at an exponential rate. Similarly to (3.46), this relation remains valid if $\delta = \delta_\varepsilon$ depends on ε and converges to 0 slowly enough as $\varepsilon \to 0$. This means that almost all the mass of the prior distribution $I\!P_\mu$ is concentrated in a small (asymptotically shrinking) neighborhood of the boundary $\{\theta : \sum_k a_k^2 \theta_k^2 = Q\}$ of the ellipsoid $\Theta(\beta, Q)$. The values θ in this neighborhood can be viewed as being the least favorable, i.e., the hardest to estimate. Since the neighborhood depends on ε, the least favorable values θ are different for different ε. Even more, one can show that there exist no fixed (that is, independent of ε) θ^* belonging to the ellipsoid $\Theta(\beta, Q)$ and such that

$$\mathbf{E}_{\theta^*} \|\hat{\theta}(\ell^*) - \theta^*\|^2 = C^* \varepsilon^{\frac{4\beta}{2\beta+1}} (1 + o(1)), \quad \varepsilon \to 0.$$

We will come back to this property in Section 3.8.

3.4 Stein's phenomenon

In this section we temporarily switch to the parametric Gaussian models, and discuss some notions related to Stein's phenomenon. This material plays an auxiliary role. It will be helpful for further constructions in the chapter. Consider the following two Gaussian models.

Model 1

This is a truncated version of the Gaussian sequence model:

$$y_j = \theta_j + \varepsilon \xi_j, \quad j = 1, \dots, d,$$

where $\varepsilon > 0$ and ξ_j are i.i.d. $\mathcal{N}(0, 1)$ random variables. In this section we will denote by y, θ, and ξ the following d-dimensional vectors:

$$y = (y_1, \dots, y_d), \quad \theta = (\theta_1, \dots, \theta_d), \quad \xi = (\xi_1, \dots, \xi_d) \sim \mathcal{N}_d(0, I),$$

where $\mathcal{N}_d(0, I)$ stands for the standard d-dimensional normal distribution. Then we can write

$$y = \theta + \varepsilon \xi, \quad \xi \sim \mathcal{N}_d(0, I). \tag{3.50}$$

The statistical problem is to estimate the unknown parameter $\theta \in \mathbf{R}^d$.

Model 2

We observe random vectors $X_1, \dots, X_n$ satisfying

$$X_i = \theta + \eta_i, \ i = 1, \dots, n,$$

with $\theta \in \mathbf{R}^d$ where η_i are i.i.d. Gaussian vectors with distribution $\mathcal{N}_d(0, I)$. The statistical problem is to estimate θ. The vector $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ is a sufficient statistic in this model. We can write

$$\bar{X} = \theta + \frac{1}{\sqrt{n}}\,\xi = \theta + \varepsilon\xi$$

with

$$\varepsilon = \frac{1}{\sqrt{n}} \quad \text{and} \quad \xi = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \eta_i \sim \mathcal{N}_d(0, I).$$

Throughout this section $\mathbf{E}_\theta$ will denote the expectation with respect to the distribution y in Model 1 or with respect to the distribution of $\bar{X}$ in Model 2, and $\|\cdot\|$ will denote the Euclidean norm in $\mathbf{R}^d$. In what follows, we will write $\|\theta\|$ to denote either the $\ell^2(\mathbf{N})$-norm or the Euclidean norm on $\mathbf{R}^d$ of the vector θ according to whether $\theta \in \ell^2(\mathbf{N})$ or $\theta \in \mathbf{R}^d$.

Model 1 with $\varepsilon = 1/\sqrt{n}$ is equivalent to Model 2 in the following sense: for any Borel function $\hat{\theta} : \mathbf{R}^d \to \mathbf{R}^d$ the squared risk $\mathbf{E}_\theta\|\hat{\theta}(y) - \theta\|^2$ of the estimator $\hat{\theta}(y)$ in Model 1 with $\varepsilon = 1/\sqrt{n}$ is equal to the risk $\mathbf{E}_\theta\|\hat{\theta}(\bar{X}) - \theta\|^2$ of the estimator $\hat{\theta}(\bar{X})$ in Model 2.

