Cluster-Seeking James-Stein Estimators

K. Pavan Srinath
University of Cambridge
pk423@cam.ac.uk

Ramji Venkataramanan
University of Cambridge
ramji.v@eng.cam.ac.uk

July 25, 2017

Abstract

This paper considers the problem of estimating a high-dimensional vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^n$ from a noisy observation. The noise vector is i.i.d. Gaussian with known variance. For a squared-error loss function, the James-Stein (JS) estimator is known to dominate the simple maximum-likelihood (ML) estimator when the dimension n exceeds two. The JS-estimator shrinks the observed vector towards the origin, and the risk reduction over the ML-estimator is greatest for $\boldsymbol{\theta}$ that lie close to the origin. JS-estimators can be generalized to shrink the data towards any target subspace. Such estimators also dominate the ML-estimator, but the risk reduction is significant only when $\boldsymbol{\theta}$ lies close to the subspace. This leads to the question: in the absence of prior information about $\boldsymbol{\theta}$, how do we design estimators that give significant risk reduction over the ML-estimator for a wide range of $\boldsymbol{\theta}$?

In this paper, we propose shrinkage estimators that attempt to infer the structure of $\boldsymbol{\theta}$ from the observed data in order to construct a good attracting subspace. In particular, the components of the observed vector are separated into clusters, and the elements in each cluster shrunk towards a common attractor. The number of clusters and the attractor for each cluster are determined from the observed vector. We provide concentration results for the squared-error loss and convergence results for the risk of the proposed estimators. The results show that the estimators give significant risk reduction over the ML-estimator for a wide range of $\boldsymbol{\theta}$, particularly for large n. Simulation results are provided to support the theoretical claims.

1 Introduction

Consider the problem of estimating a vector of parameters $\boldsymbol{\theta} \in \mathbb{R}^n$ from a noisy observation $\mathbf{y}$ of the form

$$\mathbf{y} = \boldsymbol{\theta} + \mathbf{w}.$$

The noise vector $\mathbf{w} \in \mathbb{R}^n$ is distributed as $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, i.e., its components are i.i.d. Gaussian random variables with mean zero and variance σ^2. We emphasize that $\boldsymbol{\theta}$ is deterministic, so the joint probability density function of $\mathbf{y} = [y_1, \dots, y_n]^T$ for a given $\boldsymbol{\theta}$ is

$$p_{\boldsymbol{\theta}}(\mathbf{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{\|\mathbf{y}-\boldsymbol{\theta}\|^2}{2\sigma^2}}. \tag{1}$$

The performance of an estimator $\hat{\boldsymbol{\theta}}$ is measured using the squared-error loss function given by

$$L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}(\mathbf{y})) := \|\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm. The *risk* of the estimator for a given $\boldsymbol{\theta}$ is the expected value of the loss function:

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) := \mathbb{E}\left[\|\hat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}\|^2\right],$$

where the expectation is computed using the density in (1). The *normalized risk* is $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})/n$.

Applying the maximum-likelihood (ML) criterion to (1) yields the ML-estimator $\hat{\boldsymbol{\theta}}_{ML} = \mathbf{y}$. The ML-estimator is an unbiased estimator, and its risk is $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ML}) = n\sigma^2$. The goal of this paper is to design estimators that give significant risk reduction over $\hat{\boldsymbol{\theta}}_{ML}$ for a wide range of $\boldsymbol{\theta}$, without any prior assumptions about its structure.

In 1961 James and Stein published a surprising result [1], proposing an estimator that uniformly achieves lower risk than $\hat{\boldsymbol{\theta}}_{ML}$ for any $\boldsymbol{\theta} \in \mathbb{R}^n$, for $n \geq 3$. Their estimator $\hat{\boldsymbol{\theta}}_{JS}$ is given by

$$\hat{\boldsymbol{\theta}}_{JS} = \left[1 - \frac{(n-2)\sigma^2}{\|\mathbf{y}\|^2}\right]\mathbf{y}, \tag{2}$$

and its risk is [2, Chapter 5, Thm. 5.1]

$$R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS}\right) = n\sigma^2 - (n-2)^2\sigma^4\mathbb{E}\left[\frac{1}{\|\mathbf{y}\|^2}\right]. \tag{3}$$

Hence for $n \geq 3$,

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS}) < R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{ML}) = n\sigma^2, \quad \forall \boldsymbol{\theta} \in \mathbb{R}^n. \tag{4}$$

An estimator $\hat{\boldsymbol{\theta}}_1$ is said to *dominate* another estimator $\hat{\boldsymbol{\theta}}_2$ if

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_1) \leq R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_2), \quad \forall \boldsymbol{\theta} \in \mathbb{R}^n,$$

with the inequality being strict for at least one $\boldsymbol{\theta}$. Thus (4) implies that the James-Stein estimator (JS-estimator) dominates the ML-estimator. Unlike the ML-estimator, the JS-estimator is non-linear and biased. However, the risk reduction over the ML-estimator can be significant, making it an attractive option in many situations — see, for example, [3].

By evaluating the expression in (3), it can be shown that the risk of the JS-estimator depends on $\boldsymbol{\theta}$ only via $\|\boldsymbol{\theta}\|$ [1]. Further, the risk decreases as $\|\boldsymbol{\theta}\|$ decreases. (For intuition about this, note in (3) that for large n, $\|\mathbf{y}\|^2 \approx n\sigma^2 + \|\boldsymbol{\theta}\|^2$.) The dependence of the risk on $\|\boldsymbol{\theta}\|$ is illustrated in Fig. 1, where the average loss of the JS-estimator is plotted versus $\|\boldsymbol{\theta}\|$, for two different choices of $\boldsymbol{\theta}$.

The JS-estimator in (2) shrinks each element of $\mathbf{y}$ towards the origin. Extending this idea, JS-like estimators can be defined by shrinking $\mathbf{y}$ towards any vector, or more generally, towards a target subspace $\mathbb{V} \subset \mathbb{R}^n$. Let $P_{\mathbb{V}}(\mathbf{y})$ denote the projection of $\mathbf{y}$ onto $\mathbb{V}$, so that $\|\mathbf{y} - P_{\mathbb{V}}(\mathbf{y})\|^2 = \min_{\mathbf{v}\in\mathbb{V}} \|\mathbf{y} - \mathbf{v}\|^2$. Then the JS-estimator that shrinks $\mathbf{y}$ towards the subspace $\mathbb{V}$ is

$$\hat{\boldsymbol{\theta}} = P_{\mathbb{V}}(\mathbf{y}) + \left[1 - \frac{(n-d-2)\sigma^2}{\|\mathbf{y} - P_{\mathbb{V}}(\mathbf{y})\|^2}\right](\mathbf{y} - P_{\mathbb{V}}(\mathbf{y})), \tag{5}$$

where d is the dimension of $\mathbb{V}$.[1] A classic example of such an estimator is Lindley's estimator [4], which shrinks $\mathbf{y}$ towards the one-dimensional subspace defined by the all-ones vector $\mathbf{1}$. It is given by

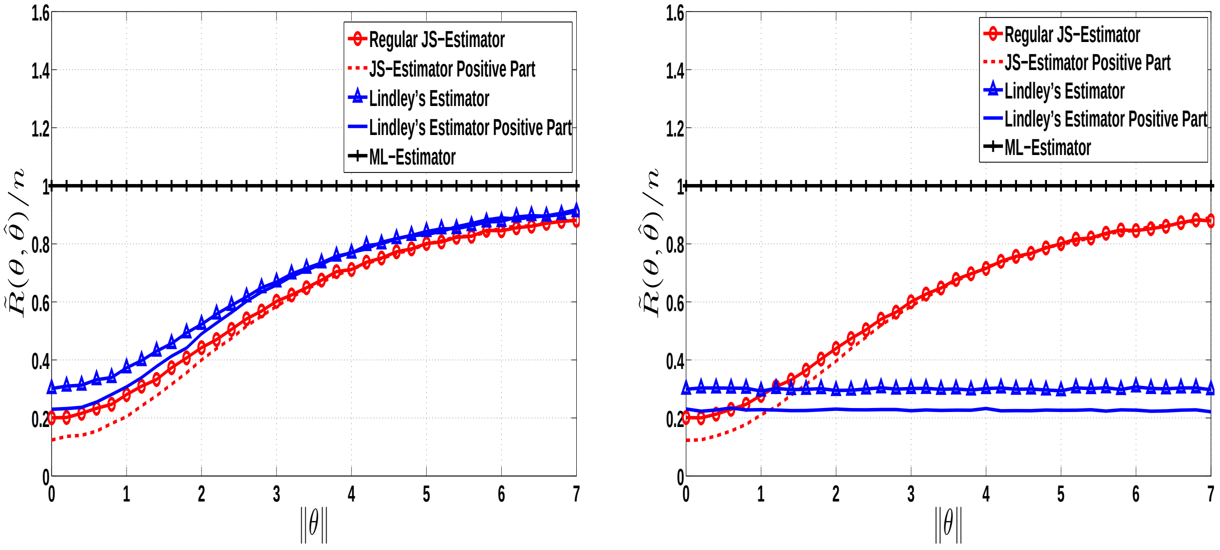
$$\hat{\boldsymbol{\theta}}_L = \bar{y}\mathbf{1} + \left[1 - \frac{(n-3)\sigma^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|^2}\right](\mathbf{y} - \bar{y}\mathbf{1}), \tag{6}$$

where $\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i$ is the empirical mean of $\mathbf{y}$.

It can be shown that the different variants of the JS-estimator such as (2),(5),(6) all dominate the ML-estimator.[2] Further, all JS-estimators share the following key property [6–8]: **the smaller the Euclidean distance between $\boldsymbol{\theta}$ and the attracting vector, the smaller the risk**.

[1] The dimension n has to be greater than $d + 2$ for the estimator to achieve lower risk than $\hat{\boldsymbol{\theta}}_{ML}$.

[2] The risks of JS-estimators of the form (5) can usually be computed using Stein's lemma [5], which states that $\mathbb{E}[Xg(X)] = E[g'(X)]$, where X is a standard normal random variable, and g a weakly differentiable function.



(a)

(b)

Figure 1: Comparison of the average normalized loss of the regular JS-estimator, Lindley's estimator, and their positive-part versions for $n = 10$ as a function of $\|\boldsymbol{\theta}\|$. The loss of the ML-estimator is $\sigma^2 = 1$. In the left panel, $\theta_i = \|\boldsymbol{\theta}\|/\sqrt{10}$, $i = 1, \cdots, 5$, and $\theta_i = -\|\boldsymbol{\theta}\|/\sqrt{10}$, $i = 6, \cdots, 10$. On the right, $\theta_i = \|\boldsymbol{\theta}\|/\sqrt{10}$, $\forall i$.

Throughout this paper, the term "attracting vector" refers to the vector that $\mathbf{y}$ is shrunk towards. For $\hat{\boldsymbol{\theta}}_{JS}$ in (2), the attracting vector is $\mathbf{0}$, and the risk reduction over $\hat{\boldsymbol{\theta}}_{ML}$ is larger when $\|\boldsymbol{\theta}\|$ is close to zero. Similarly, if the components of $\boldsymbol{\theta}$ are clustered around some value c, a JS-estimator with attracting vector $c\mathbf{1}$ would give significant risk reduction over $\hat{\boldsymbol{\theta}}_{ML}$. One motivation for Lindley's estimator in (6) comes from a guess that the components of $\boldsymbol{\theta}$ are close to its empirical mean $\bar{\theta}$ — since we do not know $\bar{\theta}$, we approximate it by $\bar{y}$ and use the attracting vector $\bar{y}\mathbf{1}$.

Fig. 1 shows how the performance of $\hat{\boldsymbol{\theta}}_{JS}$ and $\hat{\boldsymbol{\theta}}_{L}$ depends on the structure of $\boldsymbol{\theta}$. In the left panel of the figure, the empirical mean $\bar{\theta}$ is always 0, so the risks of both estimators increase monotonically with $\|\boldsymbol{\theta}\|$. In the right panel, all the components of $\boldsymbol{\theta}$ are all equal to $\bar{\theta}$. In this case, the distance from the attracting vector for $\hat{\boldsymbol{\theta}}_{L}$ is $\|\boldsymbol{\theta} - \bar{y}\mathbf{1}\| = \sqrt{(\sum_{i=1}^{n} w_i)^2/n}$, so the risk does not vary with $\|\boldsymbol{\theta}\|$; in contrast the risk of $\hat{\boldsymbol{\theta}}_{JS}$ increases with $\|\boldsymbol{\theta}\|$ as its attracting vector is $\mathbf{0}$.

The risk reduction obtained by using a JS-like shrinkage estimator over $\hat{\boldsymbol{\theta}}_{ML}$ crucially depends on the choice of attracting vector. To achieve significant risk reduction for a wide range of $\boldsymbol{\theta}$, in this paper, we infer the structure of $\boldsymbol{\theta}$ from the data $\mathbf{y}$ and choose attracting vectors tailored to this structure. The idea is to partition $\mathbf{y}$ into clusters, and shrink the components in each cluster towards a common element (attractor). Both the number of clusters and the attractor for each cluster are to be determined based on the data $\mathbf{y}$.

As a motivating example, consider a $\boldsymbol{\theta}$ in which half the components are equal to $\|\boldsymbol{\theta}\|/\sqrt{n}$ and the other half are equal to $-\|\boldsymbol{\theta}\|/\sqrt{n}$. Fig. 1(a) shows that the risk reduction of both $\hat{\boldsymbol{\theta}}_{JS}$ and $\hat{\boldsymbol{\theta}}_{L}$ diminish as $\|\boldsymbol{\theta}\|$ gets larger. This is because the empirical mean $\bar{y}$ is close to zero, hence $\hat{\boldsymbol{\theta}}_{JS}$ and $\hat{\boldsymbol{\theta}}_{L}$ both shrink $\mathbf{y}$ towards $\mathbf{0}$. An ideal JS-estimator would shrink the y_i's corresponding to $\theta_i = \|\boldsymbol{\theta}\|/\sqrt{n}$ towards the attractor $\|\boldsymbol{\theta}\|/\sqrt{n}$, and the remaining observations towards $-\|\boldsymbol{\theta}\|/\sqrt{n}$. Such an estimator would give handsome gains over $\hat{\boldsymbol{\theta}}_{ML}$ for all $\boldsymbol{\theta}$ with the above structure. On the other hand, if $\boldsymbol{\theta}$ is such that all its components are equal (to $\bar{\theta}$), Lindley's estimator $\hat{\boldsymbol{\theta}}_{L}$ is an excellent choice, with significantly smaller risk than $\hat{\boldsymbol{\theta}}_{ML}$ for all values of $\|\boldsymbol{\theta}\|$ (Fig. 1(b)).

