Chapter 2
Fundamentals of Stein's Method

We begin by giving a detailed account of the fundamentals of Stein's method, starting with Stein's characterization of the normal distribution and the basic properties of the solution to the Stein equation. Then we provide an outline of the basic Stein identities and distributional transformations which play a large role in coupling constructions, introducing first the construction of the K function for independent random variables, the exchangeable pair approach due to Stein, the zero bias transformation for random variable with mean zero and variance one, and lastly the size bias transformation for non-negative random variables with finite mean. We conclude the chapter with a framework under which a number of Stein identities can be placed, and a proposition for normal approximation using Lipschitz functions. Some of the more technical results on bounds to the Stein equation can be found in the Appendix to this chapter.

2.1 Stein's Equation

Stein's method rests on the following characterization of the distribution of a standard normal variable Z, given in Stein (1972).

Lemma 2.1 *If W has a standard normal distribution, then*

$$Ef'(W) = E\big[Wf(W)\big], \tag{2.1}$$

for all absolutely continuous functions $f : \mathbb{R} \to \mathbb{R}$ with $E|f'(Z)| < \infty$. Conversely, if (2.1) holds for all bounded, continuous and piecewise continuously differentiable functions f with $E|f'(Z)| < \infty$, then W has a standard normal distribution.

Though there is no known definitive method for the construction of a characterizing identity, of the type given in Lemma 2.1, for the distribution of a random variable Y in general, two main contenders emerge. The first one we might call the 'density approach.' If W has density $p(w)$ then in many cases one can replace the coefficient W on the right hand side of (2.1) by $-p'(W)/p(W)$; this approach is pursued in

L.H.Y. Chen et al., *Normal Approximation by Stein's Method*,
Probability and Its Applications,
DOI 10.1007/978-3-642-15007-4_2,

Chap. 13 to study approximations by non-normal distributions. In another avenue, one which we might call the 'generator approach', we seek a Markov process that has as its stationary distribution the one of interest. In this case, the generator, or some variation thereof, of such a process has expectation zero when applied to sufficiently smooth functions, giving the difference between the two sides of (2.1). In Sect. 2.3.2 we discuss the relation between the generator method and exchangeable pairs, and in Sect. 2.2 its relation to the solution of the Stein equation, the differential equation motivated by the characterization (2.1). In fact, we now prove one direction of Lemma 2.1 using the Stein equation (2.2).

Lemma 2.2 *For fixed $z \in \mathbb{R}$ and $\Phi(z) = P(Z \le z)$, the cumulative distribution function of Z, the unique bounded solution $f(w) := f_z(w)$ of the equation*

$$f'(w) - wf(w) = \mathbf{1}_{\{w \le z\}} - \Phi(z) \tag{2.2}$$

is given by

$$f_z(w) = \begin{cases} \sqrt{2\pi} e^{w^2/2} \Phi(w)[1 - \Phi(z)] & \text{if } w \le z, \\ \sqrt{2\pi} e^{w^2/2} \Phi(z)[1 - \Phi(w)] & \text{if } w > z. \end{cases} \tag{2.3}$$

Proof Multiplying both sides of (2.2) by the integrating factor $e^{-w^2/2}$ yields

$$\big(e^{-w^2/2} f(w)\big)' = e^{-w^2/2}\big(\mathbf{1}_{\{w \le z\}} - \Phi(z)\big).$$

Integration now yields

$$\begin{aligned} f_z(w) &= e^{w^2/2} \int_{-\infty}^{w} \big[\mathbf{1}_{\{x \le z\}} - \Phi(z)\big] e^{-x^2/2}\, dx \\ &= -e^{w^2/2} \int_{w}^{\infty} \big[\mathbf{1}_{\{x \le z\}} - \Phi(z)\big] e^{-x^2/2}\, dx, \end{aligned}$$

which is equivalent to (2.3). Lemma 2.3 below shows $f_z(w)$ is bounded.

The general solution to (2.2) is given by $f_z(w)$ plus some constant multiple, say $ce^{w^2/2}$, of the solution to the homogeneous equation. Hence the only bounded solution is obtained by taking $c = 0$. □

Proof of Lemma 2.1 *Necessity.* Let f be an absolutely continuous function satisfying $E|f'(Z)| < \infty$. If W has a standard normal distribution then

$$\begin{aligned} Ef'(W) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(w) e^{-w^2/2}\, dw \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} f'(w) \left(\int_{-\infty}^{w} -x e^{-x^2/2}\, dx \right) dw \\ &\quad + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} f'(w) \left(\int_{w}^{\infty} x e^{-x^2/2}\, dx \right) dw. \end{aligned}$$

By Fubini's theorem, it thus follows that

$$\begin{aligned} Ef'(W) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} \left(\int_{x}^{0} f'(w)\,dw \right) (-x) e^{-x^2/2}\,dx \\ &\quad + \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} \left(\int_{0}^{x} f'(w)\,dw \right) x e^{-x^2/2}\,dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \big[f(x) - f(0) \big] x e^{-x^2/2}\,dx \\ &= E\big[Wf(W) \big]. \end{aligned}$$

Sufficiency. The function f_z as given in (2.3) is clearly continuous and piecewise continuously differentiable; Lemma 2.3 below shows f_z is bounded as well. Hence, if (2.1) holds for all bounded, continuous and continuously differentiable functions, then by (2.2)

$$0 = E\big[f_z'(W) - W f_z(W) \big] = E\big[\mathbf{1}_{\{W \le z\}} - \Phi(z) \big] = P(W \le z) - \Phi(z).$$

Thus W has a standard normal distribution. □

When f is an absolutely continuous and bounded function, one can prove (2.1) holds for a standard normal W using integration by parts, as in this case

$$\begin{aligned} E\big[Wf(W) \big] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} w f(w) e^{-w^2/2} dw \\ &= -\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(w) d\big(e^{-w^2/2} \big) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(w) e^{-w^2/2} dw \\ &= Ef'(W). \end{aligned}$$

For a given real valued measurable function h with $E|h(Z)| < \infty$ we denote $Eh(Z)$ by Nh and call

$$f'(w) - wf(w) = h(w) - Nh \tag{2.4}$$

the Stein equation for h, or simply the Stein equation. Note that (2.2) is the special case of (2.4) for $h(w) = \mathbf{1}_{\{w \le z\}}$. By the same method of integrating factors that produced (2.3) one may show that the unique bounded solution of (2.4) is given by

$$\begin{aligned} f_h(w) &= e^{w^2/2} \int_{-\infty}^{w} \big(h(x) - Nh \big) e^{-x^2/2}\,dx \\ &= -e^{w^2/2} \int_{w}^{\infty} \big(h(x) - Nh \big) e^{-x^2/2}\,dx. \end{aligned} \tag{2.5}$$

2.2 Properties of the Solutions

We now list some properties of the solutions (2.3) and (2.5) to the Stein equations (2.2) and (2.4), respectively, that are required to determine error bounds in our various approximations to come. We defer the detailed proofs of Lemmas 2.3 and 2.4

to an Appendix since they are somewhat technical. As the arguments used to prove these bounds do not themselves figure in the methods themselves, the reader may skip them if they so choose. We begin with the solution f_z to (2.2).

Lemma 2.3 *Let $z \in \mathbb{R}$ and let f_z be given by (2.3). Then*

$$wf_z(w) \text{ is an increasing function of } w. \tag{2.6}$$

Moreover, for all real w, u and v,

$$\big|wf_z(w)\big| \le 1, \qquad \big|wf_z(w) - uf_z(u)\big| \le 1 \tag{2.7}$$

$$\big|f_z'(w)\big| \le 1, \qquad \big|f_z'(w) - f_z'(u)\big| \le 1 \tag{2.8}$$

$$0 < f_z(w) \le \min\big(\sqrt{2\pi}/4, 1/|z|\big) \tag{2.9}$$

and

$$\big|(w+u)f_z(w+u) - (w+v)f_z(w+v)\big| \le \big(|w| + \sqrt{2\pi}/4\big)\big(|u| + |v|\big). \tag{2.10}$$

We mostly use (2.8) and (2.9) for our approximations. If one does not care much about constants, the bounds

$$\big|f_z'(w)\big| \le 2 \quad \text{and} \quad 0 < f_z(w) \le \sqrt{\pi/2}$$

may be easily obtained by using the well-known inequality

$$1 - \Phi(w) \le \min\left(\frac{1}{2}, \frac{1}{w\sqrt{2\pi}}\right) e^{-w^2/2}, \quad w > 0. \tag{2.11}$$

Next, we consider (2.5), the solution f_h to the Stein equation (2.4). For any real valued function h on $\mathbb{R}^p$ let

$$\|h\| = \sup_{x \in \mathbb{R}^p} \big|h(x)\big|.$$

Lemma 2.4 *For a given function $h : \mathbb{R} \to \mathbb{R}$, let f_h be the solution (2.5) to the Stein equation (2.4). If h is bounded, then*

$$\|f_h\| \le \sqrt{\pi/2}\big\|h(\cdot) - Nh\big\| \quad \textit{and} \quad \|f_h'\| \le 2\big\|h(\cdot) - Nh\big\|. \tag{2.12}$$

If h is absolutely continuous, then

$$\|f_h\| \le 2\|h'\|, \quad \|f_h'\| \le \sqrt{2/\pi}\,\|h'\| \quad \textit{and} \quad \|f_h''\| \le 2\|h'\|. \tag{2.13}$$

Some of the results that follow are shown by letting $h(w)$ be the indicator of $(-\infty, z]$ with a linear decay to zero over an interval of length $\alpha > 0$, that is, the function

$$h(w) = \begin{cases} 1 & w \le z, \\ 1 + (z - w)/\alpha & z < w \le z + \alpha, \\ 0 & w > z + \alpha. \end{cases} \tag{2.14}$$

The following bounds for the solution to the Stein equation for the smoothed indicator appear in Chen and Shao (2004).