Model 1 is a useful building block in the context of nonparametric estimation, as we will see later. On the other hand, Model 2 is classical for parametric statistics. In this section proofs of the results are only given for Model 1. In view of the equivalence, analogous results for Model 2 are obtained as an immediate by-product.

Definition 3.2 *An estimator θ^* of the parameter θ is called* **inadmissible** *on $\Theta \subseteq \mathbf{R}^d$ with respect to the squared risk if there exists another estimator $\hat{\theta}$ such that*

$$\mathbf{E}_\theta\|\hat{\theta} - \theta\|^2 \le \mathbf{E}_\theta\|\theta^* - \theta\|^2 \quad \textit{for all } \theta \in \Theta,$$

and there exists $\theta_0 \in \Theta$ such that

$$\mathbf{E}_{\theta_0}\|\hat{\theta} - \theta_0\|^2 < \mathbf{E}_{\theta_0}\|\theta^* - \theta_0\|^2.$$

Otherwise, the estimator θ^* is called **admissible**.

The squared risk of the estimator $\bar{X}$ in Model 2 is given by

$$\mathbf{E}_\theta\|\bar{X} - \theta\|^2 = \frac{d}{n} = d\varepsilon^2, \quad \forall\ \theta \in \mathbf{R}^d.$$

This risk is therefore constant as a function of θ.

Stein (1956) considered Model 2 and showed that if $d \ge 3$, then the estimator $\bar{X}$ is inadmissible. This property is known as *Stein's phenomenon*. Moreover, Stein proposed an estimator whose risk is smaller than that of $\bar{X}$ everywhere on $\mathbf{R}^d$ if $d \ge 3$. This improved estimator is based on a shrinkage of $\bar{X}$ towards the origin with a shrinkage factor that depends on $\|\bar{X}\|$.

3.4.1 Stein's shrinkage and the James–Stein estimator

We now explain the idea of Stein's shrinkage for Model 1. The argument for Model 2 is analogous and we omit it. We start with two preliminary lemmas.

Lemma 3.6 (Stein's lemma). *Suppose that a function $f : \mathbf{R}^d \to \mathbf{R}$ satisfies:*

(i) $f(u_1, \dots, u_d)$ is absolutely continuous in each coordinate u_i for almost all values (with respect to the Lebesgue measure on $\mathbf{R}^{d-1}$) of other coordinates $(u_j, j \neq i)$,

(ii)

$$\mathbf{E}_\theta \left| \frac{\partial f(y)}{\partial y_i} \right| < \infty, \qquad i = 1, \dots, d.$$

Then

$$\mathbf{E}_\theta \left[(\theta_i - y_i) f(y) \right] = -\varepsilon^2 \mathbf{E}_\theta \left[\frac{\partial f}{\partial y_i}(y) \right], \qquad i = 1, \dots, d.$$

Proof. We will basically use integration by parts with a slight modification due to the fact that the function f is not differentiable in the standard sense.

Observe first that it is sufficient to prove the lemma for $\theta = 0$ and $\varepsilon = 1$. Indeed, the random vector $\zeta = \varepsilon^{-1}(y - \theta)$ has distribution $\mathcal{N}_d(0, I)$. Hence, for $\tilde{f}(y) = f(\varepsilon y + \theta)$ we have

$$\mathbf{E}_\theta \left[\varepsilon^{-1} (\theta_i - y_i) f(y) \right] = -\mathbf{E} \left[\zeta_i \tilde{f}(\zeta) \right], \qquad \mathbf{E} \left[\frac{\partial f}{\partial \zeta_i}(\zeta) \right] = \varepsilon \mathbf{E} \left[\frac{\partial \tilde{f}}{\partial \zeta_i}(\zeta) \right],$$

where $\zeta_1, \dots, \zeta_d$ are the coordinates of ζ. It is clear that f satisfies assumption (ii) of the lemma if and only if $\tilde{f}$ satisfies the inequality

$$\mathbf{E} \left| \frac{\partial \tilde{f}(\zeta)}{\partial \zeta_i} \right| < \infty, \qquad i = 1, \dots, d, \tag{3.51}$$

where $\zeta \sim \mathcal{N}_d(0, I)$. Therefore it is sufficient to prove that for any function $\tilde{f}$ satisfying (3.51) and assumption (i) of the lemma we have

$$\mathbf{E}[\zeta_i \tilde{f}(\zeta)] = \mathbf{E} \left[\frac{\partial \tilde{f}}{\partial \zeta_i}(\zeta) \right], \qquad i = 1, \dots, d. \tag{3.52}$$

