The Annals of Applied Probability
2017, Vol. 27, No. 1, 216–241
DOI: 10.1214/16-AAP1202
© Institute of Mathematical Statistics, 2017

DISTANCES BETWEEN NESTED DENSITIES AND A MEASURE OF THE IMPACT OF THE PRIOR IN BAYESIAN STATISTICS

BY CHRISTOPHE LEY, GESINE REINERT[1] AND YVIK SWAN[2]

Ghent University, University of Oxford and Université de Liège

In this paper, we propose tight upper and lower bounds for the Wasserstein distance between any two univariate continuous distributions with probability densities p_1 and p_2 having nested supports. These explicit bounds are expressed in terms of the derivative of the likelihood ratio p_1/p_2 as well as the Stein kernel τ_1 of p_1. The method of proof relies on a new variant of Stein's method which manipulates Stein operators.

We give several applications of these bounds. Our main application is in Bayesian statistics: we derive explicit data-driven bounds on the Wasserstein distance between the posterior distribution based on a given prior and the no-prior posterior based uniquely on the sampling distribution. This is the first finite sample result confirming the well-known fact that with well-identified parameters and large sample sizes, reasonable choices of prior distributions will have only minor effects on posterior inferences if the data are benign.

1. Introduction. A key question in Bayesian analysis is the effect of the prior on the posterior, and how this effect could be assessed. As more and more data are collected, will the posterior distributions derived with different priors be very similar? This question has a long history; see, for example, [4, 5, 29]. While asymptotic results which give conditions under which the effect of the prior wanes as the sample size tends to infinity can be found, for example, in [4, 5], here we are interested, at *fixed* sample size, in explicit bounds on some measure of the distributional distance between posteriors based on a given prior and the no-prior data-only based posterior, allowing to detect at fixed sample size the effect of the prior.

In the simple setting of prior and posterior being univariate and continuous, the basic relation that the posterior is proportional to the prior times the likelihood leads to the more general problem of comparing two distributions P_1 and P_2 whose densities p_1 and p_2 have nested support. Letting $\mathcal{I}_1$ (resp., $\mathcal{I}_2$) be the support of p_1 (resp., p_2) and assuming $\mathcal{I}_2 \subset \mathcal{I}_1$ we can write

$$p_2 = \pi_0 p_1$$

Received October 2015; revised April 2016.

[1]Supported by EPSRC Grant EP/K032402/I and by the Keble College Advanced Studies Centre.

[2]Supported by the IAP Research Network P7/06 of the Belgian State (Belgian Science Policy).

MSC2010 subject classifications. Primary 60E15; secondary 62F15.

Key words and phrases. Stein's method, Bayesian analysis, prior distribution, posterior distribution.

for $\pi_0 = p_2/p_1$ a nonnegative finite function called likelihood ratio in statistics. To assess the distance between such distributions, we choose the Wasserstein-1 distance (referred to as Wasserstein distance in the sequel) defined as

$$d_{\mathcal{W}}(P_1, P_2) = \sup_{h\in\mathcal{H}} \big|\mathbb{E}\big[h(X_2)\big] - \mathbb{E}\big[h(X_1)\big]\big| \tag{1.1}$$

for $\mathcal{H} = \mathrm{Lip}(1)$ the class of Lipschitz-1 functions, where X_1 has distribution P_1 [resp., probability density function (p.d.f.) p_1] and X_2 has distribution P_2 (resp., p.d.f. p_2). The central aim of this paper is to provide meaningful bounds on $d_{\mathcal{W}}(P_1, P_2)$ in terms of π_0.

Our approach to this problem relies on Stein's *density approach* introduced in [30, 31], as further developed in [17–20]. Let P_1 have density p_1 with interval support $\mathcal{I}_1$ with closure $[a_1, b_1]$ for some $-\infty \le a_1 < b_1 \le +\infty$. Suppose also that P_1 has mean μ. Throughout this paper, we assume that *all* random variables considered have finite mean so that $\mathbb{E}|P_1| < \infty$. Then a notion which will be of particular importance is the Stein kernel of the distribution p_1 which is the function $\tau_1 : [a_1, b_1] \to \mathbb{R}$ given by

$$\tau_1(x) = \frac{1}{p_1(x)} \int_{a_1}^{x} (\mu - y) p_1(y)\, dy$$

and whose properties we will discuss in detail in Section 2.4. Our main results assume that p_1 and p_2 are absolutely continuous densities with finite means, and that $\pi_0 = p_2/p_1$ is a differentiable function with interval support such that $\bar{\mathcal{I}}_2 = [a_2, b_2] \subset [a_1, b_1]$ and the following two assumptions are satisfied.

ASSUMPTION A. $(\pi_0(x) \int_{a_1}^{x} (h(y) - \mathbb{E}[h(X_1)]) p_1(y)\, dy)' \in L^1(dx)$ for all Lipschitz-continuous functions h. Here, $X_1 \sim P_1$.

ASSUMPTION B. $\lim_{x\downarrow a_2} \pi_0(x) \int_{a_1}^{x} (h(y) - \mathbb{E}[h(X_1)]) p_1(y)\, dy = 0 = \lim_{x\uparrow b_2} \pi_0(x) \int_{a_1}^{x} (h(y) - \mathbb{E}[h(X_1)]) p_1(y)\, dy$ for all Lipschitz-continuous functions h.

Assumptions A and B are not stringent as can be seen from the wealth of examples that we treat in this paper. Note that it is implicitly understood in Assumption B that $\mathbb{E}|h(X_1)| < \infty$ for all h Lipschitz because we suppose that X_1 has finite mean. Under these assumptions, we prove the following result; see Theorem 3.1 for a complete statement.

THEOREM. *The Wasserstein distance between P_1 with p.d.f. p_1 and P_2 with p.d.f. $p_2 = \pi_0 p_1$ satisfies the following inequalities*:

$$\big|\mathbb{E}\big[\pi_0'(X_1)\tau_1(X_1)\big]\big| \le d_{\mathcal{W}}(P_1, P_2) \le \mathbb{E}\big[\big|\pi_0'(X_1)\big|\tau_1(X_1)\big],$$

where τ_1 is the Stein kernel associated with p_1 and $X_1 \sim P_1$.

In particular if $P_1 = \mathcal{N}(\mu, \sigma^2)$ is a normal distribution, then the above result simplifies considerably because $\tau_1(x) = \sigma^2$ is constant, yielding

$$\sigma^2 \big|\mathbb{E}\big[\pi_0'(X_1)\big]\big| \le d_{\mathcal{W}}(P_1, P_2) \le \sigma^2 \mathbb{E}\big[\big|\pi_0'(X_1)\big|\big]$$

for any probability distribution P_2 with differentiable density p_2 such that $p_2(x)/p_1(x)$ vanishes at the boundary of $\mathcal{I}_2$.

The Gaussian is characterized by the fact that its Stein kernel is constant. More generally, all distributions belonging to the classical Pearson family possess a polynomial Stein kernel (see [30]). The problem of determining the Stein kernel is, in general, difficult. Even when the Stein kernel τ_1 is not available we can give the following simpler bound (Corollary 3.4).

COROLLARY. *Under the same assumptions as in the above theorem,*

$$\big|\mathbb{E}[X_1] - \mathbb{E}[X_2]\big| \le d_{\mathcal{W}}(P_1, P_2) \le \big\|\pi_0'\big\|_\infty \operatorname{Var}(X_1).$$

More generally, because the Stein kernel is always positive, the upper and lower bounds in the theorem turn out to be the same whenever the likelihood ratio π_0 is monotone, which is equivalent to requiring that P_1 and P_2 are stochastically ordered in the sense of likelihood ratios. This brings our next result (Corollary 3.5).

COROLLARY. *Let $X_1 \sim P_1$ and $X_2 \sim P_2$. If $X_1 \le_{\mathrm{LR}} X_2$ or $X_2 \le_{\mathrm{LR}} X_1$ then*

$$\begin{aligned} d_{\mathcal{W}}(P_1, P_2) &= \big|\mathbb{E}[X_2] - \mathbb{E}[X_1]\big| \\ &= \mathbb{E}\big[\big|\pi_0'(X_1)\big|\tau_1(X_1)\big] \\ &= \mathbb{E}\big[\big|\big(\log \pi_0(X_2)\big)'\big|\tau_1(X_2)\big]. \end{aligned}$$

In case of a monotone likelihood ratio between P_1 and P_2, the first of the above identities is easy to derive directly from the first of the following known alternative definitions of the Wasserstein distance:

$$\begin{aligned} d_{\mathcal{W}}(P_1, P_2) &= \int_{\mathbb{R}} \big|F_{P_1}(x) - F_{P_2}(x)\big|\,dx \\ &= \int_0^1 \big|F_{P_1}^{-1}(u) - F_{P_2}^{-1}(u)\big|\,du \\ &= \inf \mathbb{E}|\xi_1 - \xi_2| \end{aligned} \tag{1.2}$$

with F_{P_1} and $F_{P_1}^{-1}$ (resp., F_{P_2} and $F_{P_2}^{-1}$) the cumulative distribution functions and quantile functions of P_1 (resp., P_2) and where the infimum in this last expression is taken over all possible couplings (ξ_1, ξ_2) of (P_1, P_2) (see, e.g., [32, 33]).

We illustrate the effectiveness of our bounds in several examples at the end of Section 3.1, comparing, for example, Gaussian random variables or Azzalini's skew-symmetric densities with their symmetric counterparts. In Section 4,

we treat as main application the Bayes example wherein we measure explicitly the effect of priors on posterior distributions. Suppose we observe data points $x := (x_1, x_2, \ldots, x_n)$ with sampling density $f(x;\theta)$ (proportional to the likelihood), where θ is the one-dimensional parameter of interest. Let $p_0(\theta)$ be a certain prior distribution, possibly improper and let Θ_2 be the resulting posterior guess for θ perceived as a random variable. By Bayes' theorem, this has density $p_2(\theta; x) = \kappa_2(x) f(x;\theta) p_0(\theta)$ with $\kappa_2(x)$ the normalizing constant which depends on the data. Under moderate assumptions, we provide computable expressions for the Wasserstein distance $d_{\mathcal{W}}(\Theta_2, \Theta_1)$ between this posterior distribution and Θ_1, whose law is the no-prior posterior distribution with density (proportional to the likelihood) given by $p_1(\theta; x) = \kappa_1(x) f(x;\theta)$, again with normalizing constant $\kappa_1(x)$ depending on the data. The bounds we derive are expressed in terms of the data, the prior and the Stein kernel τ_1 of the sampling distribution.

