a Stein kernel for γ; the corresponding operator is

$$\mathcal{A}_\gamma g(x) = \Sigma\, \nabla g(x) - (x - \nu) g(x).$$

The only difference to the score-Stein operator from Example 3.18 is the position of the covariance matrix Σ.

We note that the definition of a Stein kernel does not make any assumption about its nullity, but in view of Proposition 3.14, often a Stein kernel with zero nullity may be desirable.

As argued in the Introduction, Stein kernels play an important role in the study of probability distributions. We defer a detailed study (existence, construction, examples, applications) to Section 4.

3.4.3 Divergence based second order operators

Finally we consider the product rule (3.12) and corresponding Stein identity (3.13), for some differentiable matrix valued function $\mathbf{F}$. We deduce the family of operators

$$g \mapsto \mathcal{A}_p g = \mathcal{T}_{\mathrm{div},p}\big(\mathbf{F}^T \nabla g\big) = \langle \mathcal{T}_{\mathrm{div},p}\mathbf{F}, \nabla g \rangle + \big\langle \mathbf{F}, \nabla^2 g \big\rangle_{\mathrm{HS}} \tag{3.17}$$

with corresponding class $\mathcal{F}(\mathcal{A}_p)$ the collection of g such that $\nabla g \in \mathrm{dom}(p, \mathbf{F})$.

In line with the previous considerations, two choices for $\mathbf{F}$ stand out:

(i) $\mathbf{F} \equiv \mathrm{I}_d$, so that

$$\mathcal{A}_p g = \langle \nabla \log p, \nabla g \rangle + \Delta g \tag{3.18}$$

with class $\mathcal{F}(\mathcal{A}_p)$ the collection of g such that $\nabla g \in \mathrm{dom}(p, I_d)$;

(ii) $\mathbf{F} = \boldsymbol{\tau}$, a Stein kernel for p, so that

$$\mathcal{A}_p g = \langle \nu - \mathrm{Id}, \nabla g \rangle + \langle \boldsymbol{\tau}, \nabla^2 g \rangle_{\mathrm{HS}} \tag{3.19}$$

with domain $\mathcal{F}(\mathcal{A}_p)$ the collection of g such that $\nabla g \in \mathrm{dom}(p, \boldsymbol{\tau})$.

Example 3.22 (Gaussian second order operator)**.** For $\mathcal{N}(\nu, \Sigma)$ with pdf γ, (3.18) yields

$$\mathcal{A}_\gamma g(x) = \Delta g(x) - \left\langle \Sigma^{-1}(x-\nu), \nabla g(x) \right\rangle$$

whereas (3.19) with $\boldsymbol{\tau} = \Sigma$ yields

$$\mathcal{A}_\gamma g(x) = \left\langle \Sigma, \nabla^2 g \right\rangle_{\mathrm{HS}} - \left\langle x-\nu, \nabla g(x) \right\rangle.$$

Example 3.23 (Generators of diffusions)**.** [Gaussian second order operator] Infinitesimal generators of multivariate diffusions as studied e.g. in [25, 19] are of the form (3.17), with $\mathbf{F}$ imposed by the properties of the underlying process. Indeed consider a measure μ with pdf p in $\mathcal{C}^2(\mathbb{R}^d)$ which is the ergodic measure of the Itô stochastic differential equation

$$\mathrm{d}Z_t = b(Z_t)\,\mathrm{d}t + \sigma(Z_t)\,\mathrm{d}B_t, \quad Z_0 = x,$$

where B_t is a standard d-dimensional Brownian motion, $b\colon\ \mathbb{R}^d \to \mathbb{R}^d$ is a sufficiently regular (typically Lipschitz) drift coefficient and $\sigma\colon\ \mathbb{R}^d \to \mathbb{R}^{d\times m}$ is a sufficiently regular (again, typically Lipschitz) diffusion coefficient. Let $\mathbf{a} = \sigma\sigma^T$ denote the covariance coefficient of the process. The Stein operators from [25, Theorem 2] are of the form

$$\mathcal{A}_p g = \frac{1}{2}\mathcal{T}_{\mathrm{div},p}\left((\mathbf{a}+\mathbf{c})\,\nabla g\right)$$

where $\mathbf{c}$ is a differentiable skew-symmetric matrix-valued function such that $\mathcal{T}_{\mathrm{div},p}\mathbf{c}(x)$ is nonreversible.

4 More about Stein kernels

In this section we study constructions of Stein kernels and give some applications towards distributional comparisons.

4.1 Existence and construction

From Definition 3.20, a matrix-valued function $\boldsymbol{\tau}$ is a Stein kernel for a continuously differentiable pdf p if and only if $\boldsymbol{\tau}\, p$ is continuously differentiable on Ω_p and (3.16) holds. By the product rule (2.3), the score function and the divergence operator are linked through

$$\mathcal{T}_{\mathrm{div},p}\mathbf{F} = \mathbf{F}\rho_p + \mathrm{div}(\mathbf{F}). \tag{4.1}$$

Hence for a pdf p with mean ν, any continuously differentiable matrix-valued function $\boldsymbol{\tau}$ satisfying Lebesgue-almost surely on Ω_p,

$$\boldsymbol{\tau}\,\rho_p + \mathrm{div}(\boldsymbol{\tau}) = \nu - \mathrm{Id}$$

is a Stein kernel for p. This leads to the following simple explicit construction of a family of Stein kernels.

Lemma 4.1. *A Stein kernel can be constructed as*

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

where $\mathbf{F}$ *is a continuously differentiable matrix-valued function such that for some constant* $\alpha, \beta \in \mathbb{R}$*,* $\alpha+\beta \neq 0$ *and a function* $\mathbf{r}\colon\ \mathbb{R}^d \to \mathbb{R}^d$*,*

$$\mathbf{F}(x)\,\rho_p(x) = \alpha(\nu - x) + \mathbf{r}(x) \tag{4.2}$$

$$\mathrm{div}\,\mathbf{F}(x) = \beta(\nu - x) - \mathbf{r}(x). \tag{4.3}$$

Proof. Inserting (4.2) and (4.3) into (4.1) gives the assertion. □

Remark 4.2. Using product rule (3.8) and a density argument, it is straightforward to see that showing some differentiable matrix valued function is a Stein kernel can be done equivalently by checking that the Stein identity

$$\mathbb{E}_p\left[\boldsymbol{\tau}\,\nabla g\right] = \mathbb{E}_p\left[(\mathrm{Id}-\nu)g\right] \tag{4.4}$$

holds at least for $g \in \mathcal{C}_c^\infty(\Omega_p)$; this concurs with the original definitions of multivariate Stein kernels as proposed e.g. in [38].

Remark 4.3. Equation (4.4) can be used to relate the moments of p with the moments of a Stein kernel. For instance, if $\boldsymbol{\tau}$ is a Stein kernel for p, if p has finite variance Σ, and if $\mathrm{Id}-\nu \in \mathrm{dom}(p,\boldsymbol{\tau})$ (a mild condition), it follows directly from (4.4) that $\mathbb{E}\left[\boldsymbol{\tau}\right] = \Sigma$.

Example 4.4 (Gaussian Stein kernel)**.** For $\mathcal{N}(\nu,\Sigma)$, (3.14) entails that $\boldsymbol{\tau}(x) = \Sigma$ is a Gaussian Stein kernel. Example 4.18 will illustrate that it is not the only one.

In the one-dimensional case, the first order differential equation corresponding to (4.4) is easy to solve. If Ω_p is an interval (possible infinite), then

$$\tau(x) = \frac{1}{p(x)}\int_{-\infty}^{x}(\nu-u)\,p(u)\,\mathrm{d}u = -\frac{1}{p(x)}\int_{x}^{\infty}(\nu-u)\,p(u)\,\mathrm{d}u$$

is the unique Stein kernel with $\tau(x)\,p(x)$ tending to zero as x approaches the boundary of Ω_p or infinity. We refer to [17, 43] for an overview and bibliography on one-dimensional Stein kernels. In higher dimensions, given that there exist infinitely many functions which share a divergence, the Stein kernel is by no means uniquely defined. Aside from the construction in Lemma 4.1, we also have a result, due to [15], which provides a characterizing Stein kernel under the assumption that the distribution admits a Poincaré constant, as follows.

Definition 4.5. *A probability distribution P satisfies a Poincaré inequality if there exists a constant $C<\infty$ such that for every locally Lipschitz function $\varphi \in L^2(P)$ with expectation $\mathbb{E}_P\varphi = 0$, we have $\mathbb{E}_P\varphi^2 \leqslant C\mathbb{E}_P\|\nabla\varphi\|^2$. The smallest constant for which this inequality holds is the Poincaré constant of P, denoted C_P; it is also referred to as P's spectral gap.*

Example 4.6 (Stein kernels under a Poincaré inequality)**.** In [15] a Stein kernel is defined as any matrix $\boldsymbol{\tau}$ such that (1.2) holds for all $g\colon\ \mathbb{R}^d \to \mathbb{R}^d$ in the Sobolev space $W_p^{1,2}$. With this definition it is shown that if p satisfies a Poincaré-type inequality then it possesses a Stein kernel such that each row is the gradient of a vector field.

Our next main result, Theorem 4.12, provides another explicit construction which is valid under very weak assumptions. We start with a formula for bivariate distributions. For a bivariate density p, denote by p_1 the marginal density of the first component (i. e., in direction e_1) and by $p_{2|1}(x_2 \mid x_1) := p(x_1,x_2)/p_1(x_1)$ the conditional density of the second component given the first one. Next, define $\partial_1 p_{2|1}$ and $\partial_2 p_{2|1}$ as $\partial_1 p_{2|1}(x_2 \mid x_1) := \partial_{x_1}(x_2 \mid x_1)$ and $\partial_2 p_{2|1}(x_2 \mid x_1) := \partial_{x_2}(x_2 \mid x_1)$.

Lemma 4.7 (Bivariate Stein kernels)**.** *Let p be a continuous pdf on $\mathbb{R}^2$ which is $\mathcal{C}^1$ on its support Ω_p. Suppose that each $x_1 \in \Omega_{p_1}$ has a neighbourhood U such that $\int_{-\infty}^{\infty}\sup_{u\in U}\left|\partial_1 p_{2|1}(v \mid u)\right|\mathrm{d}v < \infty$. Moreover, suppose that p_1 has finite mean ν_1, and let τ_1 be the corresponding univariate kernel. Set $\tau_{11}(x_1,x_2) = \tau_1(x_1)$ and*

$$\begin{aligned}\tau_{12}(x_1,x_2) &= -\frac{\tau_1(x_1)}{p_{2|1}(x_2 \mid x_1)}\int_{-\infty}^{x_2}\partial_1 p_{2|1}(v \mid x_1)\,\mathrm{d}v \\ &= \frac{\tau_1(x_1)}{p_{2|1}(x_2 \mid x_1)}\int_{x_2}^{\infty}\partial_1 p_{2|1}(v \mid x_1)\,\mathrm{d}v\,.\end{aligned} \tag{4.5}$$

Then the vector $(x_1, x_2) \mapsto (\tau_{11}(x_1, x_2), \tau_{12}(x_1, x_2))$ *is a Stein kernel for* p *in the direction* e_1*. A Stein kernel for* p *in the direction* e_2 *is defined similarly, by reversing the roles of* x_1 *and* x_2*.*

Proof. First, we observe that both expressions in the right-hand side of (4.5) agree because $\int_{-\infty}^{\infty} \partial_1 p_{2|1}(v \mid x_1)\,\mathrm{d}v = \partial_{x_1} \int_{-\infty}^{\infty} p_{2|1}(v \mid x_1)\,\mathrm{d}v = 0$: the stated conditions allow us to differentiate under the integral sign. The fact that $\sum_{j=1}^{2} \partial_{x_j}\big[\tau_{1j}(x_1, x_2)\, p(x_1, x_2)\big] = (\nu_1 - x_1)\, p(x_1, x_2)$ for all $(x_1, x_2) \in \Omega_p$ follows easily from the definitions of τ_{11} and τ_{12} and some straightforward manipulations. Finally, to check differentiability of $(\tau_{11}, \tau_{12})p$, we rewrite (4.5) as

$$\tau_{12}(x_1, x_2)\, p(x_1, x_2) = \tau_1(x_1)\, p_1(x_1) \int_{x_2}^{\infty} \partial_1 p_{2|1}(v \mid x_1)\,\mathrm{d}v\,. \qquad \square$$

Remark 4.8. The inspiration for formula (4.5) is [6, Equation (9)], where a similar quantity is introduced via a transport argument. To see the connection, assume that $\Omega_p = \Omega \times \Omega$ for an open set $\Omega \subseteq \mathbb{R}$. Fix $i = 1$ and, for each $t, t', x_2 \in \Omega$ let $x_2 \mapsto T_{t,t'}(x_2)$ be the map $\Omega \to \Omega$ transporting the conditional pdf at $x_1 = t$ to that at $x_1 = t'$. By the transformation formula, we have

$$p_{2|1}(x_2 \mid t) = p_{2|1}\big(T_{t,t'}(x_2 \mid t')\big)\, \partial_{x_2} T_{t,t'}(x_2). \tag{4.6}$$

In particular $T_{t,t}(x_2) = x_2$ and $\partial_{x_2} T_{t,t}(x_2) = 1$. Assume that the function $(t, t', x_2) \mapsto T_{t,t'}(x_2)$ is twice continuously differentiable. Taking derivatives in (4.6) with respect to t' and setting $t' = t = x_1$ we deduce that

$$\begin{aligned}\partial_2 p_{2|1}(x_2 \mid x_1)\, \partial_{t'} T_{t,t'}(x_2)\big|_{t'=t=x_1} + \partial_1 p_{2|1}(x_2 \mid x_1) \\ + p_{2|1}(x_2 \mid x_1)\, \partial_{t'} \partial_{x_2} T_{t,t'}(x_2)\big|_{t'=t=x_1} = 0\,.\end{aligned}$$

Interchanging the differentiation in the last term and applying the product rule, we obtain

$$\partial_{x_2}\Big[p_{2|1}(x_2 \mid x_1)\, \partial_{t'} T_{t,t'}(x_2)\big|_{t'=t=x_1}\Big] + \partial_1 p_{2|1}(x_2 \mid x_1) = 0\,.$$

Integrating by x_2, applying (4.5) and multiplying by $p_1(x_1)$, we find that the functions

$$p(x_1, x_2)\, \partial_{t'} T_{t,t'}(x_2)\big|_{t'=t=x_1} \quad \text{and} \quad p(x_1, x_2)\, \frac{\tau_{12}(x_1, x_2)}{\tau_1(x_1)}$$

differ only by a function of x_1. Thus, the ratio of the Stein kernel components corresponds to the direction of the transport from x_1 to x_2 at t. Another construction connected to optimal transport considerations is provided in [20].