Lemma 2.5 *For $z \in \mathbb{R}$ and $\alpha > 0$, let f be the solution (2.5) to the Stein equation (2.4) for the smoothed indicator function (2.14). Then, for all $w, v \in \mathbb{R}$,*

$$0 \le f(w) \le 1, \quad \big|f'(w)\big| \le 1, \quad \big|f'(w) - f'(v)\big| \le 1 \tag{2.15}$$

and

$$\big|f'(w+v) - f'(w)\big| \le |v|\left(1 + |w| + \frac{1}{\alpha}\int_0^1 \mathbf{1}_{[z,z+\alpha]}(w+rv)dr\right). \tag{2.16}$$

For multivariate approximations we consider an extension of the Stein equation (2.4) to $\mathbb{R}^p$. For a twice differentiable function $g : \mathbb{R}^p \to \mathbb{R}$ let ∇g and $D^2 g$ denote the gradient and second derivative, or Hessian matrix, of g respectively and let $\mathrm{Tr}(A)$ be the trace of a matrix A. Let $\mathbf{Z}$ be multivariate normal vector in $\mathbb{R}^p$ with mean zero and identity covariance matrix. For a test function $h : \mathbb{R}^p \to \mathbb{R}$ and for $u \ge 0$ define

$$(T_u h)(\mathbf{w}) = E\big\{h\big(\mathbf{w}e^{-u} + \sqrt{1 - e^{-2u}}\mathbf{Z}\big)\big\}. \tag{2.17}$$

Letting $Nh = Eh(\mathbf{Z})$, the following lemma provides bounds on the solution of the 'multivariate generator' method for solutions to the Stein equation

$$(\mathcal{A}g)(\mathbf{w}) = h(\mathbf{w}) - Nh \quad \text{where } (\mathcal{A}g)(\mathbf{w}) = \mathrm{Tr}\, D^2 g(\mathbf{w}) - \mathbf{w}\cdot\nabla g(\mathbf{w}). \tag{2.18}$$

We note that in one dimension (2.18) reduces to (2.4) with one extra derivative, that is, to

$$g''(w) - wg'(w) = h(w) - Nh. \tag{2.19}$$

For a vector $\mathbf{k} = (k_1, \ldots, k_p)$ of nonnegative integers and a function $h : \mathbb{R}^p \to \mathbb{R}$, let

$$h^{(\mathbf{k})}(\mathbf{w}) = \frac{\partial^{|\mathbf{k}|}}{\prod_{j=1}^p \partial w_{k_j}} h(\mathbf{w}) \quad \text{where } |\mathbf{k}| = \sum_{j=1}^p k_j,$$

and for a matrix $A \in \mathbb{R}^{p\times p}$, let

$$\|A\| = \max_{1\le i,j\le p} |a_{ij}|.$$

Lemma 2.6 *If $h : \mathbb{R}^p \to \mathbb{R}$ has three bounded derivatives then*

$$g(\mathbf{w}) = -\int_0^\infty \big[T_u h(\mathbf{w}) - Nh\big]\,du \tag{2.20}$$

solves (2.18), and if the $\mathbf{k}$th partial derivative of h exists then

$$\big\|g^{(\mathbf{k})}\big\| \le \frac{1}{k}\big\|h^{(\mathbf{k})}\big\|.$$

Further, for any $\boldsymbol{\mu} \in R^p$ and positive definite $p \times p$ matrix Σ, f defined by the change of variable

$$f(\mathbf{w}) = g\big(\Sigma^{-1/2}(\mathbf{w} - \boldsymbol{\mu})\big) \tag{2.21}$$

solves

$$\operatorname{Tr}\Sigma D^2 f(\mathbf{w}) - (\mathbf{w}-\boldsymbol{\mu})\cdot\nabla f(\mathbf{w}) = h\big(\Sigma^{-1/2}(\mathbf{w}-\boldsymbol{\mu})\big) - Nh, \tag{2.22}$$

and satisfies

$$\big\|f^{(\mathbf{k})}\big\| \le \frac{p^k}{k}\big\|\Sigma^{-1/2}\big\|^k\big\|h^{(\mathbf{k})}\big\|. \tag{2.23}$$

The operator $\mathcal{A}$ in (2.18) is the generator of the Ornstein–Uhlenbeck process in $\mathbb{R}^p$, whose stationary distribution is the standard normal. The operator $(T_u h)(\mathbf{w})$ in (2.17) is the expected value of h evaluated at the position of the Ornstein–Uhlenbeck process at time u, when it has initial position $\mathbf{w}$ at time 0. Equations of the form $\mathcal{A}g = h - Eh(\mathbf{Z})$ may be solved more generally by (2.20) when $\mathcal{A}$ is the generator of a Markov process with stationary distribution $\mathbf{Z}$, see Ethier and Kurtz (1986). Indeed, the generator method may be employed to solve the Stein equation for distributions other than the normal, see, for instance, Barbour et al. (1992) for the Poisson, and Luk (1994) for the Gamma distribution.

For the specific case at hand, the proof of Lemma 2.6 can be found in Barbour (1990), see equations (2.23) and (2.5), and also in Götze (1991). Essentially, following Barbour (1990) one shows that g is a solution, and that under the assumptions above, differentiating (2.20) and applying the dominated convergence yields

$$g^{(\mathbf{k})}(\mathbf{w}) = -\int_0^\infty e^{-ku} E\big\{h^{(\mathbf{k})}\big(\mathbf{w}e^{-u} + \sqrt{1-e^{-2u}}\mathbf{Z}\big)\big\}\,du.$$

The bounds then follow by straightforward calculations.

2.3 Construction of Stein Identities

Stein's equation (2.4) is the starting point of Stein's method. To prove that a mean zero, variance one random variable W can be approximated by a standard normal distribution, that is, to show that $Eh(W) - Eh(Z)$ is small for some large class of functions h, rather than estimating this difference directly, we solve (2.4) for a given h and show that $E[f'(W) - Wf(W)]$ is small instead. As we shall see, this latter quantity is often much easier to deal with than the former, as various identities and couplings may be applied to handle it.

In essence Stein's method shows that the distribution of two random variables are close by using the fact that they satisfy similar identities. For example, in Sect. 2.3.1, we demonstrate that when W is the sum of independent mean zero random variables $\xi_1, \ldots, \xi_n$ whose variances sum to 1, then

$$E\big[Wf(W)\big] = Ef'\big(W^{(I)} + \xi_I^*\big)$$

where $W^{(I)}$ is the sum W with a random summand ξ_I removed, and ξ_I^* is a random variable independent of $W^{(I)}$. Hence W satisfies an identity very much like the characterization (2.1) for the normal.

We present four different approaches, or variations, for handling the Stein equation. Sect. 2.3.1 introduces the K function method when W is a sum of independent random variables. In Sect. 2.3.2 we present the exchangeable pair approach of Stein, which works well when W has a certain dependency structure. We then discuss the zero bias distribution and the associated transformation, which, in principle, may be applied for arbitrary mean zero random variables having finite variance. We note that the K function method of Sect. 2.3.1 and the zero bias method of Sect. 2.3.3 are essentially identical in the simple context of sums of independent random variables, but these approaches will later diverge. Size bias transformations, and some associated couplings, presented in Sect. 2.3.3 are closely related to those for zero biasing; the size bias method is most naturally applied to non-negative variables such as counts.

2.3.1 Sums of Independent Random Variables

In this subsection we consider the most elementary case and apply Stein's method to justify the normal approximation of the sum W of independent random variables $\xi_1, \xi_2, \ldots, \xi_n$ satisfying

$$E\xi_i = 0, \quad 1 \le i \le n \quad \text{and} \quad \sum_{i=1}^{n} E\xi_i^2 = 1.$$

Set

$$W = \sum_{i=1}^{n} \xi_i \quad \text{and} \quad W^{(i)} = W - \xi_i,$$

and define

$$K_i(t) = E\big\{\xi_i(\mathbf{1}_{\{0\le t\le \xi_i\}} - \mathbf{1}_{\{\xi_i\le t<0\}})\big\}. \tag{2.24}$$

It is easy to check that $K_i(t) \ge 0$ for all real t, and that

$$\int_{-\infty}^{\infty} K_i(t)\,dt = E\xi_i^2 \quad \text{and} \quad \int_{-\infty}^{\infty} |t| K_i(t)\,dt = \frac{1}{2}E|\xi_i|^3. \tag{2.25}$$

Let h be a measurable function with $E|h(Z)| < \infty$, and let $f = f_h$ be the corresponding solution of the Stein equation (2.4). Our goal is to estimate

$$Eh(W) - Nh = E\big\{f'(W) - Wf(W)\big\}. \tag{2.26}$$

The argument below is fundamental to the K function approach, with many of the following tricks reappearing repeatedly in the sequel.

Since ξ_i and $W^{(i)}$ are independent for each $1 \le i \le n$, we have

$$E\big[Wf(W)\big] = \sum_{i=1}^{n} E\big[\xi_i f(W)\big]$$
$$= \sum_{i=1}^{n} E\big\{\xi_i\big[f(W) - f\big(W^{(i)}\big)\big]\big\},$$

where the last equality follows because $E\xi_i = 0$. Writing the final difference in integral form, we thus have

$$E\big[Wf(W)\big] = \sum_{i=1}^{n} E\left\{\xi_i \int_0^{\xi_i} f'\big(W^{(i)} + t\big)\, dt\right\}$$
$$= \sum_{i=1}^{n} E\left\{\int_{-\infty}^{\infty} f'\big(W^{(i)} + t\big)\xi_i\big(\mathbf{1}_{\{0\le t\le \xi_i\}} - \mathbf{1}_{\{\xi_i \le t<0\}}\big)\, dt\right\}$$
$$= \sum_{i=1}^{n} \int_{-\infty}^{\infty} E\big\{f'\big(W^{(i)} + t\big)\big\} K_i(t)\, dt, \tag{2.27}$$

from the definition of K_i and again using independence. However, from

$$\sum_{i=1}^{n} \int_{-\infty}^{\infty} K_i(t)\, dt = \sum_{i=1}^{n} E\xi_i^2 = 1, \tag{2.28}$$

it follows that

$$Ef'(W) = \sum_{i=1}^{n} \int_{-\infty}^{\infty} E\big\{f'(W)\big\} K_i(t)\, dt. \tag{2.29}$$

Thus, by (2.27) and (2.29),

$$E\big\{f'(W) - Wf(W)\big\} = \sum_{i=1}^{n} \int_{-\infty}^{\infty} E\big\{f'(W) - f'\big(W^{(i)} + t\big)\big\} K_i(t)\, dt. \tag{2.30}$$

Since $K_i(t)$ is non-negative and $\int_{-\infty}^{\infty} K_i(t)\, dt = E\xi_i^2$, the ratio $K_i(t)/E\xi_i^2$ can be regarded as a probability density function. Let ξ_i^*, $i = 1, \ldots, n$ be independent random variables, independent of ξ_j for $j \neq i$, having density function $K_i(t)/E\xi_i^2$ for each i. Let I be a random index, independent of $\{\xi_i,\ \xi_i^*,\ i = 1, \ldots, n\}$ with distribution

$$P(I = i) = E\xi_i^2.$$

Then we may rewrite (2.27) as

$$E\big[Wf(W)\big] = Ef'\big(W^{(I)} + \xi_I^*\big) \tag{2.31}$$

and (2.30) as

$$E\big\{f'(W) - Wf(W)\big\} = E\big\{f'(W) - f'\big(W^{(I)} + \xi_I^*\big)\big\}.$$

Equations (2.27), and (2.30) play a key role in proving good normal approximations. Note in particular that (2.30) is an *equality*, and that (2.27) and (2.30) hold for all bounded absolutely continuous functions f. It is easy to see that bounds on the solution f such as those furnished by Lemma 2.4 can now come into play to bound the expected difference in (2.30), and therefore the left hand side of (2.26).

2.3.2 Exchangeable Pairs

Suppose now that W is an arbitrary random variable, in particular, not necessarily a sum. A number of variations of Stein's method introduce an auxiliary random variable coupled to W possessing certain properties. In the exchangeable pair approach, (see Stein 1986) one constructs W' on the same probability space as W in such a way that (W, W') is an exchangeable pair, that is, such that $(W, W') =_d (W', W)$, where $=_d$ signifies equality in distribution. The exchangeable pair approach makes essential use of the elementary fact that, if (W, W') is an exchangeable pair, then

$$Eg(W, W') = 0 \tag{2.32}$$

for all antisymmetric measurable functions $g(x, y)$ such that the expected value above exists.

The key identities applied in the exchangeable pair approach are given in Lemma 2.7, for which we require the following definition.

Definition 2.1 If the pair (W, W') is exchangeable and satisfies the 'linear regression condition'

$$E(W'|W) = (1 - \lambda)W \tag{2.33}$$

with $\lambda \in (0, 1)$, then we call (W, W') a λ-Stein pair, or more simply, a Stein pair.