Without loss of generality, it is enough to prove (3.52) for $i = 1$ only. To do this, it suffices to show that, almost surely,

$$\mathbf{E} \left[\zeta_1 \tilde{f}(\zeta) | \zeta_2, \dots, \zeta_d \right] = \mathbf{E} \left[\frac{\partial \tilde{f}}{\partial \zeta_1}(\zeta) \Big| \ \zeta_2, \dots, \zeta_d \right]. \tag{3.53}$$

Since the variables ζ_j are mutually independent with distribution $\mathcal{N}(0, 1)$, equality (3.53) will be proved if we show that for almost all $\zeta_2, \dots, \zeta_d$ with respect to the Lebesgue measure on $\mathbf{R}^{d-1}$ we have

$$\int_{-\infty}^{\infty} u\tilde{f}(u,\zeta_2,\dots,\zeta_d)e^{-u^2/2}du = \int_{-\infty}^{\infty} \frac{\partial \tilde{f}}{\partial u}(u,\zeta_2,\dots,\zeta_d)e^{-u^2/2}du.$$

Put $h(u) = \tilde{f}(u,\zeta_2,\dots,\zeta_d)$. In order to complete the proof, it remains to show that for any absolutely continuous function $h : \mathbf{R} \to \mathbf{R}$ such that

$$\int_{-\infty}^{\infty} |h'(u)|e^{-u^2/2}du < \infty,$$

we have

$$\int_{-\infty}^{\infty} uh(u)e^{-u^2/2}du = \int_{-\infty}^{\infty} h'(u)e^{-u^2/2}du. \tag{3.54}$$

To show (3.54) note first that

$$e^{-u^2/2} = \begin{cases} \int_u^{\infty} ze^{-z^2/2}dz, & \text{if } u > 0, \\ -\int_{-\infty}^{u} ze^{-z^2/2}dz, & \text{if } u < 0. \end{cases}$$

Therefore,

$$\begin{aligned} \int_{-\infty}^{\infty} h'(u)e^{-u^2/2}du &= \int_0^{\infty} h'(u)\Big[\int_u^{\infty} ze^{-z^2/2}dz\Big]du \\ &\quad - \int_{-\infty}^{0} h'(u)\Big[\int_{-\infty}^{u} ze^{-z^2/2}dz\Big]du \\ &= \int_0^{\infty} ze^{-z^2/2}\Big[\int_0^{z} h'(u)du\Big]dz \\ &\quad - \int_{-\infty}^{0} ze^{-z^2/2}\Big[\int_z^{0} h'(u)du\Big]dz \\ &= \left(\int_0^{\infty} + \int_{-\infty}^{0}\right)\{ze^{-z^2/2}[h(z) - h(0)]\}dz \\ &= \int_{-\infty}^{\infty} zh(z)e^{-z^2/2}dz \end{aligned}$$

implying (3.54). ∎

Lemma 3.7 *Let $d \geq 3$. Then, for all $\theta \in \mathbf{R}^d$,*

$$0 < \mathbf{E}_\theta\left(\frac{1}{\|y\|^2}\right) < \infty.$$

Proof. By (3.50), we have

$$\mathbf{E}_\theta\left(\frac{1}{\|y\|^2}\right) = \frac{1}{\varepsilon^2}\mathbf{E}\left(\frac{1}{\|\varepsilon^{-1}\theta + \xi\|^2}\right),$$

where $\xi \sim \mathcal{N}_d(0, I)$ is a standard Gaussian d-dimensional vector. Since the distribution $\mathcal{N}_d(0, I)$ is spherically symmetric,

$$\forall\, v, v' \in \mathbf{R}^d :\ \|v\| = \|v'\| \implies \|\xi + v\| \stackrel{\mathcal{D}}{=} \|\xi + v'\|, \tag{3.55}$$

where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution. Indeed, since the norms of v and v' are equal, there exists an orthogonal matrix Γ such that $v' = \Gamma v$. Since $\Gamma\xi \stackrel{\mathcal{D}}{=} \xi$, we obtain (3.55). In particular,