We study the normal model with general and normal priors, the binomial model under a general prior, a conjugate prior, and the Jeffreys' prior. We also consider the Poisson model with an exponential prior, in which case we can make use of the likelihood ratio ordering. For example, with a normal $\mathcal{N}(\mu, \delta^2)$ prior and a random sample $x_1, \ldots, x_n$ from a normal $\mathcal{N}(\theta, \sigma^2)$ model with fixed σ^2, we obtain in (4.4) that

$$\frac{\sigma^2}{n\delta^2 + \sigma^2}|\bar{x} - \mu| \le d_{\mathcal{W}}(\Theta_1, \Theta_2) \le \frac{\sigma^2}{n\delta^2 + \sigma^2}|\bar{x} - \mu| + \frac{\sqrt{2}}{\sqrt{\pi}} \frac{\sigma^3}{n\delta\sqrt{\delta^2 n + \sigma^2}}.$$

Not only do we see that for $n \to \infty$, the distance becomes zero, as is well known, but we also have an explicit dependence on the difference between the sample mean $\bar{x}$ and the prior mean μ, indicating the importance of a reasonable choice for the prior. For a normal $\mathcal{N}(\theta, \sigma^2)$ model and a general prior on θ, we obtain in (4.3) that

$$\frac{\sigma^2}{n}\left|\mathbb{E}\left[\rho_0(\Theta_2)\right]\right| \le d_{\mathcal{W}}(\Theta_1, \Theta_2) \le \frac{\sigma^2}{n}\mathbb{E}\left[\left|\rho_0(\Theta_2)\right|\right]$$

with ρ_0 the score function of the prior distribution. Here, the data are hidden in the distribution of Θ_2. In the binomial case with conjugate prior, we obtain (letting $y = n\bar{x}$, see Section 4.3.1)

$$\frac{1}{n+2}\left|(y+\alpha)\frac{\alpha+\beta-2}{n+\alpha+\beta} - (\alpha - 1)\right| \le d_{\mathcal{W}}(\Theta_1, \Theta_2) \le \frac{1}{n+2}\left((y+\alpha)\frac{|\beta - \alpha|}{n+\alpha+\beta} + |\alpha - 1|\right),$$

with α and β the parameters of the conjugate (beta) prior. Finally, in the Poisson case, we obtain

$$d_{\mathcal{W}}(\Theta_1, \Theta_2) = \frac{\lambda}{n+\lambda}\bar{x} + \frac{\lambda}{n(n+\lambda)},$$

with $\lambda > 0$ the parameter of the exponential prior.

The main tool in this paper is a specification of the general approach in [17] which allows to manipulate Stein operators. Distributions can be compared through their Stein operators which are far from being unique; for a single distribution there is a whole family of operators which could serve as Stein operators; see, for example, [17]. In this paper, for probability distribution P with p.d.f. p we choose the Stein operator $\mathcal{T}_P$ as

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

with the convention that $\mathcal{T}_P f(x) = 0$ outside of the support of P; for details, see Definition 2.1 and [19]. For this choice of operator, the product structure implies a convenient connection between $\mathcal{T}_1$, the Stein operator for P_1 with p.d.f. p_1, and $\mathcal{T}_2$, the Stein operator for P_2 with p.d.f. $p_2 = \pi_0 p_1$, namely

$$\mathcal{T}_2(f) = \mathcal{T}_1(f) + f\frac{\pi_0'}{\pi_0} = \mathcal{T}_1(f) + f(\log \pi_0)';$$

see (3.2). The difference

$$\mathcal{T}_2(f) - \mathcal{T}_1(f) = f(\log \pi_0)'$$

is the cornerstone of our results.

REMARK 1.1. This paper restricts attention to the univariate case. The multivariate case is of considerable interest but our approach requires an extension of the density method to a multivariate setting, which is to date still under construction and not yet available.

Using the approach in [17], it would be possible to extend our results to more general Radon–Nikodym derivatives, at the expense of clarity of exposition.

The paper is organized as follows. In Section 2, we provide the necessary notation and definitions from Stein's method, which allows us to state our main result, Theorem 3.1, in Section 3.1. Several applications of this result are discussed in Examples 3.6 to 3.9, while Section 4 tackles our motivating Bayesian problem by providing a measure of the impact of the choice of the prior on the posterior distribution for finite sample size n.

2. A review of Stein's density approach.

2.1. *Notation and definitions.* Here, we recall some notions from [17] and [19]. Consider a probability distribution P with continuous univariate Lebesgue density p on the real line and let $L^1(p) = L^1(p(x)\,dx)$ denote the collection of $f : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}|f(X)| = \int |f(x)|p(x)\,dx < \infty$, where $X \sim P$. Let $\mathcal{I} = \{x \in \mathbb{R} \mid p(x) > 0\}$ be the support of p. In this paper, we shall use the following definition of a Stein operator; see, for example, [17] for a discussion of alternative choices.

DEFINITION 2.1 (Stein pair). The Stein class $\mathcal{F}(P)$ of P is the collection of $f : \mathbb{R} \to \mathbb{R}$ such that (i) fp is absolutely continuous, (ii) $(fp)' \in L^1(dx)$ and (iii) $\int_{\mathbb{R}} (fp)'\,dx = 0$. The Stein operator $\mathcal{T}_P$ for P is

$$\mathcal{T}_P : \mathcal{F}(P) \to L^1(p) : f \mapsto \mathcal{T}_P f = \frac{(fp)'}{p} \tag{2.1}$$

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

Here, $(fp)'$ denotes the derivative of fp which exists Lebesgue-almost surely due to the assumption of absolute continuity. Often the Stein pair $(\mathcal{F}(P), \mathcal{T}_P)$ is written as dependent on $X \sim P$ rather than on P [i.e., as $(\mathcal{F}(X), \mathcal{T}_X)$]; we use the dependence on the distribution to emphasize that the pair itself is not random.

Note that because we only consider f multiplied by p the behavior of f outside of $\mathcal{I}$ is irrelevant.

REMARK 2.2. A sufficient condition for $\mathcal{F}(P) \neq \{f = 0\}$ is that p' is integrable with integral 0 so that, for example, $f = 1 \in \mathcal{F}(P)$. Such an assumption is in general too strong (see, e.g., [31] for a discussion about the arcsine distribution) and weaker assumptions on p are permitted in our framework, although in such cases stronger constraints on the functions in $\mathcal{F}(P)$ are necessary. In particular, the nonzero constant functions may not belong to $\mathcal{F}(P)$.

All random quantities appearing in the sequel will be assumed to have nonempty Stein class (an assumption verified for all classical distributions from the literature).

It is easy to see from Definition 2.1(iii) that $\mathbb{E}[\mathcal{T}_P f(X)] = 0$ for all $f \in \mathcal{F}(P)$. More generally, one can prove that if Y is absolutely continuous with same support as X and p.d.f. q and if X has p.d.f. p such that $\mathcal{F}(P)$ is dense in $L^1(P)$ and q/p is differentiable, then $Y \overset{\mathcal{D}}{=} X$ (equality in distribution) if and only if $\mathbb{E}[\mathcal{T}_P f(Y)] = 0$ for all $f \in \mathcal{F}(P)$; see [17], Section 3.5, Equation (41) with $g = 1$. For any family of operators $\mathcal{T}$ indexed by univariate probability measures P and Q and for any class of functions $\mathcal{G}$, we say that $(\mathcal{T}_P, \mathcal{G})$ is a *Stein characterization* if

$$P = Q \quad \Longleftrightarrow \quad \mathcal{T}_Q(f) = \mathcal{T}_P(f) \qquad \forall f \in \mathcal{G}; \tag{2.2}$$

see [18, 19] for general versions. In particular, a Stein pair $(\mathcal{T}_P, \mathcal{F}(P))$ is a Stein characterization.

With our notation, the operator $\mathcal{T}_P$ also admits an inverse which is easy to write out formally at least. Let $X \sim P$ have (open, closed or half-open) interval support $\mathcal{I}$ between a and b, where $-\infty \leq a < b \leq +\infty$ and

$$\mathcal{F}^{(0)}(P) = \big\{h \in L^1(p) : \mathbb{E}\big[h(X)\big] = 0\big\}.$$

Define $\mathcal{T}_P^{-1} : \mathcal{F}^{(0)}(P) \to \mathcal{F}(P)$ by

$$\mathcal{T}_P^{-1} h(x) = \frac{1}{p(x)} \int_a^x h(y) p(y)\, dy = -\frac{1}{p(x)} \int_x^b h(y) p(y)\, dy. \tag{2.3}$$

The operator $\mathcal{T}_P^{-1}$ is the *inverse Stein operator* of P in the sense that

$$\mathcal{T}_P\big(\mathcal{T}_P^{-1} h\big) = h.$$

Note how the particular structure of the right-hand side of (2.3) ensures that $\mathcal{T}_P^{-1} h$ belongs to $\mathcal{F}(P)$ for any $h \in \mathcal{F}^{(0)}(P)$. If in addition $(fp)(a) = (fp)(b) = 0$ for all $f \in \mathcal{F}(P)$, then

$$\mathcal{T}_P^{-1}(\mathcal{T}_P f) = f$$

so that $\mathcal{T}_P^{-1}$ constitutes a bona fide inverse in this case.

2.2. *Standardizations of the operator.* Although the Stein pair $(\mathcal{T}_P, \mathcal{F}(P))$ is uniquely defined in Definition 2.1, there are many implicit conditions on $f \in \mathcal{F}(P)$ which are useful to identify before applying this construction to specific approximation problems. In particular, for favorable behavior of the inverse Stein operator it may be advantageous to consider only subclasses $\mathcal{F}_{\mathrm{sub}}(P) \subset \mathcal{F}(P)$ of functions satisfying certain target-specific and well-chosen constraints. A good choice of subclass will lead to specific forms of the resulting operator which may turn out to have a smooth inverse Stein operator, as illustrated in the next example. As long as $\mathcal{F}_{\mathrm{sub}}(P)$ is a measure-determining class, the class is informative enough to satisfy (2.2).