One heuristic explanation of why property (2.33) should be of any importance in normal approximation is that it is parallel to the conditional expectation property enjoyed by the bivariate normal distribution. That is, if Z, Z' have the bivariate normal distribution then the conditional expectation of Z' given Z is linear, specifically

$$E(Z'|Z) = \mu_1 + \sigma_1 \rho \left(\frac{Z - \mu_2}{\sigma_2} \right),$$

where σ_1^2 and σ_2^2 are the variances of Z' and Z, respectively, and ρ is the correlation coefficient. Hence, when Z and Z' have mean zero and equal variance, we obtain (2.33),

$$E(Z'|Z) = (1 - \lambda)Z,$$

with $\lambda = 1 - \rho$.

Lemma 2.7 *Let (W, W') be a Stein pair and $\Delta = W - W'$. Then*

$$EW = 0 \quad \text{and} \quad E\Delta^2 = 2\lambda EW^2 \quad \text{if } EW^2 < \infty. \tag{2.34}$$

Furthermore, when $EW^2 < \infty$, for every absolutely continuous function f satisfying $|f(w)| \le C(1 + |w|)$, we have

$$E\big[Wf(W)\big] = \frac{1}{2\lambda} E\big\{(W - W')\big(f(W) - f(W')\big)\big\}, \tag{2.35}$$

$$E\big[Wf(W)\big] = E\left\{\int_{-\infty}^{\infty} f'(W + t)\hat{K}(t)\,dt\right\}, \tag{2.36}$$

and

$$\begin{aligned} &E\big[f'(W) - EWf(W)\big] \\ &\quad = Ef'(W)\left(1 - \frac{\Delta^2}{2\lambda}\right) + E\int_{-\infty}^{\infty} \big(f'(W) - f'(W + t)\big)\hat{K}(t)\,dt, \end{aligned} \tag{2.37}$$

where

$$\hat{K}(t) = \frac{\Delta}{2\lambda}(\mathbf{1}_{\{-\Delta \le t \le 0\}} - \mathbf{1}_{\{0 < t \le -\Delta\}}) \tag{2.38}$$

satisfies

$$\int_{-\infty}^{\infty} \hat{K}(t)\,dt = \frac{\Delta^2}{2\lambda}. \tag{2.39}$$

Proof Taking expectation in (2.33) yields, by exchangeability, $EW = EW' = (1 - \lambda)EW$ so $EW = 0$. Furthermore, as

$$EW'W = E\big(E(W'W|W)\big) = E\big(WE(W'|W)\big) = (1 - \lambda)EW^2,$$

we have

$$E(W' - W)^2 = 2EW^2 - 2EW'W = 2\lambda EW^2.$$

Next we exploit (2.32) with the antisymmetric function $g(x, y) = (x - y)(f(y) + f(x))$, for which $Eg(W, W')$ exists, because of the growth assumption on f. Identity (2.32) yields

$$\begin{aligned} 0 &= E\big\{(W - W')\big(f(W') + f(W)\big)\big\} \\ &= E\big\{(W - W')\big(f(W') - f(W)\big)\big\} + 2E\big\{f(W)(W - W')\big\} \\ &= E\big\{(W - W')\big(f(W') - f(W)\big)\big\} + 2E\big\{f(W)E(W - W'|W)\big\} \\ &= E\big\{(W - W')\big(f(W') - f(W)\big)\big\} + 2\lambda E\big\{Wf(W)\big\}, \end{aligned}$$

this last by (2.33). Rearranging this equality yields (2.35), and now

$$\begin{aligned} E\big[Wf(W)\big] &= \frac{1}{2\lambda} E\big\{(W - W')\big(f(W) - f(W')\big)\big\} \\ &= \frac{1}{2\lambda} E\big\{\Delta\big(f(W) - f(W - \Delta)\big)\big\} \\ &= \frac{1}{2\lambda} E \int_{-\Delta}^{0} \Delta f'(W + t)\, dt \\ &= E \int_{-\infty}^{\infty} f'(W + t)\hat{K}(t)\, dt. \end{aligned} \tag{2.40}$$

This proves (2.36).

Now note that integrating (2.38) yields (2.39) and so to prove (2.37), we need only observe that

$$Ef'(W) = E\left\{ f'(W)\left(1 - \frac{\Delta^2}{2\lambda}\right)\right\} + E\left\{\int_{-\infty}^{\infty} f'(W)\hat{K}(t)\, dt\right\}$$

and subtract using (2.40). □

As the linear regression condition (2.33) may at times be too restrictive, it can be replaced by

$$E(W - W' \mid W) = \lambda(W - R), \tag{2.41}$$

where R is a random variable of small order. Following the proof of (2.36), if W' and W are mean zero exchangeable random variables with finite second moments, and (2.41) holds for some $\lambda \in (0, 1)$ and random variable R, then

$$E\big[Wf(W)\big] = E \int_{-\infty}^{\infty} f'(W + t)\hat{K}(t)\, dt + E\big[Rf(W)\big], \tag{2.42}$$

with $\hat{K}(t)$ given by (2.38).

We present three examples that give the flavor of the construction of exchangeable pairs; sometimes we will denote the pair by (W', W''), instead of by (W, W').

Example 2.1 (Independent random variables) Let $\{\xi_i, 1 \le i \le n\}$ be independent random variables with zero means and $\sum_{i=1}^{n} E\xi_i^2 = 1$, and put $W = \sum_{i=1}^{n} \xi_i$. Let $\{\xi_i',\ i = 1, \ldots, n\}$ be an independent copy of $\{\xi_i,\ i = 1, \ldots, n\}$, and let I have uniform distribution on $\{1, 2, \ldots, n\}$, independent of $\{\xi_i, \xi_i',\ i = 1, \ldots, n\}$. Define $W' = W - \xi_I + \xi_I'$. Then (W, W') is an exchangeable pair, and it is easy to verify

$$E(W' \mid W) = \left(1 - \frac{1}{n}\right) W,$$

so that (2.33) is satisfied with $\lambda = 1/n$.

The exchangeable pair above is a special case of the following general construction.

Example 2.2 (Exchangeable pair by substitution) Let $W = g(\xi_1, \ldots, \xi_n)$, and ξ_i' have the conditional distribution of ξ_i given ξ_j, $1 \le j \ne i \le n$. Let I be a random index uniformly distribution over $\{1, \ldots, n\}$, independent of $\{\xi_i, \xi_i', \ i = 1, \ldots, n\}$. Define $W' = g(\xi_1, \ldots, \xi_{I-1}, \xi_I', \xi_{I+1}, \ldots, \xi_n)$. That is, in the definition of W, the ξ_I is replaced by ξ_I' while the other variables remain the same. Then (W, W') is an exchangeable pair. We note that unlike Example 2.1 the linearly condition (2.33) is not automatically satisfied.

Example 2.3 (Combinatorial Central Limit Theorem) For a given array $\{a_{ij}\}_{1 \le i,j \le n}$ of real numbers and $\pi = \pi'$ a random permutation, let

$$Y' = \sum_{i=1}^{n} a_{i,\pi'(i)}. \tag{2.43}$$

Classically, π' is taken to be uniformly distributed over the symmetric group $\mathcal{S}_n$; we specialize to that case here, and study it in Sects. 4.4 and 6.1.1, but also consider alternative permutation distributions in Sect. 6.1.2.

Let

$$a_{\cdot\cdot} = \frac{1}{n^2} \sum_{i,j=1}^{n} a_{ij}, \quad a_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} a_{ij} \quad \text{and} \quad a_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} a_{ij}. \tag{2.44}$$

Using that π' is uniform one easily obtains $EY' = na_{\cdot\cdot} = \sum_i a_{i\cdot} = \sum_i a_{\cdot\pi(i)}$, and therefore

$$Y' - EY' = \sum_{i=1}^{n} (a_{i,\pi(i)} - a_{\cdot\cdot}) = \sum_{i=1}^{n} (a_{i,\pi(i)} - a_{i\cdot} - a_{\cdot\pi(i)} + a_{\cdot\cdot}). \tag{2.45}$$

As our goal is to derive bounds to the normal for the standardized variable $(Y' - EY')/\sqrt{\mathrm{Var}(Y')}$, without loss of generality we may replace a_{ij} by $a_{ij} - a_{i\cdot} - a_{\cdot j} + a_{\cdot\cdot}$, and assume

$$a_{i\cdot} = a_{\cdot j} = a_{\cdot\cdot} = 0. \tag{2.46}$$

Let τ_{ij} be the permutation that transposes i and j, $\pi'' = \pi'\tau_{ij}$ and Y'' be given by (2.43) with π' replaced by π''. Since $\pi''(k) = \pi'(k)$ for $k \notin \{i, j\}$ while $\pi''(i) = \pi'(j)$ and $\pi''(j) = \pi'(i)$, we have

$$Y'' - Y' = b\big(i, j, \pi(i), \pi(j)\big), \tag{2.47}$$

where $b(i, j, k, l) = a_{il} + a_{jk} - (a_{ik} + a_{jl})$. Taking (I, J) to be independent of π', with the uniform distribution over all pairs satisfying $1 \le I \ne J \le n$, the permutations π' and $\pi'' = \tau_{IJ}\pi'$ are exchangeable, and hence so are Y' and Y''.

To prove that the linear regression property (2.33) is satisfied, write

$$Y'' - Y' = (a_{I,\pi'(J)} + a_{J,\pi'(I)}) - (a_{I,\pi'(I)} + a_{J,\pi'(J)}). \tag{2.48}$$

Taking the conditional expectation given π', using (2.46), we obtain

$$E(Y'' - Y'|\pi') = 2\left(-\frac{1}{n}\sum_{i=1}^{n} a_{i,\pi'(i)} + \frac{1}{n(n-1)}\sum_{i\neq j} a_{i,\pi'(j)}\right)$$
$$= -2\left(\frac{1}{n}\sum_{i=1}^{n} a_{i,\pi'(i)} + \frac{1}{n(n-1)}\sum_{i=1}^{n} a_{i,\pi'(i)}\right) = -\frac{2}{n-1}Y'.$$

As the right hand side is measurable with respect to Y', we conclude that

$$E(Y''|Y') = \left(1 - \frac{2}{n-1}\right)Y',$$

demonstrating that Y', Y'' is a $2/(n-1)$-Stein pair.

One particular special case of note is when $a_{ij} = b_i c_j$ where $b_1, \ldots, b_n$ are any real numbers and the values $c_j \in \{0, 1\}$, $j = 1, \ldots, n$ satisfy

$$\sum_{i=1}^{n} c_j = m.$$

In this case, as any set of m values from $\{b_1, \ldots, b_n\}$ are as likely to be summed to yield Y' as any other set of that same size, Y' is the sum of a simple random sample of size m from a population whose numerical characteristics are given by $\{b_i,\ i = 1, \ldots, n\}$.

It is worth mentioning a connection between the exchangeable pair and the generator approach which gave the solutions and bounds to the Stein equation in Lemma 2.6. To see the connection, let (W, W') be a λ-Stein pair and rewrite

$$E(W'|W) = (1-\lambda)W \quad \text{as } E(W' - W|W) = -\lambda W.$$

If one can construct a sequence $W_1, W_2, \ldots$ such that

$$(W_t, W_{t+1}) =_d (W, W'), \quad \text{for } t = 1, 2, \ldots,$$

then $E(W_{t+1} - W_t|W_t) = -\lambda W_t$, and so, with $\Delta W_t = W_{t+1} - W_t$ we have

$$\Delta W_t = -\lambda W_t + \epsilon_t \quad \text{where } E[\epsilon_t|W_t] = 0,$$

a recursion reminiscent of the stochastic differential equation for the Ornstein–Uhlenbeck process,

$$dW_t = -\lambda W_t + \sigma dB_t$$

where B_t is a Brownian motion.