$$\mathbf{E}\left(\frac{1}{\|\varepsilon^{-1}\theta + \xi\|^2}\right) = \mathbf{E}\left(\frac{1}{\|v_0 + \xi\|^2}\right)$$

with $v_0 = (\|\theta\|/\varepsilon, 0, \ldots, 0)$. On the other hand,

$$\begin{aligned}\mathbf{E}\left(\frac{1}{\|v_0 + \xi\|^2}\right) &= \frac{1}{(\sqrt{2\pi})^d}\int_{\mathbf{R}^d} \exp\left(-\frac{\|x\|^2}{2}\right)\|v_0 + x\|^{-2}dx \\ &= \frac{1}{(\sqrt{2\pi})^d}\exp\left(-\frac{\|\theta\|^2}{2\varepsilon^2}\right)\ \times \\ &\qquad \int_{\mathbf{R}^d}\exp\left(\frac{u_1\|\theta\|}{\varepsilon} - \frac{\|u\|^2}{2}\right)\|u\|^{-2}du\end{aligned}$$

with $u = (u_1, \ldots, u_d)$. Since $xy \le 3x^2 + y^2/3$ for $x \ge 0, y \ge 0$, we have $|u_1|\|\theta\|/\varepsilon \le 3\|\theta\|^2/\varepsilon^2 + \|u\|^2/3$. Then

$$\mathbf{E}\left(\frac{1}{\|v_0 + \xi\|^2}\right) \le \frac{1}{(\sqrt{2\pi})^d}\exp\left(\frac{5\|\theta\|^2}{2\varepsilon^2}\right)\int_{\mathbf{R}^d}\exp\left(-\frac{\|u\|^2}{6}\right)\|u\|^{-2}du.$$

We complete the proof by observing that if $d \ge 3$, there exists a constant $C > 0$ such that

$$\int_{\mathbf{R}^d}\exp\left(-\frac{\|u\|^2}{6}\right)\|u\|^{-2}du = C\int_0^\infty e^{-r^2/6}r^{d-3}dr < \infty. \qquad \blacksquare$$

Stein introduced the class of estimators of the form

$$\hat\theta = g(y)y, \tag{3.56}$$

where $g : \mathbf{R}^d \to \mathbf{R}$ is a function to be chosen. The coordinates of the vector $\hat\theta = (\hat\theta_1, \ldots, \hat\theta_d)$ have the form

$$\hat\theta_j = g(y)y_j.$$

On the other hand, the random vector y is a natural estimator of θ, similar to the arithmetic mean $\bar X$ in Model 2. The risk of this estimator equals

$$\mathbf{E}_\theta\|y - \theta\|^2 = d\varepsilon^2.$$

Let us look for a function g such that the risk of the estimator $\hat{\theta} = g(y)y$ is smaller than that of y. We have

$$\begin{aligned}\mathbf{E}_\theta\|\hat{\theta}-\theta\|^2 &= \sum_{i=1}^d \mathbf{E}_\theta\left[(g(y)y_i-\theta_i)^2\right] \\ &= \sum_{i=1}^d \Big\{\mathbf{E}_\theta\left[(y_i-\theta_i)^2\right] + 2\mathbf{E}_\theta\Big[(\theta_i-y_i)(1-g(y))y_i\Big] \\ &\qquad + \mathbf{E}_\theta\Big[y_i^2(1-g(y))^2\Big]\Big\}.\end{aligned}$$

Suppose now that the function g is such that the assumptions of Lemma 3.6 hold for the functions $f = f_i$ where $f_i(y) = (1-g(y))y_i$, $i = 1,\dots,d$. Then

$$\mathbf{E}_\theta[(\theta_i-y_i)(1-g(y))y_i] = -\varepsilon^2\mathbf{E}_\theta\left[1-g(y)-y_i\frac{\partial g}{\partial y_i}(y)\right],$$

and

$$\mathbf{E}_\theta\left[(\hat{\theta}_i-\theta_i)^2\right] = \varepsilon^2 - 2\varepsilon^2\mathbf{E}_\theta\left[1-g(y)-y_i\frac{\partial g}{\partial y_i}(y)\right] + \mathbf{E}_\theta\Big[y_i^2(1-g(y))^2\Big].$$

Summing over i gives

$$\mathbf{E}_\theta\|\hat{\theta}-\theta\|^2 = d\varepsilon^2 + \mathbf{E}_\theta[W(y)] \tag{3.57}$$

with

$$W(y) = -2\varepsilon^2 d(1-g(y)) + 2\varepsilon^2\sum_{i=1}^d y_i\frac{\partial g}{\partial y_i}(y) + \|y\|^2(1-g(y))^2.$$

The above argument is summarized in the following way.