We would like an intelligent estimator that can correctly distinguish between different $\boldsymbol{\theta}$ structures (such as the two above) and choose an appropriate attracting vector, based only on $\mathbf{y}$. We propose such estimators in Sections 3 and 4. For reasonably large n, these estimators choose a good attracting subspace tailored to the structure of $\boldsymbol{\theta}$, and use an approximation of the best attracting vector within the subspace.

The main contributions of our paper are as follows.

- We construct a two-cluster JS-estimator, and provide concentration results for the squared-error loss, and asymptotic convergence results for its risk. Though this estimator does not dominate the ML-estimator, it is shown to provide significant risk reduction over Lindley's estimator and the regular JS-estimator when the components of $\boldsymbol{\theta}$ can be approximately separated into two clusters.
- We present a hybrid JS-estimator that, for any $\boldsymbol{\theta}$ and for large n, has risk close to the minimum of that of Lindley's estimator and the proposed two-cluster JS-estimator. Thus the hybrid estimator asymptotically dominates both the ML-estimator and Lindley's estimator, and gives significant risk reduction over the ML-estimator for a wide range of $\boldsymbol{\theta}$.
- We generalize the above idea to define general multiple-cluster hybrid JS-estimators, and provide concentration and convergence results for the squared-error loss and risk, respectively.
- We provide simulation results that support the theoretical results on the loss function. The simulations indicate that the hybrid estimator gives significant risk reduction over the ML-estimator for a wide range of $\boldsymbol{\theta}$ even for modest values of n, e.g. $n = 50$. The empirical risk of the hybrid estimator converges rapidly to the theoretical value with growing n.

1.1 Related work

George [7, 8] proposed a"multiple shrinkage estimator", which is a convex combination of multiple subspace-based JS-estimators of the form (5). The coefficients defining the convex combination give larger weight to the estimators whose target subspaces are closer to $\mathbf{y}$. Leung and Barron [9, 10] also studied similar ways of combining estimators and their risk properties. Our proposed estimators also seek to emulate the best among a class of subspace-based estimators, but there are some key differences. In [7, 8], the target subspaces are fixed a priori, possibly based on prior knowledge about where $\boldsymbol{\theta}$ might lie. In the absence of such prior knowledge, it may not be possible to choose good target subspaces. This motivates the estimators proposed in this paper, which use a target subspace constructed from the data $\mathbf{y}$. The nature of clustering in $\boldsymbol{\theta}$ is inferred from $\mathbf{y}$, and used to define a suitable subspace.

Another difference from earlier work is in how the attracting vector is determined given a target subspace $\mathbb{V}$. Rather than choosing the attracting vector as the projection of $\mathbf{y}$ onto $\mathbb{V}$, we use an approximation of the projection of $\boldsymbol{\theta}$ onto $\mathbb{V}$. This approximation is computed from $\mathbf{y}$, and concentration inequalities are provided to guarantee the goodness of the approximation.

The risk of a JS-like estimator is typically computed using Stein's lemma [5]. However, the data-dependent subspaces we use result in estimators that are hard to analyze using this technique. We therefore use concentration inequalities to bound the loss function of the proposed estimators. Consequently, our theoretical bounds get sharper as the dimension n increases, but may not be accurate for small n. However, even for relatively small n, simulations indicate that the risk reduction over the ML-estimator is significant for a wide range of $\boldsymbol{\theta}$.

Noting that the shrinkage factor multiplying $\mathbf{y}$ in (2) could be negative, Stein proposed the following positive-part JS-estimator [1]:

$$\hat{\boldsymbol{\theta}}_{JS_+} = \left[1 - \frac{(n-2)\sigma^2}{\|\mathbf{y}\|^2}\right]_+ \mathbf{y}, \tag{7}$$

where X_+ denotes $\max(0, X)$. We can similarly define positive-part versions of JS-like estimators such as (5) and (6). The positive-part Lindley's estimator is given by

$$\hat{\boldsymbol{\theta}}_{L_+} = \bar{y}\mathbf{1} + \left[1 - \frac{(n-3)\sigma^2}{\|\mathbf{y} - \bar{y}\mathbf{1}\|^2}\right]_+ (\mathbf{y} - \bar{y}\mathbf{1}). \tag{8}$$

Baranchik [11] proved that $\hat{\boldsymbol{\theta}}_{JS_+}$ dominates $\hat{\boldsymbol{\theta}}_{JS}$, and his result also proves that $\hat{\boldsymbol{\theta}}_{L_+}$ dominates $\hat{\boldsymbol{\theta}}_{L}$. Estimators that dominate $\hat{\boldsymbol{\theta}}_{JS_+}$ are discussed in [12,13]. Fig. 1 shows that the positive-part versions can give noticeably lower loss than the regular JS and Lindley estimators. However, for large n, the shrinkage factor is positive with high probability, hence the positive-part estimator is nearly always identical to the regular JS-estimator. Indeed, for large n, $\frac{\|\mathbf{y}\|^2}{n} \approx \frac{\|\boldsymbol{\theta}\|^2}{n} + \sigma^2$, and the shrinkage factor is

$$\left(1 - \frac{(n-2)\sigma^2}{\|\mathbf{y}\|^2}\right) \approx \left(1 - \frac{(n-2)\sigma^2}{\|\boldsymbol{\theta}\|^2 + n\sigma^2}\right) > 0.$$

We analyze the positive-part version of the proposed hybrid estimator using concentration inequalities. Though we cannot guarantee that the hybrid estimator dominates the positive-part JS or Lindley estimators for any finite n, we show that for large n, the loss of the hybrid estimator is equal to the minimum of that of the positive-part Lindley's estimator and the cluster-based estimator with high probability (Theorems 3 and 4).

The rest of the paper is organized as follows. In Section 2, a two-cluster JS-estimator is proposed and its performance analyzed. Section 3 presents a hybrid JS-estimator along with its performance analysis. General multiple-attractor JS-estimators are discussed in Section 4, and simulation results to corroborate the theoretical analysis are provided in Section 5. The proofs of the main results are given in Section 6. Concluding remarks and possible directions for future research constitute Section 7.

1.2 Notation

Bold lowercase letters are used to denote vectors, and plain lowercase letters for their entries. For example, the entries of $\mathbf{y} \in \mathbb{R}^n$ are y_i, $i = 1, \cdots, n$. All vectors have length n and are column vectors, unless otherwise mentioned. For vectors $\mathbf{y}, \mathbf{z} \in \mathbb{R}^n$, $\langle \mathbf{y}, \mathbf{z} \rangle$ denotes their Euclidean inner product. The all-zero vector and the all-one vector of length n are denoted by $\mathbf{0}$ and $\mathbf{1}$, respectively. The complement of a set A is denoted by A^c. For a finite set A with real-valued elements, $\min(A)$ denotes the minimum of the elements in A. We use $\mathbf{1}_{\{\mathcal{E}\}}$ to denote the indicator function of an event $\mathcal{E}$. A central chi-squared distributed random variable with n degrees of freedom is denoted by $\mathcal{X}_n^2$. The Q-function is given by $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})dx$, and $Q^c(x) := 1 - Q(x)$. For a random variable X, X_+ denotes $\max(0, X)$. For real-valued functions $f(x)$ and $g(x)$, the notation $f(x) = o(g(x))$ means that $\lim_{x\to 0}[f(x)/g(x)] = 0$, and $f(x) = O(g(x))$ means that $\lim_{x\to\infty}[f(x)/g(x)] = c$ for some positive constant c.

For a sequence of random variables $\{X_n\}_{n=1}^\infty$, $X_n \xrightarrow{P} X$, $X_n \xrightarrow{a.s.} X$, and $X_n \xrightarrow{\mathcal{L}^1} X$ respectively denote convergence in probability, almost sure convergence, and convergence in $\mathcal{L}^1$ norm to the random variable X.

We use the following shorthand for concentration inequalities. Let $\{X_n(\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^n\}_{n=1}^{\infty}$ be a sequence of random variables. The notation $X_n(\boldsymbol{\theta}) \doteq X$, where X is either a random variable or a constant, means that for any $\epsilon > 0$,

$$\mathbb{P}\left(|X_n(\boldsymbol{\theta}) - X| \geq \epsilon\right) \leq K e^{-\frac{nk\min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}}, \tag{9}$$

where K and k are positive constants that do not depend on n or $\boldsymbol{\theta}$. The exact values of K and k are not specified.

The shrinkage estimators we propose have the general form

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\nu} + \left[1 - \frac{n\sigma^2}{\|\mathbf{y} - \boldsymbol{\nu}\|^2}\right]_+ (\mathbf{y} - \boldsymbol{\nu}).$$

For $1 \leq i \leq n$, the ith component of the attracting vector $\boldsymbol{\nu}$ is the attractor for y_i (the point towards which it is shrunk).

2 A two-cluster James-Stein estimator

Recall the example in Section 1 where $\boldsymbol{\theta}$ has half its components equal to $\|\boldsymbol{\theta}\|/\sqrt{n}$, and the other half equal to $\|\boldsymbol{\theta}\|/\sqrt{n}$. Ideally, we would like to shrink the y_i's corresponding to the first group towards $\|\boldsymbol{\theta}\|/\sqrt{n}$, and the remaining points towards $-\|\boldsymbol{\theta}\|/\sqrt{n}$. However, without an oracle, we cannot accurately guess which point each y_i should be shrunk towards. We would like to obtain an estimator that identifies separable clusters in $\mathbf{y}$, constructs a suitable attractor for each cluster, and shrinks the y_i in each cluster towards its attractor.

We start by dividing the observed data into two clusters based on a separating point $s_{\mathbf{y}}$, which is obtained from $\mathbf{y}$. A natural choice for the $s_{\mathbf{y}}$ would be the empirical mean $\bar{\theta}$; since this is unknown we use $s_{\mathbf{y}} = \bar{y}$. Define the clusters

$$\mathcal{C}_1 := \{y_i,\ 1 \leq i \leq n \mid y_i > \bar{y}\}, \quad \mathcal{C}_2 := \{y_i,\ 1 \leq i \leq n \mid y_i \leq \bar{y}\}.$$

The points in $\mathcal{C}_1$ and $\mathcal{C}_2$ will be shrunk towards attractors $a_1(\mathbf{y})$ and $a_2(\mathbf{y})$, respectively, where $a_1, a_2 : \mathbb{R}^n \to \mathbb{R}$ are defined in (20) later in this section. For brevity, we henceforth do not indicate the dependence of the attractors on $\mathbf{y}$. Thus the attracting vector is

$$\boldsymbol{\nu}_2 := a_1 \begin{bmatrix} 1_{\{y_1 > \bar{y}\}} \\ 1_{\{y_2 > \bar{y}\}} \\ \vdots \\ 1_{\{y_n > \bar{y}\}} \end{bmatrix} + a_2 \begin{bmatrix} 1_{\{y_1 \leq \bar{y}\}} \\ 1_{\{y_2 \leq \bar{y}\}} \\ \vdots \\ 1_{\{y_n \leq \bar{y}\}} \end{bmatrix}, \tag{10}$$

with a_1 and a_2 defined in (20). The proposed estimator is

$$\hat{\boldsymbol{\theta}}_{JS_2} = \boldsymbol{\nu}_2 + \left[1 - \frac{n\sigma^2}{\|\mathbf{y} - \boldsymbol{\nu}_2\|^2}\right]_+ (\mathbf{y} - \boldsymbol{\nu}_2) = \boldsymbol{\nu}_2 + \left[1 - \frac{\sigma^2}{g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)}\right] (\mathbf{y} - \boldsymbol{\nu}_2), \tag{11}$$

where the function g is defined as

$$g(x) := \max(\sigma^2, x), \quad x \in \mathbb{R}. \tag{12}$$

The attracting vector $\boldsymbol{\nu}_2$ in (10) lies in a two-dimensional subspace defined by the orthogonal vectors $[1_{\{y_1 > \bar{y}\}}, \cdots, 1_{\{y_n > \bar{y}\}}]^T$ and $[1_{\{y_1 \leq \bar{y}\}}, \cdots, 1_{\{y_n \leq \bar{y}\}}]^T$. To derive the values of a_1 and a_2 in (10), it is useful to compare $\boldsymbol{\nu}_2$ to the attracting vector of Lindley's estimator in (6). Recall that Lindley's attracting vector lies in the one-dimensional subspace spanned by $\mathbf{1}$. The vector lying in this subspace that is closest in Euclidean distance to $\boldsymbol{\theta}$ is its projection $\bar{\theta}\mathbf{1}$. Since $\bar{\theta}$ is unknown, we use the approximation $\bar{y}$ to define the attracting vector $\bar{y}\mathbf{1}$.

Analogously, the vector in the two-dimensional subspace defined by (10) that is closest to $\boldsymbol{\theta}$ is the projection of $\boldsymbol{\theta}$ onto this subspace. Computing this projection, the desired values for a_1, a_2 are found to be

$$a_1^{des} = \frac{\sum_{i=1}^n \theta_i 1_{\{y_i > \bar{y}\}}}{\sum_{i=1}^n 1_{\{y_i > \bar{y}\}}}, \quad a_2^{des} = \frac{\sum_{i=1}^n \theta_i 1_{\{y_i \leq \bar{y}\}}}{\sum_{i=1}^n 1_{\{y_i \leq \bar{y}\}}}. \tag{13}$$

As the θ_i's are not available, we define the attractors a_1, a_2 as approximations of a_1^{des}, a_2^{des}, obtained using the following concentration results.