It is sometimes possible to produce the sequence $W_1, W_2, \ldots$ as the successive states of a reversible Markov chain in stationarity. Or, looking at this construction in another way, for a given W of interest, one may be able to create a Stein pair by constructing a reversible Markov chain with stationary distribution W. As an illustration, consider the sum Y of a simple random sample $S = \{X_1, \ldots, X_n\}$ of size n of N population characteristics $\mathcal{A} = \{a_1, \ldots, a_N\}$ which have been centered to satisfy

$$\sum_{i=1}^{N} a_i = 0. \tag{2.49}$$

Given a simple random sample S_0, one may construct a Markov chain $S_0, S_1, \ldots$, whose state space consists of all size n subsets of $\mathcal{A}$, by interchanging at time step n a randomly chosen element of S_n with one from the complement of S_n to form S_{n+1}. The chain is in equilibrium and is reversible, hence the sets $S_n, n = 0, 1, \ldots$ are identically distributed, and (S_n, S_{n+1}) is exchangeable. In particular, the sums Y_n and Y_{n+1} of S_n and S_{n+1} respectively, are exchangeable and have the same distribution as Y. This construction is, essentially, the one used in Theorem 4.10, and it is shown there that the linearity condition (2.33) holds under the centering (2.49). This method for the construction of exchangeable pairs features prominently in the analysis of the anti-voter model in Sect. 6.4.

2.3.3 Zero Bias

Stein's characterization (2.1) of the standard normal Z can be easily extended to the mean zero normal family in general. In particular, a simple change of variable in Lemma 2.1 shows that X is $\mathcal{N}(0, \sigma^2)$ if and only if

$$\sigma^2 Ef'(X) = E\big[Xf(X)\big] \tag{2.50}$$

for all absolutely continuous functions for which these expectations exist. Though the left and right hand sides of (2.50) will only be equal at the normal, one can create an identity in the same spirit that holds more generally. In particular, as introduced in Goldstein and Reinert (1997), given X with mean zero and variance σ^2, we say that X^* has the X-zero bias distribution if

$$\sigma^2 Ef'\big(X^*\big) = E\big[Xf(X)\big] \tag{2.51}$$

for all absolutely continuous functions f for which these expectations exist. It is convenient to regard (2.51) as giving rise to a transformation mapping the distribution of X to that of X^*. Indeed, the characterization in Lemma 2.1 can be restated as saying that the normal distribution is the unique fixed point of the zero bias transformation.

It is the uniqueness of the fixed point of the zero bias transformation, that is, the fact that X^* has the same distribution as X only when X is normal, that provides a probabilistic reason for a normal approximation to hold. If the distribution of a random variable X gets mapped to an X^* which is close in distribution to X, then X is close to the zero bias transformation's unique fixed point, that is, close to the normal.

This same reasoning indicates that not only should a normal approximation be justified whenever the distribution of X is close to that of X^*, but that the quality of the approximation can be measured in terms of their distance. Though this claim will later be made precise in a number of ways, for now one can see how it might

be formalized by observing that a coupling of a mean zero, variance one W to such a W^* can be used in the Stein equation (2.4) as

$$Eh(W) - Nh = E\big[f'(W) - Wf(W)\big] = E\big[f'(W) - f'(W^*)\big].$$

Hence, when W and W^* are close, the right hand side, and so also the left hand side, will be small.

While the zero bias transformation fixes the mean zero normal, for non-normal distributions, in some sense, the transformation moves them closer to normality. For example, let $\xi \in \{0, 1\}$ be a Bernoulli random variable with success probability $p \in (0, 1)$. Centering ξ to form the mean zero discrete random variable $X = \xi - p$ having variance $\sigma^2 = p(1-p)$, substitution into the right hand side of (2.51) yields

$$\begin{aligned} E\big[Xf(X)\big] &= E\big[(\xi - p)f(\xi - p)\big] \\ &= p(1-p)f(1-p) - (1-p)pf(-p) \\ &= \sigma^2\big[f(1-p) - f(-p)\big] \\ &= \sigma^2 \int_{-p}^{1-p} f'(u)\,du \\ &= \sigma^2 Ef'(U), \end{aligned}$$

for U uniformly distributed over $[-p, 1-p]$. Hence, with $=_d$ indicating the equality of two random variables in distribution, and $\mathcal{U}[a, b]$ denoting the uniform distribution on the finite interval $[a, b]$,

$$(\xi - p)^* =_d U \quad \text{where } U \sim \mathcal{U}[-p, 1-p]. \tag{2.52}$$

As hinted at by the Bernoulli example, the following lemma shows that the zero bias distribution exists and is absolutely continuous for every X having mean zero and some finite, positive variance.

Proposition 2.1 *Let X be a random variable with mean zero and finite positive variance σ^2. Then there exists a unique distribution for X^* such that*

$$Ef'(X^*) = \sigma^2 E\big[Xf(X)\big] \tag{2.53}$$

for every absolutely continuous function f for which $E|Xf(X)| < \infty$.

Moreover, the distribution of X^* is absolutely continuous with density

$$p^*(x) = E\big[X\mathbf{1}(X > x)\big]/\sigma^2 = -E\big[X\mathbf{1}(X \le x)\big]/\sigma^2 \tag{2.54}$$

and distribution function

$$G^*(x) = E\big[X(X - x)\mathbf{1}(X \le x)\big]/\sigma^2. \tag{2.55}$$

Proof We prove the claims assuming $\sigma^2 = 1$, the extension to the general case being straightforward. First, regarding (2.54), we note that the second equality holds since $EX = 0$. It follows that $p^*(x)$ is nonnegative, using the first form for $x \ge 0$, and the second for $x < 0$.

To prove that we may write $E[Xf(X)]$ as the expectation on the left hand side of (2.53), in terms of an absolutely continuous variable X^* with density $p^*(x)$, let $f(x) = \int_0^x g$ with g a nonnegative function which is integrable on compact domains. Then by Fubini's theorem,

$$\begin{aligned}\int_0^\infty f'(u)E\big[X\mathbf{1}(X > u)\big]\,du &= \int_0^\infty g(u)E\big[X\mathbf{1}(X > u)\big]\,du \\ &= E\left(X\int_0^\infty g(u)\mathbf{1}(X > u)\,du\right) \\ &= E\left(X\int_0^{X\vee 0} g(u)\,du\right) \\ &= E\big[Xf(X)\mathbf{1}(X \geq 0)\big].\end{aligned}$$

A similar argument over $(-\infty, 0]$ yields

$$\int_{-\infty}^\infty f'(u)E\big[X\mathbf{1}(X > u)\big]\,du = E\big[Xf(X)\big], \tag{2.56}$$

where both sides may be $+\infty$. If $f(x) = \int_0^x g$ with $E|Xf(X)| < \infty$, then taking the difference of the contributions from the positive and negative parts of g shows that (2.56) continues to hold over this larger class of functions, as it does for f satisfying the conditions of the theorem by writing $f(x) = \int_0^x g + f(0)$ and using that the mean of X is zero. Taking $f(x) = x$ shows that $p^*(x)$ integrates to one and is therefore a density, whence the left hand side of (2.56) may be written as $Ef'(X^*)$ for X^* with density $p^*(x)$. The distribution of X^* is clearly unique, as $Ef'(X^*) = Ef'(Y^*)$ for all, say, continuously differentiable functions f with compact support, implies $X^* =_d Y^*$.

Integrating the density p^* to obtain the distribution function G^*, we have

$$\begin{aligned}G^*(x) &= -E\left(X\int_{-\infty}^x \mathbf{1}(X \leq u)\,du\right) \\ &= -E\left(X\int_X^x du\ \mathbf{1}(X \leq x)\right) \\ &= E\big[X(X - x)\mathbf{1}(X \leq x)\big].\end{aligned}$$

□

The characterization (2.51) also specifies a relationship between the moments of X and X^*. One of the most useful of these relations is the one which results from applying (2.51) with $f(x) = (1/2)x^2\,\mathrm{sgn}(x)$, for which $f'(x) = |x|$, yielding

$$\sigma^2 E|X^*| = \frac{1}{2}E|X|^3 \quad \text{where } \sigma^2 = \mathrm{Var}(X). \tag{2.57}$$

In particular, we see that $E|X|^3 < \infty$ if and only if $E|X^*| < \infty$.

We have observed that the zero bias distribution of a mean zero Bernoulli variable with support $\{-p, 1-p\}$ is uniform on $[-p, 1-p]$, and it is easy to see from (2.54) that, more generally, if x is such that $P(X > x) = 0$, then the same holds for all $y > x$, and $p^*(y) = 0$ for all such y, while if x is such that $P(X > x) > 0$ then

$p^*(x) > 0$. As similar statements hold when considering x for which $P(X \le x) = 0$, letting support(X) be the support of the distribution of X, if

$$a = \inf \text{support}(X) \quad \text{and} \quad b = \sup \text{support}(X)$$

are finite then support$(X^*) = [a, b]$. One can verify that the support continues to be given by this relation, with any closed endpoint replaced by the corresponding open one, when any of the values of a or b are infinite. One consequence of this fact is that if X is bounded by some constant then X^* is also bounded by the same constant, that is,

$$|X| \le C \quad \text{implies} \quad |X^*| \le C. \tag{2.58}$$

The zero bias transformation enjoys the following scaling, or linearity property. If X is a mean zero random variable with finite variance, and X^* has the X-zero biased distribution, then for all $a \neq 0$

$$(aX)^* =_d aX^*. \tag{2.59}$$

The verification of this claim follows directly from (2.51), as letting $\sigma^2 = \text{Var}(X)$ and $g(x) = f(ax)$, we find

$$\begin{aligned}(a\sigma)^2 Ef'(aX^*) &= a\sigma^2 Eg'(X^*) \\ &= aE\big[Xg(X)\big] \\ &= E\big[(aX)f(aX)\big] \\ &= (a\sigma)^2 Ef'\big((aX)^*\big).\end{aligned}$$

But by far the most important properties of the zero bias transformation are those like the ones given in the following lemma.

Lemma 2.8 *Let ξ_i, $i = 1, \ldots, n$ be independent mean zero random variables with $\text{Var}(\xi_i) = \sigma_i^2$ summing to 1. Let ξ_i^* have the ξ_i-zero bias distribution with ξ_i^*, $i = 1, \ldots, n$ mutually independent, and ξ_i^* independent of ξ_i for all $j \neq i$. Further, let I be a random index, independent of ξ_i, ξ_i^*, $i = 1, \ldots, n$ with distribution*

$$P(I = i) = \sigma_i^2. \tag{2.60}$$

Then

$$W^* =_d W - \xi_I + \xi_I^*, \tag{2.61}$$

where W^* has the W-zero bias distribution.