Lemma 3.8 (Stein's unbiased risk estimator). *Consider Model 1 with $d \geq 3$ and the estimator $\hat{\theta}$ defined in (3.56). Let the assumptions of Lemma 3.6 be fulfilled for the functions $f = f_i$ where $f_i(y) = (1-g(y))y_i$, $i = 1,\dots,d$. Then an unbiased estimator of the risk $\mathbf{E}_\theta\|\hat{\theta}-\theta\|^2$ is given by the formula*

$$\mathrm{SURE} = \varepsilon^2 d(2g(y)-1) + 2\varepsilon^2\sum_{i=1}^d y_i\frac{\partial g}{\partial y_i}(y) + \|y\|^2(1-g(y))^2.$$

Here SURE stands for *Stein's unbiased risk estimator*. Note that the result of Lemma 3.8 is of the same type as those obtained in Section 1.4 for unbiased estimators of the risk of kernel density estimators.

The risk of $\hat{\theta}$ is smaller than that of y if we choose g such that

$$\mathbf{E}_\theta[W(y)] < 0.$$

In order to satisfy this inequality, Stein suggested to search for g among the functions of the form

$$g(y) = 1 - \frac{c}{\|y\|^2}$$

with an appropriately chosen constant $c > 0$. If g has this form, the functions f_i defined by $f_i(y) = (1 - g(y))y_i$ satisfy the assumptions of Lemma 3.6, and (3.57) holds with

$$\begin{aligned} W(y) &= -2\varepsilon^2 d\frac{c}{\|y\|^2} + 2\varepsilon^2 \sum_{i=1}^{d} y_i^2 \frac{2c}{\|y\|^4} + \frac{c^2}{\|y\|^2} \\ &= \frac{1}{\|y\|^2}\Big(-2dc\varepsilon^2 + 4\varepsilon^2 c + c^2\Big). \end{aligned} \tag{3.58}$$

The minimizer in c of (3.58) is equal to

$$c_{opt} = \varepsilon^2(d-2).$$

The function g and the estimator $\hat{\theta} = g(y)y$ associated to this choice of g are given by

$$g(y) = 1 - \frac{\varepsilon^2(d-2)}{\|y\|^2},$$

and

$$\hat{\theta}_{JS} = \left(1 - \frac{\varepsilon^2(d-2)}{\|y\|^2}\right) y, \tag{3.59}$$

respectively. The statistic $\hat{\theta}_{JS}$ is called the *James–Stein estimator* of θ. If the norm $\|y\|$ is sufficiently large, multiplication of y by $g(y)$ shrinks the value of y to 0. This is called the *Stein shrinkage*. If $c = c_{opt}$, then

$$W(y) = -\frac{\varepsilon^4(d-2)^2}{\|y\|^2}. \tag{3.60}$$

For this function W, Lemma 3.7 implies $-\infty < \mathbf{E}_\theta[W(y)] < 0$, provided that $d \geq 3$. Therefore, if $d \geq 3$, the risk of the James–Stein estimator satisfies

$$\mathbf{E}_\theta\|\hat{\theta}_{JS} - \theta\|^2 = d\varepsilon^2 - \mathbf{E}_\theta\left(\frac{\varepsilon^4(d-2)^2}{\|y\|^2}\right) < \mathbf{E}_\theta\|y - \theta\|^2$$

for all $\theta \in \mathbf{R}^d$.

Conclusion: If $d \geq 3$, the James–Stein estimator $\hat{\theta}_{JS}$ (which is biased) is better than the (unbiased) estimator y for all $\theta \in \mathbf{R}^d$ and therefore the estimator y is not admissible in Model 1.