Lemma 1. *We have*

$$\frac{1}{n}\sum_{i=1}^n y_i 1_{\{y_i > \bar{y}\}} \doteq \frac{1}{n}\sum_{i=1}^n \theta_i 1_{\{y_i > \bar{y}\}} + \frac{\sigma}{n\sqrt{2\pi}}\sum_{i=1}^n e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}}, \tag{14}$$

$$\frac{1}{n}\sum_{i=1}^n y_i 1_{\{y_i \leq \bar{y}\}} \doteq \frac{1}{n}\sum_{i=1}^n \theta_i 1_{\{y_i \leq \bar{y}\}} - \frac{\sigma}{n\sqrt{2\pi}}\sum_{i=1}^n e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}}, \tag{15}$$

$$\frac{1}{n}\sum_{i=1}^n \theta_i 1_{\{y_i > \bar{y}\}} \doteq \sum_{i=1}^n \theta_i Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right), \quad \frac{1}{n}\sum_{i=1}^n \theta_i 1_{\{y_i \leq \bar{y}\}} \doteq \frac{1}{n}\sum_{i=1}^n \theta_i Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right), \tag{16}$$

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n 1_{\{y_i > \bar{y}\}} - \sum_{i=1}^n Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)\right| \geq \epsilon\right) \leq Ke^{-nk\epsilon^2}, \tag{17}$$

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n 1_{\{y_i \leq \bar{y}\}} - \sum_{i=1}^n Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)\right| \geq \epsilon\right) \leq Ke^{-nk\epsilon^2}. \tag{18}$$

where $Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right) := 1 - Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)$. Recall from Section 1.2 that the symbol $\doteq$ is shorthand for a concentration inequality of the form (9).

The proof is given in Appendix B.1.

Using Lemma 1, we can obtain estimates for a_1^{des}, a_2^{des} in (13) provided we have an estimate for the term $\frac{\sigma}{n\sqrt{2\pi}}\sum_{i=1}^n e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}}$. This is achieved via the following concentration result.

Lemma 2. *Fix $\delta > 0$. Then for any $\epsilon > 0$, we have*

$$\mathbb{P}\left(\left|\frac{\sigma^2}{2n\delta}\sum_{i=0}^n 1_{\{|y_i - \bar{y}| \leq \delta\}} - \left(\frac{\sigma}{n\sqrt{2\pi}}\sum_{i=0}^n e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}} + \kappa_n\delta\right)\right| \geq \epsilon\right) \leq 10e^{-nk\epsilon^2}, \tag{19}$$

where k is some positive constant and $|\kappa_n| \leq \frac{1}{\sqrt{2\pi e}}$.

The proof is given in Appendix B.2.

Note 1. *Henceforth in this paper, κ_n is used to denote a generic bounded constant (whose exact value is not needed) that is a coefficient of δ in expressions of the form $f(\delta) = a + \kappa_n\delta + o(\delta)$ where a is some constant. As an example to illustrate its usage, let $f(\delta) = \frac{1}{a+b\delta}$, where $a > 0$ and $|b\delta| < a$. Then, we have $f(\delta) = \frac{1}{a(1+b\delta/a)} = \frac{1}{a}(1 + \frac{b}{a}\delta + o(\delta)) = \frac{1}{a} + \kappa_n\delta + o(\delta)$.*

Using Lemmas 1 and 2, the two attractors are defined to be

$$a_1 = \frac{\sum_{i=1}^n y_i 1_{\{y_i > \bar{y}\}} - \frac{\sigma^2}{2\delta}\sum_{i=0}^n 1_{\{|y_i - \bar{y}| \leq \delta\}}}{\sum_{i=1}^n 1_{\{y_i > \bar{y}\}}}, \qquad a_2 = \frac{\sum_{i=1}^n y_i 1_{\{y_i \leq \bar{y}\}} + \frac{\sigma^2}{2\delta}\sum_{i=0}^n 1_{\{|y_i - \bar{y}| \leq \delta\}}}{\sum_{i=1}^n 1_{\{y_i \leq \bar{y}\}}}. \tag{20}$$

With $\delta > 0$ chosen to be a small positive number, this completes the specification of the attracting vector in (10), and hence the two-cluster JS-estimator in (11).

Note that $\boldsymbol{\nu}_2$, defined by (10), (20), is an approximation of the projection of $\boldsymbol{\theta}$ onto the two-dimensional subspace $\mathbb{V}$ spanned by the vectors $[1_{y_1 > \bar{y}}, \cdots, 1_{y_n > \bar{y}}]^T$ and $[1_{y_1 \leq \bar{y}}, \cdots, 1_{y_n \leq \bar{y}}]^T$. We remark that $\boldsymbol{\nu}_2$, which approximates the vector in $\mathbb{V}$ that is closest to $\boldsymbol{\theta}$, is distinct from the projection of $\mathbf{y}$ onto $\mathbb{V}$. While the analysis is easier (there would be no terms involving δ) if $\boldsymbol{\nu}_2$ were chosen to be a projection of $\mathbf{y}$ (instead of $\boldsymbol{\theta}$) onto $\mathbb{V}$, our numerical simulations suggest that this choice yields significantly higher risk. The intuition behind choosing the projection of $\boldsymbol{\theta}$ onto $\mathbb{V}$ is that if all the y_i in a group are to be attracted to a common point (without any prior information), a natural choice would be the mean of the θ_i within the group, as in (13). This mean is determined by the term $\mathbb{E}(\sum_{i=1}^n \theta_i 1_{\{y_i \geq \bar{y}\}})$, which is different from $\mathbb{E}(\sum_{i=1}^n y_i 1_{\{y_i \geq \bar{y}\}})$ because

$$\mathbb{E}\left(\sum_{i=1}^n (y_i - \theta_i) 1_{\{y_i \geq \bar{y}\}}\right) = \mathbb{E}\left(\sum_{i=1}^n w_i 1_{\{y_i \geq \bar{y}\}}\right) \neq 0.$$

The term involving δ in (20) approximates $\mathbb{E}(\sum_{i=1}^n w_i 1_{\{y_i \geq \bar{y}\}})$.

Note 2. *The attracting vector $\boldsymbol{\nu}_2$ is dependent not just on $\mathbf{y}$ but also on δ, through the two attractors a_1 and a_2. In Lemma 2, for the deviation probability in* (19) *to fall exponentially in n, δ needs to be held constant and independent of n. From a practical design point of view, what is needed is $n\delta^2 \gg 1$. Indeed , for $\frac{\sigma^2}{2n\delta}\sum_{i=0}^n 1_{\{|y_i - \bar{y}| \leq \delta\}}$ to be a reliable approximation of the term $\frac{\sigma}{n\sqrt{2\pi}}\sum_{i=1}^n e^{-\frac{(\bar{\theta} - \theta_i)^2}{2\sigma^2}}$, it is shown in Appendix B.2, specifically in* (99) *that we need $n\delta^2 \gg 1$. Numerical experiments suggest a value of $5/\sqrt{n}$ for δ to be large enough for a good approximation.*

We now present the first main result of the paper.

Theorem 1. *The loss function of the two-cluster JS-estimator in* (11) *satisfies the following:*

(1) For any $\epsilon > 0$, and for any fixed $\delta > 0$ that is independent of n,

$$\mathbb{P}\left(\left|\frac{1}{n}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_2}\|^2 - \left[\min\left(\beta_n, \frac{\beta_n \sigma^2}{\alpha_n + \sigma^2}\right) + \kappa_n \delta + o(\delta)\right]\right| \geq \epsilon\right) \leq K e^{-\frac{nk\min(\epsilon^2, 1)}{\max(\|\boldsymbol{\theta}\|^2/n, 1)}}, \tag{21}$$

where α_n, β_n are given by (24) *and* (23) *below, and K is a positive constant that is independent of n and δ, while $k = \Theta(\delta^2)$ is another positive constant that is independent of n (for a fixed δ).*

(2) For a sequence of $\boldsymbol{\theta}$ with increasing dimension n, if $\limsup_{n\to\infty} \|\boldsymbol{\theta}\|^2/n < \infty$, we have

$$\lim_{n\to\infty}\left|\frac{1}{n}R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}\right) - \left[\min\left(\beta_n, \frac{\beta_n \sigma^2}{\alpha_n + \sigma^2}\right) + \kappa_n \delta + o(\delta)\right]\right| = 0. \tag{22}$$

The constants β_n, α_n are given by

$$\beta_n := \frac{\|\boldsymbol{\theta}\|^2}{n} - \frac{c_1^2}{n}\sum_{i=1}^n Q\left(\frac{\bar{\theta} - \theta_i}{\sigma}\right) - \frac{c_2^2}{n}\sum_{i=1}^n Q^c\left(\frac{\bar{\theta} - \theta_i}{\sigma}\right), \tag{23}$$

$$\alpha_n := \beta_n - \left(\frac{2\sigma}{n\sqrt{2\pi}}\right)\left(\sum_{i=1}^{n} e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}}\right)(c_1 - c_2), \tag{24}$$

where

$$c_1 := \frac{\sum_{i=1}^n \theta_i Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)}{\sum_{i=1}^n Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)}, \quad c_2 := \frac{\sum_{i=1}^n \theta_i Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)}{\sum_{i=1}^n Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)}. \tag{25}$$

The proof of the theorem is given in Section 6.2.

Remark 1. *In Theorem 1, β_n represents the concentrating value for the distance between $\boldsymbol{\theta}$ and the attracting vector $\boldsymbol{\nu}_2$. (It is shown in Sec. 6.2 that $\|\boldsymbol{\theta} - \boldsymbol{\nu}_2\|^2/n$ concentrates around $\beta_n + \kappa_n\delta$.) Therefore, the closer $\boldsymbol{\theta}$ is to the attracting subspace, the lower the normalized asymptotic risk $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2})/n$. The term $\alpha_n + \sigma^2$ represents the concentrating value for the distance between $\mathbf{y}$ and $\boldsymbol{\nu}_2$. (It is shown in Sec. 6.2 that $\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n$ concentrates around $\alpha_n + \sigma^2 + \kappa_n\delta$.)*

Remark 2. *Comparing β_n in* (23) *and α_n in* (24)*, we note that $\beta_n \geq \alpha_n$ because*

$$c_1 - c_2 = \frac{-n\sum_{i=1}^n(\theta_i - \bar{\theta})Q\left(\frac{\theta_i-\bar{\theta}}{\sigma}\right)}{\left(\sum_{i=1}^n Q\left(\frac{\theta_i-\bar{\theta}}{\sigma}\right)\right)\left(\sum_{i=1}^n Q^c\left(\frac{\theta_i-\bar{\theta}}{\sigma}\right)\right)} \geq 0. \tag{26}$$

To see (26)*, observe that in the sum in the numerator, the $Q(\cdot)$ function assigns larger weight to the terms with $(\theta_i - \bar{\theta}) < 0$ than to the terms with $(\theta_i - \bar{\theta}) > 0$.*

Furthermore, $\alpha_n \approx \beta_n$ for large n if either $|\theta_i - \bar{\theta}| \approx 0, \forall i$, or $|\theta_i - \bar{\theta}| \to \infty, \forall i$. In the first case, if $\theta_i = \bar{\theta}$, for $i = 1, \cdots, n$, we get $\beta_n = \alpha_n = \|\boldsymbol{\theta} - \bar{\theta}\mathbf{1}\|^2/n = 0$. In the second case, suppose that n_1 of the θ_i values equal p_1 and the remaining $(n - n_1)$ values equal $-p_2$ for some $p_1, p_2 > 0$. Then, as $p_1, p_2 \to \infty$, it can be verified that $\beta_n \to [\|\boldsymbol{\theta}\|^2 - n_1p_1^2 - (n-n_1)p_2^2]/n = 0$. Therefore, the asymptotic normalized risk $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2})/n$ converges to 0 in both cases.

The proof of Theorem 1 further leads to the following corollaries.

Corollary 1. *The loss function of the positive-part JS-estimator in* (7) *satisfies the following:*

(1) For any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_+}\|^2}{n} - \frac{\gamma_n\sigma^2}{\gamma_n + \sigma^2}\right| \geq \epsilon\right) \leq Ke^{-nk\min(\epsilon^2,1)},$$

where $\gamma_n := \|\boldsymbol{\theta}\|^2/n$, and K and k are positive constants.

(2) For a sequence of $\boldsymbol{\theta}$ with increasing dimension n, if $\limsup_{n\to\infty}\|\boldsymbol{\theta}\|^2/n < \infty$, we have $\lim_{n\to\infty}\left|\frac{1}{n}R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_+}) - \frac{\gamma_n\sigma^2}{\gamma_n+\sigma^2}\right| = 0$.

Note that the positive-part Lindley's estimator in (8) is essentially a single-cluster estimator which shrinks all the points towards $\bar{y}$. Henceforth, we denote it by $\hat{\boldsymbol{\theta}}_{JS_1}$.

Corollary 2. *The loss function of the positive-part Lindley's estimator in* (8) *satisfies the following:*

(1) For any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_1}\|^2}{n} - \frac{\rho_n\sigma^2}{\rho_n + \sigma^2}\right| \geq \epsilon\right) \leq Ke^{-nk\min(\epsilon^2,1)},$$

where K and k are positive constants, and

$$\rho_n := \frac{\left\|\boldsymbol{\theta} - \bar{\theta}\mathbf{1}\right\|^2}{n}. \tag{27}$$

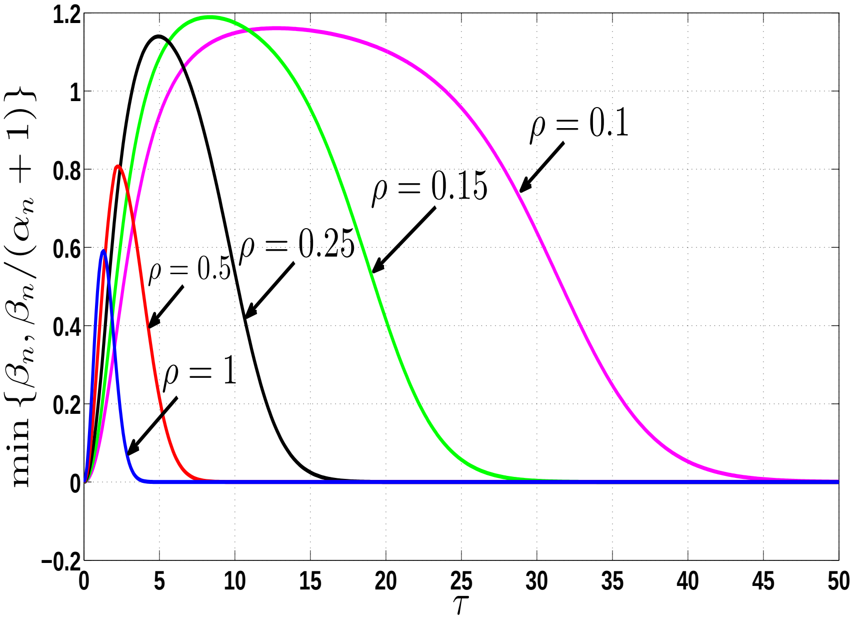


Figure 2: The asymptotic risk term $\min\{\beta_n, \beta_n/(\alpha_n+\sigma^2)\}$ for the two-cluster estimator is plotted vs τ for $n = 1000$, $\sigma = 1$, and different values of ρ. Here, the components of $\boldsymbol{\theta}$ take only two values, τ and $-\rho\tau$. The number of components taking the value τ is $\lfloor n\rho/(1+\rho)\rfloor$.