EXAMPLE 2.3. In the case of the Laplace distribution Lap with p.d.f. $p(x) \propto e^{-|x|}$ the Stein operator from Definition 2.1 is

$$\mathcal{T}_{\mathrm{Lap}} f(x) = f'(x) - \operatorname{sign}(x) f(x) \tag{2.4}$$

with $f \in \mathcal{F}(\mathrm{Lap})$, the class of functions such that $f(x)e^{-|x|}$ is differentiable almost surely with integrable derivative, and the derivative of $f(x)e^{-|x|}$ integrates to 0 over the real line. This operator does not have agreeable properties, mainly because the assumptions on $\mathcal{F}(\mathrm{Lap})$ are not explicit (see, e.g., [10] and [25]). It is indeed sufficient to consider functions of the form $f(x) = (x f_0(x) e^{|x|})'/e^{|x|}$ for certain functions f_0. Applying $\mathcal{T}_{\mathrm{Lap}}$ to such functions yields the second-order operator

$$\mathcal{T}_{\mathrm{Lap}} f(x) = \mathcal{A}_X f_0(x) = x f_0''(x) + 2 f_0'(x) - x f_0(x) \tag{2.5}$$

with $f_0 \in \mathcal{F}(\mathcal{A}_{\mathrm{Lap}})$ the class of functions which are piecewise twice continuously differentiable such that $x f_0''(x)$, $f_0'(x)$ and $x f_0(x)$ are all in $L^1(e^{-|x|}\, dx)$,

as considered, for example, in [10, 11]. In [25], functions of the form $f(x) = (-(g(x) - g(0))e^{|x|})'/e^{|x|}$ yielded the second-order operator

$$\mathcal{T}_{\mathrm{Lap,PR}}g(x) = g(x) - g(0) - g''(x)$$

for g locally absolutely continuous with $g \in L^1(e^{-|x|}\,dx)$, g' also locally absolutely continuous and $g'' \in L^1(e^{-|x|}\,dx)$. The operator $\mathcal{T}_{\mathrm{Lap,PR}}$ is also discussed in [10] but not used in [10] because it did not fit in with Malliavin calculus as well as (2.5).

Even in the straightforward situation of a normal distribution, often a standardization is applied, as explained in the next example.

EXAMPLE 2.4. For the standard normal distribution $\mathcal{N}(0, 1)$ it is easy to write out the operator (2.1) explicitly to get $\mathcal{T}_{\mathcal{N}(0,1)}(f)(x) = f'(x) - xf(x)$ acting on a wide class of functions $\mathcal{F}(\mathcal{N}(0, 1))$ which includes all absolutely continuous functions with polynomial decay at $\pm\infty$. In particular, the constant function $\mathbf{1}$ is in $\mathcal{F}(\mathcal{N}(0, 1))$. A standardization of the form $f(x) = H_n(x)f_0(x)$ with H_n the nth Hermite polynomial $[H_0(x) = 1, H_1(x) = x, H_2(x) = x^2 - 1]$ gives as operator $\mathcal{A}f_0(x) = H_n(x)f_0'(x) - H_{n+1}(x)f_0(x)$; see, for example, [12].

It is also possible to study the behavior of functions f_h under quite general conditions on h. For instance, if $\mathcal{H}$ is the set of measurable functions $h : \mathbb{R} \to [0, 1]$ (leading to the total variation measure), then $\mathcal{F}^{(1)}$ is contained in the collection of differentiable functions such that $\|f\| \le \sqrt{\pi/2}$ and $\|f'\| \le 2$; see, for instance, [21].

For the general normal distribution $\mathcal{N}(\mu, \sigma^2)$, the operator (2.1) gives

$$\mathcal{T}_{\mathcal{N}(\mu,\sigma^2)}(f)(x) = f'(x) - \frac{x - \mu}{\sigma^2}f(x). \tag{2.6}$$

The standardization $f(x) = \sigma^2 g'(x)$ yields the classical Ornstein–Uhlenbeck Stein operator $\mathcal{A}g(x) = \sigma^2 g''(x) - (x - \mu)g'(x)$; see, for example, [2].

We call the passage from a parsimonious operator $\mathcal{T}_P$ [such as (2.4)] acting on the implicit class $\mathcal{F}(P)$ to a specific operator $\mathcal{A}_P$ [such as (2.5)] acting on a generic class $\mathcal{F}(\mathcal{A}_P)$ a *standardization* of $(\mathcal{T}_P, \mathcal{F}(P))$. Given P there are infinitely many different possible standardizations.

2.3. *The Stein transfer principle.* Suppose that we aim to assess the discrepancy between the laws of two random quantities X with distribution P and W with distribution Q, say, in terms of some probability distance of the form

$$d_{\mathcal{H}}(P, Q) = d_{\mathcal{H}}(X, W) = \sup_{h\in\mathcal{H}} \big|\mathbb{E}\big[h(W)\big] - \mathbb{E}\big[h(X)\big]\big|, \tag{2.7}$$

for $\mathcal{H}$ some measure-determining class; many common distances can be written in the form (2.7), including the Kolmogorov distance (with $\mathcal{H}$ the collection of indicators of half-lines), the total variation distance (with $\mathcal{H}$ the collection of indicators of Borel sets) and the Wasserstein distance [see (1.1)]. Here, writing $d_{\mathcal{H}}(X, W)$ is a shorthand for (2.7): this distance is not random.

Let P have Stein pair $(\mathcal{T}_P, \mathcal{F}(P))$ and consider a standardization $(\mathcal{A}_P, \mathcal{F}(\mathcal{A}_P))$ as described in Section 2.2. The first key idea in Stein's method is to relate the test functions h of interest to a function $f = f_h \in \mathcal{F}(\mathcal{A}_P)$ through the so-called *Stein equation*

$$h(x) - \mathbb{E}\big[h(X)\big] = \mathcal{A}_P f(x), \qquad x \in \mathcal{I}, \tag{2.8}$$

so that, for f_h solving (2.8), we get $h(W) - \mathbb{E}[h(X)] = \mathcal{A}_P f_h(W)$ and, in particular,

$$\sup_{h\in\mathcal{H}} \big|\mathbb{E}\big[h(W)\big] - \mathbb{E}\big[h(X)\big]\big| = \sup_{f\in\mathcal{F}^{(1)}} \big|\mathbb{E}\big[\mathcal{A}_P f(W)\big]\big|, \tag{2.9}$$

where $\mathcal{F}^{(1)} = \mathcal{F}^{(1)}(\mathcal{A}_P, \mathcal{H}) = f \in \mathcal{F}(\mathcal{A}_P) \mid \mathcal{A}_P f = h - \mathbb{E}[h(X)]$ for some $h \in \mathcal{H}$. The first step in Stein's method thus consists in some form of transfer principle whereby one transforms the problem of bounding the distance $d_{\mathcal{H}}(P, Q)$ into that of bounding the expectations of the operators $\mathcal{A}_P$ over a specific class of functions.

EXAMPLE 2.5. For the standard normal distribution, the operators (2.1) and (2.6) give $\mathcal{T}_{\mathcal{N}(0,1)}(f)(x) = f'(x) - xf(x)$. Bounding expressions of the form $|\mathbb{E}[f'(W) - Wf(W)]|$ as occurring in the right-hand side of (2.9) is a potent starting point for Gaussian approximation problems. Prominent examples include $W = \sum_i \xi_i$ a standardized sum of weakly dependent variables, and $W = F(X)$ a functional of a Gaussian process; see, for example, [2, 21, 26] for an overview. It is also possible to study the behavior of functions f_h under quite general conditions on h. For instance, if $\mathcal{H}$ is the set of measurable functions $h : \mathbb{R} \to [0, 1]$ (leading to the total variation distance), then $\mathcal{F}^{(1)}$ is contained in the collection of differentiable functions such that $\|f\| \le \sqrt{\pi/2}$ and $\|f'\| \le 2$; see [21], Section 3.

In general, the success of Stein's method for a particular target relies on the positive combination of three factors:

(i) the functions in $\mathcal{F}^{(1)}$ need to have "good" properties (e.g., be bounded with bounded derivatives as in Example 2.5),

(ii) the operator $\mathcal{A}_P$ needs to be amenable to computations (e.g., its expression should only involve polynomial functions),

(iii) there must be some "handle" on the expressions $\mathbb{E}[\mathcal{A}_P f(W)]$ (e.g., allowing for Taylor-type expansions or the application of couplings).

Conditions (i) to (iii) are satisfied for a great variety of target distributions (including the exponential, chi-squared, gamma, semi-circle, variance gamma and many others; see, e.g., https://sites.google.com/site/steinsmethod/).

2.4. *The Stein kernel.* One of the many keys to a successful application of Stein's method for a given target distribution P lies in the properties of P's *Stein kernel* introduced in [30], Lecture VI, Lemma 1. We now review some properties of this quantity which will play a central role in our analysis.

DEFINITION 2.6. Let $X \sim P$ an absolutely continuous probability distribution with p.d.f. p and mean μ. Suppose that p has interval support with closure $[a, b]$. Then, letting Id denote the identity function, the *Stein kernel* of P is the function $x \mapsto \tau_P(x)$ defined by

$$\tau_P(x) = \mathcal{T}_P^{-1}(\mu - \mathrm{Id})(x) = \frac{1}{p(x)} \int_a^x (\mu - y) p(y)\, dy. \tag{2.10}$$

By metonymy, we also call the random variable $\tau_P(X)$ a Stein kernel for P.

Following Lecture VI, Lemma 1 in [30], one can show that the Stein kernel satisfies the integration by parts identity

$$\mathbb{E}\big[\tau_P(X)\varphi'(X)\big] = \mathbb{E}\big[(X - \mu)\varphi(X)\big] \tag{2.11}$$

for all continuous and piecewise continuous functions $\varphi : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[|\varphi'(X)|\tau_P(X)] < \infty$. If P is such that τ_P is well defined and finite on $[a, b]$, then (2.11) also uniquely defines the Stein kernel (up to zero-Lebesgue measure sets) so that this equation is sometimes used as a definition; see [23]. A careful study of integrals of the form (2.10) is performed in [6], Section 3.

The following properties of the Stein kernel are immediate consequences of its definition:

$$\text{for all } x \in \mathbb{R} \text{ we have that } \tau_P(x) \geq 0 \quad \text{and} \quad \mathbb{E}\big[\tau_P(X)\big] = \mathrm{Var}(X). \tag{2.12}$$

The Stein kernels for a wide variety of classical distributions (all members of the Pearson family, as it turns out) bear agreeable expressions; see [6], [8], Table 1, [22, 23]. Multivariate extensions have been proposed; see [13, 14]. For Stein operators in Hilbert spaces, see, for example, [3] and [24].