In other words, upon replacing the variable ξ_I by ξ_I^* in the sum $W = \sum_{i=1}^n \xi_i$ we obtain a variable with the W-zero bias distribution. The distributional identity (2.61) indicates that a normal approximation is justified when the difference $\xi_I - \xi_I^*$ is small, since then the distribution of W will be close to that of W^*. To prepare for the proof, note that we may write the variables ξ_I and ξ_I^* selected by I using indicators as follows

$$\xi_I = \sum_{i=1}^{n} \mathbf{1}\{I = i\}\xi_i \quad \text{and} \quad \xi_I^* = \sum_{i=1}^{n} \mathbf{1}\{I = i\}\xi_i^*,$$

from which it is clear, writing $\mathcal{L}$ for the distribution, or law of a random variable, that the distributions of ξ_I and ξ_I^* are the mixtures

$$\mathcal{L}(\xi_I) = \sum_{i=1}^{n} \mathcal{L}(\xi_i)\sigma_i^2 \quad \text{and} \quad \mathcal{L}\big(\xi_I^*\big) = \sum_{i=1}^{n} \mathcal{L}\big(\xi_i^*\big)\sigma_i^2.$$

Proof Let W^* have the W-zero bias distribution. Then for all absolutely continuous functions f for which the following expectations exist,

$$\begin{aligned}
E\big[f'\big(W^*\big)\big] &= E\big[Wf(W)\big] \\
&= E\left[\sum_{i=1}^{n} \xi_i f(W)\right] \\
&= \sum_{i=1}^{n} E\big[\xi_i f(W - \xi_i + \xi_i)\big] \\
&= \sum_{i=1}^{n} E\big[\sigma_i^2 f'\big(W - \xi_i + \xi_i^*\big)\big] \\
&= E\left[\sum_{i=1}^{n} f'\big(W - \xi_i + \xi_i^*\big)\mathbf{1}(I = i)\right] \\
&= E\big[f'\big(W - \xi_I + \xi_I^*\big)\big],
\end{aligned}$$

where independence is used in the fourth and fifth equalities. The equality of the expectations of W^* and $W - \xi_I + \xi_I^*$ over this class of functions is sufficient to guarantee (2.61), that is, that these two random variables have the same distribution, as in the proof of Proposition 2.1. □

When handling the sum of independent random variables, the zero bias method and the K function approach of Sect. 2.3.1 are essentially equivalent, with the former providing a probabilistic formulation of the latter. To begin to see the connection, note that by (2.54) and (2.24) the zero bias density $p^*(t)$ and the $K(t)$ function are almost sure multiplies,

$$p^*(t) = K(t)/\sigma^2.$$

In particular, by Lemma 2.8, with $K_i(t)$ the function (2.24) corresponding to ξ_i, integrating against the density of ξ_i^* yields

$$Ef\big(W^*\big) = \sum_{i=1}^{n} Ef\big(W^{(i)} + \xi_i^*\big)\sigma_i^2 = \sum_{i=1}^{n} \int_{-\infty}^{\infty} Ef\big(W^{(i)} + t\big)K_i(t)\,dt. \quad (2.62)$$

Likewise, that $p^*(x)$ is a density function, and the moment identity (2.57), are probabilistic interpretations of the two equalities in (2.25), respectively, in terms of

random variables. In addition, we note the correspondence between Lemma 2.8 and identity (2.31). To later explore the relationship between the zero bias method and the general Stein identity in Sect. 2.4, note now that if W and W^* are defined on the same space then trivially from the defining zero bias identity (2.51) we have

$$E\big[Wf(W)\big] = Ef'(W+\Delta) \quad \text{where } \Delta = W^* - W.$$

Though the K function approach and zero biasing are essentially completely parallel when dealing with sums of independent variables, these two views each give rise to useful, and separate, ways of handling different classes of examples. In addition to its ties to the K function approach, we will see in Proposition 4.6 that zero biasing is also connected to the exchangeable pair.

2.3.4 Size Bias

The size bias and zero bias transformations are close relatives, and as such, size bias and zero bias couplings can be used in the Stein equation in somewhat similar manners. The size bias transformation is defined on the class of non-negative random variables X with finite non-zero means. For such an X with mean $EX = \mu$, we say X^s has the X-size biased distribution if for all functions f for which $E[Xf(X)]$ exists,

$$E\big[Xf(X)\big] = \mu Ef\big(X^s\big). \tag{2.63}$$

We note that this characterization for size biasing is of the same form as (2.51) for zero biasing, but with the mean replacing the variance, and f replacing f' for the evaluation of the biased variable.

To place size biasing in the framework of Sect. 2.4 to follow, we note that when $\mathrm{Var}(X) = \sigma^2$ and $W = (X-\mu)/\sigma$, and, with a slight abuse of notation, $W^s = (X^s - \mu)/\sigma$, if X and X^s are defined on the same space, identity (2.63) can be written

$$E\big[Wf(W)\big] = \frac{\mu}{\sigma} E\big[f\big(W^s\big) - f(W)\big] = E\int_{-\infty}^{\infty} f'(W+t)\hat{K}(t)\,dt, \tag{2.64}$$

where

$$\hat{K}(t) = \frac{\mu}{\sigma}(\mathbf{1}_{\{0\le t\le W^s - W\}} - \mathbf{1}_{\{W^s - W\le t<0\}}). \tag{2.65}$$

The characterization (2.63) is easily seen to be the same as the more common specification of the size bias distribution $F^s(x)$ as the one which is absolutely continuous with respect to the distribution $F(x)$ of X with Radon Nikodym derivative

$$\frac{dF^s(x)}{dF(x)} = \frac{x}{\mu}. \tag{2.66}$$

Hence, parallel to property (2.58) for zero bias, here we have

$$0 \le X \le C \quad \text{implies} \quad 0 \le X^s \le C. \tag{2.67}$$

Moreover, if X is absolutely continuous with density $p(x)$, then X^s is also absolutely continuous, and has density $xp(x)/\mu$. Size biasing also enjoys a scaling property. If X^s has the X-size bias distribution, then for $a > 0$

$$(aX)^s = aX^s$$

by an argument nearly identical to the one that proves (2.59).

Size biasing can occur, possibly unwanted, when applying various sampling designs where items associated with larger outcomes are more likely to be chosen. For instance, when sampling an individual in a population at random, their report of the number of siblings in their family is size biased. Size biasing is also responsible for the well known waiting time paradox (see Feller 1968b), but can also be used to advantage, in particular, to form unbiased ratio estimates (Midzuno 1951).

Lemma 2.8 carries over with only minor changes when replacing zero biasing by size biasing, though the variable replaced is now selected proportional to its mean, rather its variance. Moreover, the size bias construction generalizes easily to the case where the sum is of dependent random variables. In particular, let $\mathbf{X} = \{X_\alpha, \alpha \in \mathcal{A}\}$ be a collection of nonnegative random variables with finite, nonzero means $\mu_\alpha = EX_\alpha$. For $\alpha \in \mathcal{A}$, we say that $\mathbf{X}^\alpha$ has the $\mathbf{X}$ distribution biased in direction, or coordinate, α if

$$EX_\alpha f(\mathbf{X}) = \mu_\alpha Ef\left(\mathbf{X}^\alpha\right) \tag{2.68}$$

for all real valued functions f for which the expectation of the left hand side exists.

Parallel to (2.66), if $F(\mathbf{x})$ is the distribution of $\mathbf{X}$, then the distribution $F^\alpha(\mathbf{x})$ of $\mathbf{X}^\alpha$ satisfies

$$\frac{dF^\alpha(\mathbf{x})}{dF(\mathbf{x})} = \frac{x_\alpha}{\mu_\alpha}. \tag{2.69}$$

By considering functions f which depend only on x_α, it is easy to verify that $X_\alpha^\alpha =_d X_\alpha^s$, that is, that X_α^α has the X_α-size biased distribution.

A consequence of the following proposition is a method for size biasing sums of dependent variables.

Proposition 2.2 *Let $\mathcal{A}$ be an arbitrary index set, and let $\mathbf{X} = \{X_\alpha, \alpha \in \mathcal{A}\}$ be a collection of nonnegative random variables with finite means. For any subset $B \subset \mathcal{A}$, set*

$$X_B = \sum_{\beta \in B} X_\beta \quad \text{and} \quad \mu_B = EX_B.$$

Suppose $B \subset \mathcal{A}$ with $0 < \mu_B < \infty$, and for $\beta \in B$ let $\mathbf{X}^\beta$ have the $\mathbf{X}$-size biased distribution in coordinate β as in Definition 2.68. Let I be a random index, independent of $\mathbf{X}$, with distribution

$$P(I = \beta) = \frac{\mu_\beta}{\mu_B}.$$

Then $\mathbf{X}^B = \mathbf{X}^I$, *that is, the collection* $\mathbf{X}^B$ *which is equal to* $\mathbf{X}^\beta$ *with probability* μ_β/μ_B, *satisfies*

$$E\big[X_B f(\mathbf{X})\big] = \mu_B E f\big(\mathbf{X}^B\big) \tag{2.70}$$

for all real valued functions f *for which these expectations exist.*

If f *is a function of* $X_A = \sum_{\alpha\in A} X_\alpha$ *only, then*

$$E\big[X_B f(X_A)\big] = \mu_B E f\big(X_A^B\big) \quad \textit{where } X_A^B = \sum_{\alpha\in A} X_\alpha^B,$$

and when $A = B$ *we have* $EX_A f(X_A) = \mu_A E f(X_A^A)$, *and that* X_A^A *has the* X_A*-size biased distribution.*

Proof Without loss of generality, assume $\mu_\beta > 0$ for all $\beta \in \mathcal{A}$. By (2.68) we have

$$E\big[X_\beta f(\mathbf{X})\big]/\mu_\beta = E f\big(\mathbf{X}^\beta\big).$$

Multiplying by μ_β/μ_B, summing over $\beta \in B$ and recalling $\mathbf{X}^B$ is a mixture yields (2.70). The remainder of the lemma now follows as special cases. □

By the last claim of the lemma, to achieve the size bias distribution of the sum $X_\mathcal{A} = \sum_{\alpha\in\mathcal{A}} X_\alpha$ of all the variables in the collection, one mixes over the distributions of $X_\mathcal{A}^\beta = \sum_{\alpha\in\mathcal{A}} X_\alpha^\beta$ using the random index with distribution

$$P(I = \beta) = \frac{\mu_\beta}{\sum_{\alpha\in\mathcal{A}} \mu_\alpha}. \tag{2.71}$$

Hence, by randomization over $\mathcal{A}$, a construction of $\mathbf{X}^\beta$ for every coordinate β leads to a construction of $X_\mathcal{A}^s$.

We may size bias in coordinates by applying the following procedure. Let $\mathcal{A} = \{1, \ldots, n\}$ for notational ease. For given $i \in \{1, \ldots, n\}$, write the joint distribution of $\mathbf{X}$ as a product of the marginal distribution of X_i times the conditional distribution of the remaining variables given X_i,

$$dF(\mathbf{x}) = dF_i(x_i)\, dF(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n | x_i), \tag{2.72}$$

which gives a factorization of (2.69) as

$$dF^i(\mathbf{x}) = dF_i^i(x_i)\, dF(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n | x_i), \tag{2.73}$$

where $dF_i^i(x_i) = (x_i/\mu_i)\, dF_i(x_i)$.