(2) For a sequence of $\boldsymbol{\theta}$ with increasing dimension n, if $\limsup_{n\to\infty} \|\boldsymbol{\theta}\|^2/n < \infty$, we have

$$\lim_{n\to\infty} \left| \frac{1}{n} R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1}\right) - \frac{\rho_n \sigma^2}{\rho_n + \sigma^2} \right| = 0. \tag{28}$$

Remark 3. *Statement* (2) *of Corollary 1, which is known in the literature [14], implies that $\hat{\boldsymbol{\theta}}_{JS_+}$ is asymptotically minimax over Euclidean balls. Indeed, if Θ_n denotes the set of $\boldsymbol{\theta}$ such that $\gamma_n = \|\boldsymbol{\theta}\|^2/n \le c^2$, then Pinsker's theorem [15, Ch. 5] implies that the minimax risk over Θ_n is asymptotically (as $n \to \infty$) equal to $\frac{\sigma^2 c^2}{c^2+\sigma^2}$.*

Statement (1) *of Corollary 1 and both the statements of Corollary 2 are new, to the best of our knowledge. Comparing Corollaries 1 and 2, we observe that $\rho_n \le \gamma_n$ since $\|\boldsymbol{\theta} - \bar{\theta}\mathbf{1}\| \le \|\boldsymbol{\theta}\|$ for all $\boldsymbol{\theta} \in \mathbb{R}^n$ with strict inequality whenever $\bar{\theta} \ne 0$. Therefore the positive-part Lindley's estimator asymptotically dominates the positive part JS-estimator.*

It is well known that both $\hat{\boldsymbol{\theta}}_{JS_+}$ and $\hat{\boldsymbol{\theta}}_{JS_1}$ dominate the ML-estimator [11]. From Corollary 1, it is clear that asymptotically, the normalized risk of $\hat{\boldsymbol{\theta}}_{JS_+}$ is small when γ_n is small, i.e., when $\boldsymbol{\theta}$ is close to the origin. Similarly, from Corollary 2, the asymptotic normalized risk of $\hat{\boldsymbol{\theta}}_{JS_1}$ is small when ρ_n is small, which occurs when the components of $\boldsymbol{\theta}$ are all very close to the mean $\bar{\theta}$. It is then natural to ask if the two-cluster estimator $\hat{\boldsymbol{\theta}}_{JS_2}$ dominates $\hat{\boldsymbol{\theta}}_{ML}$, and when its asymptotic normalized risk is close to 0. To answer these questions, we use the following example, shown in Fig. 2. Consider $\boldsymbol{\theta}$ whose components take one of two values, τ or $-\rho\tau$, such that $\bar{\theta}$ is as close to zero as possible. Hence the number of components taking the value τ is $\lfloor n\rho/(1+\rho)\rfloor$. Choosing $\sigma = 1$, $n = 1000$, the key asymptotic risk term $\min\{\beta_n, \beta_n/(\alpha_n+\sigma^2)\}$ in Theorem 1 is plotted as a function of τ in Fig. 2 for various values of ρ.

Two important observations can be made from the plots. Firstly, $\min\{\beta_n, \beta_n/(\alpha_n+\sigma^2)\}$ exceeds $\sigma^2 = 1$ for certain values of ρ and τ. Hence, $\hat{\boldsymbol{\theta}}_{JS_2}$ does not dominate $\hat{\boldsymbol{\theta}}_{ML}$. Secondly, for any ρ, the normalized risk of $\hat{\boldsymbol{\theta}}_{JS_2}$ goes to zero for large enough τ. Note that when τ is large, both $\gamma_n = \|\boldsymbol{\theta}\|^2/n$ and $\rho_n = \|\boldsymbol{\theta} - \bar{\theta}1\|^2/n$ are large and hence, the normalized risks of both $\hat{\boldsymbol{\theta}}_{JS_+}$ and $\hat{\boldsymbol{\theta}}_{JS_1}$ are close to 1. So, although $\hat{\boldsymbol{\theta}}_{JS_2}$ does not dominate $\hat{\boldsymbol{\theta}}_{ML}$, $\hat{\boldsymbol{\theta}}_{JS_+}$ or $\hat{\boldsymbol{\theta}}_{JS_2}$, there is a range of $\boldsymbol{\theta}$ for which $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2})$ is much lower than both $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_+})$ and $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1})$. This serves as motivation for designing a hybrid estimator that attempts to pick the better of $\hat{\boldsymbol{\theta}}_{JS_1}$ and $\hat{\boldsymbol{\theta}}_{JS_2}$ for the $\boldsymbol{\theta}$ in context. This is described in the next section.

In the example of Fig. 2, it is worth examining why the two-cluster estimator performs poorly for a certain range of τ, while giving significantly risk reduction for large enough τ. First consider an ideal case, where it is known which components of theta are equal to τ and which ones are equal to $-\rho\tau$ (although the values of ρ, τ may not be known). In this case, we could use a James-Stein estimator $\hat{\boldsymbol{\theta}}_{JS_{\mathbb{V}}}$ of the form (5) with the target subspace $\mathbb{V}$ being the two-dimensional subspace with basis vectors

$$\mathbf{u}_1 := \begin{bmatrix} 1_{\{\theta_1 = \tau\}} \\ 1_{\{\theta_2 = \tau\}} \\ \vdots \\ 1_{\{\theta_n = \tau\}} \end{bmatrix}, \quad \mathbf{u}_2 := \begin{bmatrix} 1_{\{\theta_1 = -\rho\tau\}} \\ 1_{\{\theta_2 = -\rho\tau\}} \\ \vdots \\ 1_{\{\theta_n = -\rho\tau\}} \end{bmatrix}.$$

Since $\mathbb{V}$ is a fixed subspace that does not depend on the data, it can be shown that $\hat{\boldsymbol{\theta}}_{JS_{\mathbb{V}}}$ dominates the ML-estimator [7, 8]. In the actual problem, we do not have access to the ideal basis vectors $\mathbf{u}_1, \mathbf{u}_2$, so we cannot use $\hat{\boldsymbol{\theta}}_{JS_{\mathbb{V}}}$. The two-cluster estimator $\hat{\boldsymbol{\theta}}_{JS_2}$ attempts to approximate $\hat{\boldsymbol{\theta}}_{JS_{\mathbb{V}}}$ by choosing the target subspace from the data. As shown in (10), this is done using the basis vectors:

$$\widehat{\mathbf{u}}_1 := \begin{bmatrix} 1_{\{y_1 \geq \bar{y}\}} \\ 1_{\{y_2 \geq \bar{y}\}} \\ \vdots \\ 1_{\{y_n \geq \bar{y}\}} \end{bmatrix}, \quad \widehat{\mathbf{u}}_2 := \begin{bmatrix} 1_{\{y_1 < \bar{y}\}} \\ 1_{\{y_2 < \bar{y}\}} \\ \vdots \\ 1_{\{y_n < \bar{y}\}} \end{bmatrix}.$$

Since $\bar{y}$ is a good approximation for $\bar{\theta} = 0$, when the separation between θ_i and $\bar{\theta}$ is large enough, the noise term w_i is unlikely to pull $y_i = \theta_i + w_i$ into the wrong region; hence, the estimated basis vectors $\widehat{\mathbf{u}}_1, \widehat{\mathbf{u}}_2$ will be close to the ideal ones $\mathbf{u}_1, \mathbf{u}_2$. Indeed, Fig. 2 indicates that when the minimum separation between θ_i and $\bar{\theta}$ (here, equal to $\rho\tau$) is at least 4.5σ, then $\widehat{\mathbf{u}}_1, \widehat{\mathbf{u}}_2$ approximate the ideal basis vectors very well, and the normalized risk is close to 0. On the other hand, the approximation to the ideal basis vectors turns out to be poor when the components of $\boldsymbol{\theta}$ are neither too close to nor too far from $\bar{\theta}$, as evident from Remark 1.

3 Hybrid James-Stein estimator with up to two clusters

Depending on the underlying $\boldsymbol{\theta}$, either the positive-part Lindley estimator $\hat{\boldsymbol{\theta}}_{JS_1}$ or the two-cluster estimator $\hat{\boldsymbol{\theta}}_{JS_2}$ could have a smaller loss (cf. Theorem 1 and Corollary 2). So we would like an estimator that selects the better among $\hat{\boldsymbol{\theta}}_{JS_1}$ and $\hat{\boldsymbol{\theta}}_{JS_2}$ for the $\boldsymbol{\theta}$ in context. To this end, we estimate the loss of $\hat{\boldsymbol{\theta}}_{JS_1}$ and $\hat{\boldsymbol{\theta}}_{JS_2}$ based on $\mathbf{y}$. Based on these loss estimates, denoted by $\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1})$ and $\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2})$ respectively, we define a hybrid estimator as

$$\hat{\boldsymbol{\theta}}_{JS_H} = \gamma_{\mathbf{y}} \hat{\boldsymbol{\theta}}_{JS_1} + (1 - \gamma_{\mathbf{y}}) \hat{\boldsymbol{\theta}}_{JS_2}, \tag{29}$$

where $\hat{\boldsymbol{\theta}}_{JS_1}$ and $\hat{\boldsymbol{\theta}}_{JS_2}$ are respectively given by (8) and (11), and $\gamma_{\mathbf{y}}$ is given by

$$\gamma_{\mathbf{y}} = \begin{cases} 1 & \text{if} \quad \frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1}) \leq \frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}), \\ 0 & \text{otherwise.} \end{cases} \tag{30}$$

The loss function estimates $\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1})$ and $\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2})$ are obtained as follows. Based on Corollary 2, the loss function of $\hat{\boldsymbol{\theta}}_{JS_1}$ can be estimated via an estimate of $\rho_n\sigma^2/(\rho_n + \sigma^2)$, where ρ_n is given by (27). It is straightforward to check, along the lines of the proof of Theorem 1, that

$$g\left(\frac{\|\mathbf{y} - \bar{y}\mathbf{1}\|^2}{n}\right) \doteq g\left(\rho_n + \sigma^2\right) = \rho_n + \sigma^2. \tag{31}$$

Therefore, an estimate of the normalized loss $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1})/n$ is

$$\frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1}) = \sigma^2\left(1 - \frac{\sigma^2}{g\left(\|\mathbf{y} - \bar{y}\mathbf{1}\|^2/n\right)}\right). \tag{32}$$

The loss function of the two-cluster estimator $\hat{\boldsymbol{\theta}}_{JS_2}$ can be estimated using Theorem 1, by estimating β_n and α_n defined in (24) and (23), respectively. From Lemma 13 in Section 6.2, we have

$$\frac{1}{n}\|\mathbf{y} - \boldsymbol{\nu}_2\|^2 \doteq \alpha_n + \sigma^2 + \kappa_n\delta + o(\delta). \tag{33}$$

Further, using the concentration inequalities in Lemmas 1 and 2 in Section 2, we can deduce that

$$\frac{1}{n}\|\mathbf{y} - \boldsymbol{\nu}_2\|^2 - \sigma^2 + \frac{\sigma^2}{n\delta}\left(\sum_{i=0}^{n}\mathbf{1}_{\{|y_i - \bar{y}| \leq \delta\}}\right)(a_1 - a_2) \doteq \beta_n + \kappa_n\delta + o(\delta), \tag{34}$$

where a_1, a_2 are defined in (20). We now use (33) and (34) to estimate the concentrating value in (21), noting that

$$\min\left(\beta_n, \frac{\beta_n\sigma^2}{\alpha_n + \sigma^2}\right) = \frac{\beta_n\sigma^2}{g(\alpha_n + \sigma^2)},$$

where $g(x) = \max(x, \sigma^2)$. This yields the following estimate of $L(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2})/n$:

$$\frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}) = \frac{\sigma^2\left(\frac{1}{n}\|\mathbf{y} - \boldsymbol{\nu}_2\|^2 - \sigma^2 + \frac{\sigma^2}{n\delta}\left(\sum_{i=0}^{n}\mathbf{1}_{\{|y_i - \bar{y}| \leq \delta\}}\right)(a_1 - a_2)\right)}{g(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n)}. \tag{35}$$

The loss function estimates in (32) and (35) complete the specification of the hybrid estimator in (29) and (30). The following theorem characterizes the loss function of the hybrid estimator, by showing that the loss estimates in (32) and (35) concentrate around the values specified in Corollary 2 and Theorem 1, respectively.

Theorem 2. *The loss function of the hybrid JS-estimator in* (29) *satisfies the following:*

(1) For any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_H}\|^2}{n} - \min\left(\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_1}\|^2}{n}, \frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_2}\|^2}{n}\right) \geq \epsilon\right) \leq Ke^{-\frac{nk\min(\epsilon^2, 1)}{\max(\|\boldsymbol{\theta}\|^2/n, 1)}},$$

where K and k are positive constants.