2.5. *Stein standardizations.* Let P have a continuous density p with mean μ and support $\mathcal{I}$ such that the closure of $\mathcal{I}$ is the interval $[a, b]$ (possibly with infinite endpoints). Let $(\mathcal{T}_P, \mathcal{F}(P))$ be the Stein pair of P and suppose that P admits a Stein kernel $\tau_P(x)$, as defined in Definition 2.6. We introduce the standardized Stein pair $(\mathcal{A}_P, \mathcal{F}(\mathcal{A}_P))$ with

$$\mathcal{A}_P f(x) = \mathcal{T}_P(\tau_P f)(x) = \tau_P(x) f'(x) + (\mu - x) f(x), \qquad x \in \mathcal{I}, \tag{2.13}$$

and

$$\mathcal{F}(\mathcal{A}_P) = \Big\{ f : \mathbb{R} \to \mathbb{R} \text{ absolutely continuous such that}$$
$$\lim_{x\to a} f(x) \int_a^x (\mu - u)\, p(u)\, du = \lim_{x\to b} f(x) \int_x^b (\mu - u)\, p(u)\, du = 0$$
$$\text{and } \left(f(x) \int_a^x (\mu - u)\, p(u)\, du \right)' \in L^1(dx) \Big\}.$$

When P admits a Stein kernel then the operator $\mathcal{A}_P$ in (2.13) has all the properties required in order for the Stein transfer principle from Section 2.3 to pan out. In particular, the next lemma shows that, whenever applicable, the standardization in (2.13) satisfies requirement (i).

LEMMA 2.7. *Let $\mathcal{H} = \mathrm{Lip}(1)$ be the collection of Lipschitz functions $h : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant 1 and let $\mathcal{F}^{(1)}$ be the collection of $f \in \mathcal{F}(\mathcal{A}_P)$ such that $\mathcal{A}_P f = h - \mathbb{E}[h(X)]$ for some $h \in \mathcal{H}$. Then $\mathcal{F}^{(1)}$ is contained in the collection of functions f such that $\|f\|_\infty \le 1$.*

Lemma 2.7 is a consequence of [6], Proposition 3.13(a) and Corollary 3.15, adapted to our framework. The key to our approach lies in the fact that the bound in Lemma 2.7 *does not depend* on the standardization of the target P; it is in particular independent of the mean and variance of $X \sim P$ or of any normalizing constant that might appear in the expression of the density of P.

3. Comparing univariate continuous densities. For $i = 1, 2$, let P_i be a probability distribution with an absolutely continuous density p_i having support $\mathcal{I}_i$ with closure $\bar{\mathcal{I}}_i = [a_i, b_i]$, for some $-\infty \le a_i < b_i \le +\infty$. Suppose that $\mathcal{I}_2 \subset \mathcal{I}_1$ and define π_0 through

$$p_2 = \pi_0 p_1. \tag{3.1}$$

Associate with both distributions the Stein pairs $(\mathcal{T}_i, \mathcal{F}_i)$ for $i = 1, 2$, as well as the resulting construction from the previous section.

The product structure (3.1) implies a key connection between $\mathcal{T}_1$ and $\mathcal{T}_2$, namely

$$\mathcal{T}_2(f) = \mathcal{T}_1(f) + f\frac{\pi_0'}{\pi_0} = \mathcal{T}_1(f) + f(\log \pi_0)' \tag{3.2}$$

for all $f \in \mathcal{F}_1 \cap \mathcal{F}_2$.

3.1. *Bounds on the Wasserstein distance between univariate continuous densities.* Our main objective in this section is to provide computable and meaningful bounds on the Wasserstein distance $d_{\mathcal{W}}(P_1, P_2)$, defined in (1.1), in terms of π_0 and P_1, under the product structure (3.1).

THEOREM 3.1. *For $i = 1, 2$, let P_i be a probability distribution with an absolutely continuous density p_i having interval support $\mathcal{I}_i$ with closure $\bar{\mathcal{I}}_i = [a_i, b_i]$, for some $-\infty \le a_i < b_i \le +\infty$; suppose that $\mathcal{I}_2 \subset \mathcal{I}_1$ and let $X_i \sim P_i$ have finite means μ_i for $i = 1, 2$. Assume that $\pi_0 = \frac{p_2}{p_1}$, defined on $\mathcal{I}_2$, is differentiable on $\mathcal{I}_2$ and satisfies $\mathbb{E}|(X_1 - \mu_1)\pi_0(X_1)| < \infty$ and*

$$\left(\pi_0(x)\int_{a_1}^{x}\big(h(y) - \mathbb{E}\big[h(X_1)\big]\big)p_1(y)\,dy\right)' \in L^1(dx), \tag{3.3}$$

as well as

$$\begin{aligned}&\lim_{x\downarrow a_2}\pi_0(x)\int_{a_1}^{x}\big(h(y) - \mathbb{E}\big[h(X_1)\big]\big)p_1(y)\,dy\\ &\quad= \lim_{x\uparrow b_2}\pi_0(x)\int_{a_1}^{x}\big(h(y) - \mathbb{E}\big[h(X_1)\big]\big)p_1(y)\,dy = 0\end{aligned} \tag{3.4}$$

for all $h \in \mathcal{H}$, the set of Lipschitz-1 functions on $\mathbb{R}$. Then

$$\big|\mathbb{E}\big[\pi_0'(X_1)\tau_1(X_1)\big]\big| \le d_{\mathcal{W}}(P_1, P_2) \le \mathbb{E}\big[\big|\pi_0'(X_1)\big|\tau_1(X_1)\big], \tag{3.5}$$

where τ_1 is the Stein kernel of P_1.

PROOF. We first prove the lower bound. Let $X_2 \sim P_2$. Start by noting that $d_{\mathcal{W}}(P_1, P_2) \ge |\mathbb{E}[X_2] - \mathbb{E}[X_1]|$ because $\mathrm{Id} \in \mathrm{Lip}(1)$. With (3.1), we get that

$$\begin{aligned}\mathbb{E}[X_2] - \mathbb{E}[X_1] &= \mathbb{E}\big[X_1\pi_0(X_1)\big] - \mu_1\\ &= \mathbb{E}\big[(X_1 - \mu_1)\pi_0(X_1)\big] + \mu_1\big(\mathbb{E}\big[\pi_0(X_1)\big] - 1\big)\\ &= \mathbb{E}\big[\tau_1(X_1)\pi_0'(X_1)\big],\end{aligned} \tag{3.6}$$

where we used the fact that $\mathbb{E}[\pi_0(X_1)] = 1$ and the definition (2.11) of $\tau_1(X_1)$ in the last line. Taking absolute values gives the lower bound.

Next, we prove the upper bound. By (2.3), $f_h = \mathcal{T}_1^{-1}(h - \mathbb{E}[h(X_1)]) \in \mathcal{F}_1$. On the other hand, Conditions (3.3) and (3.4) guarantee that $f_h \in \mathcal{F}_2$ for all h because

$$p_2(x)f_h(x) = \pi_0(x)\int_{a_1}^{x}\big(h(y) - \mathbb{E}\big[h(X_1)\big]\big)p_1(y)\,dy$$

is necessarily absolutely continuous. We conclude that all functions $f_h = \mathcal{T}_1^{-1}(h - \mathbb{E}[h(X_1)])$ belong to the intersection $\mathcal{F}_1 \cap \mathcal{F}_2$. Hence,

$$\begin{aligned}\mathbb{E}\big[h(X_2)\big] - \mathbb{E}\big[h(X_1)\big] &= \mathbb{E}\big[\mathcal{T}_1(f_h)(X_2)\big]\\ &= \mathbb{E}\big[\mathcal{T}_1(f_h)(X_2)\big] - \mathbb{E}\big[\mathcal{T}_2(f_h)(X_2)\big]\\ &= -\mathbb{E}\big[f_h(X_2)(\log\pi_0)'(X_2)\big].\end{aligned} \tag{3.7}$$

Equality (3.7) follows from the assumption that $f_h \in \mathcal{F}_2$ so that $\mathcal{T}_2 f_h$ cancels when integrated with respect to p_2, whereas the last equality follows from Equation

(3.2). Now we define $g_h = f_h/\tau_1$ and recall that $\tau_1 \geq 0$ to get

$$\begin{aligned}\bigl|\mathbb{E}\bigl[h(X_2)\bigr] - \mathbb{E}\bigl[h(X_1)\bigr]\bigr| &= \bigl|\mathbb{E}\bigl[g_h(X_2)(\log\pi_0)'(X_2)\tau_1(X_2)\bigr]\bigr| \\ &\leq \|g_h\|_\infty \mathbb{E}\bigl[\bigl|(\log\pi_0)'(X_2)\bigr|\tau_1(X_2)\bigr].\end{aligned}$$

It follows from Lemma 2.7 that $\|g_h\|_\infty \leq 1$ for all $h \in \mathrm{Lip}(1)$, yielding

$$\begin{aligned}d_{\mathcal{W}}(P_1, P_2) &\leq \mathbb{E}\bigl[\bigl|(\log\pi_0)'(X_2)\bigr|\tau_1(X_2)\bigr] \\ &= \mathbb{E}\bigl[|\pi_0'(X_1)|\tau_1(X_1)\bigr],\end{aligned}$$

the last equality again following from (3.1). □

Conditions (3.3) and (3.4) are crucial. Condition (3.4) is in a sense innocuous because $\mathcal{I}_2 \subset \mathcal{I}_1$. For instance, if $a_1 < a_2 < b_2 < b_1$, then Condition (3.4) is satisfied for h Lipschitz if $\pi_0(a_2+) = \pi_0(b_2-) = 0$ because

$$\begin{aligned}\left|\int_{a_1}^x \bigl(h(y) - \mathbb{E}\bigl[h(X_1)\bigr]\bigr)p_1(y)\,dy\right| &= \left|\int_{a_1}^x \int_{a_1}^{b_1} \bigl(h(y) - h(z)\bigr)p_1(z)\,dz p_1(y)\,dy\right| \\ &\leq 2\|h'\|\mathbb{E}|X_1|.\end{aligned}$$

Condition (3.3) is quite stringent yet hard to verify in practice. Our next result provides explicit and easy to verify sufficient conditions on p for these, and hence Theorem 3.1 to hold.