Next we compute this bivariate kernel for several examples.

Example 4.9 (Bivariate Gaussian)**.** For the bivariate Gaussian distribution $\mathcal{N}(\nu, \Sigma)$ direct computations of the kernel in Lemma 4.7 lead to $\boldsymbol{\tau}(x) = \Sigma$.

Example 4.10 (Bivariate Student)**.** For the bivariate Student distribution $t_k(\nu, \Sigma)$ direct computations of the kernel in Lemma 4.7 give $\boldsymbol{\tau}(x) = \frac{1}{k-1}\big((x-\nu)(x-\nu)^T + k\Sigma\big)$ which we shall encounter again in Example 4.19. See also example 3.6. Note that this Stein kernel cannot be written as a gradient.

Example 4.11 (Bivariate normal-gamma)**.** Direct computations of the kernel in Lemma 4.7 for the bivariate normal-gamma distribution $NG(\mu, \lambda, \alpha, \beta)$ with pdf

$$p(x_1, x_2) = \frac{\beta^\alpha \sqrt{\lambda}}{\Gamma(\alpha)\sqrt{2\pi}} x_2^{\alpha - \frac{1}{2}} e^{-x_2\beta - \frac{1}{2}x_2\lambda(x_1-\mu)^2}$$

on $(x_1, x_2) \in \mathbb{R} \times \mathbb{R}^+$ give

$$\boldsymbol{\tau}(x_1, x_2) = \begin{pmatrix} \frac{\kappa(x_1-\mu)^2+2\beta}{\kappa(2\alpha-1)} & \frac{2(x_1-\mu)x_2}{2\alpha-1} \\ \frac{x_1-\mu}{2\beta} & \frac{x_2}{\beta} \end{pmatrix}$$

where, for the second line, we simply exchange the roles of x_1 and x_2 in Lemma 4.7. Notably, this Stein kernel is not symmetric.

Next, we state and prove Theorem 4.12, which, inspired by [4], gives a mechanism extending the bivariate construction from Lemma 4.7 to arbitrary dimensions.

Theorem 4.12. *Let $p\colon\ \mathbb{R}^d \to (0,\infty)$ be a continuously twice differentiable pdf on $\mathbb{R}^d$ with*

$$\int \frac{\|\nabla p\|^2}{p} < \infty \quad \text{and} \quad \int \left\|\nabla^2(p)\right\| < \infty$$

and such that p has finite variance. Denote by $\tau_i^{(1)}, i = 1, \dots, d$ the marginal Stein kernels. Then, for any direction $e_i, i = 1, \dots, d$ there exists a Stein kernel $\tau_{p,i}^{(d)}(x)$ for p in direction e_i,

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

such that $\tau_i^{(d)}(x \mid x_i) = \left(\tau_{i,1}^{(d)}(x \mid x_i), \dots, \tau_{i,d}^{(d)}(x \mid x_i)\right)^T$ as a function of x solves the equation

$$\mathcal{T}_{\mathrm{div},p}(\tau_i^{(d)}(x \mid x_i)) = \rho_i(x_i). \tag{4.7}$$

Here $\rho_i(x_i) = p_i'(x_i)/p_i(x_i)$ is the score function of the marginal of p in direction e_i and $x = (x_1, \dots, x_d)$. Moreover, $\tau_{i,i}^{(d)}(x \mid x_i) = \tau_i^{(1)}(x_i)$.

Proof. Let e_i be a unit vector and p_i the marginal of p in direction e_i. The result is almost immediate from [4, Theorem 4], where it is proved (see middle of page 978) that, under the stated conditions, there exist continuously differentiable vector fields $\tau_i^{(d)}(x \mid x_i)$ as functions of x such that

$$\frac{\mathrm{div}_x\left(\tau_i^{(d)}(x \mid x_i)(x)\, p(x)\right)}{p(x)} = \frac{p_i'(x_i)}{p_i(x_i)}$$

and such that the component of $\tau_i^{(d)}(x \mid x_i)$ in direction e_i equals 1: $\langle \tau_i^{(d)}(x \mid x_i), e_i \rangle = 1$ for all x. Thus, (4.7) holds. To see the connection with Stein kernels, write $\tau_{ij}^{(d)}(x) = \tau_i^{(1)}(x_i)\, \tau_{ij}^{(d)}(x \mid x_i)$. Then

$$\begin{aligned}
\sum_{j=1}^d \partial_{x_j}\left(\tau_{ij}^{(d)}(x)\, p(x)\right) &= \sum_{j=1}^d \partial_{x_j}\left(\tau_{ij}^{(d)}(x \mid x_i)\, p(x)\, \tau_i^{(1)}(x_i)\right) \\
&= \sum_{j=1}^d \partial_{x_j}\left(\tau_{ij}^{(d)}(x \mid x_i)\, p(x)\right) \tau_i^{(1)}(x_i) \\
&\qquad + \sum_{j=1}^d \tau_{ij}^{(d)}(x \mid x_i)\, p(x)\, \partial_{x_j}\left(\tau_i^{(1)}(x_i)\right) \\
&= \rho_i(x_i)\, p(x)\, \tau_i^{(1)}(x_i) + p(x)\, \partial_i \tau_i^{(1)}(x_i)\,,
\end{aligned}$$

where in the last line we use (4.7) in the first sum and $\partial_{x_j}(\tau_i^{(1)}(x_i)) = 0$ for all $j \neq i$ in the second sum. By the definition of the univariate Stein kernel,

$$\partial_i {\tau_i}^{(1)}(x_i) = -\rho_i(x_i) {\tau_i}^{(1)}(x_i) + \mathbb{E}[X_i] - x_i\,.$$

The claim follows. □

Remark 4.13. The proof of [4, Theorem 4] provides an explicit solution of (4.7), allowing us to generalize the bivariate construction from Lemma 4.7 to the d-variate case under the same conditions. A Stein kernel in direction e_1 can be constructed in terms of the conditional densities

$$p_{j,j+1,\dots,d|1}(x_j, x_{j+1}, \dots, x_d \mid x_1) = \frac{p_{1,j,j+1,\dots,d}(x_1, x_j, x_{j+1}, \dots, x_d)}{p_1(x_1)}$$

and marginal cumulative distribution functions $P_i(x_i) = \int_{-\infty}^{x_i} p_i(v)\,\mathrm{d}v$ as follows: firstly, set $\tau_{1,1}^{(d)}(x_1, x_2, \dots, x_d) = \tau_1(x_1)$. For $j = 2, 3, \dots, d-1$, set

$$\begin{aligned}\tau_{1,j}^{(d)}(x_1, x_2, \dots, x_d) = \tau_1(x_1)\, &\frac{p_1(x_1)\, p_2(x_2) \cdots p_{j-1}(x_{j-1})}{p(x_1, x_2, \dots, x_d)} \\ &\times \int_{x_j}^{\infty} \partial_1 p_{j,j+1,\dots,d|1}(v, x_{j+1}, x_{j+2}, \dots, x_d \mid x_1)\,\mathrm{d}v \\ + \tau_1(x_1)\, &\frac{p_1(x_1)\, p_2(x_2) \cdots p_{j-1}(x_{j-1})\, P_j(x_j)}{p(x_1, x_2, \dots, x_d)} \\ &\times \partial_1 p_{j+1,j+2,\dots,d|1}(x_{j+1}, x_{j+2}, \dots, x_d \mid x_1)\,.\end{aligned}$$

Finally, set

$$\tau_{1,d}^{(d)}(x_1, x_2, \dots, x_d) = \tau_1(x_1)\, \frac{p_1(x_1)\, p_2(x_2) \cdots p_{d-1}(x_{d-1})}{p(x_1, x_2, \dots, x_d)} \int_{x_d}^{\infty} \partial_1 p_{d|1}(v \mid x_1)\,\mathrm{d}v$$

(the partial derivative ∂_1 is defined as in the bivariate case). A straightforward, though somewhat involved calculation shows that $\left(\tau_{1,1}^{(d)}, \dots, \tau_{1,d}^{(d)}\right)^T$ is indeed a Stein kernel in direction e_1. Stein kernels in other directions can be obtained analogously by rotating the indices.

4.2 Stein kernels for elliptical distributions

In this subsection we construct a family of Stein kernels for any member of the family of elliptical distributions.

Definition 4.14. *The multivariate elliptical distribution $E_d(\nu, \Sigma, \phi)$ on $\mathbb{R}^d$ has pdf*

$$p(x) = \kappa \left[\det(\Sigma)\right]^{-1/2} \phi\left(\frac{1}{2}(x-\nu)^T \Sigma^{-1} (x-\nu)\right), \quad x \in \mathbb{R}^d, \tag{4.8}$$

for $\phi\colon\ \mathbb{R}^+ \to \mathbb{R}^+$ a measurable function, $\nu \in \mathbb{R}^d$, $\Sigma = (\sigma_{ij})$ a symmetric positive definite $d \times d$ matrix, and κ the normalizing constant.

Note that the matrix Σ in (4.8) is not necessarily the covariance matrix; also not all choices of ϕ lead to well-defined densities, see [29] for a discussion and references.

Some prominent members of the elliptical family are the Gaussian distribution $\mathcal{N}_d(\nu, \Sigma)$, with $\phi(t) = e^{-t}$; the power exponential distribution, with $\phi(t) = \exp(-b_{p,\zeta} t^{\zeta})$ for $\zeta > 0$ and $b_{p,\zeta}$ a scale factor; the multivariate Student-t distribution, with $\phi(t) = (1 + 2t/k)^{-(k+d)/2}$; and the spherical distributions $E_d(0, \mathrm{I}_d, \phi)$. For simplicity, we assume that $\phi(t) > 0$ for all $t \geq 0$ so that $\Omega_p = \mathbb{R}^d$. Example 2.1 in [21] shows that in order to find Stein kernels for elliptical distributions, it suffices to consider spherical distributions, so that $\Sigma = \mathrm{I}_d$ and $\nu = 0$. However as the notion of Stein kernel is not as broad in [21] we provide a proof here.

Proposition 4.15. *The application $\boldsymbol{\tau} \mapsto \left[x \mapsto \Sigma^{1/2} \boldsymbol{\tau}(\Sigma^{-1/2}(x-\nu))\Sigma^{1/2}\right]$ maps Stein kernels of $E_d(0, \mathrm{I}_d, \phi)$ to Stein kernels of $E_d(\nu, \Sigma, \phi)$ and is a bijection.*

Proof. Fix a matrix A and a vector b of a proper dimension, define $\tilde{\mathrm{A}}(x) := \mathrm{A}x + b$. By the chain rule, we have $\operatorname{div}(\mathbf{F} \circ \tilde{\mathrm{A}}) = \big(\operatorname{div}(\mathbf{F}\mathrm{A}^T)\big) \circ \tilde{\mathrm{A}}$ for any suitable matrix-valued function $\mathbf{F}$. Alternatively, one can write $\operatorname{div}\big((\mathbf{F}\mathrm{A}^{-T}) \circ \tilde{\mathrm{A}}\big) = (\operatorname{div}\mathbf{F}) \circ \tilde{\mathrm{A}}$. Letting $\mathbf{F} = p\boldsymbol{\tau}$, where p is a scalar-valued function, we obtain $\mathcal{T}_{\operatorname{div},p\circ\tilde{\mathrm{A}}}\big((\boldsymbol{\tau}\mathrm{A}^{-T}) \circ \tilde{\mathrm{A}}\big) = (\mathcal{T}_{\operatorname{div},p}\boldsymbol{\tau}) \circ \tilde{\mathrm{A}}$. Now let $\tilde{\mathrm{A}}(x) := \Sigma^{-1/2}(x-\nu)$ and let p be the density of $E_d(0, \mathrm{I}_d, \phi)$. Clearly, $q := [\det(\Sigma)]^{-1/2}(p\circ\tilde{\mathrm{A}})$ is the density of $E_d(\nu, \Sigma, \phi)$. If $\boldsymbol{\tau}$ is a Stein kernel for $E_d(0, \mathrm{I}_d, \phi)$, then $\mathcal{T}_{\operatorname{div},p}\boldsymbol{\tau}(z) = -z$. Letting $z = \tilde{\mathrm{A}}(x)$, we then have $\mathcal{T}_{\operatorname{div},q}\big((\boldsymbol{\tau}\Sigma^{1/2}) \circ \tilde{\mathrm{A}}\big)(x) = \mathcal{T}_{\operatorname{div},p\circ\tilde{\mathrm{A}}}\big((\boldsymbol{\tau}\Sigma^{1/2}) \circ \tilde{\mathrm{A}}\big)(x) = (\mathcal{T}_{\operatorname{div},p}\boldsymbol{\tau})\big(\tilde{\mathrm{A}}(x)\big) = \Sigma^{-1/2}(\nu - x)$. Multiplying by $\Sigma^{1/2}$ from the left, we conclude that $x \mapsto \Sigma^{1/2}\boldsymbol{\tau}\big(\Sigma^{-1/2}(x-\nu)\big)\Sigma^{1/2}$ is a Stein kernel for $E_d(\nu, \Sigma, \phi)$. □

The score function for $E_d(\nu, \Sigma, \phi)$ is

$$\rho_p(x) = \Sigma^{-1}(x-\nu)\,\frac{\phi'\big((x-\nu)^T\Sigma^{-1}(x-\nu)/2\big)}{\phi\big((x-\nu)^T\Sigma^{-1}(x-\nu)/2\big)}\,, \quad x \in \mathbb{R}^d.$$

Combining this special form with Lemma 4.1 yields a family of Stein kernels for members of the elliptical distributions.

Proposition 4.16. *For $d \geq 2$, letting $t = (x-\nu)^T\Sigma^{-1}(x-\nu)/2$, the matrix-valued functions*

$$\boldsymbol{\tau}_\delta(x) = \frac{\frac{\phi''(t)/\phi'(t)}{\phi'(t)/\phi(t)} - \delta}{(2-\delta)(d-1)}\left(\left(\frac{d-1}{\delta\frac{\phi'(t)}{\phi(t)} - \frac{\phi''(t)}{\phi'(t)}} + 2t\right)\Sigma - (x-\nu)(x-\nu)^T\right) \tag{4.9}$$

are Stein kernels for $E_d(\nu, \Sigma, \phi)$ for all $\delta \neq 2$, as long as $\boldsymbol{\tau}_\delta \in \mathrm{ACL}^1_{\mathrm{loc}}(\mathbb{R}^d)$.