(2) For a sequence of $\boldsymbol{\theta}$ with increasing dimension n, if $\limsup_{n\to\infty} \|\boldsymbol{\theta}\|^2/n < \infty$, we have

$$\limsup_{n\to\infty} \frac{1}{n}\left(R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_H}\right) - \min\left[R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1}\right), R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}\right)\right]\right) \leq 0.$$

The proof of the theorem in given in Section 6.3. The theorem implies that the hybrid estimator chooses the better of the $\hat{\boldsymbol{\theta}}_{JS_1}$ and $\hat{\boldsymbol{\theta}}_{JS_2}$ with high probability, with the probability of choosing the worse estimator decreasing exponentially in n. It also implies that asymptotically, $\hat{\boldsymbol{\theta}}_{JS_H}$ dominates both $\hat{\boldsymbol{\theta}}_{JS_1}$ and $\hat{\boldsymbol{\theta}}_{JS_2}$, and hence, $\hat{\boldsymbol{\theta}}_{ML}$ as well.

Remark 4. *Instead of picking one among the two (or several) candidate estimators, one could consider a hybrid estimator which is a weighted combination of the candidate estimators. Indeed, George [8] and Leung and Barron [9, 10] have proposed combining the estimators using exponential mixture weights based on Stein's unbiased risk estimates (SURE) [5]. Due to the presence of indicator functions in the definition of the attracting vector, it is challenging to obtain a SURE for $\hat{\boldsymbol{\theta}}_{JS_2}$. We therefore use loss estimates to choose the better estimator. Furthermore, instead of choosing one estimator based on the loss estimate, if we were to follow the approach in [9] and employ a combination of the estimators using exponential mixture weights based on the un-normalized loss estimates, then the weight assigned to the estimator with the smallest loss estimate is exponentially larger (in n) than the other. Therefore, when the dimension is high, this is effectively equivalent to picking the estimator with the smallest loss estimate.*

4 General multiple-cluster James-Stein estimator

In this section, we generalize the two-cluster estimator of Section 2 to an L-cluster estimator defined by an arbitrary partition of the real line. The partition is defined by $L-1$ functions $s_j : \mathbb{R}^n \to \mathbb{R}$, such that

$$s_j(\mathbf{y}) := s_j(y_1, \cdots, y_n) \doteq \mu_j, \quad \forall j = 1, \cdots, (L-1), \tag{36}$$

with constants $\mu_1 > \mu_2 > \cdots > \mu_{L-1}$. In words, the partition can be defined via any $L-1$ functions of $\mathbf{y}$, each of which concentrates around a deterministic value as n increases. In the two-cluster estimator, we only have one function $s_1(\mathbf{y}) = \bar{y}$, which concentrates around $\bar{\theta}$. The points in (36) partition the real line as

$$\mathbb{R} = (-\infty, s_{L-1}(\mathbf{y})] \cup (s_{L-1}(\mathbf{y}), s_{L-2}(\mathbf{y})] \cup \cdots \cup (s_1(\mathbf{y}), \infty).$$

The clusters are defined as $\mathcal{C}_j = \{y_i,\ 1 \leq i \leq n \mid y_i \in (s_j(\mathbf{y}), s_{j-1}(\mathbf{y})]\}$, for $1 \leq j \leq L$, with $s_0(\mathbf{y}) = \infty$ and $s_L(\mathbf{y}) = -\infty$. In Section 4.2, we discuss one choice of partitioning points to define the L clusters, but here we first construct and analyse an estimator based on a general partition satisfying (36).

The points in $\mathcal{C}_j$ are all shrunk towards the same point a_j, defined in (40) later in this section. The attracting vector is

$$\boldsymbol{\nu}_L := \sum_{j=1}^{L} a_j \begin{bmatrix} 1_{\{y_1 \in \mathcal{C}_j\}} \\ \vdots \\ 1_{\{y_n \in \mathcal{C}_j\}} \end{bmatrix}, \tag{37}$$

and the proposed L-cluster JS-estimator is

$$\hat{\boldsymbol{\theta}}_{JS_L} = \boldsymbol{\nu}_L + \left[1 - \frac{\sigma^2}{g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)}\right](\mathbf{y} - \boldsymbol{\nu}_L), \tag{38}$$

where $g(x) = \max(\sigma^2, x)$.

The attracting vector $\boldsymbol{\nu}_L$ lies in an L-dimensional subspace defined by the L orthogonal vectors $[1_{\{y_1 \in \mathcal{C}_1\}}, \cdots, 1_{\{y_n \in \mathcal{C}_1\}}]^T, \ldots, [1_{\{y_1 \in \mathcal{C}_L\}}, \cdots, 1_{\{y_n \leq \mathcal{C}_L\}}]^T$. The desired values for $a_1, \ldots, a_L$ in (37) are such that the attracting vector $\boldsymbol{\nu}_L$ is the projection of $\boldsymbol{\theta}$ onto the L-dimensional subspace. Computing this projection, we find the desired values to be the means of the θ_i's in each cluster:

$$a_1^{des} = \frac{\sum_{i=1}^n \theta_i 1_{\{y_i \in \mathcal{C}_1\}}}{\sum_{i=1}^n 1_{\{y_i \in \mathcal{C}_1\}}}, \quad \cdots \quad , a_L^{des} = \frac{\sum_{i=1}^n \theta_i 1_{\{y_i \in \mathcal{C}_L\}}}{\sum_{i=1}^n 1_{\{y_i \in \mathcal{C}_L\}}}. \tag{39}$$

As the θ_i's are unavailable, we set $a_1, \ldots, a_L$ to be approximations of $a_1^{des}, \ldots, a_L^{des}$, obtained using concentration results similar to Lemmas 1 and 2 in Section 2. The attractors are given by

$$a_j = \frac{\sum_{i=1}^n y_i 1_{\{y_i \in \mathcal{C}_j\}} - \frac{\sigma^2}{2\delta} \sum_{i=0}^n \left[1_{\{|y_i - s_j(\mathbf{y})| \leq \delta\}} - 1_{\{|y_i - s_{j-1}(\mathbf{y})| \leq \delta\}} \right]}{\sum_{i=1}^n 1_{\{y_i \in \mathcal{C}_j\}}}, \quad 1 \leq j \leq L. \tag{40}$$

With $\delta > 0$ chosen to be a small positive number as before, this completes the specification of the attracting vector in (37), and hence the L-cluster JS-estimator in (38).

Theorem 3. *The loss function of the L-cluster JS-estimator in* (38) *satisfies the following:*

(1) For any $\epsilon > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n} \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_L}\|^2 - \left[\min\left(\beta_{n,L}, \frac{\beta_{n,L}\sigma^2}{\alpha_{n,L} + \sigma^2} \right) + \kappa_n \delta + o(\delta) \right] \right| \geq \epsilon \right) \leq K e^{-\frac{nk \min(\epsilon^2, 1)}{\max(\|\boldsymbol{\theta}\|^2/n, 1)}},$$

where K and k are positive constants, and

$$\begin{aligned} \alpha_{n,L} := \frac{\|\boldsymbol{\theta}\|^2}{n} &- \sum_{j=0}^{L} \frac{c_j^2}{n} \sum_{i=1}^{n} \left[Q\left(\frac{\mu_j - \theta_i}{\sigma} \right) - Q\left(\frac{\mu_{j-1} - \theta_i}{\sigma} \right) \right] \\ &- \left(\frac{2\sigma}{n\sqrt{2\pi}} \right) \left(\sum_{j=1}^{L} c_j \sum_{i=1}^{n} \left[e^{-\frac{(\mu_j - \theta_i)^2}{2\sigma^2}} - e^{-\frac{(\mu_{j-1} - \theta_i)^2}{2\sigma^2}} \right] \right), \end{aligned} \tag{41}$$

$$\beta_{n,L} := \frac{\|\boldsymbol{\theta}\|^2}{n} - \sum_{j=0}^{L} \frac{c_j^2}{n} \sum_{i=1}^{n} \left[Q\left(\frac{\mu_j - \theta_i}{\sigma} \right) - Q\left(\frac{\mu_{j-1} - \theta_i}{\sigma} \right) \right], \tag{42}$$

with

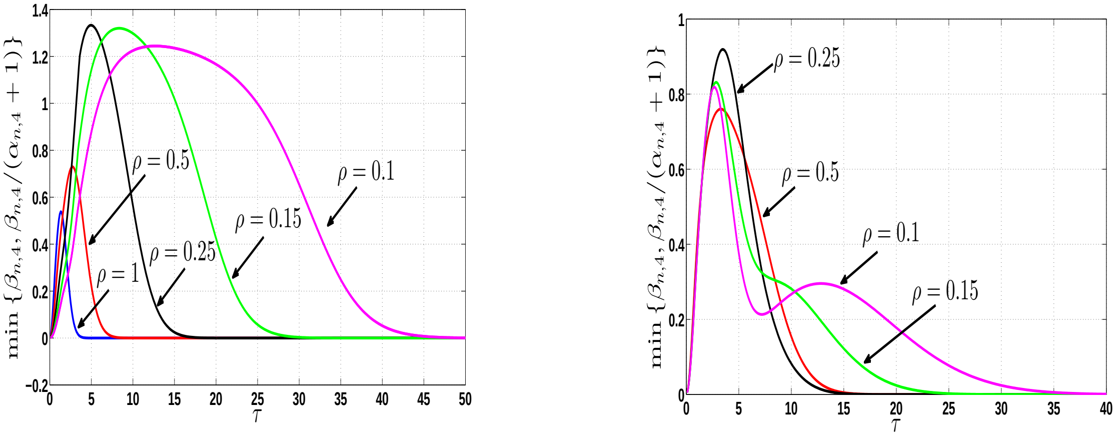
$$c_j := \frac{\sum_{i=1}^n \theta_i \left[Q\left(\frac{\mu_j - \theta_i}{\sigma} \right) - Q\left(\frac{\mu_{j-1} - \theta_i}{\sigma} \right) \right]}{\sum_{i=1}^n \left[Q\left(\frac{\mu_j - \theta_i}{\sigma} \right) - Q\left(\frac{\mu_{j-1} - \theta_i}{\sigma} \right) \right]}, \quad 1 \leq j \leq L. \tag{43}$$

(2) For a sequence of $\boldsymbol{\theta}$ with increasing dimension n, if $\limsup_{n \to \infty} \|\boldsymbol{\theta}\|^2/n < \infty$, we have

$$\lim_{n \to \infty} \left| \frac{1}{n} R\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_L} \right) - \left[\min\left(\beta_{n,L}, \frac{\sigma^2 \beta_{n,L}}{\alpha_{n,L} + \sigma^2} \right) + \kappa_n \delta + o(\delta) \right] \right| = 0.$$

The proof is similar to that of Theorem 1, and we provide its sketch in Section 6.4.

To get intuition on how the asymptotic normalized risk $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_L})/n$ depends on L, consider the four-cluster estimator with $L = 4$. For the same setup as in Fig. 2, i.e., the components of $\boldsymbol{\theta}$ take one of two values: τ or $-\rho\tau$, Fig. 3a plots the asymptotic risk term $\min\{\beta_n, \beta_{n,4}/(\alpha_{n,4} + 1)\}$



(a)

(b)

Figure 3: The asymptotic risk term $\min\{\beta_n, \beta_{n,4}/(\alpha_{n,4}+1)\}$ for the four-cluster estimator is plotted vs τ, for $n = 1000, \sigma = 1$, and different values of ρ. In (a), the components of $\boldsymbol{\theta}$ take only two values τ and $-\rho\tau$, with $\lfloor 1000\rho/(1+\rho) \rfloor$ components taking the value τ. In (b), they take values from the set $\{\tau, \rho\tau, -\rho\tau, -\tau\}$ with equal probability.

versus τ for the four-cluster estimator. Comparing Fig. 3a and Fig. 2, we observe that the four-cluster estimator's risk $\min\{\beta_n, \beta_{n,4}/(\alpha_{n,4}+1)\}$ behaves similarly to the two-cluster estimator's risk $\min\{\beta_n, \beta_{n,2}/(\alpha_{n,2}+1)\}$ with the notable difference being the magnitude. For $\rho = 0.5, 1$, the peak value of $\beta_{n,4}/(\alpha_{n,4}+1)$ is smaller than that of $\beta_{n,2}/(\alpha_{n,2}+1)$. However, for the smaller values of ρ, the reverse is true. This means that $\hat{\boldsymbol{\theta}}_{JS_4}$ can be better than $\hat{\boldsymbol{\theta}}_{JS_2}$, even in certain scenarios where the θ_i take only two values. In the two-value example, $\hat{\boldsymbol{\theta}}_{JS_4}$ is typically better when two of the four attractors of $\hat{\boldsymbol{\theta}}_{JS_4}$ are closer to the θ_i values, while the two attracting points of $\hat{\boldsymbol{\theta}}_{JS_2}$ are closer to $\bar{\theta}$ than the respective θ_i values.

Next consider an example where θ_i take values from $\{\tau, \rho\tau, -\rho\tau, -\tau\}$ with equal probability. This is the scenario favorable to $\hat{\boldsymbol{\theta}}_{JS_4}$. Figure 3b shows the plot of $\min\{\beta_n, \beta_{n,4}/(\alpha_{n,4}+1)\}$ as a function of τ for different values of ρ. Once again, it is clear that when the separation between the points is large enough, the asymptotic normalized risk approaches 0.

4.1 L-hybrid James-Stein estimator

Suppose that we have estimators $\hat{\boldsymbol{\theta}}_{JS_1}, \dots, \hat{\boldsymbol{\theta}}_{JS_L}$, where $\hat{\boldsymbol{\theta}}_{JS_\ell}$ is an ℓ-cluster JS-estimator constructed as described above, for $\ell = 1, \dots, L$. (Recall that $\ell = 1$ corresponds to Lindley's positive-part estimator in (8).) Depending on $\boldsymbol{\theta}$, any one of these L estimators could achieve the smallest loss. We would like to design a hybrid estimator that picks the best of these L estimators for the $\boldsymbol{\theta}$ in context. As in Section 3, we construct loss estimates for each of the L estimators, and define a hybrid estimator as

$$\hat{\boldsymbol{\theta}}_{JS_{H,L}} = \sum_{\ell=1}^{L} \gamma_\ell \, \hat{\boldsymbol{\theta}}_{JS_\ell} \tag{44}$$

where

$$\gamma_\ell = \begin{cases} 1 & \text{if } \frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_\ell}) = \min_{1 \le k \le L} \ \frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_k}), \\ 0 & \text{otherwise} \end{cases}$$

with $\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_\ell})$ denoting the loss function estimate of $\hat{\boldsymbol{\theta}}_{JS_\ell}$.