The representation (2.73) says that one may form $\mathbf{X}^i$ by first generating X_i^i having the X_i-sized biased distribution, and then the remaining variables from their original distribution, conditioned on x_i taking on its newly chosen sized biased value. For $\mathbf{X}$ already given, a coupling between the sum of $Y = X_1 + \cdots + X_n$ and Y^s can be generated by first constructing, for every i, the biased variable X_i^i and then 'adjusting' the remaining variables X_j, $j \neq i$ as necessary so that they have the correct conditional distribution. Mixing then yields Y^s. Typically the goal

is to adjust the variables as little as possible in order to have the resulting bounds to normality small.

The following important corollary of Proposition 2.2 handles the case where the variables $X_1, \ldots, X_n$ are independent, so that (2.72) reduces to

$$dF(\mathbf{x}) = dF_i(x_i)dF(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n).$$

The following result is parallel to Lemma 2.8.

Corollary 2.1 *Let $Y = \sum_{i=1}^n X_i$, where $X_1, \ldots, X_n$ are independent, nonnegative random variables with means $EX_i = \mu_i$, $i = 1, \ldots, n$. Let I be a random index with distribution given by (2.71), independent of all other variables. Then, upon replacing the summand X_I selected by I with a variable X_I^s having its size biased distribution, independent of X_j for $j \neq I$, we obtain*

$$Y^I =_d Y - X_I + X_I^s,$$

a variable having the Y-size bias distribution.

Proof Letting $\mathbf{X} = (X_1, \ldots, X_n)$, the vector

$$\mathbf{X}^i = \big(X_1, \ldots, X_{i-1}, X_i^s, X_{i+1}, \ldots, X_n\big)$$

has the $\mathbf{X}$-size biased distribution in coordinate i, as the conditional distribution in (2.73) is the same as the unconditional one. Now apply Proposition 2.2. □

In other words, when the variables are independent and X_i is replaced by its size biased version, here is no need to change any of the remaining variables X_j, $j \neq i$ in order for them to have their original conditional distribution given the new value X_i^s.

As shown in Goldstein and Reinert (2005), size biasing and zero biasing are both special cases of a general form of distributional biasing, where given a 'biasing function' $P(x)$ with $m \in \{0, 1, \ldots\}$ sign changes, and a distribution X which satisfies the $m-1$ orthogonality relations $EX^i P(X) = 0$, $i = 0, \ldots, m-1$, there exists a distribution $X^{(P)}$ satisfying

$$E\big[P(X)f(X)\big] = \alpha E f^{(m)}\big(X^{(P)}\big) \tag{2.74}$$

when $\alpha = EP(X)X^m/m! > 0$.

For example, for zero biasing the function $P(x) = x$ has $m = 1$ sign change, so the identity involves the first derivative f', and we require that the distribution of X satisfies the single orthogonality relation $E(1 \cdot X) = EX = 0$, and set $\alpha = EX^2 = \sigma^2$. For size biasing $P(x) = \max\{x, 0\}$, which has $m = 0$ sign changes, so no derivatives of f are involved, and neither are there any orthogonality relations to be satisfied, and $\alpha = EX$. Letting $X^\square$ be characterized by (2.74) with $P(x) = x^2$, since $P(x)$ has no sign changes the distribution $X^\square$ exists for any distribution X with finite second moment, and $\alpha = EX^2$. In this particular case, where

$$E\big[X^2 f(X)\big] = EX^2 E f\big(X^\square\big) \tag{2.75}$$

for all functions f for which $E|Xf(X)| < \infty$, we say that $X^\square$ has the X-square bias distribution. As in (2.69) for the size biased distribution, the distribution of $X^\square$ can also be characterized by its Radon–Nikodym derivative with respect to the distribution of X, as we do in Proposition 2.3, below.

By comparing Lemma 2.8 with Corollary 2.1 one can already see that zero and size biasing are closely related. Another relation between the two is given by the following proposition.

Proposition 2.3 *Let X be a symmetric random variable with finite, non-zero variance σ^2, and let $X^\square$ have the X-square bias distribution, that is,*

$$dF^\square(x) = \frac{x^2 dF(x)}{\sigma^2}.$$

Then, with $U \sim \mathcal{U}[-1,1]$ independent of $X^\square$, the variable

$$X^* \stackrel{d}{=} UX^\square$$

has the X-zero bias distribution.

Proof Since X is symmetric with finite second moment, $EX = 0$ and $EX^2 = \sigma^2$. For an absolutely continuous function f with derivative $g \in C_c$, the collection of continuous functions having compact support, using the characterization (2.75) for the fourth equality below, we have

$$\begin{aligned}
\sigma^2 Eg\big(UX^\square\big) &= \sigma^2 Ef'\big(UX^\square\big) \\
&= \frac{\sigma^2}{2} E \int_{-1}^{1} f'\big(uX^\square\big) du \\
&= \frac{\sigma^2}{2} E\left(\frac{f(X^\square) - f(-X^\square)}{X^\square} \right) \\
&= \frac{1}{2} E\left(X^2 \frac{f(X) - f(-X)}{X} \right) \\
&= \frac{1}{2} E\big(X\big(f(X) - f(-X)\big)\big) \\
&= \frac{1}{2}\big(EXf(X) + E(-X)f(-X)\big) \\
&= E\big[Xf(X)\big].
\end{aligned}$$

Hence, if X^* has the X-zero bias distribution,

$$\sigma^2 Eg\big(UX^\square\big) = E\big[Xf(X)\big] = \sigma^2 E\big[f'\big(X^*\big)\big] = \sigma^2 Eg\big(X^*\big).$$

As the expectation of $g(UX^\square)$ and $g(X^*)$ agree for all $g \in C_c$, the random variables $UX^\square$ and X^* must be equal in distribution. □

2.4 A General Framework for Stein Identities and Normal Approximation for Lipschitz Functions

Identity (2.42)

$$E\big[Wf(W)\big] = E\int_{-\infty}^{\infty} f'(W+t)\hat{K}(t)dt + E\big[Rf(W)\big], \tag{2.76}$$

arose when allowing for the possibility that a given exchangeable pair may not satisfy the linearity condition (2.33) exactly. The function $\hat{K}(t)$ may be random, and, to obtain a good bound, R should be a random variable so that the second term $E[Rf(W)]$ is of smaller order than the first. The exchangeable pair and size bias identities, (2.36) and (2.64), respectively, are both the special case of (2.76) when $R = 0$. For the first case, the function $\hat{K}(t)$ is given by (2.38), and by (2.65) in the second.

Though the zero bias identity (2.51) with $\sigma^2 = 1$ does not fit the mold of (2.76) precisely, in somewhat the same spirit, with $\Delta = W^* - W$ we have

$$EWf(W) = Ef'(W+\Delta), \tag{2.77}$$

holding for all absolutely continuous functions f for which the expectations above exist. The following proposition provides a general bound for normal approximation for smooth functions when (2.76) or (2.77) holds.

Proposition 2.4 *Let h be an absolutely continuous function with $\|h'\| < \infty$ and $\mathcal{F}$ any σ-algebra containing $\sigma\{W\}$.*

(i) *If (2.76) holds, then*

$$\big|Eh(W) - Nh\big| \le \|h'\|\left(\sqrt{\frac{2}{\pi}}E|1-\hat{K}_1| + 2E\hat{K}_2 + 2E|R|\right), \tag{2.78}$$

where

$$\hat{K}_1 = E\left\{\int_{-\infty}^{\infty}\hat{K}(t)\,dt|\mathcal{F}\right\} \quad \text{and} \quad \hat{K}_2 = \int_{-\infty}^{\infty}\big|t\hat{K}(t)\big|\,dt.$$

(ii) *If (2.77) holds, then*

$$\big|Eh(W) - Nh\big| \le 2\|h'\|E|\Delta|. \tag{2.79}$$

Proof Let f_h be the solution (2.5) to the Stein equation (2.4). We note that by (2.13), both f_h and f_h' are bounded. We may assume the expectations on the right hand side of (2.78) are finite, as otherwise the result is trivial. By (2.4) and (2.76),

$$\begin{aligned} Eh(W) - Nh &= E\big[f_h'(W) - Wf_h(W)\big] \\ &= Ef_h'(W) - E\int_{-\infty}^{\infty} f_h'(W+t)\hat{K}(t)\,dt - E\big[Rf_h(W)\big] \\ &= Ef_h'(W)(1-\hat{K}_1) + E\int_{-\infty}^{\infty}\big\{f_h'(W) - f_h'(W+t)\big\}\hat{K}(t)\,dt \\ &\quad - E\big[Rf_h(W)\big]. \end{aligned}$$

By the properties of the Stein solution f_h given in (2.13) and the mean value theorem, we have

$$\bigl|Ef_h'(W)(1-\hat{K}_1)\bigr| \le \|h'\|\sqrt{\frac{2}{\pi}}E|1-\hat{K}_1|,$$

$$\Bigl|E\int_{-\infty}^{\infty}\bigl\{f_h'(W)-f_h'(W+t)\bigr\}\hat{K}(t)\,dt\Bigr| \le E\int_{-\infty}^{\infty}2\|h'\|\bigl|t\hat{K}(t)\bigr|\,dt = 2\|h'\|E\hat{K}_2$$

and

$$\bigl|E\bigl[Rf_h(W)\bigr]\bigr| \le 2\|h'\|E|R|.$$

This proves (2.78).

Next, (2.79) follows from (2.13) and

$$\begin{aligned}\bigl|Eh(W)-Nh\bigr| &= \bigl|E\bigl(f_h'(W)-Wf_h(W)\bigr)\bigr| \\ &= \bigl|E\bigl(f_h'(W)-f_h'(W+\Delta)\bigr)\bigr| \\ &\le \|f_h''\|E|\Delta|.\end{aligned}$$

□

We will explore smooth function bounds extensively in Chap. 4.

Appendix

Here we prove Lemmas 2.3 and 2.4, giving the basic properties of the solutions to the Stein equations (2.2) and (2.4). The proof of Lemma 2.3, and part of Lemma 2.4, follow Stein (1986), while parts of the proof of Lemma 2.4 are due to Stroock (2000) and Raič (2004) (see also Chatterjee 2008).

Before beginning, note that from (2.2) and (2.3) it follows that

$$\begin{aligned}f_z'(w) &= wf_z(w)+\mathbf{1}_{\{w\le z\}}-\Phi(z) \\ &= \begin{cases} wf_z(w)+1-\Phi(z) & \text{for } w<z, \\ wf_z(w)-\Phi(z) & \text{for } w>z, \end{cases} \\ &= \begin{cases} (\sqrt{2\pi}we^{w^2/2}\Phi(w)+1)(1-\Phi(z)) & \text{for } w<z, \\ (\sqrt{2\pi}we^{w^2/2}(1-\Phi(w))-1)\Phi(z) & \text{for } w>z, \end{cases}\end{aligned} \tag{2.80}$$

and

$$\bigl(wf_z(w)\bigr)' = \begin{cases} \sqrt{2\pi}(1-\Phi(z))((1+w^2)e^{w^2/2}\Phi(w)+\frac{w}{\sqrt{2\pi}}) & \text{if } w<z, \\ \sqrt{2\pi}\Phi(z)((1+w^2)e^{w^2/2}(1-\Phi(w))-\frac{w}{\sqrt{2\pi}}) & \text{if } w>z. \end{cases} \tag{2.81}$$

Proof of Lemma 2.3 Since $f_z(w) = f_{-z}(-w)$, we need only consider the case $z \ge 0$. Note that for $w > 0$

$$\int_w^{\infty} e^{-x^2/2}\,dx \le \int_w^{\infty}\frac{x}{w}e^{-x^2/2}\,dx = \frac{e^{-w^2/2}}{w},$$

and that

$$\int_w^\infty e^{-x^2/2}\,dx \geq \frac{we^{-w^2/2}}{1+w^2},$$

by comparing the derivatives of the two functions and their values at $w=0$. Thus

$$\frac{we^{-w^2/2}}{(1+w^2)\sqrt{2\pi}} \leq 1-\Phi(w) \leq \frac{e^{-w^2/2}}{w\sqrt{2\pi}}. \tag{2.82}$$

Applying the lower bound in inequality (2.82) to the form $(wf_z(w))'$ for $w>z$ in (2.81), we see that this derivative is nonnegative, thus yielding (2.6). Now, in view of (2.82) and the fact that $wf_z(w)$ is increasing, taking limits using (2.3) we have,

$$\lim_{w\to-\infty} wf_z(w) = \Phi(z)-1 \quad \text{and} \quad \lim_{w\to\infty} wf_z(w) = \Phi(z), \tag{2.83}$$

and (2.7) follows.