The James–Stein estimator for Model 2 is obtained in a similar way; we just need to replace y by $\bar{X}$ and ε by $1/\sqrt{n}$ in (3.59):

$$\hat{\theta}_{JS} = \left(1 - \frac{d-2}{n\|\bar{X}\|^2}\right)\bar{X}. \tag{3.61}$$

Since Models 1 and 2 are equivalent, (3.61) is better than the estimator $\bar{X}$ for all $\theta \in \mathbf{R}^d$ when $d \geq 3$. Therefore we have proved the following result.

Theorem 3.3 (Stein's phenomenon). *Let $d \geq 3$. Then the estimator $\hat{\theta} = y$ is inadmissible on $\mathbf{R}^d$ in Model 1 and the estimator $\hat{\theta} = \bar{X}$ is inadmissible on $\mathbf{R}^d$ in Model 2.*

It is interesting to analyze the improvement given by $\hat{\theta}_{JS}$ with respect to y. For $\theta = 0$ the risk of the James–Stein estimator is

$$\mathbf{E}_0\|\hat{\theta}_{JS}\|^2 = d\varepsilon^2 - \varepsilon^4(d-2)^2\mathbf{E}\left(\frac{1}{\|\varepsilon\xi\|^2}\right) = 2\varepsilon^2,$$

since $\mathbf{E}\left(\|\xi\|^{-2}\right) = 1/(d-2)$ (check this as an exercise). Therefore, for $\theta = 0$ the improvement is characterized by the ratio

$$\frac{\mathbf{E}_0\|\hat{\theta}_{JS}\|^2}{\mathbf{E}_0\|y\|^2} = \frac{2}{d}, \tag{3.62}$$

which is a constant independent of ε. On the contrary, for all $\theta \neq 0$ the ratio of the squared risks of $\hat{\theta}_{JS}$ and y tends to 1 as $\varepsilon \to 0$ (cf. Lehmann and Casella (1998), p. 407) making the improvement asymptotically negligible.

3.4.2 Other shrinkage estimators

It follows from (3.58) that there exists a whole family of estimators that are better than y in Model 1 when the dimension d is large enough: It is sufficient to take the constant c in the definition of g so that $-2dc\varepsilon^2 + 4\varepsilon^2 c + c^2 < 0$. For example, if $c = \varepsilon^2 d$, we obtain the *Stein estimator* :

$$\hat{\theta}_S \triangleq \left(1 - \frac{\varepsilon^2 d}{\|y\|^2}\right) y.$$

This estimator is better than y for $d \geq 5$. However, it is worse than $\hat{\theta}_{JS}$ for $d \geq 3$.

Estimators performing even better correspond to nonnegative functions g:

$$g(y) = \left(1 - \frac{c}{\|y\|^2}\right)_+$$

with $c > 0$. For example, taking here $c = \varepsilon^2(d-2)$ and $c = \varepsilon^2 d$ we obtain the *positive part James–Stein estimator* and the *positive part Stein estimator*:

$$\hat{\theta}_{JS+} = \left(1 - \frac{\varepsilon^2(d-2)}{\|y\|^2}\right)_+ y,$$

and

$$\hat{\theta}_{S+} = \left(1 - \frac{\varepsilon^2 d}{\|y\|^2}\right)_+ y$$

respectively.

Lemma 3.9 *For all $d \geq 1$ and all $\theta \in \mathbf{R}^d$,*

$$\mathbf{E}_\theta \|\hat{\theta}_{JS+} - \theta\|^2 < \mathbf{E}_\theta \|\hat{\theta}_{JS} - \theta\|^2, \qquad \mathbf{E}_\theta \|\hat{\theta}_{S+} - \theta\|^2 < \mathbf{E}_\theta \|\hat{\theta}_S - \theta\|^2.$$

A proof of this lemma is given in the Appendix (Lemma A.6).

Thus, the estimators $\hat{\theta}_{JS+}$ and $\hat{\theta}_{S+}$ are better than $\hat{\theta}_{JS}$ and $\hat{\theta}_S$, respectively. Though the four estimators are better than y, they are all inadmissible (since $\hat{\theta}_{JS+}$ and $\hat{\theta}_{S+}$ are inadmissible; see, for example, Lehmann and Casella (1998), p. 357). However, it can be shown that the estimator $\hat{\theta}_{JS+}$ can be improved in the smaller order terms only, so that it is "quite close" to being admissible. We mention also that there exists an admissible estimator of θ, though its construction is more cumbersome than that of $\hat{\theta}_{JS+}$.