Proof. We first set $\nu = 0$ and $\Sigma = \mathrm{I}_d$. Using the temporary notation $\psi(t) = \phi(t)/\phi'(t)$, the score function is $\rho_p(x) = x/\psi(t)$ with $t = x^Tx/2$. To use (4.2) with $r = 0$, the equation

$$\mathbf{F}(x)x = -\alpha\psi(t)x$$

is solved for example by $\mathbf{F}(x) = -\alpha\frac{\psi(t)}{2t}xx^T$. More generally a family of solutions of (4.2) with $r = 0$ is given by matrix-valued functions of the form

$$\mathbf{F}(x) = -\alpha\,\frac{\psi(t)}{2(t+f(t))}\left(xx^T + 2f(t)\,\mathrm{I}_d\right)$$

for some $f\colon\ \mathbb{R} \to \mathbb{R}$: it is easy to check that $\mathbf{F}(x)\rho_p(x) = -\alpha x$. For (4.3) with $r = 0$ the flexibility in the choice of f enters: we introduce $b(t) = -\alpha\psi(t)/\big(2(t+f(t))\big)$ so that $\mathbf{F}(x) = b(t)\left(xx^T + 2f(t)\,\mathrm{I}_d\right)$. By (2.3) and straightforward calculation, we obtain

$$\operatorname{div}_x\Big[b(t)\left(xx^T + 2f(t)\,\mathrm{I}_d\right)\Big] = \Big[2tb'(t) + 2\big(b'(t)f(t) + b(t)f'(t)\big) + (d+1)b(t)\Big]x\,.$$

For (4.3) with $r = 0$ to hold, it suffices to choose f such that, for all $t \in \mathbb{R}$,

$$2\big(t+f(t)\big)b'(t) + \big(2f'(t) + (d+1)\big)b(t) = -\beta \tag{4.10}$$

for some β. Since $b(t) = -\alpha\psi(t)/\big(2(t+f(t))\big)$, simple calculations lead to the requirement that

$$t + f(t) = -\frac{(d-1)\psi(t)}{2\psi'(t) - 2\beta/\alpha}$$

at all t. With this in hand, we easily obtain

$$\mathbf{F}(x) = \frac{-\beta + \alpha\psi'(t)}{d-1}\left(xx^T - \left(\frac{(d-1)\psi(t)}{\psi'(t) - \beta/\alpha} + 2t\right)\mathrm{I}_d\right)$$

Plugging in $\psi(t) = \phi(t)/\phi'(t)$ whose derivative is $\psi'(t) = 1 - \phi(t)\phi''(t)/(\phi'(t))^2$, and dividing by $\alpha + \beta$, then setting $\delta = 1 - \beta/\alpha$, Proposition 4.16 ensues for $\nu = 0, \Sigma = \mathrm{I}_d$; the general formula follows from Proposition 4.15. □

The equation (4.10) can be reparameterized as follows.

Corollary 4.17. *Let $a, b\colon \mathbb{R}^d \to \mathbb{R}$ be two continuously differentiable functions such that, for all $t \geqslant 0$,*

$$\frac{(a(t)\phi(t))'}{\phi(t)} + 2t\frac{(b(t)\phi(t))'}{\phi(t)} + (d+1)b(t) + 1 = 0. \tag{4.11}$$

Then

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

with $t = \frac{1}{2}(x-\nu)^T\Sigma^{-1}(x-\nu)$ is a Stein kernel for $E_d(\nu, \Sigma, \phi)$.

Proof. In (4.10) take $\beta = -1$ and require that $a(t)$ satisfies

$$(a\phi)'(t) = 2\phi(t)(bf)'(t) + \frac{t\phi(t)}{t + f(t)}.$$

Then the assertion follows from Proposition 4.16. □

Many options in (4.11) are possible. For instance, setting $b \equiv 0$ gives $a(t) = \frac{1}{\phi(t)}\int_t^{+\infty}\phi(u)\,\mathrm{d}u$, and the matrix-valued function

$$\boldsymbol{\tau}(x) = \left(\frac{1}{\phi((x-\nu)^T\Sigma^{-1}(x-\nu)/2)}\int_{(x-\nu)^T\Sigma^{-1}(x-\nu)/2}^{+\infty}\phi(u)\,\mathrm{d}u\right)\Sigma \tag{4.12}$$

is a Stein kernel for $E_d(\nu, \Sigma, \phi)$ (provided it is continuously differentiable). This recovers [29, Theorem 2]. Setting $a \equiv 0$ leads to:

$$\boldsymbol{\tau}(x) = \left(\frac{t^{-\frac{d+1}{2}}}{2\phi(t)}\int_t^{+\infty} u^{\frac{d-1}{2}}\phi(u)\,\mathrm{d}u\right)(x-\nu)(x-\nu)^T,$$

is a Stein kernel for $E_d(\nu, \Sigma, \phi)$; we have so far not found any use for this formula.

Example 4.18 (Stein kernels for the multivariate Gaussian distribution)**.** As the multivariate Gaussian has $\phi(t) = e^{-t}$ and $\phi'(t)/\phi(t) = -1$, we recover that

$$\rho_\gamma(x) = -\Sigma^{-1}(x-\nu)$$

is the score function of γ. Since $\frac{1}{\phi(t)}\int_t^{\infty}\phi(u)\,\mathrm{d}u = 1$ for all t, (4.12) shows that $\boldsymbol{\tau}_1 = \Sigma$ is a Stein kernel for γ. Moreover, (4.9) yields, after some simplifications, the following family $\boldsymbol{\tau}_\delta(x)$ given for $\delta \neq 2$ by

$$\boldsymbol{\tau}_\delta(x) = \frac{1}{(2-\delta)(d-1)}\left((d-1+2t(1-\delta))\,\Sigma - (1-\delta)(x-\nu)(x-\nu)^T\right) \tag{4.13}$$

are all Stein kernels for γ. In particular, the choice $\delta = 1$ recovers $\boldsymbol{\tau}_0(x) = \Sigma$, although many other choices are possible. First explorations indicate that in this example the freedom of choice in $\boldsymbol{\tau}_\delta$ does not provide improvement over the most natural choice $\boldsymbol{\tau}(x) = \Sigma$, for comparison of normal distributions in Wasserstein distance, see Example 4.20.

Example 4.19 (Stein kernels for the multivariate Student t-distribution)**.** This distribution is an elliptical distribution with $\phi(t) = (1 + 2t/k)^{-(k+d)/2}$ and hence $\phi'(t)/\phi(t) = -(d+k)/(k+2t)$. From $k > 1$ it follows that $d + k > 2$ and

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

Hence (4.12) gives that

$$\boldsymbol{\tau}_1(x) = \frac{(x-\nu)^T\Sigma^{-1}(x-\nu)+k}{d+k-2}\,\Sigma \tag{4.14}$$

is a Stein kernel for the multivariate Student distribution for $k>2$; here $k>2$ guarantees that t_k has finite variance. Also, we note that $\boldsymbol{\tau}_1 \in \mathcal{F}(t_k)$. Similarly, using that $\phi''(t)/\phi'(t) = -(d+k+2)/(k+2t)$, (4.9) gives a family of Stein kernels which are indexed by $\delta \neq 2$:

$$\begin{aligned}\boldsymbol{\tau}_\delta(x) = \frac{1}{(2-\delta)(d+k)(d-1)}\Big[&\{(d-1)(k+2t)+2t((1-\delta)(d+k)+2)\}\,\Sigma \\ &-\{(1-\delta)(d+k)+2\}\,(x-\nu)(x-\nu)^T\Big].\end{aligned} \tag{4.15}$$

We note that $\boldsymbol{\tau}_\delta \in \mathcal{F}(t_k)$ for $k>2$. The particular choice of δ such that $d-1+(1-\delta)(d+k)+2=0$, i.e. $\delta = 1-(d+1)/(d+k)$, eliminates t from (4.15) and, after simplifications, we obtain for $k>2$

$$\boldsymbol{\tau}_2(x) = \frac{1}{k-1}\left((x-\nu)(x-\nu)^T + k\Sigma\right). \tag{4.16}$$

This agrees with the kernel already identified in Example 4.10. When $d=1$, then both $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ simplify to $\tau(x) = (x^2+k\sigma^2)/(k-1)$, the univariate kernel for the Student-t distribution with k degrees of freedom and centrality parameter σ^2, see e.g. [33, page 30]. We will see in Example 4.21 that, when comparing with a Gaussian pdf, neither $\boldsymbol{\tau}_1$ nor $\boldsymbol{\tau}_2$ are "better" choices; in fact optimizing over δ in (4.15) provides possibilities for strict improvement.

4.3 Stein kernel discrepancies

Stein kernels provide a natural means for comparing distributions through the discrepancy

$$S(p_2 \mid p_1) = \inf_{\boldsymbol{\tau}_1,\boldsymbol{\tau}_2} \mathbb{E}_{p_2}[\|\boldsymbol{\tau}_2 - \boldsymbol{\tau}_1\|_{\mathrm{HS}}^2]^{1/2}$$

where the infimum is taken over all Stein kernels for p_1 and for p_2. In, fact, the specific case $p_1=\gamma$ the standard normal distribution has been studied in detail e.g. in [38, 30, 15] where connections with various classical probability metrics as well as information-type discrepancies are identified. In particular it is shown in [30] that

$$\mathcal{W}_2(\gamma, p_2) \leqslant S(p_2 \mid \gamma)$$

(here $\mathcal{W}_2$ denotes the classical 2-Wasserstein distance). We shall see in Example 5.4 in the next section that, outside a Gaussian context, $S(p_2 \mid p_1)$ also bounds 1-Wasserstein distance, under some additional assumptions on p_1.

Example 4.20 (Comparison between Gaussians)**.** For $i=1,2$ let p_i be a centred Gaussian pdf with covariance Σ_i. Then

$$S(p_2 \mid p_1)^2 \leqslant \|\Sigma_2 - \Sigma_1\|_{\mathrm{HS}}. \tag{4.17}$$

If the Σ_i are of the form $\Sigma_i = \begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix}$ for some $\rho_i \in [0,1]$, $i=1,2$, then

$$\|\Sigma_2 - \Sigma_1\|_{\mathrm{HS}} = \sqrt{2}|\rho_1 - \rho_2|$$

which is exactly the value of the 2-Wasserstein distance in this case, see [47, Theorem 2.4]. A legitimate question in this context is whether there is some optimization to be reaped

from the freedom of choice in the Gaussian Stein kernels from (4.13), and considering $\mathbb{E}_{p_j}\|\boldsymbol{\tau}_{1,\delta} - \boldsymbol{\tau}_{2,\delta}\|_{\mathrm{HS}}$ with $j = 1, 2$ and optimizing over δ. Explicit numeric computations with the various kernels indicate that the answer is negative; they show that the smallest bound is attained at $\delta = 1$ and $\boldsymbol{\tau}_{i,1} = \Sigma_i$, $i = 1, 2$. The covariance matrix is therefore, in this sense, the "best" Stein kernel for the Gaussian pdf. In the next example we will exhibit a situation where optimization is in fact possible.

Example 4.21 (Student vs Gaussian)**.** Let $p_1 = \gamma$ be the standard Gaussian pdf for which we fix $\boldsymbol{\tau}_1 = \mathrm{I}_d$, and $p_2 = t_k$ the centred Student pdf with k degrees of freedom and shape $\Sigma = \mathrm{I}_d$. The Stein kernels for this distribution from Example 4.19 provide a variety of possible differences $\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2$. For instance we get

$$\boldsymbol{\tau}_1(x) - \boldsymbol{\tau}_2(x) = \left(1 - \frac{x^T x + k}{d + k - 2}\right)\mathrm{I}_d \quad \text{and} \quad \boldsymbol{\tau}_1(x) - \boldsymbol{\tau}_2(x) = -\frac{1}{k-1}\left(xx^T + \mathrm{I}_d\right)$$

where the first is obtained by taking $\boldsymbol{\tau}_2$ as given in (4.14) and the second by taking $\boldsymbol{\tau}_2$ as given in (4.16). These expressions lead to

$$S_1(\gamma \mid t_k) = \frac{\sqrt{2d(d+2)}}{d+k-2} \quad \text{and} \quad S_2(\gamma \mid t_k) = \frac{\sqrt{d(5+d)}}{k-1}\,,$$

respectively. When $d = 1$ we get $S_1(\gamma \mid t_k) = S_2(\gamma \mid t_k) = \sqrt{6}/(k-1)$ which concurs with [33, Section 6.3]. In dimension $d \geq 2$, both bounds tell a similar story, although for fixed d the bound $S_2(\gamma \mid t_k)$ is slightly smaller for large k than $S_1(\gamma \mid t_k)$. However dependence of $S_1(\gamma \mid t_k)$ on the dimension is more informative, as this last bound does not explode as d goes to infinity. For the sake of illustration, in the case $d = 2$, we also computed the discrepancy provided by comparing $\boldsymbol{\tau}_1 = \mathrm{I}_d$ with the kernel $\boldsymbol{\tau}_\delta$ given in (4.15). Then one can see that an optimal choice of parameters is $\delta = 1 - 4(2k-3)/(3k^2+4k-4)$ leading to a third discrepancy given by

$$S_3(\gamma \mid t_k) = \sqrt{\frac{40}{8 + k(3k-4)}}$$

which improves $S_1(\gamma \mid t_k)$ and $S_2(\gamma \mid t_k)$ for all $k > 1$. Similar manipulations will be possible in higher dimensions; this may be of independent interest. Finally we remark that we have also computed $S_j(t_k \mid \gamma)$, $j = 1, 2, 3$; however this results in bounds which are hard to read and which we do not reproduce here.

Thanks to the results in Sections 4.1 and 4.2, manipulation of Stein kernels can be surprisingly easy even for densities which are not elliptical and do not satisfy the usual regularity assumptions (such as being log-concave, stationary distributions of diffusions, having a spectral gap, etc.). This final example serves as an illustration.