PROPOSITION 3.2. *We use the notation of Theorem 3.1. Suppose that π_0, p_1 and p_2 are differentiable over their support and that their derivatives are integrable. Suppose that*

$$\lim_{x\to a_2,b_2} \pi_0(x)p_1(x)\tau_1(x) = \lim_{x\to a_2,b_2} p_2(x)\tau_1(x) = 0$$

(*these limits are to be interpreted as left or right hand limits if necessary*). *Let $\rho_1 = p_1'/p_1$ and suppose also that*

$$\pi_0' p_1 \tau_1 = p_2'\tau_1 - \rho_1\tau_1 p_2 \in L^1(dx).$$

Then Theorem 3.1 applies.

PROOF. Conditions (3.3) and (3.4) are equivalent to requiring that $f_h \in \mathcal{F}_2$, in other words $(f_h p_2)$ needs to be differentiable, $(f_h p_2)'$ needs to be integrable with integral on $\mathcal{I}_2$ (the support of p_2) equal to 0. By definition,

$$f_h(x)p_2(x) = \pi_0(x)\int_{a_1}^x \bigl(h(y) - \mathbb{E}\bigl[h(X_1)\bigr]\bigr)p_1(y)\,dy$$

is differentiable if π_0 is differentiable. Next, differentiating,

$$\begin{aligned}(f_h p_2)'(x) = \pi_0'(x)&\int_{a_1}^x \bigl(h(y) - \mathbb{E}\bigl[h(X_1)\bigr]\bigr)p_1(y)\,dy \\ &+ \pi_0(x)\bigl(h(x) - \mathbb{E}\bigl[h(X_1)\bigr]\bigr)p_1(x).\end{aligned}$$

For the second summand, the Lipschitz property of h gives the bound

$$\bigl|h(x)-\mathbb{E}\bigl[h(X_1)\bigr]\bigr| \le \int_{a_1}^{b_1} |h(x)-h(y)|p_1(y)\,dy \le \int_{a_1}^{b_1} |x-y|p_1(y)\,dy,$$

so that

$$\begin{aligned}\int_{a_1}^{b_1} \bigl|\pi_0(x)\bigl(h(x)-\mathbb{E}\bigl[h(X_1)\bigr]\bigr)p_1(x)\bigr|\,dx &\le \int_{a_1}^{b_1} p_2(x)\int_{a_1}^{b_1} |x-y|p_1(y)\,dy\,dx\\ &\le \mathbb{E}|X_1|+\mathbb{E}|X_2|,\end{aligned}$$

and the latter expectations are assumed to exist. Hence, in order to guarantee (3.3) it is sufficient to impose that

$$\pi_0'(x)\int_{a_1}^{x}\bigl(h(y)-\mathbb{E}\bigl[h(X_1)\bigr]\bigr)p_1(y)\,dy \in L^1(dx). \tag{3.8}$$

We can write

$$\int_{a_1}^{x}\bigl(h(y)-\mathbb{E}\bigl[h(X_1)\bigr]\bigr)p_1(y)\,dy = p_1(x)\tau_1(x)g_h(x) \tag{3.9}$$

with

$$g_h(x)=\frac{1}{\tau_1(x)p_1(x)}\int_{a_1}^{x}\bigl(h(y)-\mathbb{E}\bigl[h(X_1)\bigr]\bigr)p_1(y)\,dy$$

a function which we know from Lemma 2.7 to be bounded uniformly by 1. Hence, (3.8) [and therefore (3.3)] boils down to a condition on $\pi_0'(x)p_1(x)\tau_1(x)$. Similarly, using (3.9), Condition (3.4) is satisfied when $\pi_0(x)p_1(x)\tau_1(x)$ vanishes at the boundaries. □

REMARK 3.3. Our upper bounds are not restricted to the Wasserstein case only. Indeed, mimicking large parts of the proof of Theorem 3.1, we obtain the general bound

$$d_{\mathcal{H}}(P_1,P_2) \le \kappa_{\mathcal{H}}\mathbb{E}\bigl[\bigl|\pi_0'(X_1)\bigr|\tau_1(X_1)\bigr] \tag{3.10}$$

with $\kappa_{\mathcal{H}}=\sup_{h\in\mathcal{H}}\|\mathcal{T}_1^{-1}(h-E_1h)/\tau_1\|_\infty$ and $\mathcal{H}$ a measure-determining class of functions (the Kolmogorov distance corresponds to the class of indicators of half-lines, the total variation distance to the indicators of Borel sets). Usefulness of (3.10) hinges around availability of bounds similar to Lemma 2.7 on the more general constant $\kappa_{\mathcal{H}}$.

Unravelling the lower bound and using (2.12) in the upper bound of (3.5), we also obtain the following weaker but perhaps more transparent result.

COROLLARY 3.4. *Under the same assumptions as for Theorem 3.1, with* $X_2\sim P_2$,

$$\bigl|\mathbb{E}[X_2]-\mathbb{E}[X_1]\bigr| \le d_{\mathcal{W}}(P_1,P_2) \le \bigl\|\pi_0'\bigr\|_\infty \operatorname{Var}(X_1). \tag{3.11}$$

We shall use Corollary 3.4 in Section 4. We stress the fact that there is *no* normalizing constant appearing in the bounds (3.5) and (3.11). Also, the absence of a Stein kernel in (3.11) is in some cases an advantage because the Stein kernel is not always easy to compute.

Even in situations where the c.d.f.'s are available, exact computable expressions of Wasserstein distances via (1.2) tend to be difficult to obtain, as can be seen, for example, in Example 3.6. The similarity between the upper and lower bounds in (3.5) encourages us to formulate the next result.

COROLLARY 3.5. *If $X_i \sim P_i, i = 1, 2$ are as in Theorem 3.1 and if π_0 is monotone increasing or decreasing, then*

$$\begin{aligned} d_{\mathcal{W}}(P_1, P_2) &= \big|\mathbb{E}[X_2] - \mathbb{E}[X_1]\big| \\ &= \mathbb{E}\big[|\pi_0'(X_1)|\tau_1(X_1)\big] \\ &= \mathbb{E}\big[\big|(\log \pi_0)'(X_2)\big|\tau_1(X_2)\big]. \end{aligned} \tag{3.12}$$

Note how the second expression in (3.12) can be immediately obtained from the first by applying the same argument as in (3.6). Now while the second expression in (3.12) is new, the first is in fact not. Indeed the condition that π_0 be monotone in Corollary 3.5 is equivalent to requiring $X_1 \geq_{\mathrm{LR}} X_2$ or $X_1 \leq_{\mathrm{LR}} X_2$ (stochastically ordered in the sense of likelihood ratio, see, e.g., [27], Section 9.4, or Example 3.8). If $X_1 \leq_{\mathrm{LR}} X_2$, then $F_{P_2} \leq F_{P_1}$ (see, e.g., [28], Theorem 1.C.4), so that $d_{\mathcal{W}}(X_1, X_2) = \int_{\mathbb{R}} (F_{P_1}(x) - F_{P_2}(x))\,dx = \mathbb{E}[X_1] - \mathbb{E}[X_2]$.

EXAMPLE 3.6 (Distance between Gaussians). To compare two Gaussian distributions, $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$, order them so that $\sigma_2^2 \geq \sigma_1^2$, and if $\sigma_1 = \sigma_2$ then assume that $\mu_1 > \mu_2$. If P_1 is $\mathcal{N}(\mu_1, \sigma_1^2)$, then $\tau_1(x) = \sigma_1^2$ is constant (see, e.g., [30]). With P_2 being $\mathcal{N}(\mu_2, \sigma_2^2)$, all conditions in Proposition 3.2 are satisfied. Applying Theorem 3.1 and noting that $(\log \pi_0(x))' = x(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}) + (\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2})$, we obtain that

$$\begin{aligned} |\mu_2 - \mu_1| &\leq d_{\mathcal{W}}(P_1, P_2) \\ &\leq \sigma_1^2 \mathbb{E}\left| X_2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) + \left(\frac{\mu_2}{\sigma_2^2} - \frac{\mu_1}{\sigma_1^2}\right)\right| \\ &\leq \left|\frac{\sigma_1^2}{\sigma_2^2}\mu_2 - \mu_1\right| + \left(\frac{\sigma_1^2}{\sigma_2^2} - 1\right)\mathbb{E}|X_2|. \end{aligned} \tag{3.13}$$

Inequality (3.13) generalizes [21], Proposition 3.6.1, to the case of nonzero means. In the special case $\mu_2 = \mu_1 = 0$, we compute $\mathbb{E}|X_2| = \sqrt{2/\pi}\sigma_2$ to get

$$d_{\mathcal{W}}(P_1, P_2) \leq \sqrt{\frac{2}{\pi}}\frac{\sigma_1^2 - \sigma_2^2}{\sigma_2},$$

which is exactly the same as in [21], Proposition 3.6.1.

If $\mu_2 \neq 0$, then the general expression for $\mathbb{E}|X_2|$ is not agreeable, which is why we suggest using the inequality $\mathbb{E}|X_2| \leq (\mathbb{E}[X_2^2])^{1/2} = \sqrt{\sigma_2^2 + \mu_2^2}$, leading to

$$|\mu_2 - \mu_1| \leq d_{\mathcal{W}}(P_1, P_2) \leq \left|\frac{\sigma_1^2}{\sigma_2^2}\mu_2 - \mu_1\right| + \left(\frac{\sigma_1^2}{\sigma_2^2} - 1\right)\sqrt{\sigma_2^2 + \mu_2^2}.$$

With $\mu_1 = \mu_2 = \mu$, the upper bound becomes $(|\mu| + \sqrt{\sigma_2^2 + \mu_2^2})(\frac{\sigma_1^2}{\sigma_2^2} - 1)$. We have not found a similar result in the literature (outside of the centered case) and computing the Wasserstein distance directly using (1.2) is prohibitive as the c.d.f.'s are not available in closed form.

EXAMPLE 3.7 (Distance between Azzalini-type skew-symmetric distributions). Consider a symmetric density p_1 on the real line. The so-called *Azzalini-type skew-symmetric distributions* are constructed from such a p.d.f. p_1 by considering the densities $p_2(x) = 2p_1(x)G(\lambda x)$ with G the c.d.f. of a univariate symmetric distribution with p.d.f. g and $\lambda \in \mathbb{R}$ a parameter (called skewness parameter); see [15] for an overview of these skewing mechanisms and of their applications. The founding example is Azzalini's *skew-normal density* $2\phi(x)\Phi(\lambda x)$ (denoted $\mathcal{SN}(0, 1, \lambda)$, see [1]), where ϕ and Φ respectively stand for the standard normal density and cumulative distribution function.