Lemma 3.10 *Let $\theta \in \mathbf{R}^d$. For all $d \geq 4$,*

$$\mathbf{E}_\theta \|\hat{\theta}_S - \theta\|^2 \leq \frac{d\varepsilon^2 \|\theta\|^2}{\|\theta\|^2 + d\varepsilon^2} + 4\varepsilon^2 \tag{3.63}$$

and, for all $d \geq 1$,

$$\mathbf{E}_\theta \|\hat{\theta}_{S+} - \theta\|^2 \leq \frac{d\varepsilon^2 \|\theta\|^2}{\|\theta\|^2 + d\varepsilon^2} + 4\varepsilon^2. \tag{3.64}$$

Proof. We first prove (3.63). From (3.57) and (3.58) with $c = \varepsilon^2 d$ we obtain

$$\begin{aligned}\mathbf{E}_\theta \|\hat{\theta}_S - \theta\|^2 &= d\varepsilon^2 + (-2dc\varepsilon^2 + 4\varepsilon^2 c + c^2)\mathbf{E}_\theta \left(\frac{1}{\|y\|^2}\right) \\ &= d\varepsilon^2 - (d^2 - 4d)\varepsilon^4 \mathbf{E}_\theta \left(\frac{1}{\|y\|^2}\right).\end{aligned}$$

By Jensen's inequality,

$$\mathbf{E}_\theta \left(\frac{1}{\|y\|^2}\right) \geq \frac{1}{\mathbf{E}_\theta \|y\|^2} = \frac{1}{\|\theta\|^2 + \varepsilon^2 d}\,.$$

Therefore

$$\mathbf{E}_\theta \|\hat{\theta}_S - \theta\|^2 \leq d\varepsilon^2 - \frac{\varepsilon^4 d(d-4)}{\|\theta\|^2 + \varepsilon^2 d} = \frac{d\varepsilon^2 \|\theta\|^2}{\|\theta\|^2 + \varepsilon^2 d} + \frac{4\varepsilon^4 d}{\|\theta\|^2 + \varepsilon^2 d}$$

implying (3.63).

We now prove (3.64). By Lemma 3.9 and (3.63), it is sufficient to show (3.64) for $d \le 3$. Observe that the function $f(y) = (1 - g(y))y_i$ satisfies the assumptions of Lemma 3.6 if $g(y) = (1 - \varepsilon^2 d/\|y\|^2)_+$. In particular,

$$\frac{\partial g(y)}{\partial y_i} = \frac{2\varepsilon^2 d y_i}{\|y\|^4} I(\|y\|^2 > \varepsilon^2 d).$$

Hence, by formula (3.57),

$$\mathbf{E}_\theta \|\hat\theta_{S+} - \theta\|^2 = d\varepsilon^2 + \mathbf{E}_\theta[W(y)],$$

where

$$\begin{aligned} W(y) &= \Big(\|y\|^2 - 2\varepsilon^2 d\Big) I(\|y\|^2 \le \varepsilon^2 d) + \frac{\varepsilon^4 d(4-d)}{\|y\|^2} I(\|y\|^2 > \varepsilon^2 d) \\ &\le \frac{\varepsilon^4 d(4-d)}{\|y\|^2} I(\|y\|^2 > \varepsilon^2 d). \end{aligned}$$

If $d \le 3$, the last expression is less than or equal to $\varepsilon^2(4-d)$. Therefore, for $d \le 3$,

$$\mathbf{E}_\theta \|\hat\theta_{S+} - \theta\|^2 \le 4\varepsilon^2,$$

implying (3.64). ∎

Two other important types of shrinkage are hard and soft thresholding. If we choose the shrinkage factor in the form $g(y) = I(\|y\| > \tau)$ with some $\tau > 0$, we obtain the *global hard thresholding estimator* of θ in Model 1:

$$\hat\theta_{GHT} = I(\|y\| > \tau) y.$$

At first sight, this thresholding seems very rough: We either keep or "kill" all the observations. Nevertheless, some important properties of the Stein shrinkage are preserved. In particular, if $\tau = c\varepsilon\sqrt{d}$ for a suitably chosen absolute constant $c > 0$, a result similar to Lemma 3.10 remains valid for $\hat\theta_{GHT}$, though with coarser constants (cf. Exercise 3.7). Analogous properties can be proved for the *global soft thresholding estimator*

$$\hat\theta_{GST} = \left(1 - \frac{\tau}{\|y\|}\right)_+ y.$$

One can also consider coordinate-wise rather than global shrinkage of y. The main examples are: the *hard thresholding estimator* $\hat\theta_{HT}$ whose components are equal to

$$\hat\theta_{j,HT} = I(|y_j| > \tilde\tau) y_j\,;$$

the *soft thresholding estimator* $\hat\theta_{ST}$ with the components

$$\hat\theta_{j,ST} = \operatorname{sign}(y_j)\big(|y_j| - \tilde\tau\big)_+ = \left(1 - \frac{\tilde\tau}{|y_j|}\right)_+ y_j\,;$$

and the *nonnegative garotte estimator* $\hat{\theta}_G$ with the components

$$\hat{\theta}_{j,G} = \left(1 - \frac{\tilde{\tau}^2}{y_j^2}\right)_+ y_j \,.$$

Here $\tilde{\tau} > 0$ is a threshold, which usually has the form $\tilde{\tau} = c\varepsilon\sqrt{\log(1/\varepsilon)}$, for a suitable absolute constant $c > 0$.

In either case, the coordinate-wise shrinkage keeps large observations (perhaps, slightly transforming them) and sets others equal to 0. Note that the nonnegative garotte is a particular case of the positive part Stein shrinkage corresponding to $d = 1$.

Finally, the coordinate-wise *linear shrinkage* is equivalent to the Tikhonov regularization:

$$\hat{\theta}_j^{TR} = \frac{y_j}{1 + b_j}$$

where $b_j > 0$ (cf. Section 1.7.3).

3.4.3 Superefficiency

The estimator $\bar{X}$ is asymptotically efficient on $(\mathbf{R}^d, \|\cdot\|)$ in Model 2 in the sense of Definition 2.2 and the estimator y is asymptotically efficient on $(\mathbf{R}^d, \|\cdot\|)$ in Model 1 for $\varepsilon = 1/\sqrt{n}$. In fact, these estimators are not only asymptotically efficient, but also minimax in the nonasymptotic sense for all fixed n (or ε) (cf. Lehmann and Casella (1998), p. 350). In particular, the minimax risk associated to Model 1 is equal to the maximal risk of y:

$$\inf_{\hat{\theta}_\varepsilon} \sup_{\theta \in \mathbf{R}^d} \mathbf{E}_\theta \|\hat{\theta}_\varepsilon - \theta\|^2 = \sup_{\theta \in \mathbf{R}^d} \mathbf{E}_\theta \|y - \theta\|^2 = d\varepsilon^2,$$

where the infimum is over all estimators. So, the maximal risk of any asymptotically efficient estimator in Model 1 is $d\varepsilon^2(1 + o(1))$ as $\varepsilon \to 0$. Estimators with smaller asymptotic risk can be called superefficient. More precisely, the following definition is used.

Definition 3.3 *We say that an estimator θ_ε^* is* **superefficient** *in Model 1 if*

$$\limsup_{\varepsilon \to 0} \frac{\mathbf{E}_\theta \|\theta_\varepsilon^* - \theta\|^2}{d\varepsilon^2} \le 1, \quad \forall\ \theta \in \mathbf{R}^d, \tag{3.65}$$

and if there exists $\theta = \overline{\theta} \in \mathbf{R}^d$ such that the inequality in (3.65) is strict. The points $\overline{\theta}$ satisfying the strict inequality are called **superefficiency points** *of θ_ε^*.*

The remarks after Theorem 3.3 imply that $\hat{\theta}_{JS}$ is superefficient with the only superefficiency point $\overline{\theta} = 0$ for $d \ge 3$. In a similar way, it can be shown that $\hat{\theta}_S$ is superefficient if $d \ge 5$. Using Lemma 3.9 and the remarks preceding it we obtain the following result.