Example 4.22 (Normal-gamma prior)**.** Now consider a Bayesian model where the prior distribution of θ is bivariate normal-gamma $\mathrm{NG}(\mu_0, \lambda_0, \alpha_0, \beta_0)$ (see Example 4.11). Let $\xi = 1/\sigma^2$ denote the precision and $\boldsymbol{\theta} = (\nu, \xi) \in \mathbb{R} \times \mathbb{R}^+$ be the parameter of interest. For θ_2 following the posterior distribution given $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ (independently sampled from a univariate Gaussian $\mathcal{N}(\nu, \sigma^2)$ distribution) it is known that the resulting posterior distribution P_2 of $\boldsymbol{\theta}_2$ is

$$\boldsymbol{\theta}_2 \sim \mathrm{NG}\left(\frac{\lambda_0\mu_0 + n\bar{x}}{n+\lambda_0}, n + \lambda_0, \alpha_0 + \frac{n}{2}, \beta_0 + \frac{n}{2}s^2 + \frac{1}{2}\frac{n\lambda_0}{n+\lambda_0}(\bar{x} - \mu_0)^2\right)$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ denotes the sample mean and $s^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2$. Similarly if $\boldsymbol{\theta}$ has (improper) prior the uniform distribution then it is easy to see that the corresponding

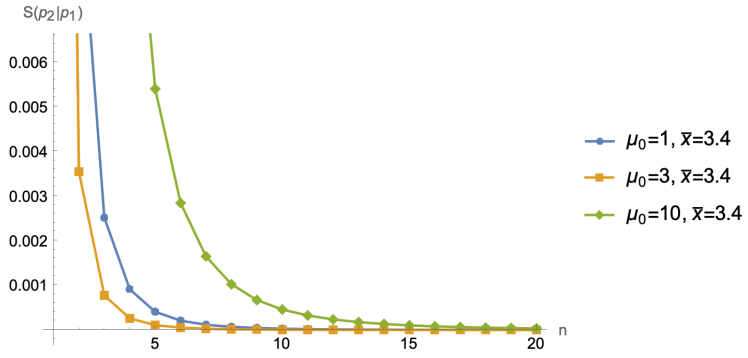

Figure 1: The Stein discrepancy $S(p_2 \mid p_1)$ comparing the effect of normal-gamma $\mathrm{NG}(\mu_0, \lambda_0, \alpha_0, \beta_0)$ prior vs uniform prior on the posterior in a normal $\mathcal{N}(\nu, \sigma^2)$ model for $\mu_0 = 1$ (blue circles), 3 (orange squares), 10 (green diamonds) and the fixed arbitrary choice $\lambda_0 = 1, \alpha_0 = 2, \beta_0 = 3$, and $s^2 = 1.2$. The Stein discrepancy to the normal decreases superlinearly with increasing n, and with increasing distance $|\bar{x} - \mu_0|$.

posterior distribution $\boldsymbol{\theta}_1$ is

$$\boldsymbol{\theta}_1 \sim \mathrm{NG}\left(\bar{x}, n, \frac{n+1}{2}, \frac{n}{2}s^2\right).$$

Example 4.11 gives a family of Stein kernels, each of which results in a bound on the Stein discrepancy which we do not detail here but but illustrate its behavior numerically for certain choices of parameter in Figure 1. The Stein discrepancy to the normal decreases superlinearly with increasing n, and with increasing distance $|\bar{x} - \mu_0|$.

5 Comparing distributions in Wasserstein distance

The main motivation behind Stein's method is the quantitative comparison of probability distributions. Consider two probability measures P_i on the same probability space, with Stein operator $\mathcal{A}_i$ and Stein class $\mathcal{F}(\mathcal{A}_i)$, for $i = 1, 2$. Given some class $\mathcal{G}$ of suitable test functions, Stein's method uses $\sup_{g\in\mathcal{G}} \min\{|\mathbb{E}_{P_2}\mathcal{A}_1 g|, |\mathbb{E}_{P_1}\mathcal{A}_2 g|\}$ as a measure of difference between P_1 and P_2. Given this premise, there are a variety of possible routes, including couplings, exchangeable pairs, and comparison of operators. Here we explore the latter approach; we focus on $\sup_{g\in\mathcal{G}} |\mathbb{E}_{P_2}\mathcal{A}_1|$, the other bound following by exchanging the roles of P_1 and P_2.

5.1 Comparing Stein operators with nested support

Recalling that $\mathbb{E}_{p_2}\mathcal{A}_2 g = 0$ for all $g \in \mathcal{F}(\mathcal{A}_2)$, for any $\mathcal{G} \subseteq \mathcal{F}(\mathcal{A}_2)$ it holds that

$$\sup_{g\in\mathcal{G}} \left|\mathbb{E}_{P_2}\mathcal{A}_1 g\right| = \sup_{g\in\mathcal{G}} \left|\mathbb{E}_{P_2}(\mathcal{A}_1 - \mathcal{A}_2) g\right|$$

so that the difference $\mathcal{A}_1 - \mathcal{A}_2$ controls the Stein discrepancy and, consequently, any metric controlled by the latter. There is much freedom in the choice of operators and classes for this purpose. For transparency of exposition we focus here on using operators (3.17) to control Wasserstein distance $\mathcal{W}_1(P_1, P_2)$, recalling (1.5). The main general result of the section then follows from previous developments.

Theorem 5.1. *Let P_1 and P_2 be two probability measures on $\mathbb{R}^d$ with respective pdfs p_1 and p_2 having nested support $\Omega_{p_2} \subseteq \Omega_{p_1}$. Assume that $\mathbb{E}_{p_1}|h| < \infty$ for every $h \in$*

$\mathrm{Lip}(\Omega_{p_1}, 1)$. *Associate to* p_i, $i = 1, 2$ *the second order operator* (3.17)

$$\mathcal{A}_i g = \langle \mathcal{T}_{\mathrm{div},p_i} \mathbf{F}_i, \nabla g \rangle + \big\langle \mathbf{F}_i, \nabla^2 g \big\rangle_{\mathrm{HS}}\,; \quad i = 1, 2$$

for some matrix-valued functions $\mathbf{F}_i \in \mathcal{F}_{\mathrm{loc}}(p_i)$, $i = 1, 2$*; let* $\mathcal{F}(\mathcal{A}_i)$, $i = 1, 2$ *be the corresponding Stein classes. Consider a collection* $\mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_2) \subseteq \mathcal{F}(\mathcal{A}_1)$ *such that*

(i) $\mathbb{E}_{p_2}|\mathcal{A}_i g| < \infty$ *for* $i = 1, 2$*,*

(ii) $\int_{\Omega_{p_2}} \big|\mathrm{div}\big(\mathbf{F}_2^T \nabla g\, p_2\big)\big| < \infty$

for all $g \in \mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_2)$*. Suppose that the matrix-valued functions* $\mathbf{F}_1$ *and* $\mathbf{F}_2$ *are such that, for every* $h \in \mathrm{Lip}(\Omega_{p_1}, 1)$*, we can find a solution* $g \in \mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_2)$ *of the* Stein equation

$$\langle \mathcal{T}_{\mathrm{div},p_1} \mathbf{F}_1, \nabla g \rangle + \big\langle \mathbf{F}_1, \nabla^2 g \big\rangle_{\mathrm{HS}} = h - \mathbb{E}_{p_1} h\,. \tag{5.1}$$

Then

$$\mathcal{W}_1(P_1, P_2) \leqslant \sup_{g \in \mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_2)} \Big| \mathbb{E}_{p_2}\big[\langle \mathcal{T}_{\mathrm{div},p_1} \mathbf{F}_1 - \mathcal{T}_{\mathrm{div},p_2} \mathbf{F}_2, \nabla g \rangle + \big\langle \mathbf{F}_1 - \mathbf{F}_2, \nabla^2 g \big\rangle_{\mathrm{HS}}\big] \Big| + \kappa_2(\mathbf{F}_1, \mathbf{F}_2)$$

where $\kappa_2(\mathbf{F}_1, \mathbf{F}_2) = \sup_{g \in \mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_2)} \Big| \int_{\Omega_{p_2}} \mathrm{div}\big(\mathbf{F}_2^T \nabla g\, p_2\big) \Big|$.

Proof. We use (3.12) to calculate

$$\mathbb{E}_{p_2}\Big[\langle \mathcal{T}_{\mathrm{div},p_2} \mathbf{F}_2, \nabla g \rangle + \big\langle \mathbf{F}_2, \nabla^2 g \big\rangle_{\mathrm{HS}}\Big] = \mathbb{E}_{p_2}\big[\mathcal{T}_{\mathrm{div},p_2}\big(\mathbf{F}_2^T \nabla g\big)\big] = \int_{\Omega_{p_2}} \mathrm{div}\big(\mathbf{F}_2^T \nabla g\, p_2\big)\,.$$

If $h \in \mathrm{Lip}(\Omega, 1)$ and $g \in \mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_2)$ is a solution of (5.1), then the assumptions of the theorem guarantee that the $\mathbb{E}_{p_2} h$ is well defined and

$$\mathbb{E}_{p_2} h - \mathbb{E}_{p_1} h = \mathbb{E}_{p_2}\Big[\langle \mathcal{T}_{\mathrm{div},p_1} \mathbf{F}_1 - \mathcal{T}_{\mathrm{div},p_2} \mathbf{F}_2, \nabla g \rangle\Big] + \mathbb{E}_{p_2}\big[\big\langle \mathbf{F}_1 - \mathbf{F}_2, \nabla^2 g \big\rangle_{\mathrm{HS}}\big] + \int_{\Omega_{p_2}} \mathrm{div}\big(\mathbf{F}_2^T \nabla g\, p_2\big)\,.$$

Taking suprema leads to the claims. □

Remark 5.2.

- Taking g as in (5.1), it is not guaranteed that $g \in \mathcal{F}(\mathcal{A}_2)$, so that $\mathbb{E}_{p_2} \mathcal{A}_2 g$ need not be zero. It is assured that g is in $\mathrm{dom}(\mathcal{A}_2)$, the set of functions for which $\mathcal{A}_2$ is defined.
- As we shall see, in many cases of interest, it will be easy to verify that the assumptions of Theorem 5.1 are satisfied and, moreover, the solutions g belong to $\mathcal{F}(\mathcal{A}_2)$, i.e. $\mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_2) \subseteq \mathcal{F}(\mathcal{A}_2)$. Then $\kappa_2(\mathbf{F}_1, \mathbf{F}_2) = 0$.

There is considerable flexibility in the bounds that can be obtained from Theorem 5.1; two particular cases are illustrated in the next examples.

Example 5.3 (Wasserstein distance and Fisher information distance)**.** Suppose that the assumptions of Theorem 5.1 are satisfied for some $\mathbf{F}_1 = \mathbf{F}_2$ then

$$\mathcal{W}_1(P_1, P_2) \leqslant \sup_{g \in \mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_1)} \Big| \mathbb{E}_{p_2}\Big[\big\langle \big(\mathcal{T}_{\mathrm{div},p_1} - \mathcal{T}_{\mathrm{div},p_2}\big)\, \mathbf{F}_1, \nabla g \big\rangle\Big] \Big| + \kappa_2(\mathbf{F}_1, \mathbf{F}_1)\,. \tag{5.2}$$

Suppose that $\mathbf{F}_1 = \mathrm{I}_d$ is allowed in (5.2), then

$$\mathcal{W}_1(P_1, P_2) \leqslant \sup_{g \in \mathcal{G}_2(\mathrm{I}_d, \mathrm{I}_d)} \big(\mathbb{E}_{p_2} \|\nabla g\|^2\big)^{1/2} I(p_2 \mid p_1) + \kappa_2(\mathrm{I}_d, \mathrm{I}_d) \tag{5.3}$$

where $I(p_2 \mid p_1) = \left(\mathbb{E}_{p_2}\left[\|\nabla \log p_1 - \nabla \log p_2\|^2\right]\right)^{1/2}$ is a *Fisher information distance* which is well-known to metrize convergence in distribution, see e.g. [28]. Thus bounds on the Fisher information (which scales well over convolutions, see e.g. in [6, 28, 38]) translate immediately into bounds on Wasserstein distance. We will study inequality (5.2) in more detail in Section 5.1.

Example 5.4 (Wasserstein distance and Stein kernel discrepancy)**.** Suppose that the assumptions of Theorem 5.1 are satisfied for some $\mathbf{F}_1, \mathbf{F}_2$ such that $\mathcal{T}_{\mathrm{div},p_1}\mathbf{F}_1 = \mathcal{T}_{\mathrm{div},p_2}\mathbf{F}_2$. Then

$$\mathcal{W}_1(P_1, P_2) \leqslant \sup_{g \in \mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_2)} \left|\mathbb{E}_{p_2}\left[\left\langle \mathbf{F}_1 - \mathbf{F}_2, \nabla^2 g\right\rangle_{\mathrm{HS}}\right]\right| + \kappa_2(\mathbf{F}_1, \mathbf{F}_2). \tag{5.4}$$

If p_1 and p_2 share a common mean and if the Stein kernels $\mathbf{F}_1 = \boldsymbol{\tau}_1$ and $\mathbf{F}_2 = \boldsymbol{\tau}_2$ are allowed in (5.4), then

$$\mathcal{W}_1(P_1, P_2) \leqslant \sup_{g \in \mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_2)} \left|\mathbb{E}_{p_2}\left[\left\langle \boldsymbol{\tau}_1 - \boldsymbol{\tau}_2, \nabla^2 g\right\rangle_{\mathrm{HS}}\right]\right| + \kappa_2(\mathbf{F}_1, \mathbf{F}_2) \tag{5.5}$$

which provides a direct connection between Wasserstein distance, Stein's method and the *Stein kernel discrepancies* studied in Section 4.3.

Applicability of Theorem 5.1 (or inequalities (5.2) and (5.4)) rests on a good understanding of the properties of solutions $g \in \mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_2)$ of Stein equations, and more particularly on ∇g and $\nabla^2 g$. Bounds on these quantities are called *Stein factors*. For ease of use here we collect relevant estimates from the literature; a new result, under the assumption of Poincaré constant, will be provided in Section 5.2.

Example 5.5 (Gaussian Stein factors)**.** Let Σ be a $d \times d$ positive-definite matrix. The Stein equation resulting from (1.1) for $\mathcal{N}(0, \Sigma)$ is

$$\left\langle \Sigma, \nabla^2 g(x)\right\rangle_{\mathrm{HS}} - \langle x, \nabla g(x)\rangle = h(x) - \mathbb{E}[h(\Sigma^{1/2} Z)]; \qquad x \in \mathbb{R}^d, \tag{5.6}$$

with Z a standard normal random vector. Letting $Z_{x,t} = e^{-t}x + \sqrt{1 - e^{-2t}}\Sigma^{1/2} Z$, a solution of (5.6) is identified in [9] as $g(x) = -\int_0^\infty \mathbb{E}[\bar{h}(Z_{x,t})]\,\mathrm{d}t$ where $\bar{h} = h - \mathbb{E}[\Sigma^{1/2} Z]$, see also [26]. In [23, 40] it is shown that if h is n times differentiable then g is also n times differentiable, and

$$\left|\frac{\partial^k g(x)}{\prod_{j=1}^k \partial x_{i_j}}\right| \leq \frac{1}{k}\left|\frac{\partial^k h(x)}{\prod_{j=1}^k \partial x_{i_j}}\right|$$

for all $k = 1, \ldots, n$ and all $x \in \mathbb{R}^d$. Moreover, if $h \in \mathrm{Lip}(\mathbb{R}^d, 1)$ then $g \in \mathcal{F}(\mathcal{A}_1)$ and

$$\sup_{x \in \mathbb{R}^d} \|\nabla g(x)\| \leqslant 1 \quad \text{and} \quad \sup_{x \in \mathbb{R}^d} \left\|\nabla^2 g(x)\right\|_{\mathrm{HS}} \leqslant \sqrt{\frac{2}{\pi}}\left\|\Sigma^{-1/2}\right\|_{\mathrm{op}}. \tag{5.7}$$

Much more is known on the properties of these solutions, and we refer to [36] for an overview.