For $\ell \geq 2$, we estimate the loss of $\hat{\boldsymbol{\theta}}_{JS_\ell}$ using Theorem 3, by estimating $\beta_{n,\ell}$ and $\alpha_{n,\ell}$ which are defined in (41) and (42), respectively. From (77) in Section 6.4, we obtain

$$\frac{1}{n}\left\|\mathbf{y}-\boldsymbol{\nu}_\ell\right\|^2 \doteq \alpha_{n,\ell}+\sigma^2+\kappa_n\delta+o(\delta). \tag{45}$$

Using concentration inequalities similar to those in Lemmas 1 and 2 in Section 2, we deduce that

$$\frac{1}{n}\left\|\mathbf{y}-\boldsymbol{\nu}_\ell\right\|^2-\sigma^2+\frac{\sigma^2}{n\delta}\left(\sum_{j=1}^{\ell} a_j \sum_{i=0}^{n}\left[1_{\{|y_i-s_j(\mathbf{y})|\leq\delta\}}-1_{\{|y_i-s_{j-1}(\mathbf{y})|\leq\delta\}}\right]\right) \doteq \beta_{n,\ell}+\kappa_n\delta+o(\delta), \tag{46}$$

where $a_1,\ldots,a_\ell$ are as defined in (40). We now use (45) and (46) to estimate the concentrating value in Theorem 3, and thus obtain the following estimate of $L(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_{JS_\ell})/n$:

$$\frac{1}{n}\hat{L}(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_{JS_\ell})=\frac{\sigma^2\left(\frac{1}{n}\left\|\mathbf{y}-\boldsymbol{\nu}_\ell\right\|^2-\sigma^2+\frac{\sigma^2}{n\delta}\left(\sum_{j=1}^{\ell} a_j\sum_{i=0}^{n}\left[1_{\{|y_i-s_j(\mathbf{y})|\leq\delta\}}-1_{\{|y_i-s_{j-1}(\mathbf{y})|\leq\delta\}}\right]\right)\right)}{g(\|\mathbf{y}-\boldsymbol{\nu}_l\|^2/n)}. \tag{47}$$

The loss function estimator in (47) for $2 \leq \ell \leq L$, together with the loss function estimator in (32) for $\ell = 1$, completes the specification of the L-hybrid estimator in (44). Using steps similar to those in Theorem 2, we can show that

$$\frac{1}{n}\hat{L}\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_{JS_\ell}\right) \doteq \min\left(\beta_{n,\ell},\frac{\sigma^2\beta_{n,\ell}}{\alpha_{n,\ell}+\sigma^2}\right)+\kappa_n\delta+o(\delta), \quad 2\leq\ell\leq L. \tag{48}$$

Theorem 4. *The loss function of the L-hybrid JS-estimator in* (44) *satisfies the following:*

(1) For any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{JS_{H,L}}\|^2}{n}-\min_{1\leq\ell\leq L}\left(\frac{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{JS_\ell}\|^2}{n}\right)\geq\epsilon\right)\leq Ke^{-\frac{nk\min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}},$$

where K and k are positive constants.

(2) For a sequence of $\boldsymbol{\theta}$ with increasing dimension n, if $\limsup_{n\to\infty}\|\boldsymbol{\theta}\|^2/n<\infty$, we have

$$\limsup_{n\to\infty}\frac{1}{n}\left[R\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_{JS_{H,L}}\right)-\min_{1\leq\ell\leq L}R\left(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}_{JS_\ell}\right)\right]\leq 0.$$

The proof of the theorem is omitted as it is along the same lines as the proof of Theorem 3. Thus with high probability, the L-hybrid estimator chooses the best of the $\hat{\boldsymbol{\theta}}_{JS_1},\ldots,\hat{\boldsymbol{\theta}}_{JS_L}$, with the probability of choosing a worse estimator decreasing exponentially in n.

4.2 Obtaining the clusters

In this subsection, we present a simple method to obtain the $(L-1)$ partitioning points $s_j(\mathbf{y})$, $1 \leq j \leq (L-1)$, for an L-cluster JS-estimator when $L = 2^a$ for an integer $a > 1$. We do this recursively, assuming that we already have a 2^{a-1}-cluster estimator with its associated partitioning points $s'_j(\mathbf{y})$, $j = 1,\cdots,2^{a-1}-1$. This means that for the 2^{a-1}-cluster estimator, the real line is partitioned as

$$\mathbb{R}=\left(-\infty,s'_{2^{a-1}-1}(\mathbf{y})\right]\cup\left(s'_{2^{a-1}-1}(\mathbf{y}),s'_{2^{a-1}-2}(\mathbf{y})\right]\cup\cdots\cup\left(s'_1(\mathbf{y}),\infty\right).$$

Recall that Section 2 considered the case of $a = 1$, with the single partitioning point being $\bar{y}$.

The new partitioning points $s_k(\mathbf{y})$, $k = 1, \cdots, (2^a - 1)$, are obtained as follows. For $j = 1, \cdots, (2^{a-1} - 1)$, define

$$s_{2j}(\mathbf{y}) = s'_j(\mathbf{y}), \quad s_{2j-1}(\mathbf{y}) = \frac{\sum_{i=1}^n y_i \mathbf{1}_{\left\{s'_j(\mathbf{y}) < y_i \leq s'_{j-1}(\mathbf{y})\right\}}}{\sum_{i=1}^n \mathbf{1}_{\left\{s'_j(\mathbf{y}) < y_i \leq s'_{j-1}(\mathbf{y})\right\}}}, \quad s_{2^a-1}(\mathbf{y}) = \frac{\sum_{i=1}^n y_i \mathbf{1}_{\left\{-\infty < y_i \leq s'_{2^{a-1}-1}(\mathbf{y})\right\}}}{\sum_{i=1}^n \mathbf{1}_{\left\{-\infty < y_i \leq s'_{2^{a-1}-1}(\mathbf{y})\right\}}}$$

where $s'_0(\mathbf{y}) = \infty$. Hence, the partition for the L-cluster estimator is

$$\mathbb{R} = (-\infty, s_{2^a-1}(\mathbf{y})] \cup (s_{2^a-1}(\mathbf{y}), s_{2^a-2}(\mathbf{y})] \cup \cdots \cup (s_1(\mathbf{y}), \infty).$$

We use such a partition to construct a 4-cluster estimator for our simulations in the next section.

5 Simulation Results

In this section, we present simulation plots that compare the average normalized loss of the proposed estimators with that of the regular JS-estimator and Lindley's estimator, for various choices of $\boldsymbol{\theta}$. In each plot, the normalized loss, labelled $\frac{1}{n}\tilde{R}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ on the Y-axis, is computed by averaging over 1000 realizations of $\mathbf{w}$. We use $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I})$, i.e., the noise variance $\sigma^2 = 1$. Both the regular JS-estimator $\hat{\boldsymbol{\theta}}_{JS}$ and Lindley's estimator $\hat{\boldsymbol{\theta}}_{JS_1}$ used are the positive-part versions, respectively given by (7) and (8). We choose $\delta = 5/\sqrt{n}$ for our proposed estimators.

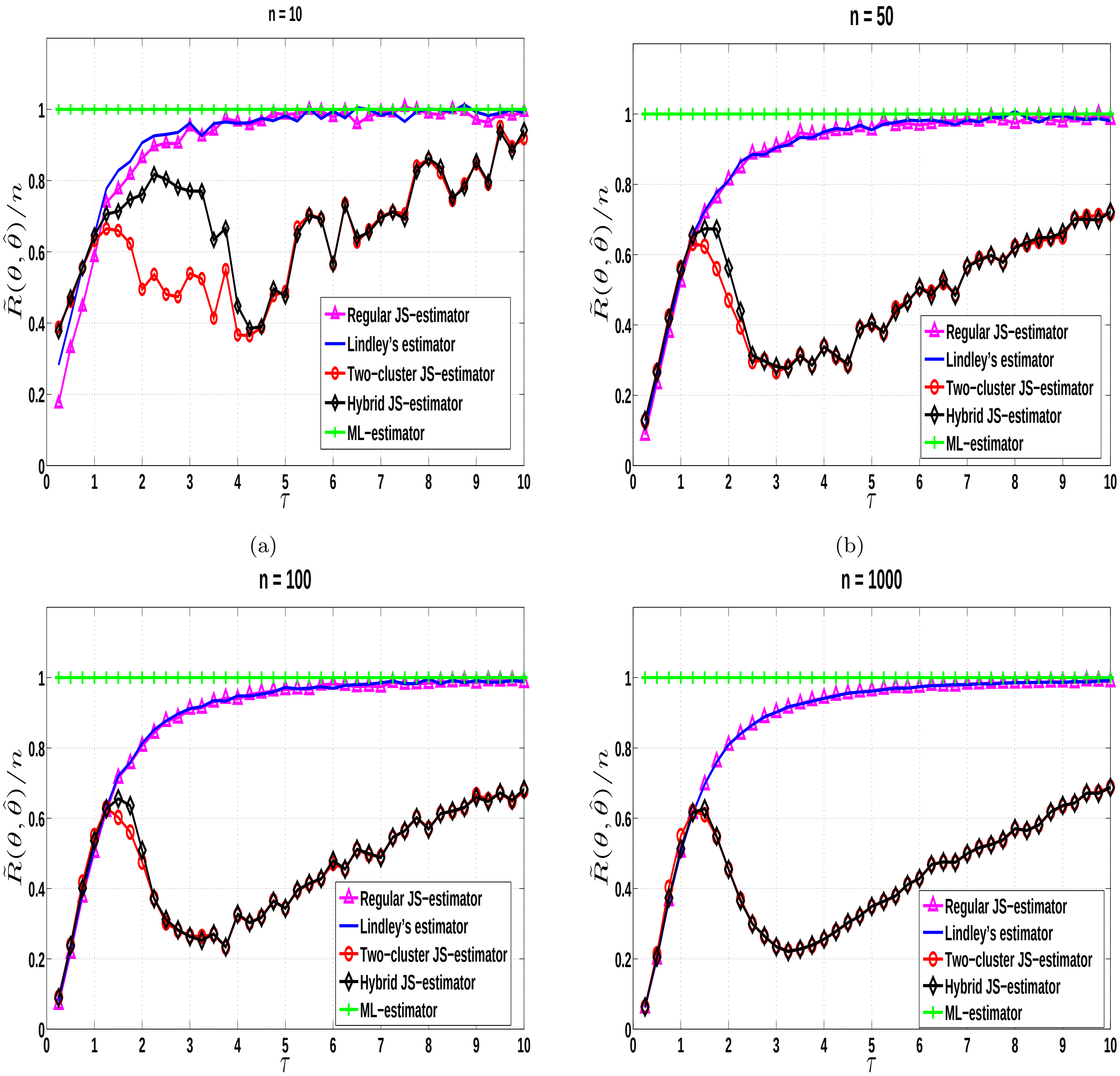
In Figs. 4–7, we consider three different structures for $\boldsymbol{\theta}$, representing varying degrees of clustering. In the first structure, the components $\{\theta_i\}_{i=1}^n$ are arranged in two clusters. In the second structure for $\boldsymbol{\theta}$, $\{\theta_i\}_{i=1}^n$ are uniformly distributed within an interval whose length is varied. In the third structure, $\{\theta_i\}_{i=1}^n$ are arranged in four clusters. In both clustered structures, the locations and the widths of the clusters as well as the number of points within each cluster are varied; the locations of the points within each cluster are chosen uniformly at random. The captions of the figures explain the details of each structure.

In Fig. 4, $\{\theta_i\}_{i=1}^n$ are arranged in two clusters, one centred at $-\tau$ and the other at τ. The plots show the average normalized loss as a function of τ for different values of n, for four estimators: $\hat{\boldsymbol{\theta}}_{JS}$, $\hat{\boldsymbol{\theta}}_{JS_1}$, the two-attractor JS-estimator $\hat{\boldsymbol{\theta}}_{JS_2}$ given by (11), and the hybrid JS-estimator $\hat{\boldsymbol{\theta}}_{JS_H}$ given by (29). We observe that as n increases, the average loss of $\hat{\boldsymbol{\theta}}_{JS_H}$ gets closer to the minimum of that of $\hat{\boldsymbol{\theta}}_{JS_1}$ and $\hat{\boldsymbol{\theta}}_{JS_2}$;

Fig. 5 shows the the average normalized loss for different arrangements of $\{\theta_i\}$, with n fixed at 1000. The plots illustrate a few cases where $\hat{\boldsymbol{\theta}}_{JS_2}$ has significantly lower risk than $\hat{\boldsymbol{\theta}}_{JS_1}$, and also the strength of $\hat{\boldsymbol{\theta}}_{JS_H}$ when n is large.

Fig. 6 compares the average normalized losses of $\hat{\boldsymbol{\theta}}_{JS_1}$, $\hat{\boldsymbol{\theta}}_{JS_2}$, and $\hat{\boldsymbol{\theta}}_{JS_H}$ with their asymptotic risk values, obtained in Corollary 2, Theorem 1, and Theorem 2, respectively. Each subfigure considers a different arrangement of $\{\theta_i\}_{i=1}^n$, and shows how the average losses converge to their respective theoretical values with growing n.

Fig. 7 demonstrates the effect of choosing four attractors when $\{\theta_i\}_{i=1}^n$ form four clusters. The four-hybrid estimator $\hat{\boldsymbol{\theta}}_{JS_{H,4}}$ attempts to choose the best among $\hat{\boldsymbol{\theta}}_{JS_1}$, $\hat{\boldsymbol{\theta}}_{JS_2}$ and $\hat{\boldsymbol{\theta}}_{JS_4}$ based on the data $\mathbf{y}$. It is clear that depending on the values of $\{\theta_i\}$, $\hat{\boldsymbol{\theta}}_{JS_{H,4}}$ reliably tracks the best of these. and can have significantly lower loss than both $\hat{\boldsymbol{\theta}}_{JS_1}$ and $\hat{\boldsymbol{\theta}}_{JS_2}$, especially for large values of n.