Corollary 3.5 provides the Wasserstein distance between P_1 with p.d.f. p_1 and its skew-symmetric counterpart P_2 with p.d.f. p_2 (assuming that G is such that the assumptions of Theorem 3.1 are satisfied) since in this case $\pi_0(x) = 2G(\lambda x)$ is necessarily monotone and thus

$$d_{\mathcal{W}}(p_1, p_2) = 2|\lambda|\mathbb{E}\big[\tau_1(X_1)g(\lambda X_1)\big]. \tag{3.14}$$

Perhaps the most interesting instance of the above is the comparison of the standard normal with the skew-normal which does satisfy the conditions of Theorem 3.1:

$$d_{\mathcal{W}}\big(\mathcal{N}(0, 1), \mathcal{SN}(0, 1, \lambda)\big) = \sqrt{\frac{2}{\pi}}\frac{|\lambda|}{\sqrt{1 + \lambda^2}}$$

[recall that $\tau_1(x) = 1$]. Letting $\lambda \to \infty$ we obtain that the distance between the half-normal with density $2\phi(x)\mathcal{I}_{x\geq 0}$ and the normal is $\sqrt{2/\pi}$, see also [7]. As in the previous example, such results are not easy to obtain directly from (1.2).

Likelihood ratio orderings have a natural role in comparing parametric densities. Let $p(x; \theta)$ be a parametric family of densities with parameter of interest $\theta \in \mathbb{R}$ (see, e.g., [20] for discussion and references). Set $p_1(\cdot) = p(\cdot; \theta_1)$ and $p_2(\cdot) = p(\cdot; \theta_2)$. The family $p(x; \theta)$ is said to have *monotone likelihood ratio* if $x \mapsto p(x; \theta_2)/p(x; \theta_1)$ is nondecreasing as soon as $\theta_2 > \theta_1$ (and vice versa). If P_1 has p.d.f. p_1 and if P_2 has p.d.f. p_2, then under monotone likelihood ratio, $P_1 \leq P_2$. The property of monotone likelihood ratio is intrinsically linked with the validity of one-sided tests in statistics; see [16].

EXAMPLE 3.8 (Distances within the exponential family). A noteworthy class of parametric distributions which satisfy the property of monotone likelihood ratio is the *canonical regular exponential family* $p(x;\theta) = h(x)e^{\theta x - A(\theta)}$ for some scalar functions h and A, with the range of the distribution being independent of θ; see, for example, [16], page 639. If $\theta_1 > \theta_2$, then we can apply Corollary 3.5: $(\log \pi_0)'(x) = (\log \frac{p_2(x)}{p_1(x)})' = \theta_2 - \theta_1 < 0$ for all $x \in \mathbb{R}$, and thus from (3.12) we find with $X_2 \sim P_2$ that $d_{\mathcal{W}}(P_1, P_2) = |\theta_2 - \theta_1|\mathbb{E}[\tau_1(X_2)]$ under mild and easy-to-check conditions on P_1 and P_2.

EXAMPLE 3.9 (Distances between "tilted" distributions). Fix a density p_1 with mean μ_1 and consider, among all other densities g with same support and fixed but different mean $\mu_2 \neq \mu_1$, the density that minimizes the Kullback–Leibler divergence

$$\mathrm{KL}(g \parallel p_1) = \int g(x) \log\left(\frac{g(x)}{p_1(x)}\right) dx.$$

The Euler–Lagrange equation for the constrained variational problem is $\log g(x) = \log p_1(x) + \lambda_1 x + \lambda_2$, which is solved by

$$p_2(x) = p_1(x)\frac{e^{\lambda_1 x}}{M_1(\lambda_1)} \tag{3.15}$$

with $M_1(t) = \mathbb{E}[e^{tX_1}]$ the moment generating function of $X_1 \sim p_1$ and λ_1 a solution to

$$\frac{d}{dt}\big(\log M_1(t)\big)_{t=\lambda_1} = \mu_2$$

in order to guarantee $\mathbb{E}[X_2] = \mu_2$. We call (3.15) a "tilted" version of p_1 (following the classical notion of exponential tilting see, e.g., [9]). It is easy to compute

$$\mathrm{KL}(p_2 \parallel p_1) = \lambda_1\mu_2 - \log M_1(\lambda_1).$$

Setting $\pi_0(x) = e^{\lambda_1 x}/M_1(\lambda_1)$, we have $\log(\pi_0)'(x) = \lambda_1$, and hence, if the conditions of Theorem 3.1 are satisfied, then

$$d_{\mathcal{W}}(p_1, p_2) = |\lambda_1|\mathbb{E}\big[\tau_1(X_2)\big]. \tag{3.16}$$

For the sake of illustration, take p_1 the Gamma distribution on the positive half-line with density $p_1(x;\lambda,k) = \frac{1}{\Gamma(k)}e^{-x/\lambda}x^{k-1}\lambda^{-k}$. Then $M_1(t) = (1-\lambda t)^{-k}$ for $t < \frac{1}{\lambda}$ and $\lambda_1 = \frac{1}{\lambda} - \frac{k}{\mu_2}$. Moreover, $\tau_1(x) = \lambda x$. It is thus easy to check in this case that all conditions in Proposition 3.2 are satisfied. This allows us to deduce from (3.16) that

$$d_{\mathcal{W}}(p_1, p_2) = |\mu_2 - \lambda k|$$

which nicely complements $\mathrm{KL}(p_2 \parallel p_1) = \frac{\mu_2}{\lambda} - k + k\log(\frac{k\lambda}{\mu_2})$ as an alternative comparison statistic.

4. On the influence of the prior in Bayesian statistics. We now tackle the problem that motivated Theorem 3.1: assessing the impact of the choice of the prior distribution on the resulting posterior distribution in Bayesian statistics. In all examples, the conditions in Proposition 3.2 are easy to verify explicitly.

We first fix the notation. Assume that the observation x comes from a parametric model with p.d.f. $f(x;\theta)$ with $\theta \in \Theta$—$f(x;\theta)$ is often called the *likelihood* or the *sampling density*. We turn this model into a p.d.f. for θ through

$$p_1(\theta;x) = \kappa_1(x) f(x;\theta),$$

where $\kappa_1(x) = (\int f(x;\theta)\,d\theta)^{-1}$, and we assume that $\kappa_1 < \infty$. Let P_1 have p.d.f. p_1 and call its Stein kernel τ_1. Choose a possibly improper prior density $\pi_0(\theta)$, and let

$$p_2(\theta;x) = \pi_0(\theta;x)\, p_1(\theta;x),$$

where

$$\pi_0(\theta;x) = \kappa_2(x)\pi_0(\theta) \quad \text{such that} \quad \int p_2(\theta;x)\,d\theta = 1.$$

Then

$$1 = \int p_2(\theta;x)\,d\theta = \kappa_2(x) \int \pi_0(\theta) p_1(\theta;x)\,d\theta = \kappa_2(x)\mathbb{E}\big[\pi_0(\Theta_1)\big],$$

where Θ_1 has distribution P_1 which gives an expression for the normalizing constant. Let $P_2 = P_2(\cdot;x)$ be the probability distribution on Θ with p.d.f. $p_2(\cdot;x)$. Then P_2 is the posterior distribution of θ under the prior π_0 and the data x; moreover, P_1 can be seen as the distribution of θ under a uniform prior and the data x.

Now we extract from (3.5) of Theorem 3.1 the first bounds on the impact of a prior on the posterior distribution:

$$\frac{|\mathbb{E}[\tau_1(\Theta_1)\pi_0'(\Theta_1)]|}{\mathbb{E}[\pi_0(\Theta_1)]} \le d_{\mathcal{W}}(P_2, P_1) \le \frac{\mathbb{E}[\tau_1(\Theta_1)|\pi_0'(\Theta_1)|]}{\mathbb{E}[\pi_0(\Theta_1)]} \tag{4.1}$$

which can also be rewritten as

$$\begin{aligned} \big|\mathbb{E}[\Theta_2] - \mathbb{E}[\Theta_1]\big| = \big|\mathbb{E}\big[\tau_1(\Theta_2)\rho_0(\Theta_2)\big]\big| &\le d_{\mathcal{W}}(P_2, P_1) \\ &\le \mathbb{E}\big[\tau_1(\Theta_2)\big|\rho_0(\Theta_2)\big|\big] \end{aligned} \tag{4.2}$$

with $\Theta_2 \sim P_2$ and

$$\rho_0(\theta) = \frac{\pi_0'(\theta)}{\pi_0(\theta)},$$

the score function of $\pi_0(\theta;x)$ with respect to θ, which does not depend on the data x. As we shall see in the forthcoming subsections which treat some classical examples in Bayesian statistics, (4.2) often turns out to be handier for computations than (4.1).

4.1. *A normal model.* Consider the simple setting where $x = (x_1, \dots, x_n)$ is a random sample from a $\mathcal{N}(\theta, \sigma^2)$ population, where the scale σ is known and the location θ is the parameter of interest, and assume that the prior $\pi_0(\theta) > 0$ for all $\theta \in \Theta$ is differentiable. The likelihood $f(x;\theta)$ of the normal model can be factorized into

$$\begin{aligned} f(x;\theta) &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i-\theta)^2}{\sigma^2}\right\} \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right)\right\} \exp\left\{-\frac{1}{2}\frac{(\theta-\bar{x})^2}{\sigma^2/n}\right\} \\ &\propto \exp\left\{-\frac{1}{2}\frac{(\theta-\bar{x})^2}{\sigma^2/n}\right\} \qquad \text{when viewed as a function of } \theta, \end{aligned}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Thus, $P_1 = \mathcal{N}(\bar{x}, \sigma^2/n)$. Since τ_1 is constant, equal to σ^2/n, the variance of $\Theta_1 \sim P_1$, the bound (4.1) becomes

$$\frac{\sigma^2}{n}\frac{|\mathbb{E}[\pi_0'(\Theta_1)]|}{\mathbb{E}[\pi_0(\Theta_1)]} \le d_{\mathcal{W}}(P_2, P_1) \le \frac{\sigma^2}{n}\frac{\mathbb{E}[|\pi_0'(\Theta_1)|]}{\mathbb{E}[\pi_0(\Theta_1)]}$$

and (4.2) becomes

$$|\mathbb{E}[\Theta_2] - \bar{x}| = \frac{\sigma^2}{n}\left|\mathbb{E}\left[\rho_0(\Theta_2)\right]\right| \le d_{\mathcal{W}}(P_1, P_2) \le \frac{\sigma^2}{n}\mathbb{E}\left[\left|\rho_0(\Theta_2)\right|\right]. \tag{4.3}$$

Both inequalities are equalities in the case that π_0 is monotone.