Example 5.6 (Log-concave Stein factors, [35])**.** Let P_1 have pdf p_1 with full support $\Omega_{p_1} = \mathbb{R}^d$, and consider the Stein equation

$$\Delta g + \langle \nabla \log p_1, \nabla g\rangle = h - \mathbb{E}_{p_1} h \tag{5.8}$$

with $h \in \mathrm{Lip}(\mathbb{R}^d, 1)$. For a function $g\colon\ \mathbb{R}^d \to \mathbb{R}$ introduce $M_j, j = 2, 3$ as

$$M_j(g) := \sup_{x, y \in \mathbb{R}^d, x \neq y} \frac{\|\nabla^{k-1} g(x) - \nabla^{k-1} g(y)\|_{\mathrm{op}}}{\|x - y\|}.$$

Recall that a function $f \in \mathcal{C}^2(\mathbb{R}^d)$ is *k-strongly concave* for $k > 0$ if for all $x, y \in \mathbb{R}^d$, $y^T \nabla^2 f(x) y \leqslant -k\|y\|^2$. Suppose that $\log p_1 \in \mathcal{C}^4(\mathbb{R}^d)$ is k-strongly concave with $M_3(\log p) \leqslant L_3$. Then (5.8) has a solution $g \in \mathcal{F}(\mathcal{A}_1)$ which satisfies

$$\sup_{x\in\mathbb{R}^d} \|\nabla g(x)\| \leqslant \frac{2}{k} \quad \text{and} \quad \sup_{x\in\mathbb{R}^d} \left\|\nabla^2 g(x)\right\|_{\mathrm{HS}} \leqslant \frac{2L_3}{k^2} + \frac{1}{k} M_2(h)\,.$$

Example 5.7 (More general Stein factors, [25, 19])**.** Let P_1 with support $\Omega_{p_1} = \mathbb{R}^d$ be the invariant measure of an Itô diffusion as in Example 3.23 and consider the strong Stein equation (using the notations from that example)

$$\big\langle a(x) + c(x), \nabla^2 g(x)\big\rangle_{\mathrm{HS}} - \langle \mathcal{T}_{\mathrm{div},p_1} a(x) + f(x), \nabla g(x)\rangle = h(x) - \mathbb{E}_{p_1} h \tag{5.9}$$

with $h \in \mathrm{Lip}(\mathbb{R}^d,1)$. In [25], it is assumed that the transition semigroup of the diffusion satisfies a condition of *Wasserstein decay rate* (see their Definition 4), while [19] make more analytical assumptions (see their Assumption 2.1); under these assumptions, Equation (5.9) has a solution $g \in \mathcal{F}(\mathcal{A}_1)$ which is twice continuously differentiable and Stein factors are available (see [25, Theorem 5] and [19, Theorem 3.1]).

For the remainder of the section we focus on the case where p_1 is the pdf of $\mathcal{N}(0, \Sigma_1)$ with positive definite covariance Σ_1. Then the bounds on ∇g and $\nabla^2 g$ can be read from Example 5.5 and the following two observations follow directly. First, by (5.3), if P_2 has full support $\mathbb{R}^d$ then $\kappa_2(\mathrm{I}_d, \mathrm{I}_d) = 0$ and the first bound in (5.7) gives

$$\mathcal{W}_1(P_1, P_2) \leqslant I(p_2 \mid p_1).$$

Hence, in particular, for P_2 with support $\mathbb{R}^d$ then any central limit theorem (CLT) in Fisher information directly translates to one in Wasserstein distance; if the support is not $\mathbb{R}^d$ then some adaptations are necessary in order to incorporate the integration constants. Second, if P_2 is centred with a Stein kernel $\boldsymbol{\tau}_2$ which is allowed as $\mathbf{F}_2$ in (5.4), choosing $\boldsymbol{\tau}_1 = \Sigma_1$ in (5.5) the Cauchy–Schwarz inequality and the second bound in (5.7) give

$$\mathcal{W}_1(p_1, p_2) \leqslant \sqrt{\frac{2}{\pi}} \left\|\Sigma_1^{-1/2}\right\|_{\mathrm{op}} \mathbb{E}_{p_2} \|\boldsymbol{\tau}_2 - \Sigma_1\|_{\mathrm{HS}} + \kappa_2(\Sigma_1, \boldsymbol{\tau}_2)\,. \tag{5.10}$$

Thus, any CLT in Stein kernel discrepancy leads to a CLT in Wasserstein distance. For instance, using known properties of Stein kernels, the following result is straightforward.

Example 5.8 (CLT in Wasserstein)**.** Suppose that p_2 is the pdf of the random variable $W_n = (\sum_{i=1}^n X_i)/\sqrt{n}$ where the $X_i \sim p$ are i.i.d. random vectors with mean 0, variance I_d and (common) Stein kernel $\boldsymbol{\tau}$. Then, following [15, Theorem 3.2], one shows that

$$\boldsymbol{\tau}_2(x) = n^{-1} \sum_{i=1}^n \mathbb{E}\big[\boldsymbol{\tau}(X_i) \mid W_n = x\big]$$

is a Stein kernel for p_2 and $\mathbb{E}_{p_2}\|\boldsymbol{\tau}_2 - \mathrm{I}_d\|_{\mathrm{HS}} \leqslant \frac{1}{\sqrt{n}} \mathbb{E}_p \|\boldsymbol{\tau} - \mathrm{I}_d\|_{\mathrm{HS}}$. If furthermore p has support $\mathbb{R}^d$ then $\kappa_2(\Sigma_1, \boldsymbol{\tau}_2) = 0$ and (5.10) yields

$$\mathcal{W}_1(\gamma, p_2) \leqslant \frac{1}{\sqrt{n}} \sqrt{\frac{2}{\pi}}\, \mathbb{E}_p \|\boldsymbol{\tau} - \mathrm{I}_d\|_{\mathrm{HS}}\,.$$

Under the same conditions as in [15, Corollary 2.5] this leads to a CLT in Wasserstein distance with correct dependence on the dimension and on the sample size n.

The next example serves to compare our approach with standard results.

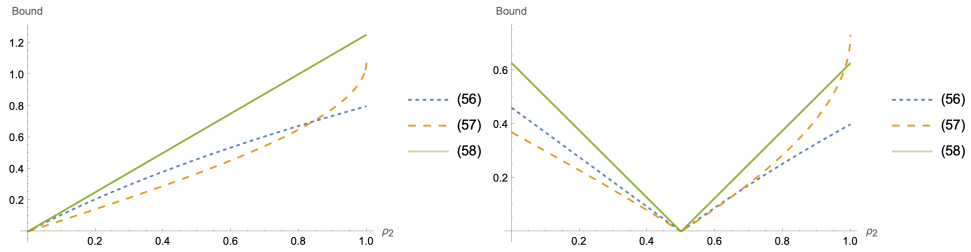

Figure 2: Comparison between Gaussians. We fix $d = 2$ and $\Sigma_i = \begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix}$ for $i = 1, 2$. Both plots report the bounds (5.11) (blue dots), (5.12) (orange dashes), and (5.13) (a green line) for $\rho_2 \in (0, 1)$. For the left plot, $\rho_1 = 0$; for the right plot, $\rho_1 = 0.5$. The new bound (5.11) outperforms both (5.13) and (5.12) when $\rho_1 = 0.5$ and ρ_2 is large.

Example 5.9 (Comparison between Gaussians)**.** For $i = 1, 2$ let p_i be a centred Gaussian pdf with covariance Σ_i. Inequality (5.10) applies with $\kappa_2(\Sigma_1, \Sigma_2) = 0$ so that

$$\mathcal{W}_1(p_1, p_2) \leqslant \sqrt{\frac{2}{\pi}} \min\left\{ \left\| \Sigma_1^{-1/2} \right\|_{\mathrm{op}}, \left\| \Sigma_2^{-1/2} \right\|_{\mathrm{op}} \right\} \|\Sigma_1 - \Sigma_2\|_{\mathrm{HS}} \,.$$

A general bound for Wasserstein distance between Gaussians is also given in [13, Lemma 2.4]. For the sake of illustration, take $d = 2$ and $\Sigma_i = \begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix}$ for some $\rho_i \in [0, 1]$. Then our bound reads

$$\mathcal{W}_1(p_1, p_2) \leqslant \frac{2}{\sqrt{\pi}} \min\left\{ \frac{1}{\sqrt{1+\rho_1}}, \frac{1}{\sqrt{1+\rho_2}} \right\} |\rho_2 - \rho_1| \tag{5.11}$$

whereas that from [13, Lemma 2.4] is

$$\mathcal{W}_1(p_1, p_2) \leqslant \sqrt{4 - 2\sqrt{1-\rho_1}\sqrt{1-\rho_2} - 2\sqrt{1+\rho_1}\sqrt{1+\rho_2}} \,. \tag{5.12}$$

Finally, an efficient coupling can be constructed directly (simply write $X_j = \Sigma_j^{1/2} N$ for N standard normal), proving that

$$\mathcal{W}_1(p_1, p_2) \leqslant \sqrt{\frac{\pi}{2}} \, |\rho_1 - \rho_2| \,. \tag{5.13}$$

We provide illustrations of the comparison of these bounds in Figure 2. In this scenario, the bound (5.13) is outperformed by the bounds from (5.11) and (5.12), with our new bound (5.11) outperforming (5.12) when $\rho_1 = 0.5$ and ρ_2 is large.

We now concentrate on (5.2) in the special case $\mathbf{F}_2 = \mathbf{F}_1$ in Theorem 5.1. Similar arguments as for Theorem 5.1 lead to the following result.

Proposition 5.10. *Suppose that P_1 and P_2 have pdfs p_1 and p_2 with nested supports $\Omega_{p_2} \subseteq \Omega_{p_1}$. Instate the notations and assumptions from Theorem 5.1, but with $\mathbf{F}_2 = \mathbf{F}_1$. Then, letting $\pi_0 = p_2/p_1$, it holds that*

$$\mathcal{W}_1(P_1, P_2) \leqslant \sup_{g \in \mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_1)} |\mathbb{E}_{p_1} \langle \mathbf{F}_1 \nabla \pi_0, \nabla g \rangle| + \kappa_2(\mathbf{F}_1, \mathbf{F}_1). \tag{5.14}$$

In particular,

$$\mathcal{W}_1(P_1, P_2) \leqslant \sup_{g \in \mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_1)} \sup_{x \in \mathbb{R}^d} \|\nabla g(x)\| \, \mathbb{E}_{p_1} \|\mathbf{F}_1 \nabla \pi_0\| + \kappa_2(\mathbf{F}_1, \mathbf{F}_1) \tag{5.15}$$

$$\mathcal{W}_1(P_1, P_2) \leqslant \sup_{g \in \mathcal{G}_2(\mathbf{F}_1, \mathbf{F}_1)} (\mathbb{E}_{p_1} \|\nabla g\|^2)^{1/2} \sqrt{\mathbb{E}_{p_1} \|\mathbf{F}_1 \nabla \pi_0\|^2} + \kappa_2(\mathbf{F}_1, \mathbf{F}_1). \tag{5.16}$$

If, furthermore, $\boldsymbol{\tau}_1$ is a Stein kernel for p_1 then

$$\mathcal{W}_1(P_1, P_2) \geqslant \|\mathbb{E}_{p_1}[\boldsymbol{\tau}_1 \nabla \pi_0]\|. \tag{5.17}$$

Proof. Since p_1 and p_2 have nested supports, it follows from (3.8) that

$$\mathcal{T}_{\mathrm{div},p_2}\mathbf{F} = \mathcal{T}_{\mathrm{div},p_1}\mathbf{F} + \mathbf{F}\,\nabla \log(p_2/p_1) \tag{5.18}$$

(over Ω_{p_2}) for any $\mathbf{F}$. The bound (5.14) now follows directly from Theorem 5.1. Inequalities (5.15) and (5.16) follow from different applications of the Cauchy–Schwarz inequality. For (5.15), one bounds $\sum_{i=1}^d |\partial_i g|$ by $\|\nabla g\|$ and for (5.16) one uses the Cauchy–Schwarz inequality inside the expectation.