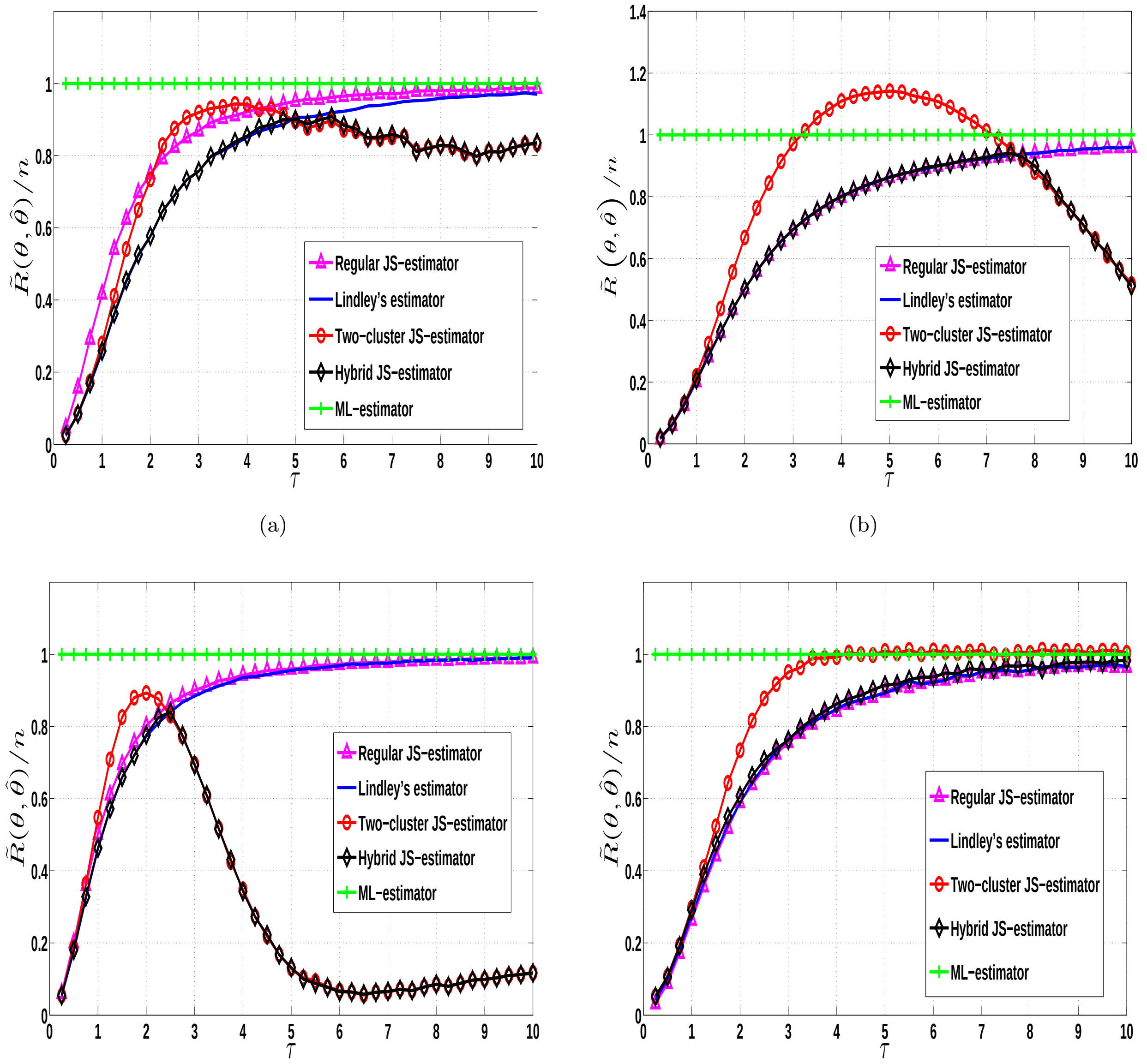
(a)

(b)

(c)

(d)

Figure 4: Average normalized loss of various estimators for different values of n. The $\{\theta_i\}_{i=1}^n$ are placed in two clusters, one centred at τ and another at $-\tau$. Each cluster has width 0.5τ and $n/2$ points.



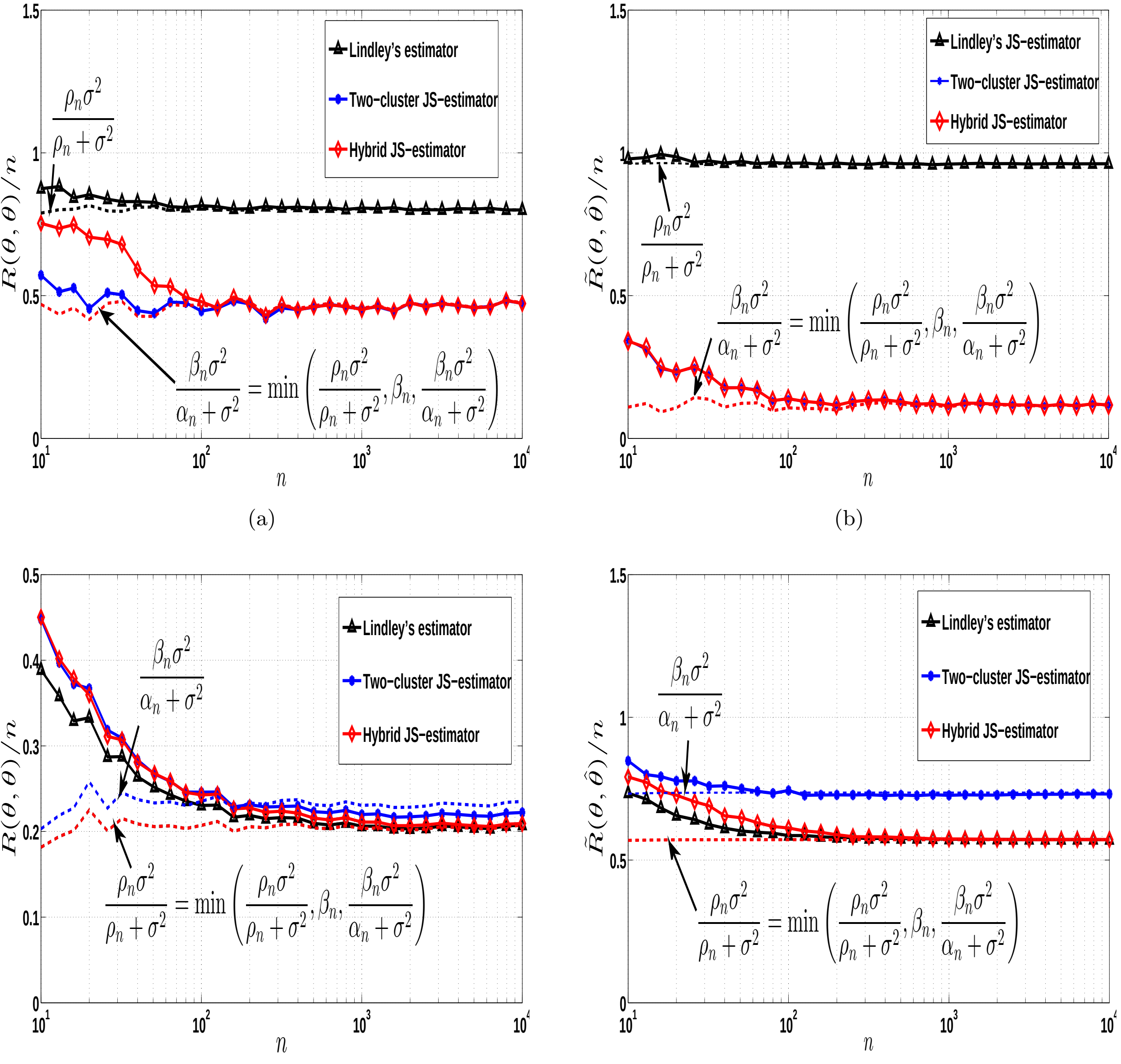
(a)

(b)

(c)

(d)

Figure 5: Average normalized loss of various estimators for different arrangements of the samples $\{\theta_i\}_{i=1}^n$, with $n = 1000$. In (a), there are 2 clusters each of width 0.5τ, one around 0.25τ containing 300 points, and the other around $-\tau$ and containing 700 points. In (b), $\boldsymbol{\theta}$ consists of 200 components taking the value τ and the remaining 800 taking the value -0.25τ. In (c), there are two clusters of width 0.125τ, one around τ containing 300 points and another around $-\tau$ containing 700 points. In (d), $\{\theta_i\}_{i=1}^n$ are arranged uniformly from $-\tau$ to τ.



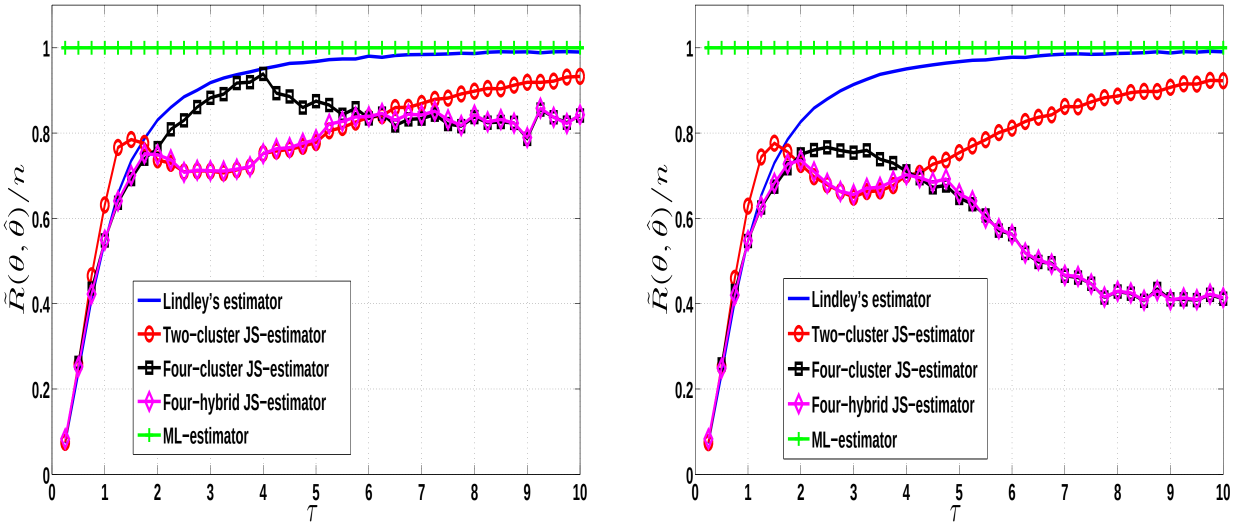
(a)

(b)

(c)

(d)

Figure 6: Average normalized loss of various estimators versus $\frac{\rho_n}{\rho_n+\sigma^2}$, $\frac{\beta_n}{\alpha_n+\sigma^2}$, $\min(\frac{\rho_n}{\rho_n+\sigma^2}, \beta_n, \frac{\beta_n}{\alpha_n+\sigma^2})$ as a function of n for different arrangements of $\{\theta_i\}_{i=1}^n$. In (a), the $\{\theta_i\}_{i=1}^n$ are placed in two clusters of width 1, one around 2 and the other around -2, each containing an equal number of points. In (b), $\{\theta_i\}_{i=1}^n$ are placed in two clusters of width 1.25, one around 5 and the other around -5, each containing an equal number of points. In (c), $\{\theta_i\}_{i=1}^n$ are placed in two clusters of width 0.25, one around 0.5 and the other around -0.5, each containing an equal number of points. In (d), $\{\theta_i\}_{i=1}^n$ are placed uniformly between -2 and 2.



(a) (b)

Figure 7: Average normalized loss of various estimators for $n = 1000$ and for different arrangements of the samples $\{\theta_i\}_{i=1}^n$. In (a), $\{\theta_i\}_{i=1}^n$ are placed in four equal-sized clusters of width 0.5τ and in (b), the clusters are of width 0.25τ. In both cases, the clusters are centred at 1.5τ, 0.9τ, -0.5τ and -1.25τ.

6 Proofs

6.1 Mathematical preliminaries

Here we list some concentration results that are used in the proofs of the theorems.

Lemma 3. *Let $\{X_n(\boldsymbol{\theta}), \boldsymbol{\theta} \in \mathbb{R}^n\}_{n=1}^{\infty}$ be a sequence of random variables such that $X_n(\boldsymbol{\theta}) \doteq 0$, i.e., for any $\epsilon > 0$,*

$$\mathbb{P}(|X_n(\boldsymbol{\theta})| \geq \epsilon) \leq K e^{-\frac{nk \min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}},$$

where K and k are positive constants. If $C := \limsup_{n\to\infty} \|\boldsymbol{\theta}\|^2/n < \infty$, then $X_n \xrightarrow{a.s.} 0$.

Proof. For any $\tau > 0$, there exists a positive integer M such that $\forall n \geq M$, $\|\boldsymbol{\theta}\|^2/n < C + \tau$. Hence, we have, for any $\epsilon > 0$, and for some $\tau > 1$,

$$\sum_{n=1}^{\infty} \mathbb{P}(|X_n(\boldsymbol{\theta})| \geq \epsilon) \leq \sum_{n=1}^{\infty} K e^{-\frac{nk \min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}} \leq C_0 + \sum_{n=M}^{\infty} K e^{-\frac{nk \min(\epsilon^2,1)}{C+\tau}} < \infty.$$

Therefore, we can use the Borel-Cantelli lemma to conclude that $X_n \xrightarrow{a.s.} 0$. □

Lemma 4. *For two sequences of random variables $\{X_n\}_{n=1}^{\infty}$, $\{Y_n\}_{n=1}^{\infty}$ such that $X_n \doteq 0$, $Y_n \doteq 0$, it follows that $X_n + Y_n \doteq 0$.*

Proof. This is an application of the triangle inequality: if for $\epsilon > 0$, $\mathbb{P}(|X_n| \geq \epsilon) \leq K_1 e^{-\frac{nk_1 \min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}}$ and $\mathbb{P}(|Y_n| \geq \epsilon) \leq K_2 e^{-\frac{nk_2 \min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}}$ for positive constants K_1, K_2, k_1 and k_2, then

$$\mathbb{P}(|X_n + Y_n| \geq \epsilon) \leq \mathbb{P}\left(|X_n| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(|Y_n| \geq \frac{\epsilon}{2}\right) \leq K e^{-\frac{nk \min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}}$$

where $K = K_1 + K_2$ and $k = \min\left(\frac{k_1}{4}.\frac{k_2}{4}\right)$. □

Lemma 5. *Let X and Y be random variables such that for any $\epsilon > 0$,*

$$\mathbb{P}\left(|X - a_X| \geq \epsilon\right) \leq K_1 e^{-nk_1 \min\left(\epsilon^2, 1\right)},$$
$$\mathbb{P}\left(|Y - a_Y| \geq \epsilon\right) \leq K_2 e^{-nk_2 \min\left(\epsilon^2, 1\right)}$$

where k_1, k_2 are positive constants, and K_1, K_2 are positive integer constants. Then,

$$\mathbb{P}\left(|XY - a_X a_Y| \geq \epsilon\right) \leq K e^{-nk \min\left(\epsilon^2, 1\right)}$$

where $K = 2(K_1 + K_2)$, and k is a positive constant depending on k_1 and k_2.