4.2. *Normal prior and normal model.* Consider the same setting as in the previous section with the additional information that the prior π_0 is the density of a $\mathcal{N}(\mu, \delta^2)$, where μ and $\delta^2 > 0$ are known. Then the posterior P_2 is also normal, since

$$p_2(\theta; x) \propto \exp\left\{-\frac{1}{2}\left(\frac{(\theta-\bar{x})^2}{\sigma^2/n} + \frac{(\theta-\mu)^2}{\delta^2}\right)\right\}.$$

Defining $a = \frac{n}{\sigma^2} + \frac{1}{\delta^2}$ and $b(x) = \frac{\bar{x}}{\sigma^2/n} + \frac{\mu}{\delta^2}$, we see that $P_2 = \mathcal{N}(\frac{b(x)}{a}, \frac{1}{a})$.

Since the prior π_0 is not monotone, we cannot exactly evaluate the Wasserstein distance between P_1 and P_2. However, then we can write $\rho_0(\theta) = -(\theta-\mu)/\delta^2$ to obtain

$$\begin{aligned} \frac{\sigma^2}{n\delta^2+\sigma^2}|\bar{x}-\mu| &\le d_{\mathcal{W}}(P_1, P_2) \\ &\le \frac{\sigma^2}{n\delta^2+\sigma^2}|\bar{x}-\mu| + \frac{\sqrt{2}}{\sqrt{\pi}}\frac{\sigma^3}{n\delta\sqrt{\delta^2 n+\sigma^2}}. \end{aligned} \tag{4.4}$$

To see this, the lower bound follows directly from simplifying the difference of the expectations,

$$\left|\frac{b(x)}{a} - \bar{x}\right| = \frac{\sigma^2}{n\delta^2 + \sigma^2}|\bar{x} - \mu|.$$

For the upper bound, using $\rho_0(\theta) = -(\theta - \mu)/\delta^2$ in (4.3) gives

$$\begin{aligned}
d_{\mathcal{W}}(P_1, P_2) &\le \frac{\sigma^2}{n}\mathbb{E}\big[|\rho_0(\Theta_2)|\big] \\
&= \frac{\sigma^2}{n\delta^2}\mathbb{E}\big[|\Theta_2 - \mu|\big] \\
&\le \frac{\sigma^2}{n\delta^2}\left(\mathbb{E}\left[\left|\Theta_2 - \frac{b(x)}{a}\right|\right] + \left|\frac{b(x)}{a} - \mu\right|\right) \\
&= \frac{\sqrt{2}}{\sqrt{\pi}}\sqrt{\frac{1}{a}}\frac{\sigma^2}{n\delta^2} + \frac{\sigma^2}{n\delta^2}\left|\frac{b(x)}{a} - \mu\right| \\
&= \frac{\sqrt{2}}{\sqrt{\pi}}\frac{\delta\sigma}{\sqrt{\delta^2 n + \sigma^2}}\frac{\sigma^2}{n\delta^2} + \frac{\sigma^2}{n\delta^2}\frac{\delta^2}{\delta^2 + \frac{\sigma^2}{n}}|\bar{x} - \mu| \\
&= \frac{\sqrt{2}}{\sqrt{\pi}}\frac{\sigma^3}{n\delta\sqrt{\delta^2 n + \sigma^2}} + \frac{\sigma^2}{n\delta^2 + \sigma^2}|\bar{x} - \mu|,
\end{aligned}$$

which yields the upper bound in (4.4).

Inequality (4.4) provides a quite concrete and intuitive idea of the impact of the prior. First, we see that, for $n \to \infty$, the distance becomes zero, as is well known. The prior variance δ^2 has the same influence, which is also natural given that the prior then tends toward an improper prior, too. If the data are unfavorable so that $|\bar{x} - \mu|$ is large compared to n, then the Wasserstein distance between the two posterior distributions will be large. Due to the law of large numbers, for large n the probability that $|\bar{x} - \mu| > \delta^2 n + \sigma^2$ is small; but in contrast to such asymptotic considerations, the bound (4.4) makes the influence of the data on the distance explicit. Further, the upper and lower bounds only differ by an $O(n^{-3/2})$ term, hence at a $1/n$ precision, we have an exact expression for the Wasserstein distance. Finally, the $O(1/n)$ term in both bounds perfectly reflects the intuition that the better the guess of the prior mean μ (w.r.t. the data), the smaller the influence of the prior.

4.3. *The binomial model.* As the next example, we treat the case of n independent and identically distributed Bernoulli random variables with parameter of interest $\theta \in [0, 1]$; alternatively, we may say we have a single observation

$y \in \{0, 1, \dots, n\}$ from a Binomial distribution with known n and parameter of interest θ. The corresponding sampling density is

$$f(y;\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}$$

and $p_1(\theta; y) = \kappa_1(y)\theta^y(1-\theta)^{n-y}$ is a Beta density with

$$\kappa_1(y) = \frac{1}{B(y+1, n-y+1)},$$

where $B(\cdot,\cdot)$ denotes the Beta function, and $P_1 = P_1(\cdot; y) = \text{Beta}(y+1, n-y+1)$ is a Beta distribution.

Recall that, if $X \sim p(x) = \frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}$ then

$$\mathbb{E}[X] = \frac{\alpha}{\alpha+\beta}, \qquad \mathbb{E}\big[X^2\big] = \frac{\alpha(1+\alpha)}{(\alpha+\beta)(\alpha+\beta+1)}$$

and

$$\text{Var}[X] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

The Stein kernel is $\tau(x) = \frac{x(1-x)}{\alpha+\beta}$ and in particular $\tau_1(\theta) = \frac{\theta(1-\theta)}{n+2}$. Corollary 3.4 gives that, for any differentiable prior π_0 on $\mathcal{I} = [0, 1]$,

$$d_{\mathcal{W}}(P_1, P_2) \le \sup_{0\le\theta\le1} |\pi_0'(\theta)| \frac{(y+1)(n-y+1)}{(n+2)^2(n+3)}.$$

For y close to $\frac{n}{2}$, this bound is of order n^{-1}. In particular, for any $0 \le y \le n$, for a prior with bounded derivative, the Wasserstein distance converges to zero as $n \to \infty$ no matter which data are observed, but the data may affect the rate of convergence. Next, we consider some choices of prior densities which may not have bounded derivatives.

4.3.1. *Beta prior.* For a Beta prior,

$$\pi_0(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{4.5}$$

the assumptions of Theorem 3.1 are satisfied but $\sup_{0\le\theta\le1} |\pi_0'(\theta)|$ is infinite unless $\alpha, \beta \in \{1\} \cup [2, +\infty]$. Let P_1 denote the Beta$(y+1, n-y+1)$ distribution and P_2 the posterior distribution using the prior (4.5). It is well known that P_2 is again Beta distributed: the Beta distributions are *conjugate priors* for the Binomial distribution; in fact, it is easy to see that P_2 is the Beta$(\alpha+y, \beta+n-y)$ distribution.

We shall show that

$$\begin{aligned}\frac{1}{n+2}\left|(y+\alpha)\left(\frac{\alpha+\beta-2}{n+\alpha+\beta}\right)-(\alpha-1)\right| &\\ \le d_{\mathcal{W}}(P_1,P_2) &\\ \le \frac{1}{n+2}\left\{(y+\alpha)\frac{|\beta-\alpha|}{n+\alpha+\beta}+|\alpha-1|\right\}.&\end{aligned} \tag{4.6}$$

Note how both the upper and the lower bound vanish when $\alpha=\beta=1$. Also, unless $\alpha=1$, the upper bound is of order $O(n^{-1})$, no matter how favorable the data y are.

To this end, let $\Theta_1\sim P_1$ and $\Theta_2\sim P_2$. With (4.2) we have the immediate lower bound on the Wasserstein distance, namely

$$\begin{aligned}d_{\mathcal{W}}(P_1,P_2)&\ge\left|\mathbb{E}[\Theta_2]-\mathbb{E}[\Theta_1]\right|\\ &=\left|\frac{y+1}{n+2}-\frac{y+\alpha}{n+\alpha+\beta}\right|\\ &=\left|(y+\alpha)\left(\frac{1}{n+2}-\frac{1}{n+\alpha+\beta}\right)-\frac{\alpha-1}{n+2}\right|\end{aligned}$$

which leads to (4.6) after simplifications.

For the upper bound, we calculate that

$$\rho_0(\theta)=\frac{(\alpha-1)(1-\theta)+(\beta-1)\theta}{\theta(1-\theta)}$$

and hence

$$\tau_1(\theta)\rho_0(\theta)=\frac{1}{n+2}\big\{(\alpha-1)(1-\theta)+(\beta-1)\theta\big\}.$$

Using (4.2), we obtain the upper bound

$$\begin{aligned}d_{\mathcal{W}}(P_1,P_2)&\le\frac{1}{n+2}\mathbb{E}\big|(\alpha-1)(1-\Theta_2)-(\beta-1)\Theta_2\big|\\ &\le\frac{1}{n+2}\big\{|\alpha-1|+|\beta-\alpha|\mathbb{E}\Theta_2\big\}\\ &=\frac{1}{n+2}\left\{|\alpha-1|+\frac{y+\alpha}{n+\alpha+\beta}|\beta-\alpha|\right\}.\end{aligned}$$

4.3.2. *The Jeffreys prior.* An alternative popular prior is

$$\pi_0(\theta)=\frac{1}{\sqrt{\theta(1-\theta)}},$$

the so-called Jeffreys prior obtained for $\alpha = \beta = 1/2$ in (4.5). This is an improper prior which satisfies the assumptions of Theorem 3.1. The posterior distribution P_2 is Beta$(y + \frac{1}{2}, n - y + \frac{1}{2})$. Moreover,

$$\rho_0(\theta) = \frac{2\theta - 1}{2\theta(1-\theta)}$$

and

$$\tau_1(\theta)\rho_0(\theta) = \frac{1}{2(n+2)}(2\theta - 1).$$

Using (4.2), we obtain that

$$\frac{1}{(n+1)}\left|\frac{y+1}{n+2} - \frac{1}{2}\right| \le d_{\mathcal{W}}(P_1, P_2)$$

and

$$d_{\mathcal{W}}(P_1, P_2) \le \frac{1}{n+2}\left\{\sqrt{\frac{(y+\frac{1}{2})(n-y+\frac{1}{2})}{(n+2)(n+1)^2}} + \left|\frac{y+\frac{1}{2}}{n+1} - \frac{1}{2}\right|\right\}.$$

The upper bound follows from the Cauchy–Schwarz inequality via

$$\begin{aligned} d_{\mathcal{W}}(P_1, P_2) &\le \frac{1}{2(n+2)}\mathbb{E}\left|(2\Theta_2 - 1)\right| \\ &\le \frac{1}{n+2}\left\{\mathbb{E}\left|\Theta_2 - \mathbb{E}[\Theta_2]\right| + \left|\mathbb{E}[\Theta_2] - \frac{1}{2}\right|\right\} \\ &\le \frac{1}{n+2}\left\{\sqrt{\operatorname{Var}[\Theta_2]} + \left|\mathbb{E}[\Theta_2] - \frac{1}{2}\right|\right\} \\ &= \frac{1}{n+2}\left\{\sqrt{\frac{(y+\frac{1}{2})(n-y+\frac{1}{2})}{(n+2)(n+1)^2}} + \left|\frac{y+\frac{1}{2}}{n+1} - \frac{1}{2}\right|\right\}. \end{aligned}$$

In contrast to (4.6), the Jeffreys prior can achieve a bound of order $O(n^{-\frac{3}{2}})$ if the data y is close to $\frac{n}{2}$.