For (5.17), let e be a unit vector. Let $\nu_1 = \mathbb{E}_{p_1}\mathrm{Id}$ be the mean of p_1, recalling that Id is the identity function and that $\mathrm{Id}_e = \langle \mathrm{Id}, e\rangle$ denotes the marginal projection in direction e. Noting that Id_e is 1-Lipschitz and using the nested structure we have with $\nu_e = \mathbb{E}_{p_1}[\mathrm{Id}_e]$:

$$\begin{aligned}\mathcal{W}_1(P_1, P_2) &\geqslant |\mathbb{E}_{p_1}[\mathrm{Id}_e] - \mathbb{E}_{p_2}[\mathrm{Id}_e]| = |\nu_{e,1} - \mathbb{E}_{p_1}[\mathrm{Id}_e \pi_0]| \\ &= |\mathbb{E}_{p_1}[(\nu_{e,1} - \mathrm{Id}_e)\pi_0)]| = |\mathbb{E}_{p_1}[\langle(\nu - \mathrm{Id})\pi_0, e\rangle]| = |\langle \mathbb{E}_{p_1}[\boldsymbol{\tau}_1 \nabla \pi_0], e\rangle|\end{aligned}$$

where the last identity follows from (4.4). Taking $e = \mathbb{E}_{p_1}[\boldsymbol{\tau}_1 \nabla \pi_0]/\|\mathbb{E}_{p_1}[\boldsymbol{\tau}_1 \nabla \pi_0]\|$ gives the claim. □

For evaluating expectations it may be useful to note that the expectations $\mathbb{E}_{p_1}$ in Proposition 5.10 can be expressed in terms of $\mathbb{E}_{p_2}$, because

$$\mathbb{E}_{p_1}\left[\langle \mathbf{F}_1 \nabla \pi_0, \nabla g\rangle\right] = \mathbb{E}_{p_2}\left[\langle \mathbf{F}_1 \nabla \log \pi_0, \nabla g\rangle\right].$$

Remark 5.11. Let $\boldsymbol{\tau}_1$ be a Stein kernel for p_1 and suppose, for simplicity, that $\mathcal{G}_2(\boldsymbol{\tau}_1, \boldsymbol{\tau}_1) \subseteq \mathcal{F}(\mathcal{A}_2)$ so that $\kappa_2(\boldsymbol{\tau}_1, \boldsymbol{\tau}_1) = 0$. When $d = 1$, in [32, Equation (4.2)], the following bound is provided:

$$|\mathbb{E}_{p_1}[\boldsymbol{\tau}_1 \pi_0']| \leqslant \mathcal{W}_1(P_1, P_2) \leqslant \mathbb{E}_{p_1}[\boldsymbol{\tau}_1 |\pi_0'|].$$

If $\mathbf{F} = \boldsymbol{\tau}_1$ satisfies the assumptions of Theorem 5.1, then, combining (5.17) and (5.15),

$$\|\mathbb{E}_{p_1}[\boldsymbol{\tau}_1 \nabla \pi_0]\| \leqslant \mathcal{W}_1(P_1, P_2) \leqslant C\mathbb{E}_{p_1}\left[\|\boldsymbol{\tau}_1 \nabla \pi_0\|\right]. \tag{5.19}$$

with $C = \sup_{g \in \mathcal{G}_2(\boldsymbol{\tau}_1, \boldsymbol{\tau}_1)} \sup_{x \in \mathbb{R}^d} \|\nabla g(x)\|$, yielding a similar bound as in the case $d = 1$. In particular, if p_1 is the Gaussian distribution then $C = 1$ (recall Example 5.5) which leads to the same bound as in the univariate case, whereas if p_1 is only assumed to be k-strongly log-concave with $\log p_1 \in \mathcal{C}^4(\mathbb{R}^d)$ then the best available bound is, to the best of our knowledge, $C \leqslant 2/k$ (recall Example 5.6). Bounds in more general cases are also available (recall Example 5.7).

Example 5.12 (Azzalini–Dalla Valle skew-normal distributions vs multivariate normal)**.** The pdf of the centred Azzalini–Dalla Valle type r.v. $X \in \mathbb{R}^d$ is given by

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

where $\omega_d(x; \Sigma)$ is the pdf of the d-dimensional normal distribution $\mathcal{N}(0, \Sigma)$, Φ the c.d.f. of the standard normal on $\mathbb{R}$, and $\alpha \in \mathbb{R}^d$ is a skewness parameter, see [5]. In [33], an exact expression for the Wasserstein distance between $p_2 = p_\alpha$ and $p_1 = \omega_d$ is given for $d = 1$; here we extend this result to general d. We aim to apply Proposition 5.10 with $\mathbf{F}_1 = \Sigma$ the covariance (a Stein kernel for the Gaussian) and $\pi_0(x) = 2\Phi(\alpha^T x)$. Solutions of the Gaussian Stein equations exist, recall Example 5.5; these solutions also belong to $\mathcal{F}(\mathcal{A}_2)$ because π_0 is bounded, hence $\kappa_2(\Sigma, \Sigma) = 0$. It is easy to see

that $2\Phi(\alpha^T x) - \mathbb{E}_{\mathcal{N}(0,\Sigma)}[2\Phi(\alpha^T x)] \in W_0^{1,2}(\mathcal{N}(0,\Sigma))$. Then (5.19) can be applied with $C = \sup_{g\in\mathcal{G}_2(\Sigma,\Sigma)} \sup_{x\in\mathbb{R}^d} \|\nabla g(x)\| = 1$ (recall Example 5.5). Since $\nabla\pi_0(x) = 2\alpha\gamma(\alpha^T x)$ with γ the one-dimensional standard normal pdf, the upper and lower bounds coincide in (5.19), leading to

$$\begin{aligned}\mathcal{W}_1(p_\alpha,\omega_d) &= 2\|\Sigma\alpha\|\mathbb{E}_{\omega_d}\left[\gamma(\alpha^T\mathrm{Id})\right]\\ &= 2\|\Sigma\alpha\|(2\pi)^{-(d+1)/2}\frac{1}{\sqrt{\det(\Sigma)}}\int \exp\left(-\frac{1}{2}\big(x^T(\alpha\alpha^T+\Sigma^{-1})x\big)\right)\mathrm{d}x\\ &= \frac{2\|\Sigma\alpha\|}{\sqrt{2\pi(1+\alpha^T\Sigma\alpha)}}.\end{aligned}$$

where the second equation follows from the matrix determinant lemma. In particular we recover the univariate result from [33] when $d = 1$.

The nestedness assumption is naturally satisfied in a Bayesian setting when comparing the effect of a specific prior against the uniform prior on the posterior distribution. We use this context to illustrate our bounds on two last examples.

Example 5.13 (Normal versus uniform prior in normal model)**.** Consider a normal $\mathcal{N}(\theta,\Sigma)$ model with mean $\theta\in\mathbb{R}^d$ and positive definite covariance matrix Σ. Let P_1, with pdf p_1, denote the posterior distribution of θ with uniform prior and P_2, with pdf p_2, the posterior distribution with prior $\mathcal{N}(\nu,\Sigma_2)$; Σ_2 is assumed positive definite. Suppose that a sample $(x_1,\dots,x_n)$ with $x_i\in\mathbb{R}^d$ for all i is observed. Inequality (5.19) applies with $P_1 = \mathcal{N}(\bar{x}, n^{-1}\Sigma)$, with $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, and $P_2 = \mathcal{N}(\tilde{\mu}_n,\tilde{\Sigma}_n)$ with

$$\tilde{\mu}_n = \nu + n\tilde{\Sigma}_n\Sigma^{-1}(\bar{x}-\nu) \quad \text{and} \quad \tilde{\Sigma}_n = (\Sigma_2^{-1}+n\Sigma^{-1})^{-1}$$

and $C_2(\boldsymbol{\tau}_1,p_1) = 1$. Since $p_2(\theta)\propto p_1(\theta)\exp\big((\theta-\nu)^T\Sigma_2^{-1}(\theta-\nu)\big)$, it holds that $\nabla\log\pi_0(\theta) = -\Sigma_2^{-1}(\theta-\nu)$. Since $\boldsymbol{\tau}_1 = n^{-1}\Sigma$ is a Stein kernel for p_1, inequality (5.19) (re-expressed in terms of p_2) becomes

$$\left\|n^{-1}\Sigma\Sigma_2^{-1}\mathbb{E}_{p_2}\left[\mathrm{Id}-\nu\right]\right\| \leqslant \mathcal{W}_1(P_1,P_2) \leqslant \mathbb{E}_{p_2}\left[\left\|n^{-1}\Sigma\Sigma_2^{-1}(\mathrm{Id}-\nu)\right\|\right].$$

These expectations can be evaluated and after some calculations we find

$$\begin{aligned}\left\|(\mathrm{I}_d+n\Sigma^{-1}\Sigma_2)^{-1}(\bar{x}-\nu)\right\| \leqslant \mathcal{W}_1(P_1,P_2) &\leqslant \left\|(\mathrm{I}_d+n\Sigma^{-1}\Sigma_2)^{-1}(\bar{x}-\nu)\right\|\\ &\quad + n^{-1}\left\|\Sigma\Sigma_2^{-1}(\Sigma_2^{-1}+n\Sigma^{-1})^{-1/2}\right\|_{\mathrm{op}}\frac{\sqrt{2}\Gamma(d/2+1/2)}{\Gamma(d/2)}.\end{aligned}$$

The multivariate bound indeed simplifies to the univariate bound from [32] when $d = 1$.

Example 5.14 (Bayesian logistic regression with Gaussian prior)**.** We follow [25, Example 1]. Consider the log pdf of a Bayesian logistic regression posterior based on a dataset of L observations $\mathbf{x}_\ell = (v_\ell, y_\ell)$, $\ell = 1,\dots,L$, with $v_\ell\in\mathbb{R}^d$ a vector of covariates and $y_\ell\in\{0,1\}$ and a d-dimensional $\mathcal{N}(\nu,\Sigma)$ prior on the parameter $\beta\in\mathbb{R}^d$ of the logistic regression:

$$\log p_2(\beta) = \kappa(\mathbf{x}) - \frac{1}{2}\left\|\Sigma^{-1/2}(\beta-\nu)\right\|^2 - \sum_{\ell=1}^L \log\Big(1+\exp\big(-y_\ell\,\langle v_\ell,\beta\rangle\big)\Big)$$

where $\kappa(\mathbf{x})$ is an irrelevant normalizing constant, the first summand is the multivariate Gaussian prior on β and the second term is the logistic regression likelihood. Treating the Gaussian prior as the target p_1 (and therefore the logistic likelihood as π_0), it follows from (5.19) and $\mathbb{E}_{p_1}(g\,\nabla\pi_0) = \mathbb{E}_{p_2}(g\,\nabla\log\pi_0)$, with $C_2(\Sigma,p_1)\leq 1$, $\boldsymbol{\tau}_1 = \Sigma$ and

$$\nabla\log\pi_0 = \sum_{\ell=1}^L \frac{-y_\ell}{1+\exp\big(-y_\ell\,\langle v_\ell,\beta\rangle\big)}\,v_\ell\exp\big(-y_\ell\,\langle v_\ell,\beta\rangle\big),$$

that

$$\left\| \mathbb{E}_{p_2}\left[\sum_{\ell=1}^{L} \frac{-y_\ell \Sigma v_\ell \exp\left(-y_\ell \left\langle v_\ell, \beta \right\rangle\right)}{1+\exp\left(-y_\ell \left\langle v_\ell, \beta \right\rangle\right)} \right] \right\| \leqslant \mathcal{W}_1(P_1, P_2)$$
$$\leqslant \mathbb{E}_{p_2}\left[\left\| \sum_{\ell=1}^{L} \frac{-y_\ell \Sigma v_\ell \exp\left(-y_\ell \left\langle v_\ell, \beta \right\rangle\right)}{1+\exp\left(-y_\ell \left\langle v_\ell, \beta \right\rangle\right)} \right\| \right]$$

(expectation is over $\beta \sim p_2$). Similarly the roles can be reversed, with p_1 now the logistic regression likelihood and π_0 the Gaussian prior; here a Stein kernel is available as well, see [25, Example 1].

5.2 Weak Stein equations and a bound under a Poincaré condition

Unlike in the 1-dimensional case, bounds such as those in Examples 5.5, 5.6 or 5.7 are not available for general target p_1, and even existence, let alone regularity, of a solution of the Stein equation is often out of reach. One way to bypass the necessity for solving Stein equations is to work directly with Fisher information distance $I(p_2 \mid p_1)$ or Stein kernel discrepancies $S(p_2 \mid p_1)$ and then use non-Stein's method related connections with classical discrepancies such as Total Variation, 1-Wasserstein or 2-Wasserstein. This approach was used e.g. in [38, 30, 15, 20]. We conclude the paper by introducing an alternative route via the notion of *weak Stein equations* (and corresponding *weak Stein factors*), as follows.

In this subsection we instate the notation and assumptions of Proposition 5.10. In the proof of Theorem 5.1, the Stein equation (5.1) is only used in a weak form, namely

$$\mathbb{E}_{p_2}\left[\mathcal{A}_{p_1} g\right] = \mathbb{E}_{p_2}\left[\left\langle \mathcal{T}_{\mathrm{div},p_1} \mathbf{F}_1, \nabla g \right\rangle + \left\langle \mathbf{F}_1, \nabla^2 g \right\rangle_{\mathrm{HS}}\right] = \mathbb{E}_{p_2}\left[h - \mathbb{E}_{p_1} h\right]. \tag{5.20}$$

Now recall that $p_2 = \pi_0 p_1$ and that the Stein operator has actually been obtained by (3.17), leading to

$$\mathbb{E}_{p_2}\left[\mathcal{A}_{p_1} g\right] = \mathbb{E}_{p_2}\left[\mathcal{T}_{\mathrm{div},p_1}\left(\mathbf{F}_1^T \nabla g\right)\right] = \mathbb{E}_{p_1}\left[\mathcal{T}_{\mathrm{div},p_1}\left(\mathbf{F}_1^T \nabla g\right)\pi_0\right].$$

Applying (3.11), we find that

$$\mathbb{E}_{p_2}\left[\mathcal{A}_{p_1} g\right] = -\mathbb{E}_{p_1}\left[\left\langle \mathbf{F}_1^T \nabla g, \nabla \pi_0 \right\rangle\right] = -\mathbb{E}_{p_1}\left[\left\langle \mathbf{F}_1 \nabla \pi_0, \nabla g \right\rangle\right].$$

Thus (5.20) is equivalent to

$$-\mathbb{E}_{p_1}\left[\left\langle \mathbf{F}_1 \nabla \pi_0, \nabla g \right\rangle\right] = \mathbb{E}_{p_1}\left[(h - \mathbb{E}_{p_1} h)\pi_0\right]. \tag{5.21}$$

Classical results from functional analysis provide handles on expressions such as (5.21). We follow [15] and introduce $W^{1,2}(p)$, the natural (weighted) Sobolev space of weakly differentiable functions $u\colon\ \Omega \to \mathbb{R}$ with finite (squared) Sobolev norm

$$\|u\|^2_{W^{1,2}(p)} := \|u\|^2_{L^2(p)} + \|\nabla u\|^2_{L^2(p)}, \tag{5.22}$$

see also [21]. *In the rest of the paper, we shall assume that p is continuous on Ω_p.* Then $W^{1,2}(p)$ is complete under $\|\cdot\|_{W^{1,2}(p)}$. We also introduce the space $W_0^{1,2}(p)$ of all functions $u \in W^{1,2}(p)$ such that $\int_\Omega \partial_e u = 0$ for all $e \in S^{d-1}$.

Next, motivated by (5.21), following intuition from [15], under the condition that P admits a Poincaré constant, in the sense of definition 4.5, the following holds.