Now, using that $wf_z(w)$ is an increasing function of w, (2.83) and (2.80),

$$0 < f_z'(w) \leq zf_z(z)+1-\Phi(z) < 1 \quad \text{for } w<z \tag{2.84}$$

and

$$-1 < zf_z(z)-\Phi(z) \leq f_z'(w) < 0 \quad \text{for } w>z, \tag{2.85}$$

proving the first inequality of (2.8). For the second, note that for any w and u we therefore have

$$\big|f_z'(w)-f_z'(u)\big| \leq zf_z(z)+1-\Phi(z)-\big(zf_z(z)-\Phi(z)\big) = 1.$$

Next, observe that by (2.84) and (2.85), $f_z(w)$ attains its maximum at z. Thus

$$0 < f_z(w) \leq f_z(z) = \sqrt{2\pi}e^{z^2/2}\Phi(z)\big(1-\Phi(z)\big).$$

By (2.82), $f_z(z) \leq 1/z$. To finish the proof of (2.9), let

$$g(z) = \Phi(z)\big(1-\Phi(z)\big) - \frac{e^{-z^2/2}}{4} \quad \text{and} \quad g_1(z) = \frac{1}{\sqrt{2\pi}} + \frac{z}{4} - \frac{2\Phi(z)}{\sqrt{2\pi}}.$$

Observe that $g'(z) = e^{-z^2/2}g_1(z)$ and that

$$g_1(0)=0,\ g_1'(0)<0,\ g_1''(z) = \frac{z}{\pi}e^{-z^2/2} \quad \text{and} \quad \lim_{z\to\infty} g_1(z) = \infty.$$

Hence g_1 is convex on $[0,\infty)$, and there exists $z_1>0$ such that $g_1(z)<0$ for $z<z_1$ and $g_1(z)>0$ for $z>z_1$. In particular, on $[0,\infty)$ the function $g(z)$ decreases for $z<z_1$ and increases for $z>z_1$, so its supremum must be attained at either $z=0$ or $z=\infty$, that is,

$$g(z) \leq \max\big(g(0), g(\infty)\big) = 0 \quad \text{for all } z\in[0,\infty),$$

which is equivalent to $f_z(z) \leq \sqrt{2\pi}/4$. This completes the proof of (2.9).

To verify the last inequality (2.10), write

$$\begin{aligned}&(w+u)f_z(w+u)-(w+v)f_z(w+v)\\&\quad=w\big(f_z(w+u)-f_z(w+v)\big)+uf_z(w+u)-vf_z(w+v)\end{aligned}$$

and apply the mean value theorem and (2.8) on the first term, and (2.9) on the second. □

Proof of Lemma 2.4 Let $\tilde{h}(w)=h(w)-Nh$ and put $c_0=\|\tilde{h}\|$ and let $c_1=\|h'\|$ if h is absolutely continuous, and $c_1=\infty$ otherwise. Since $\tilde{h}$ and f_h are unchanged when h is replaced by $h-h(0)$, we may assume that $h(0)=0$. Therefore $|h(t)|\le c_1|t|$ and $|Nh|\le c_1E|Z|=c_1\sqrt{2/\pi}$.

We first prove the two bounds on f_h itself. From the expression (2.5) for f_h it follows that

$$\begin{aligned}\big|f_h(w)\big|&\le\begin{cases}e^{w^2/2}\int_{-\infty}^{w}|\tilde{h}(x)|e^{-x^2/2}\,dx & \text{if } w\le 0,\\ e^{w^2/2}\int_{w}^{\infty}|\tilde{h}(x)|e^{-x^2/2}\,dx & \text{if } w\ge 0\end{cases}\\&\le e^{w^2/2}\min\bigg(c_0\int_{|w|}^{\infty}e^{-x^2/2}\,dx,\,c_1\int_{|w|}^{\infty}\big(|x|+\sqrt{2/\pi}\,\big)e^{-x^2/2}\,dx\bigg)\\&\le\min(\sqrt{\pi/2}c_0,2c_1),\end{aligned}$$

where in the last inequality we obtain

$$e^{w^2/2}\int_{|w|}^{\infty}e^{-x^2/2}\,dx\le\sqrt{\pi/2}$$

by applying (2.82) to show that the function on the left hand side above has a negative derivative for $w\ge 0$, and therefore that its maximum is achieved at $w=0$. We note that the first bound in the minimum applies if h is only bounded, thus yielding the first claim in (2.12), while if h is only absolutely continuous the second bound holds, yielding the first claim in (2.13).

Moving to bounds on f_h', by (2.4) for $w\ge 0$,

$$\begin{aligned}\big|f_h'(w)\big|&\le\big|h(w)-Nh\big|+we^{w^2/2}\int_{w}^{\infty}\big|h(x)-Nh\big|e^{-x^2/2}\,dx\\&\le c_0+c_0we^{w^2/2}\int_{w}^{\infty}e^{-x^2/2}\,dx\le 2c_0,\end{aligned}$$

using (2.82). A similar argument may be applied for $w<0$, proving the remaining claim in (2.12).

To prove the second claim in (2.13), when h is absolutely continuous write

$$\begin{aligned}h(x)-Nh&=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\big[h(x)-h(u)\big]e^{-u^2/2}\,du\\&=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\int_{u}^{x}h'(t)e^{-u^2/2}\,dt\,du-\frac{1}{\sqrt{2\pi}}\int_{x}^{\infty}\int_{x}^{u}h'(t)e^{-u^2/2}\,dt\,du\\&=\int_{-\infty}^{x}h'(t)\Phi(t)\,dt-\int_{x}^{\infty}h'(t)\big(1-\Phi(t)\big)\,dt,\end{aligned}\tag{2.86}$$

from which it follows that

$$\begin{aligned} f_h(w) &= e^{w^2/2}\int_{-\infty}^{w}\big[h(x)-Nh\big]e^{-x^2/2}\,dx \\ &= e^{w^2/2}\int_{-\infty}^{w}\left(\int_{-\infty}^{x}h'(t)\Phi(t)\,dt-\int_{x}^{\infty}h'(t)\big(1-\Phi(t)\big)\,dt\right)e^{-x^2/2}\,dx \\ &= -\sqrt{2\pi}\,e^{w^2/2}\big(1-\Phi(w)\big)\int_{-\infty}^{w}h'(t)\Phi(t)\,dt \\ &\quad -\sqrt{2\pi}\,e^{w^2/2}\Phi(w)\int_{w}^{\infty}h'(t)\big[1-\Phi(t)\big]\,dt. \end{aligned} \tag{2.87}$$

Now, from (2.4), (2.87) and (2.86),

$$\begin{aligned} f_h'(w) &= wf_h(w)+h(w)-Nh \\ &= \big(1-\sqrt{2\pi}\,we^{w^2/2}\big(1-\Phi(w)\big)\big)\int_{-\infty}^{w}h'(t)\Phi(t)\,dt \\ &\quad -\big(1+\sqrt{2\pi}\,we^{w^2/2}\Phi(w)\big)\int_{w}^{\infty}h'(t)\big(1-\Phi(t)\big)\,dt. \end{aligned}$$

Hence

$$\begin{aligned} \|f_h'\| \le \|h'\|\sup_{w\in\mathbb{R}}\Bigg(&\big|1-\sqrt{2\pi}\,we^{w^2/2}\big(1-\Phi(w)\big)\big|\int_{-\infty}^{w}\Phi(t)\,dt \\ &+\big|1+\sqrt{2\pi}\,we^{w^2/2}\Phi(w)\big|\int_{w}^{\infty}\big(1-\Phi(t)\big)\,dt\Bigg). \end{aligned}$$

By integration by parts,

$$\begin{aligned} \int_{-\infty}^{w}\Phi(t)\,dt &= w\Phi(w)+\frac{e^{-w^2/2}}{\sqrt{2\pi}} \quad \text{and} \\ \int_{w}^{\infty}\big(1-\Phi(t)\big)\,dt &= -w\big(1-\Phi(w)\big)+\frac{e^{-w^2/2}}{\sqrt{2\pi}}. \end{aligned} \tag{2.88}$$

Thus,

$$\begin{aligned} \|f_h'\| \le \|h'\|\sup_{w\in\mathbb{R}}\Bigg(&\big|1-\sqrt{2\pi}\,we^{w^2/2}\big(1-\Phi(w)\big)\big|\left(w\Phi(w)+\frac{e^{-w^2/2}}{\sqrt{2\pi}}\right) \\ &+\big|1+\sqrt{2\pi}\,we^{w^2/2}\Phi(w)\big|\left(-w\big(1-\Phi(w)\big)+\frac{e^{-w^2/2}}{\sqrt{2\pi}}\right)\Bigg). \end{aligned}$$

One may now verify that the term inside the brackets attains its maximum value of $\sqrt{2/\pi}$ at $w=0$.

Now we prove the final claim of (2.13). Differentiating (2.4) gives

$$\begin{aligned} f_h''(w) &= wf_h'(w)+f_h(w)+h'(w) \\ &= \big(1+w^2\big)f_h(w)+w\big(h(w)-Nh\big)+h'(w). \end{aligned} \tag{2.89}$$

From (2.89), (2.87), (2.86), (2.82) and (2.88) we obtain

$$
\begin{aligned}
\big|f_h''(w)\big| &\le \big|h'(w)\big| + \big|\big(1+w^2\big)f_h(w) + w\big(h(w) - Nh\big)\big| \\
&\le \big|h'(w)\big| + \Big|\big(w - \sqrt{2\pi}\big(1+w^2\big)e^{w^2/2}\big(1-\Phi(w)\big)\big)\int_{-\infty}^{w} h'(t)\Phi(t)\,dt\Big| \\
&\quad + \Big|\big(-w - \sqrt{2\pi}\big(1+w^2\big)e^{w^2/2}\Phi(w)\big)\int_{w}^{\infty} h'(t)\big(1-\Phi(t)\big)\,dt\Big| \\
&\le \big|h'(w)\big| + c_1\big(-w + \sqrt{2\pi}\big(1+w^2\big)e^{w^2/2}\big(1-\Phi(w)\big)\big)\int_{-\infty}^{w} \Phi(t)\,dt \\
&\quad + c_1\big(w + \sqrt{2\pi}\big(1+w^2\big)e^{w^2/2}\Phi(w)\big)\int_{w}^{\infty} \big(1-\Phi(t)\big)\,dt \\
&= \big|h'(w)\big| \\
&\quad + c_1\big(-w + \sqrt{2\pi}\big(1+w^2\big)e^{w^2/2}\big(1-\Phi(w)\big)\big)\Big(w\Phi(w) + \frac{e^{-w^2/2}}{\sqrt{2\pi}}\Big) \\
&\quad + c_1\big(w + \sqrt{2\pi}\big(1+w^2\big)e^{w^2/2}\Phi(w)\big)\Big(-w\big(1-\Phi(w)\big) + \frac{e^{-w^2/2}}{\sqrt{2\pi}}\Big) \\
&= \big|h'(w)\big| + c_1 \le 2c_1,
\end{aligned}
$$

as desired. □

We now present the proof of Lemma 2.5 for bounds on the solution $f(w)$ to the Stein equation for the linearly smoothed indicator function (2.14). For this case Bolthausen (1984) proved the inequalities $|f(w)| \le 1$, $|f'(w)| \le 2$, and, through use of the latter, the bound (2.16) with the factor of $|w|$ replaced by $2|w|$.