4.4. *A Poisson model.* The last case we tackle is the Poisson model with data $x = (x_1, \ldots, x_n)$ from a Poisson distribution with sampling density

$$f(x;\theta) = e^{-n\theta}\frac{\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}.$$

When $\sum_{i=1}^n x_i \neq 0$, which we shall now assume, then we obtain that P_1, the posterior distribution under a uniform prior, has p.d.f.

$$p_1(\theta; x) \propto \exp(-\theta n)\theta^{\sum_{i=1}^n x_i + 1 - 1}$$

a gamma density with parameters $1/n$ and $\sum_{i=1}^n x_i + 1$; its Stein kernel is simply $\tau_1(\theta) = \theta/n$ (see Example 3.9). The general bound (3.11) from Corollary 3.4 becomes

$$d_{\mathcal{W}}(P_1, P_2) \le \sup_{\theta \ge 0} \left|\pi_0'\Big(\theta; \sum x_i\Big)\right| \frac{\bar{x} + \frac{1}{n}}{n}, \tag{4.7}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i \ge \frac{1}{n}$.

Taking for θ a negative exponential prior Exp(λ) with $\lambda > 0$,

$$\pi_0(\theta) = \lambda e^{-\lambda\theta}$$

over $\mathbb{R}^+$ yields that the posterior P_2 has density $p_2(\theta; x) \propto \exp(-\theta(n+\lambda)) \times \theta^{\sum_{i=1}^n x_i + 1 - 1}$, again a gamma density where the first parameter is updated to $1/(n+\lambda)$. Here, the prior is monotone decreasing, hence we can exactly calculate the effect of the prior to obtain

$$\begin{aligned} d_{\mathcal{W}}(P_1, P_2) &= \mathbb{E}\left[\left|(\log \pi_0)'(\Theta_2)\right| \frac{\Theta_2}{n}\right] \\ &= \lambda \frac{\mathbb{E}[\Theta_2]}{n} \\ &= \lambda \frac{\bar{x} + \frac{1}{n}}{n+\lambda} \\ &= \frac{\lambda}{n+\lambda}\bar{x} + \frac{\lambda}{n(n+\lambda)}. \end{aligned}$$

We note that the exact distance differs from the general bound (4.7) here only through a multiplicative factor $\frac{n}{\lambda(n+\lambda)}$ [since $\sup_{\theta\ge 0} |\pi_0'(\theta; \sum x_i)| = \lambda^2$]. The distance increases with $\bar{x}$ but will always be at least as large as $\frac{\lambda}{n(n+\lambda)}$. As we assume that $\bar{x} \ge \frac{1}{n}$, the data-dependent part of the Wasserstein distance will always be at least as large as the part which stems solely from the prior. Finally, from the strong law of large numbers, $\bar{x}$ will almost surely converge to a constant as $n \to \infty$, so that the Wasserstein distance will converge to 0 almost surely.

Acknowledgments. The authors thank the three referees for their very careful reading of the paper and the many pertinent remarks and corrections they suggested. Christophe Ley and Yvik Swan also wish to thank Keble College for logistic support during an important writing stage of this paper.

REFERENCES

[1] AZZALINI, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* **12** 171–178. MR0808153

[2] CHEN, L. H. Y., GOLDSTEIN, L. and SHAO, Q.-M. (2011). *Normal Approximation by Stein's Method. Probability and Its Applications* (*New York*). Springer, Heidelberg. MR2732624

[3] CHWIALKOWSKI, K., STRATHMANN, H. and GRETTON, A. (2016). A kernel test of goodness of fit. Preprint. Available at arXiv:1602.02964v3.
[4] DIACONIS, P. and FREEDMAN, D. (1986). On inconsistent Bayes estimates of location. *Ann. Statist.* **14** 68–87. MR0829556
[5] DIACONIS, P. and FREEDMAN, D. (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14** 1–67. MR0829555
[6] DÖBLER, C. (2015). Stein's method of exchangeable pairs for the beta distribution and generalizations. *Electron. J. Probab.* **20** 109. MR3418541
[7] DÖBLER, C. (2015). Stein's method for the half-normal distribution with applications to limit theorems related to the simple symmetric random walk. *ALEA Lat. Am. J. Probab. Math. Stat.* **12** 171–191. MR3343481
[8] EDEN, R. and VÍQUEZ, J. (2015). Nourdin–Peccati analysis on Wiener and Wiener–Poisson space for general distributions. *Stochastic Process. Appl.* **125** 182–216. MR3274696
[9] EFRON, B. (1981). Nonparametric standard errors and confidence intervals. *Canad. J. Statist.* **9** 139–172. MR0640014
[10] EICHELSBACHER, P. and THÄLE, C. (2015). Malliavin–Stein method for variance-gamma approximation on Wiener space. *Electron. J. Probab.* **20** 123. MR3425543
[11] GAUNT, R. E. (2014). Variance-gamma approximation via Stein's method. *Electron. J. Probab.* **19** 38. MR3194737
[12] GOLDSTEIN, L. and REINERT, G. (2005). Distributional transformations, orthogonal polynomials, and Stein characterizations. *J. Theoret. Probab.* **18** 237–260. MR2132278
[13] GORHAM, J. and MACKEY, L. (2015). Measuring sample quality with Stein's method. *Adv. Neural Inf. Process. Syst.* 226–234.
[14] GORHAM, J. and MACKEY, L. (2016). Multivariate Stein factors for strongly log-concave distributions. *Electron. Commun. Probab.* **21**.
[15] HALLIN, M. and LEY, C. (2014). Skew-symmetric distributions and Fisher information: The double sin of the skew-normal. *Bernoulli* **20** 1432–1453. MR3217449
[16] KARLIN, S. and RUBIN, H. (1956). Distributions possessing a monotone likelihood ratio. *J. Amer. Statist. Assoc.* **51** 637–643. MR0104303
[17] LEY, C., REINERT, G. and SWAN, Y. (2016). Stein's method for comparison of univariate distributions. Preprint. Available at arXiv:1408.2998.
[18] LEY, C. and SWAN, Y. (2013). Local Pinsker inequalities via Stein's discrete density approach. *IEEE Trans. Inform. Theory* **59** 5584–5591. MR3096942
[19] LEY, C. and SWAN, Y. (2013). Stein's density approach and information inequalities. *Electron. Commun. Probab.* **18** 7. MR3019670
[20] LEY, C. and SWAN, Y. (2016). Parametric Stein operators and variance bounds. *Braz. J. Probab. Stat.* **30** 171–195. MR3481100
[21] NOURDIN, I. and PECCATI, G. (2012). *Normal Approximations with Malliavin Calculus: From Stein's Method to Universality*. *Cambridge Tracts in Mathematics* **192**. Cambridge Univ. Press, Cambridge. MR2962301
[22] NOURDIN, I., PECCATI, G. and SWAN, Y. (2014). Entropy and the fourth moment phenomenon. *J. Funct. Anal.* **266** 3170–3207. MR3158721
[23] NOURDIN, I., PECCATI, G. and SWAN, Y. (2014). Integration by parts and representation of information functionals. 2014 *IEEE International Symposium on Information Theory* (*ISIT*) 2217–2221.
[24] OATES, C. J., GIROLAMI, M. and CHOPIN, N. (2016). Control funtionals for Monte Carlo integration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* To appear. DOI:10.1111/rssb.12185.
[25] PIKE, J. and REN, H. (2014). Stein's method and the Laplace distribution. *ALEA Lat. Am. J. Probab. Math. Stat.* **11** 571–587. MR3283586
[26] ROSS, N. (2011). Fundamentals of Stein's method. *Probab. Surv.* **8** 210–293. MR2861132

[27] ROSS, S. M. (1996). *Stochastic Processes*, 2nd ed. *Wiley Series in Probability and Statistics*: *Probability and Statistics*. Wiley, New York. MR1373653

[28] SHAKED, M. and SHANTHIKUMAR, J. G. (2007). *Stochastic Orders*. *Springer Series in Statistics*. Springer, New York. MR2265633

[29] STEIN, C. (1965). Approximation of improper prior measures by prior probability measures. In *Proc. Internat. Res. Sem., Statist. Lab., Univ. California, Berkeley, Calif.*, 1963 217–240. Springer, New York. MR0199937

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

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

[32] VALLENDER, S. (1974). Calculation of the Wasserstein distance between probability distributions on the line. *Theory Probab. Appl.* **18** 784–786.

[33] VILLANI, C. (2009). *Optimal Transport*: *Old and New*. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Springer, Berlin. MR2459454

C. LEY
DEPARTMENT OF APPLIED MATHEMATICS,
COMPUTER SCIENCE AND STATISTICS
GHENT UNIVERSITY
KRIJGSLAAN 281, S9
9000 GHENT
BELGIUM
E-MAIL: christophe.ley@ugent.be

G. REINERT
DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
24–29 ST GILES
OXFORD OX1 3LB
UNITED KINGDOM
E-MAIL: reinert@stats.ox.ac.uk

Y. SWAN
DÉPARTEMENT DE MATHÉMATIQUE
UNIVERSITÉ DE LIÈGE
12 ALLÉE DE LA DÉCOUVERTE
BÂT. B37 PKG 33A
4000 LIÈGE
BELGIUM
E-MAIL: yswan@ulg.ac.be