Proposition 5.15 (Weak Stein equation and factor)**.** *Let P_1 be as in Theorem 5.1. Suppose furthermore that P_1 has finite variance and Poincaré constant C_{P_1}. Let $h \in \mathrm{Lip}(\Omega_{p_1}, 1)$. Then there exists a function $g \in W_0^{1,2}(p_1)$, which solves*

$$\mathbb{E}_{p_1}\left[\left\langle \nabla v, \nabla g \right\rangle\right] = \mathbb{E}_{p_1}\left[(h - \mathbb{E}_{p_1}(h))v\right] \tag{5.23}$$

for all $v \in W_0^{1,2}(p_1)$ *and is such that*

$$\sqrt{\mathbb{E}_{p_1}\|\nabla g\|^2} \leqslant C_{P_1}\,. \tag{5.24}$$

Proof. Let $h \in \mathrm{Lip}(\Omega_{p_1},1)$. The set $\mathcal{H}_1 = \{v \in L^2(P_1)\ ;\ \mathbb{E}_{p_1} v = 0\}$ with inner product $\langle u, v\rangle_{p_1} = \mathbb{E}_{p_1}[uv]$ is a Hilbert space and $\bar{h} = h - \mathbb{E}_{p_1} h \in \mathcal{H}_1$ as P_1 admits a variance. Moreover,

$$\|h\|_{p_1} = \sqrt{\mathbb{E}_{p_1} h^2} \leqslant \sqrt{C_{P_1}\mathbb{E}_{p_1}\|\nabla h\|^2} \leqslant \sqrt{C_{P_1}}.$$

In particular we have that $W_0^{1,2}(p_1) \subseteq \mathcal{H}_1$. We endow the space $W_0^{1,2}(p_1)$ with the symmetric bilinear form $\mathcal{E}_{p_1}(u,v) = \mathbb{E}_{p_1}[\langle \nabla u, \nabla v\rangle]$. This bilinear form is coercive as due to the Poincaré inequality

$$\mathcal{E}_{p_1}(u,u) = \mathbb{E}_{p_1}[\langle \nabla u, \nabla u\rangle] \geqslant \frac{1}{C_{P_1}}\mathbb{E}_{p_1}[u^2] = \frac{1}{C_{P_1}}\|u\|_{p_1}^2. \tag{5.25}$$

From (5.25) it follows that $\langle \cdot,\cdot\rangle_{\mathcal{E}_p}$ is also an inner product on $W_0^{1,2}(p)$ and $\|\cdot\|_{\mathcal{E}_p}$ is a norm which is uniformly equivalent to $\|\cdot\|_{W^{1,2}(p)}$. Therefore, $W_0^{1,2}(p)$ is complete under $\|\cdot\|_{\mathcal{E}_p}$, so that $\langle u, v\rangle_{\mathcal{E}_p}$ makes it a Hilbert space.

Now let $\phi_{p_1}(v) := \mathbb{E}_{p_1}\big[(h - \mathbb{E}_{p_1} h)v\big]$. By the Cauchy–Schwarz inequality and the Poincaré inequality and using that $h \in \mathrm{Lip}(\Omega_{p_1},1)$ we can bound

$$|\phi_{p_1}(v)| \leq \|h - \mathbb{E}_{p_1} h\|_{L^2(p_1)}\|v\|_{L^2(p_1)} \leq C_{P_1}\|h\|_{\mathcal{E}_{p_1}}\|v\|_{\mathcal{E}_{p_1}} \leq C_{P_1}\|v\|_{\mathcal{E}_{p_1}}\,.$$

Hence, by the Riesz representation theorem, there exists $g \in W_0^{1,2}(p_1)$, such that $\phi_{p_1}(v) = \langle v, g\rangle_{\mathcal{E}_{p_1}}$ for all $v \in W_0^{1,2}(p_1)$, and $\|g\|_{\mathcal{E}_{p_1}} \leq C_{P_1}$. This completes the proof. □

Obviously if g is a solution of the strong Stein equation satisfying all the assumptions from Proposition 5.10, then it will also satisfy (5.23). However, in Corollary 5.16 we need not solve the equation, nor require any form of regularity. Equations (5.21) and (5.23) are akin to *"weak Stein equations"* (and thus would encourage us to consider *"weak Stein operators"*) and inequality (5.24) is a form of *"weak Stein factor"*. This approach seems promising for tackling Stein's method in more general multivariate settings, without needing to solve Stein equations explicitly. For instance, the next result follows easily.

Corollary 5.16. *Let* P_1, P_2 *be as in Theorem 5.1 and suppose furthermore that* P_1 *satisfies the conditions of Proposition 5.15. Also assume that* $p_2 = \pi_0\, p_1$ *with* $\pi_0 - \mathbb{E}_{p_1}\pi_0 \in W_0^{1,2}(p_1)$. *Then*

$$\mathcal{W}_1(P_1, P_2) \leqslant C_{P_1}\sqrt{\mathbb{E}_{p_1}\|\nabla \pi_0\|^2}. \tag{5.26}$$

Proof. Let $h \in \mathrm{Lip}(\Omega_{p_1},1)$. From Proposition 5.15 applied with $v = \pi_0$, it follows that there exists $g \in W_0^{1,2}(p_1)$ which is solution of the weak Stein equation

$$\mathbb{E}_{p_2}[h] - \mathbb{E}_{p_1}[h] = \mathbb{E}_{p_1}\left[\langle \nabla \pi_0, \nabla g\rangle\right].$$

Since, furthermore, g satisfies (5.24) which does not depend on h, the claim follows by the Cauchy–Schwarz inequality. □

There is a large literature regarding Poincaré inequalities and their optimal associated Poincaré constant. For example, when P has k-log-concave pdf p such that $\log p \in \mathcal{C}^4(\mathbb{R}^d)$ then $C_P \leqslant 2/k$; for P the uniform distribution on $[0,1]^2$ it is known that $C_P = 2/\pi^2$ is an optimal Poincaré constant, see [39]. These bounds allow to obtain bounds on 1-Wasserstein distance via Stein's method even in cases where Stein factors are not available.

Example 5.17 (Comparing copulas)**.** Let $V = (V_1, V_2)$ be a 2-dimensional random vector such that the marginals V_1 and V_2 have a uniform distribution on $(0,1)$. This distribution is described by the copula of V defined as $C(x_1, x_2) = \mathbb{P}[V_1 \leqslant x_1, V_2 \leqslant x_2]$, $(x_1, x_2) \in (0,1)^2$. We want to bound the Wasserstein distance between P^V, the law of V and P^U the law of $U = (U_1, U_2)$, where U_1 and U_2 are uniform on $(0,1)$ and independent. First we recall that an optimal Poincaré constant for the uniform distribution on $(0,1)^2$ is $C_P = 2/\pi^2$. Assume that V has a pdf $c = \partial^2_{x_1 x_2} C$. Then the densities of V and U have nested supports, with $\pi_0 = c$. If $c - \mathbb{E}_{U(0,1)^2} c \in W_0^{1,2}(U(0,1)^2)$ then, a simple application of (5.26) with $C_P = 2/\pi^2$ yields

$$\mathcal{W}_1(P^V, P^U) \leqslant \frac{2}{\pi^2} \sqrt{\mathbb{E}_{U(0,1)^2} \|\nabla c\|^2}.$$

In some cases, one can compute the gradient of c in a closed form. For instance, the Ali–Mikhail–Haq copula [1] has pdf

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

Then $\mathbb{E}_{U(0,1)^2}[c] = 1$ and

$$\mathbb{E}_{U(0,1)^2} \|\nabla c\|^2 = \frac{2}{105} \frac{\theta}{(1-\theta)^2} \left(\theta(2-\theta)(52(1-\theta) + 17\theta^2) - 36\log(1-\theta)\right)$$

so that $c - \mathbb{E}_{U(0,1)^2} c \in W_0^{1,2}(U(0,1)^2)$. 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 5.16 we can assess the Wasserstein distance between the Ali–Mikhail–Haq copula and the uniform copula in terms of θ. For $-1 < \theta < 1$ we bound $\mathbb{E}_{U(0,1)^2} \|\nabla c\|^2 \leqslant \frac{8}{3} \frac{\theta^2}{(1-|\theta|)^4}$ and

$$\mathcal{W}_1(P^V, P^U) \leqslant \frac{2\sqrt{8}}{\pi^2 \sqrt{3}} |\theta| \{1 - |\theta|\}^{-2}.$$

This bound decreases to 0 for $\theta \to 0$, indicating agreement with the uniform copula for $\theta = 0$. To our knowledge there is no explicit formula for $\mathcal{W}_1(P^V, P^U)$ available; [27] simulate the Wasserstein distance under Gaussian marginals and discuss simulation strategies.

Acknowledgments. We are indebted to Andreas Kyprianou for his patient and considerate handling of this paper. MR would like to thank Oliver Dragičević for his help with Sobolev spaces. We also thank Max Fathi, Christophe Ley, Gilles Mordant and Guillaume Poly for interesting discussions, as well as Lester Mackey and Steven Vanduffel for suggesting some references which we had overlooked.

References

[1] M. M. Ali, N. Mikhail, and M. S. Haq. A class of bivariate distributions including the bivariate logistic. *Journal of Multivariate Analysis*, 8(3):405–412, 1978. MR0512610

[2] B. Arras and C. Houdré. *On Stein's method for infinitely divisible laws with finite first moment.* Springer International Publishing, 2019. MR3931309

[3] B. Arras and C. Houdré. On Stein's method for multivariate self-decomposable laws with finite first moment. *Electronic Journal of Probability*, 24, 2019. MR3933208

[4] S. Artstein, K. 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, 2004. MR2083473

[5] A. Azzalini and A. Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996. MR1440039

[6] K. Ball, F. Barthe, and A. Naor. Entropy jumps in the presence of a spectral gap. *Duke Mathematical Journal*, 119(1):41–63, 2003. MR1991646

[7] A. D. Barbour, M. 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. MR3785590

[8] A. D. Barbour, M. J. Luczak, and A. Xia. Multivariate approximation in total variation, II: Discrete normal approximation. *The Annals of Probability*, 46(3):1405–1440, 2018. MR3785591

[9] A. D. Barbour. Stein's method for diffusion approximations. *Probability Theory and Related Fields*, 84(3):297–322, 1990. MR1035659

[10] S. Betsch and B. Ebner. Testing normality via a distributional fixed point property in the Stein characterization. *TEST*, 29(1):105–138, 2020. MR4063385

[11] T. Bonis. Stein's method for normal approximation in Wasserstein distances with application to the multivariate Central Limit Theorem. *Probability Theory and Related Fields*, 178(3):827–860, 2020. MR4168389

[12] 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. MR0659549

[13] D. Chafaï and F. Malrieu. On fine properties of mixtures with respect to concentration of measure and Sobolev type inequalities. *Annales de l'IHP Probabilités et Statistiques*, 46(1):72–96, 2010. MR2641771

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

[15] T. A. Courtade, M. Fathi, and A. Pananjady. Existence of Stein kernels under a spectral gap, and discrepancy bound. *Annales de l'IHP Probabilités et Statistiques*, 55(2):777–790, 2019. MR3949953

[16] C. Döbler. Stein's method of exchangeable pairs for the Beta distribution and generalizations. *Electronic Journal of Probability*, 20(109):1–34, 2015. MR3418541

[17] M. Ernst, G. Reinert, and Y. Swan. First-order covariance inequalities via Stein's method. *Bernoulli*, 26(3):2051–2081, 2020. MR4091101

[18] L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*. CRC Press, 1992. MR1158660

[19] 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*, 174(3–4):945–979, 2019. MR3980309

[20] M. Fathi. Stein kernels and moment maps. *The Annals of Probability*, 47(4):2172–2185, 2019. MR3980918

[21] M. Fathi, L. Goldstein, G. Reinert and A. Saumard. Relaxing the Gaussian assumption in shrinkage and SURE in high dimension. *The Annals of Statistics*, 50(5):2737–2766, 2022. MR4500623

[22] 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. MR3694323

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

[24] 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. MR1484792

[25] J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability* 29(5):2884–2928, 2019. MR4019878

[26] F. Götze. On the rate of convergence in the multivariate CLT. *The Annals of Probability*, 19(2):724–739, 1991. MR1106283

[27] M. Hallin, G. Mordant, and J. Segers. Multivariate goodness-of-Fit tests based on Wasserstein distance. *arXiv preprint* arXiv:2003.06684, 2020. MR4255302

[28] O. Johnson. *Information theory and the Central Limit Theorem*. Imperial College Press, London, 2004. MR2109042

[29] Z. Landsman and J. Nešlehová. Stein's lemma for elliptical random vectors. *Journal of Multivariate Analysis*, 99(5):912–927, 2008. MR2405098

[30] M. Ledoux, I. Nourdin, and G. Peccati. Stein's method, logarithmic Sobolev and transport inequalities. *Geometric and Functional Analysis*, 25(1):256–306, 2015. MR3320893

[31] G. Leoni. *A First Course in Sobolev Spaces*. Second Edition. Graduate Texts in Mathematics, Vol. 181. AMS, 2017. MR3726909

[32] C. Ley, G. Reinert, and Y. Swan. Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. *The Annals of Applied Probability*, 27(1):216–241, 2016. MR3619787

[33] C. Ley, G. Reinert, and Y. Swan. Stein's method for comparison of univariate distributions. *Probability Surveys*, 14:1–52, 2017. MR3595350

[34] W.-L. Loh. Stein's method and multinomial approximation. *The Annals of Applied Probability*, pages 536–554, 1992. MR1177898

[35] L. Mackey and J. Gorham. Multivariate Stein factors for a class of strongly log-concave distributions. *Electronic Communications in Probability*, 21, 2016. MR3548768

[36] E. Meckes. On Stein's method for multivariate normal approximation. In *High dimensional probability V: The Luminy volume*, pages 153–178. Institute of Mathematical Statistics, 2009. MR2797946

[37] I. Nourdin and G. Peccati. *Normal approximations with Malliavin calculus : from Stein's method to universality*. Cambridge Tracts in Mathematics. Cambridge University Press, 2012. MR2962301

[38] 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.

[39] 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. MR0117419

[40] 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. MR2573554

[41] G. Reinert, N. Ross. Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs. *The Annals of Applied Probability*, 29(5):3201–3229, 2019. MR4019886

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

[43] A. Saumard. Weighted Poincaré inequalities, concentration inequalities and tail bounds related to the behavior of the Stein kernel in dimension one. *Bernoulli*, 25(4B):3978–4006, 2019. MR4010979

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

[45] C. Stein, P. Diaconis, S. Holmes, and G. Reinert. Use of exchangeable pairs in the analysis of simulations. In P. Diaconis and S. 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. MR2118600

[46] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151, 1981. MR0630098

[47] A. Takatsu. Wasserstein geometry of Gaussian measures. *Osaka J. Math.* 48(4):1005–1026, 2011. MR2871291

[48] N. Upadhye, V. Čekanavičius, and P. Vellaisamy. On Stein operators for discrete approximations. *Bernoulli*, 23(4A):2828–2859, 2017. MR3648047

A Some more remarks about our function spaces

As indicated in Remark 3.4, here is a more detailed discussion about our choices of ACL-type function classes, with ACL standing for 'absolutely continous on lines'. In particular, we discuss relationships to the Sobolev spaces.