Proof of Lemma 2.5 As in (2.87) in the proof of Lemma 2.3, letting

$$\eta(w) = \sqrt{2\pi}e^{w^2/2}\Phi(w),$$

we have

$$f(w) = -\eta(-w)\int_{-\infty}^{w} h'(t)\Phi(t)\,dt - \eta(w)\int_{w}^{\infty} h'(t)\Phi(-t)\,dt. \tag{2.90}$$

For $z \le w \le z+\alpha$, we therefore have

$$
\begin{aligned}
f(w) &= \eta(-w)\int_{z}^{w} \frac{\Phi(t)}{\alpha}dt + \eta(w)\int_{w}^{z+\alpha} \frac{\Phi(-t)}{\alpha}\,dt \\
&\le \frac{\eta(-w)\Phi(w)(w-z)}{\alpha} + \frac{\eta(w)\Phi(-w)(z+\alpha-w)}{\alpha} \\
&= \eta(w)\Phi(-w) \\
&= \sqrt{2\pi}e^{w^2/2}\Phi(w)\Phi(-w).
\end{aligned}
\tag{2.91}
$$

By symmetry we may take $w \ge 0$ without loss of generality. Then, using the fact that $\Phi(w)/w$ is decreasing, and straightforward inequalities, we derive

$$\sqrt{2\pi}\,e^{w^2/2}\Phi(w)\Phi(-w) \le \min\left(\frac{\sqrt{2\pi}}{2}, \frac{1}{w}\right)\Phi(w) \le \frac{\sqrt{2\pi}}{2}\Phi\left(\sqrt{\frac{2}{\pi}}\right) < 1,$$

showing $f(w) \le 1$ for $w \in [z, z+\alpha]$.

Next, note that $\Phi(z) \le Nh \le \Phi(z+\alpha)$, and let $f_z(w)$ be the solution to the Stein equation for the function $\mathbf{1}_{\{w \le z\}}$. For $w < z$, since $e^{w^2/2}\Phi(w)$ is increasing, we obtain

$$\begin{aligned} f(w) &= \sqrt{2\pi}(1 - Nh)e^{w^2/2}\Phi(w) \\ &\le \sqrt{2\pi}\big(1 - \Phi(z)\big)e^{w^2/2}\Phi(w) \\ &\le \sqrt{2\pi}\big(1 - \Phi(z)\big)e^{z^2/2}\Phi(z) \\ &= f_z(z) \le \sqrt{2\pi}/4, \end{aligned} \tag{2.92}$$

using (2.9). Similarly, for $w > z + \alpha$,

$$\begin{aligned} f(w) &= \sqrt{2\pi}Nhe^{w^2/2}\big(1 - \Phi(w)\big) \\ &\le \sqrt{2\pi}\Phi(z+\alpha)e^{w^2/2}\big(1 - \Phi(w)\big) \\ &\le \sqrt{2\pi}\Phi(w)e^{w^2/2}\big(1 - \Phi(w)\big) \\ &= f_w(w) \le \sqrt{2\pi}/4, \end{aligned} \tag{2.93}$$

showing that $f(w) \le 1$ for all $w \in \mathbb{R}$. The proof of the first claim of (2.15) is completed by showing the lower bound, which follows from the three expressions (2.91), (2.92) and (2.93), proving that $f(w) \ge 0$ over the three intervals $(-\infty, z)$, $[z, z+\alpha]$ and (α, ∞), respectively.

For the second claim, starting again from (2.5), we have

$$\begin{aligned} e^{-w^2/2}f(w) &= \int_{-\infty}^{w} h(x)e^{-x^2/2}dx - \int_{-\infty}^{w} e^{-x^2/2}Nh\,dx \\ &= \int_{-\infty}^{w} h(x)e^{-x^2/2}dx - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{w} e^{-x^2/2}\int_{-\infty}^{\infty} h(t)e^{-t^2/2}\,dt\,dx \\ &= \int_{-\infty}^{w} h(x)e^{-x^2/2}dx - \Phi(w)\int_{-\infty}^{\infty} h(t)e^{-t^2/2}\,dt \\ &= \big(1 - \Phi(w)\big)\int_{-\infty}^{w} h(x)e^{-x^2/2}dx - \Phi(w)\int_{w}^{\infty} h(t)e^{-t^2/2}\,dt. \end{aligned}$$

Hence,

$$\begin{aligned} f(w) &= e^{w^2/2}\big(1 - \Phi(w)\big)\int_{-\infty}^{w} h(x)e^{-x^2/2}dx - e^{w^2/2}\Phi(w)\int_{w}^{\infty} h(t)e^{-t^2/2}\,dt \\ &= \frac{1}{\sqrt{2\pi}}\eta(-w)\int_{-\infty}^{w} h(x)e^{-x^2/2}dx - \frac{1}{\sqrt{2\pi}}\eta(w)\int_{w}^{\infty} h(t)e^{-t^2/2}\,dt, \end{aligned}$$

and taking the derivative, we obtain

$$
\begin{aligned}
f'(w) &= -\frac{1}{\sqrt{2\pi}}\eta'(-w)\int_{-\infty}^{w} h(x)e^{-x^2/2}dx + \frac{1}{\sqrt{2\pi}}\eta(-w)e^{-w^2/2}h(w) \\
&\quad - \frac{1}{\sqrt{2\pi}}\eta'(w)\int_{w}^{\infty} h(t)e^{-t^2/2}\,dt + \frac{1}{\sqrt{2\pi}}\eta(w)e^{-w^2/2}h(w) \\
&= h(w)\big(\Phi(w)+\Phi(-w)\big) \\
&\quad - \frac{1}{\sqrt{2\pi}}\left(\eta'(-w)\int_{-\infty}^{w} h(x)e^{-x^2/2}dx + \eta'(w)\int_{w}^{\infty} h(t)e^{-t^2/2}\,dt\right) \\
&= h(w) - g(w),
\end{aligned}
$$

where we have set

$$
g(w) = \frac{1}{\sqrt{2\pi}}\left(\eta'(-w)\int_{-\infty}^{w} h(x)e^{-x^2/2}dx + \eta'(w)\int_{w}^{\infty} h(x)e^{-x^2/2}dx\right).
$$

Since $\eta'(w) \geq 0$, we have

$$
\begin{aligned}
&\inf_x h(x)\big(\eta'(-w)\Phi(w) + \eta'(w)\Phi(-w)\big) \\
&\quad \leq g(w) \leq \sup_x h(x)\big(\eta'(-w)\Phi(w) + \eta'(w)\Phi(-w)\big).
\end{aligned}
$$

However, noting

$$
\eta'(-w)\Phi(w) + \eta'(w)\Phi(-w) = 1, \tag{2.94}
$$

it follows that

$$
\inf_x h(x) - \sup_x h(x) \leq f'(w) \leq \sup_x h(x) - \inf_x h(x),
$$

that is, $|f'(w)| \leq 1$, proving the second claim in (2.15).

For the third claim in (2.15), differentiating (2.90) yields

$$
f'(w) = \eta'(-w)\int_{-\infty}^{w} h'(t)\Phi(t)\,dt - \eta'(w)\int_{w}^{\infty} h'(t)\Phi(-t)\,dt.
$$

For $w < z$ we have

$$
f'(w) = \eta'(w)\int_{z}^{z+\alpha} \frac{\Phi(-t)}{\alpha}\,dt,
$$

for $w \in [z, z+\alpha]$,

$$
f'(w) = -\eta'(-w)\int_{z}^{w} \frac{\Phi(t)}{\alpha}dt + \eta'(w)\int_{w}^{z+\alpha} \frac{\Phi(-t)}{\alpha}dt,
$$

and for $w > z+\alpha$

$$
f'(w) = -\eta'(-w)\int_{z}^{z+\alpha} \frac{\Phi(t)}{\alpha}\,dt.
$$

Hence, we may write

$$
f'(w) = -\frac{1}{\alpha}\int_{z}^{z+\alpha} G(w,t)\,dt,
$$

where

$$G(w,t) = \begin{cases} -\eta'(w)\Phi(-t) & \text{when } w \le t, \\ \eta'(-w)\Phi(t) & \text{when } w > t. \end{cases}$$

Now writing

$$\eta(w) = \sqrt{2\pi}e^{w^2/2}\Phi(w) = \int_{-\infty}^{0} e^{-s^2/2-sw}ds,$$

applying the dominated convergence theorem to differentiate under the integral, we obtain

$$\eta''(w) = \int_{-\infty}^{0} s^2 e^{-s^2/2-sw}ds,$$

and therefore

$$\frac{\partial G(w,t)}{\partial w} = \begin{cases} -\eta''(w)\Phi(-t) & \text{when } w < t, \\ -\eta''(-w)\Phi(t) & \text{when } w > t. \end{cases}$$

Hence, for any fixed t, the function $G(w,t)$ is decreasing in w for $w < t$ and $w > t$, and, moreover, satisfies

$$\lim_{w\to-\infty} G(w,t) = 0, \qquad \lim_{w\to\infty} G(w,t) = 0,$$

and

$$\lim_{w\uparrow t} G(w,t) = -\eta(t)\Phi(-t) < 0 \quad \text{and} \quad \lim_{w\downarrow t} G(w,t) = \eta'(-t)\Phi(t) > 0.$$

Now, from (2.94), it follows that

$$\big|G(w,t) - G(v,t)\big| \le \eta'(t)\Phi(-t) + \eta'(-t)\Phi(t) = 1,$$

and hence

$$\begin{aligned} \big|f'(w) - f'(v)\big| &= \left|\frac{1}{\alpha}\int_z^{z+\alpha} \big[G(w,t) - G(v,t)\big]dt\right| \\ &\le \frac{1}{\alpha}\int_z^{z+\alpha} \big|G(w,t) - G(v,t)\big|\,dt \\ &\le \frac{1}{\alpha}\int_z^{z+\alpha} 1\,dt = 1. \end{aligned}$$

Lastly, to demonstrate (2.16), we apply the mean value theorem and the first two bounds in (2.15) to write

$$\begin{aligned} &\big|f'(w+v) - f'(w)\big| \\ &\quad = \big|vf(w+v) + w\big(f(w+v) - f(w)\big) + h(w+v) - h(w)\big| \\ &\quad \le |v|\left(1 + |w| + \frac{1}{\alpha}\int_0^1 \mathbf{1}_{[z,z+\alpha]}(w+rv)dr\right). \end{aligned}$$

□