Proposition A.1. *Let $\Omega \subseteq \mathbb{R}^d$ be an open set and let $u \in \mathrm{ACL}(\Omega)$ or $u \in W^{1,1}_{\mathrm{loc}}(\Omega)$. Suppose that u is constant on some set $A \subseteq \Omega$. Then for each $e \in S^{d-1}$, $\partial_e u(x) = 0$ for Lebesgue-almost all $x \in A$.*

Proof. Let $u(x) = c$ for all $x \in A$. Taking a line L parallel to e, observe that almost all $x \in A \cap L$ (with respect to the one-dimensional Lebesgue measure) are points of Lebesgue density 1 for the set $A \cap L$ and consequently for the set $\{x \in \Omega \cap L \,;\, u(x) = c\}$, considering the one-dimensional Lebesgue density in direction e (see, e. g., Corollary B.121 of [31] or Corollary 3 in Section 1.7.1 of [18]). However, if x is such a point and u is differentiable at x in direction e, we have $\partial_e u(x) = 0$. The proof is now completed by Fubini's theorem. □

The next result, which is a counterpart of Proposition A.1, shows that the ACL functions 'almost' preserve a fundamental property of the $\mathcal{C}^1$ functions. This property is crucial to the proof of Proposition 3.14.

Proposition A.2. *Let $\Omega \subseteq \mathbb{R}^d$ be a connected open set and let $u \in \mathrm{ACL}(\Omega)$ be such that $\partial_i u = 0$ Lebesgue-almost everywhere on Ω for all $i = 1, 2, \ldots, n$. Then u is Lebesgue-almost everywhere constant on Ω.*

Proof. First, consider the case where Ω is a cube parallel to the coordinate directions. Then the result can be proved by induction. The case $d = 1$ is immediate. For the induction step from $d-1$ to d, observe that by Fubini's theorem, there exists a hyperplane H perpendicular to e_d and intersecting Ω, such that for all $i = 1, 2, \ldots, d-1$, $\partial_i u = 0$ almost everywhere on H with respect to the $(d-1)$-dimensional Hausdorff measure. By the induction hypothesis, u is constant Lebesgue-almost everywhere on H. Next, for almost all lines L parallel to e_d and intersecting Ω, u is absolutely continuous and therefore constant on L. The proof is now completed by Fubini's theorem.

For the general case, observe that Ω can be covered by a countable family $\mathcal{Q}$ of open cubes contained in Ω. By the above, there exists a Lebesgue-null set N, such that for any $Q \in \mathcal{Q}$, u is constant on $Q \setminus N$. Since Ω is connected, any two points $x, y \in \Omega \setminus N$ are linked by a finite sequence of overlapping cubes belonging to $\mathcal{Q}$. Clearly, if two open cubes overlap, they also contain a common point which is not in N. Therefore, $u(x) = u(y)$. □

Proposition A.3. *Let $\Omega \subseteq \mathbb{R}^d$ be an open set and let $u \in W^{1,1}_{\mathrm{loc}}(\Omega)$ be compactly supported. Then we have $u \in W^{1,1}(\Omega)$ and $\int_\Omega \partial_e u = 0$ for all $e \in S^{d-1}$.*

Proof. Since u has a compact support, so does $\partial_e u$. Therefore, $u \in W^{1,1}(\Omega)$. By Exercise C.23 in [31], there exists a cut-off function $v \in \mathcal{C}^\infty_{\mathrm{c}}(\Omega)$, which equals 1 on the supports of u and $\partial_e u$. By the very definition of the weak derivative, we then have

$$\int_\Omega \partial_e u = \int_\Omega (\partial_e u) v = -\int_\Omega u(\partial_e v) = 0\,. \qquad \square$$

In the case that $\Omega = \mathbb{R}^d$ we have the following result.

Proposition A.4. *For all $u \in W^{1,1}(\mathbb{R}^d)$ and all $e \in S^{d-1}$, we have $\int \partial_e u = 0$.*

Proof. By Proposition A.3, the assertion is true for all $u \in \mathcal{C}^\infty_{\mathrm{c}}(\mathbb{R}^d)$. However, by Theorem 11.35 of [31], $\mathcal{C}^\infty_{\mathrm{c}}(\mathbb{R}^d)$ is dense in $W^{1,1}(\mathbb{R}^d)$ with respect to the appropriate Sobolev norm. This completes the proof. □

The next group of results clarifies relationships between the ACL-type spaces and the Sobolev spaces.

Proposition A.5. *For any open set $\Omega \subseteq \mathbb{R}^d$, we have $\mathrm{ACL}^1_{\mathrm{loc}}(\Omega) \subseteq L^1_{\mathrm{loc}}(\Omega)$.*

Proof. We need to prove that each $x \in \Omega$ has an open neighbourhood U, such that the restriction of u to U belongs to $L^1(U)$. Now choose U to be a bounded rectangle and apply the basic Poincaré inequality for rectangles (see, e. g., Exercise 13.34 in [31]), noting that the restriction of u to U belongs to $\mathrm{ACL}^1(U)$. □

Proposition A.6. *For an open set $\Omega \subseteq \mathbb{R}^d$, $\dot{W}^{1,1}(\Omega)$ is precisely the set of all Borel measurable functions $u\colon\ \Omega \to \mathbb{R}$ which have a version $\bar{u} \in \mathrm{ACL}^1(\Omega)$. Moreover, the classical directional derivatives of $\bar{u}$ are also its corresponding weak derivatives.*

Proof. The result is essentially Theorem 11.45 of [31], but we need to modify it from the usual Sobolev space $W^{1,1}$ to the homogeneous Sobolev space $\dot{W}^{1,1}$ and, most important, from functions which are absolutely continuous on almost all lines *parallel to the coordinate axes* to ACL functions.

Firstly, suppose that u has a version $\bar{u} \in \mathrm{ACL}^1(\Omega)$. By Proposition A.5, $\bar{u} \in L^1_{\mathrm{loc}}(\Omega)$, so that each $x \in \Omega$ has an open neighbourhood U such that the restriction of $\bar{u}$ to U belongs to $L^1(U)$. By Theorem 11.45 of [31], the classical directional derivatives of $\bar{u}$ are then also its corresponding weak derivatives considered on U. However, as the weak derivative is a local concept, the weak derivatives can be considered on the whole Ω. By assumption, all directional derivatives belong to $L^1(\Omega)$. Therefore, $\bar{u}$, and therefore u, belongs to $\dot{W}^{1,1}(\Omega)$.

Now we turn to the non-trivial part, the converse. Thus, take $u \in \dot{W}^{1,1}(\Omega)$. As in Step 1 of the proof of Theorem 11.45 of [31], the function $\bar{u}$ is constructed by means of the *standard mollifier*:

$$\eta(z) := \begin{cases} a_d \exp\left(-\frac{1}{1-\|z\|^2}\right) & \text{if } \|z\| \le 1, \\ 0 & \text{if } \|z\| \ge 1, \end{cases}$$

where c_d is chosen so that $\int_{\mathbb{R}^d} \eta(z)\,\mathrm{d}z = 1$. As usual, the mollified functions are then defined as $u_\varepsilon(x) := \int_{\mathbb{R}^d} u(x+\varepsilon z)\,\eta(z)\,\mathrm{d}z$.

In Step 1 of the proof of Theorem 11.45 of [31], a function $\bar{u}$, which is locally absolutely continuous on Lebesgue-almost all lines parallel to the coordinate directions, is constructed as the limit of the sequence u_{ε_n} for some sequence $\{\varepsilon_n\}_n$. However, we claim that the function

$$\bar{u}(x) := \begin{cases} \lim\limits_{\varepsilon \downarrow 0} u_\varepsilon(x) & \text{if this limit exists,} \\ 0 & \text{otherwise} \end{cases}$$

is already locally absolutely continuous on Lebesgue-almost all lines parallel to e for all $e \in S^{d-1}$. Before turning to the proof of this claim, notice that $\bar{u}$ is Borel measurable, as the limit can be taken only over rational ε (the map $\varepsilon \mapsto \eta_\varepsilon$ is continuous in the supremum norm).

Now fix $e \in S^{d-1}$. Without loss of generality, we may assume that $e = e_d$. The crucial reason why the function $\bar{u}$ works for all directions is that the set $\{x \in \Omega\,;\, \bar{u}(x) \ne u(x)\}$ has $(d-1)$-dimensional Hausdorff measure zero. This essentially follows from Theorem 1 in Section 4.8 of [18] and Theorem 3 in Section 5.6.3 ibidem. However, the first one cannot be applied directly and some additional thought is needed:

- Theorem 1 in Section 4.8 of [18] concerns convergence of the functions

$$(u)_{x,r} := b_d \int_B u(x+ry)\,\mathrm{d}y$$

rather than the functions $u_\varepsilon(x)$ (here, B denotes the unit ball in $\mathbb{R}^d$ and b_d its reciprocal volume). However, a routine application of Fubini's theorem yields

$$u_\varepsilon(x) = \frac{2a_d}{b_d} \int_0^1 (u)_{x,t\varepsilon}\, \frac{t^{d+1}}{(1-t^2)^2} \exp\left(-\frac{1}{1-t^2}\right) \mathrm{d}t\,.$$

Now if $\lim_{r\downarrow 0}(u)_{x,r} = u(x)$ for a fixed x, the values $(u)_{x,r}$ must be bounded for sufficiently small r, so that the values $(u)_{x,t\varepsilon}$ are uniformly bounded in $t \in [0,1]$ for sufficiently small ε. Since $\int_0^1 \frac{t^{d+1}}{(1-t^2)^2} \exp\left(-\frac{1}{1-t^2}\right) \mathrm{d}t < \infty$, the desired convergence follows from the dominated convergence theorem.

- Theorem 1 in Section 4.8 of [18] assumes that $u \in W^{1,1}(\mathbb{R}^d)$, but here, we only have $u \in \dot{W}^{1,1}(\Omega)$. Although the convergence of u_ε towards u is an entirely local property, we still need to construct appropriate functions which are, along with their weak first-order partial derivatives, globally integrable. Thus, expressing Ω as a countable union of open balls U_n, it suffices to prove that for each n, the $(d-1)$-dimensional Hausdorff measure of the set $\{x \in U_n\,;\, \bar{u}(x) \neq u(x)\}$ equals zero. Clearly, the restriction of u to U_n belongs to $W^{1,1}(U_n)$. By Theorem 1 in Section 4.4 of [18] or Theorem 13.17 of [31], the latter restriction can be extended to a function in $W^{1,1}(\mathbb{R}^d)$, so that, finally, by Theorem 1 in Section 4.8 of [18] and Theorem 3 in Section 5.6.3 ibidem, the $(d-1)$-dimensional Hausdorff measure of the set $\{x \in U_n\,;\, \bar{u}(x) \neq u(x)\}$ vanishes.

Once we know that $(d-1)$-dimensional Hausdorff measure of the set $\{x \in \Omega\,;\, \bar{u}(x) \neq u(x)\}$ equals zero, Corollary 1 in Section 2.4.1 of [18] yields that there exists a Borel set $E_1 \subseteq \mathbb{R}^{d-1}$, such that the Lebesgue measure of $\mathbb{R}^{d-1} \setminus E_1$ vanishes and that $\bar{u}(x', x_d) = u(x', x_d)$ for all $x' \in E_1$ and all $x_d \in \mathbb{R}$ with $(x', x_d) \in \Omega$.

Following Step 1 of the proof of Theorem 11.45 of [31], we find that there exist a sequence $\{\varepsilon_n\}_n$ converging to zero and a Borel set $E_2 \subseteq \mathbb{R}^{d-1}$, such that the Lebesgue measure of $\mathbb{R}^{d-1} \setminus E_2$ vanishes and such that

$$\int_{(x',x_d)\in\Omega} \big\|\nabla u(x', x_d)\big\|\,\mathrm{d}x_d < \infty$$

and

$$\lim_{n\to\infty} \int_{(x',x_d)\in\Omega_{\varepsilon_n}} \big\|\nabla u_{\varepsilon_n}(x', x_d) - \nabla u(x', x_d)\big\|\,\mathrm{d}x_d = 0$$

for all $x' \in E_2$; here, Ω_{ε_n} is the set of all points Ω with the distance from $\mathbb{R}^d \setminus \Omega$ strictly greater than ε (for $\Omega = \mathbb{R}^d$, we take $\Omega_{\varepsilon_n} = \mathbb{R}^d$).

Now take $x' \in E_1 \cap E_2$ and claim that the function $x_d \mapsto \bar{u}(x', x_d)$ is absolutely continuous on all compact intervals I with $\{(x', x_d) \in \Omega\}$ for all $x_d \in I$. Take $a, b \in I$. Since the functions u_{ε_n} are in $\mathcal{C}^\infty(\Omega_{\varepsilon_n})$, we have, by the fundamental theorem of calculus,

$$u_{\varepsilon_n}(x', b) - u_{\varepsilon_n}(x', a) = \int_a^b \partial_d u_{\varepsilon_n}(x', t)\,\mathrm{d}t$$

for all n with $\{x'\} \times [a, b] \subseteq \Omega_{\varepsilon_n}$ (which is true for all sufficiently large n). Taking the limit, we find that

$$\bar{u}(x', b) - \bar{u}(x', a) = \int_a^b \partial_d u(x', t)\,\mathrm{d}t\,.$$

As $\int \left|\partial_d u(x',t)\right| \mathrm{d}t < \infty$, the function $x_d \mapsto \bar{u}(x', x_d)$ is absolutely continuous on I and its classical derivative (extended arbitrarily to the whole set $\{x_d \,;\, (x', x_d) \in \Omega\}$ and possibly altered on a Lebesgue-null set) are also its corresponding weak derivatives. This proves the result. □

Corollary A.7. *For an open set $\Omega \subseteq \mathbb{R}^d$, $W^{1,1}_{\mathrm{loc}}(\Omega)$ is precisely the set of all Borel measurable functions $\Omega \to \mathbb{R}$ which have a version in $\mathrm{ACL}^1_{\mathrm{loc}}(\Omega)$.* □

Corollary A.8. *For an open set $\Omega \subseteq \mathbb{R}^d$, $\dot{W}^{1,1}_0(\Omega)$ is precisely the set of all Borel measurable functions $\Omega \to \mathbb{R}$ which have a version in $\mathrm{ACL}^1_0(\Omega)$.* □