Proof. We have

$$\begin{aligned}
&\mathbb{P}\left(|XY - a_X a_Y| \geq \epsilon\right) = \mathbb{P}\left(|(X - a_X)(Y - a_Y) + X a_Y + Y a_X - 2 a_X a_Y| \geq \epsilon\right)\\
&\leq \mathbb{P}\left(|(X - a_X)(Y - a_Y)| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(|X a_Y + Y a_X - 2 a_X a_Y| \geq \frac{\epsilon}{2}\right)\\
&\leq \mathbb{P}\left(|(X - a_X)(Y - a_Y)| \geq \frac{\epsilon}{2}\right) + \mathbb{P}\left(|(X - a_X) a_Y| \geq \frac{\epsilon}{4}\right) + \mathbb{P}\left(|(Y - a_Y) a_X| \geq \frac{\epsilon}{4}\right)\\
&\leq \mathbb{P}\left(|(X - a_X)| \geq \sqrt{\frac{\epsilon}{2}}\right) + \mathbb{P}\left(|(Y - a_Y)| \geq \sqrt{\frac{\epsilon}{2}}\right) + \mathbb{P}\left(|(X - a_X) a_Y| \geq \frac{\epsilon}{4}\right) + \mathbb{P}\left(|(Y - a_Y) a_X| \geq \frac{\epsilon}{4}\right)\\
&\leq K_1 e^{-nk_1' \min(\epsilon, 1)} + K_2 e^{-nk_2' \min(\epsilon, 1)} + K_1 e^{-nk_1'' \min\left(\epsilon^2, 1\right)} + K_2 e^{-nk_2'' \min\left(\epsilon^2, 1\right)} \leq K e^{-nk \min\left(\epsilon^2, 1\right)}
\end{aligned}$$

where $k_1' = \frac{k_1}{2}$, $k_2' = \frac{k_2}{2}$, $k_1'' = \frac{k_1}{16(a_Y)^2}$, $k_2'' = \frac{k_2}{16(a_Y)^2}$, and $k = \frac{\min(k_1, k_2)}{16(a_Y)^2}$. □

Lemma 6. *Let Y be a non-negative random variable such that there exists $a_Y > 0$ such that for any $\epsilon > 0$,*

$$\mathbb{P}\left(Y - a_Y \leq -\epsilon\right) \leq K_1 e^{-nk_1 \min\left(\epsilon^2, 1\right)}, \quad \mathbb{P}\left(Y - a_Y \geq \epsilon\right) \leq K_2 e^{-nk_2 \min\left(\epsilon^2, 1\right)}$$

where k_1, k_2 are positive constants, and K_1, K_2 are positive integer constants. Then, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{Y} - \frac{1}{a_Y}\right| \geq \epsilon\right) \leq K e^{-nk \min\left(\epsilon^2, 1\right)}$$

where $K = K_1 + K_2$, and k is a positive constant.

Proof. We have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{Y} - \frac{1}{a_Y} \geq \epsilon\right) &= \mathbb{P}\left(\frac{1}{Y} \geq \frac{1}{a_Y} + \epsilon\right) = \mathbb{P}\left(Y \leq \frac{1}{\epsilon + 1/a_Y}\right)\\
&= \mathbb{P}\left(Y - a_Y \leq \frac{1}{\epsilon + 1/a_Y} - a_Y\right) = \mathbb{P}\left(Y - a_Y \leq -a_Y\left(\frac{\epsilon a_Y}{1 + \epsilon a_Y}\right)\right)\\
&\leq K_1 e^{-nk_1 \min\left(\left(\frac{\epsilon (a_Y)^2}{1 + \epsilon a_Y}\right)^2, 1\right)} \leq K_1 e^{-nk_1' \min\left(\epsilon^2, 1\right)} \qquad (49)
\end{aligned}$$

where $k_1' = k_1 \min\left(\left(\frac{(a_Y)^2}{1 + a_Y}\right)^2, 1\right)$. Similarly

$$\mathbb{P}\left(\frac{1}{Y} - \frac{1}{a_Y} \leq -\epsilon\right) = \mathbb{P}\left(\frac{1}{Y} \leq \frac{1}{a_Y} - \epsilon\right).$$

Note that when $\epsilon > \frac{1}{a_Y}$, $\mathbb{P}\left(\frac{1}{Y} - \frac{1}{a_Y} \leq -\epsilon\right) = 0$ because $Y > 0$. Therefore,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{Y} - \frac{1}{a_Y} \leq -\epsilon\right) = \mathbb{P}\left(Y \geq \frac{1}{\frac{1}{a_Y} - \epsilon}\right) = \mathbb{P}\left(Y - a_Y \geq \frac{1}{\frac{1}{a_Y} - \epsilon} - a_Y\right) \\
&= \mathbb{P}\left(Y - a_Y \geq a_Y \left(\frac{\epsilon a_Y}{1 - \epsilon a_Y}\right)\right) \\
&\leq K_2 e^{-nk_2 \min\left(\left(\frac{\epsilon (a_Y)^2}{1-\epsilon a_Y}\right)^2, 1\right)} \leq K_2 e^{-nk_2 \min\left(\epsilon^2 (a_Y)^4, 1\right)} \leq K_2 e^{-nk_2' \min\left(\epsilon^2, 1\right)} \qquad (50)
\end{aligned}$$

where $k_2' = k_2 \min\left((a_Y)^4, 1\right)$. Using (49) and (50), we obtain, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{Y} - \frac{1}{a_Y}\right| \geq \epsilon\right) \leq K e^{-nk \min\left(\epsilon^2, 1\right)}$$

where $k = \min(k_1', k_2')$. □

Lemma 7. *Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables and X be another random variable (or a constant) such that for any $\epsilon > 0$, $\mathbb{P}\left(|X_n - X| \geq \epsilon\right) \leq K e^{-nk \min\left(\epsilon^2, 1\right)}$ for positive constants K and k. Then, for the function $g(x) := \max(\sigma^2, x)$, we have*

$$\mathbb{P}\left(|g(X_n) - g(X)| \geq \epsilon\right) \leq K e^{-nk \min\left(\epsilon^2, 1\right)}.$$

Proof. We have

$$\begin{aligned}
\mathbb{P}\left(|g(X_n) - g(X)| \geq \epsilon\right) &= \mathbb{P}\left(|X_n - X| \geq \epsilon\right) \mathbb{P}\left(|g(X_n) - g(X)| \geq \epsilon \,\middle|\, |X_n - X| \geq \epsilon\right) \\
&\quad + \mathbb{P}\left(|X_n - X| < \epsilon\right) \mathbb{P}\left(|g(X_n) - g(X)| \geq \epsilon \,\middle|\, |X_n - X| < \epsilon\right) \\
&\leq K e^{-nk \min\left(\epsilon^2, 1\right)} + \mathbb{P}\left(|g(X_n) - g(X)| \geq \epsilon \,\middle|\, |X_n - X| < \epsilon\right). \qquad (51)
\end{aligned}$$

Now, when $X_n \geq \sigma^2$ and $X \geq \sigma^2$, it follows that $g(X_n) - g(X) = X_n - X$, and the second term of the RHS of (51) equals 0, as it also does when $X_n < \sigma^2$ and $X < \sigma^2$. Let us consider the case where $X_n \geq \sigma^2$ and $X < \sigma^2$. Then, $g(X_n) - g(X) = X_n - \sigma^2 < X_n - X < \epsilon$, as we condition on the fact that $|X_n - X| < \epsilon$; hence in this case $\mathbb{P}\left(|g(X_n) - g(X)| \geq \epsilon \,\middle|\, |X_n - X| < \epsilon\right) = 0$. Finally, when $X_n < \sigma^2$ and $X \geq \sigma^2$, we have $g(X) - g(X_n) = X - \sigma^2 < X - X_n < \epsilon$; hence in this case also we have $\mathbb{P}\left(|g(X_n) - g(X)| \geq \epsilon \,\middle|\, |X_n - X| < \epsilon\right) = 0$. This proves the lemma. □

Lemma 8. *(Hoeffding's Inequality [16, Thm. 2.8]). Let $X_1, \cdots, X_n$ be independent random variables such that $X_i \in [a_i, b_i]$ almost surely, for all $i \leq n$. Let $S_n = \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])$. Then for any $\epsilon > 0$, $\mathbb{P}\left(\frac{|S_n|}{n} \geq \epsilon\right) \leq 2e^{-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$.*

Lemma 9. *(Chi-squared concentration [16]). For i.i.d. Gaussian random variables $w_1, \ldots, w_n \sim \mathcal{N}(0, \sigma^2)$, we have for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} w_i^2 - \sigma^2\right| \geq \epsilon\right) \leq 2e^{-nk \min(\epsilon, \epsilon^2)},$$

where $k = \min\left(\frac{1}{4\sigma^4}, \frac{1}{2\sigma^2}\right)$.

Lemma 10. *For $i = 1, \cdots, n$, let $w_i \sim (0, \sigma^2)$ be independent, and a_i be real-valued and finite constants. We have for any $\epsilon > 0$,*

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} w_i 1_{\{w_i > a_i\}} - \frac{\sigma}{\sqrt{2\pi}}\sum_{i=1}^{n} e^{-\frac{a_i^2}{2\sigma^2}}\right| \geq \epsilon\right) \leq 2e^{-nk_1 \min(\epsilon, \epsilon^2)}, \tag{52}$$

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} w_i 1_{\{w_i \leq a_i\}} + \frac{\sigma}{\sqrt{2\pi}}\sum_{i=1}^{n} e^{-\frac{a_i^2}{2\sigma^2}}\right| \geq \epsilon\right) \leq 2e^{-nk_2 \min(\epsilon, \epsilon^2)} \tag{53}$$

where k_1 and k_2 are positive constants.

The proof is given in Appendix A.1.

Lemma 11. *Let $\mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2\mathbf{I}\right)$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a function such that for any $\epsilon > 0$,*

$$\mathbb{P}\left(|f(\mathbf{y}) - a| \geq \epsilon\right) \leq 2e^{-nk\epsilon^2}$$

for some constants a, k, such that $k > 0$. Then for any $\epsilon > 0$, we have

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} 1_{\{y_i > f(\mathbf{y})\}} - \sum_{i=1}^{n} 1_{\{y_i > a\}}\right| \geq \epsilon\right) \leq 4e^{-nk_1\epsilon^2}, \tag{54}$$

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} \theta_i 1_{\{y_i > f(\mathbf{y})\}} - \sum_{i=1}^{n} \theta_i 1_{\{y_i > a\}}\right| \geq \epsilon\right) \leq 4e^{-\frac{nk_1\epsilon^2}{\|\boldsymbol{\theta}\|^2/n}}, \tag{55}$$

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} w_i 1_{\{y_i > f(\mathbf{y})\}} - \sum_{i=1}^{n} w_i 1_{\{y_i > a\}}\right| \geq \epsilon\right) \leq 4e^{-nk_1 \min(\epsilon^2, \epsilon)} \tag{56}$$

where k_1 is a positive constant.

The proof is given in Appendix A.2.

Lemma 12. *With the assumptions of Lemma 11, let $h : \mathbb{R}^n \to \mathbb{R}$ be a function such that $b > a$ and $\mathbb{P}\left(|h(\mathbf{y}) - b| \geq \epsilon\right) \leq 2e^{-nl\epsilon^2}$ for some $l > 0$. Then for any $\epsilon > 0$, we have*

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} 1_{\{h(\mathbf{y}) \geq y_i > f(\mathbf{y})\}} - \sum_{i=1}^{n} 1_{\{b \geq y_i > a\}}\right| \geq \epsilon\right) \leq 8e^{-nk\epsilon^2}.$$

Proof. The result follows from Lemma 11 by noting that $1_{\{h(\mathbf{y}) \geq y_i > f(\mathbf{y})\}} = 1_{\{y_i > f(\mathbf{y})\}} - 1_{\{y_i > h(\mathbf{y})\}}$, and $1_{\{b \geq y_i > a\}} = 1_{\{y_i > a\}} - 1_{\{y_i > b\}}$. □

6.2 Proof of Theorem 1

We have,

$$\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_2}\|^2}{n} = \frac{1}{n}\left\|\boldsymbol{\theta} - \boldsymbol{\nu}_2 - \left[1 - \frac{\sigma^2}{g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)}\right](\mathbf{y} - \boldsymbol{\nu}_2)\right\|^2$$

$$= \frac{1}{n}\left\|\mathbf{y} - \boldsymbol{\nu}_2 - \left[1 - \frac{\sigma^2}{g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)}\right](\mathbf{y} - \boldsymbol{\nu}_2) - \mathbf{w}\right\|^2 = \frac{1}{n}\left\|\left(\frac{\sigma^2}{g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)}\right)(\mathbf{y} - \boldsymbol{\nu}_2) - \mathbf{w}\right\|^2$$

$$= \frac{\sigma^4\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n}{\left(g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)\right)^2} + \frac{\|\mathbf{w}\|^2}{n} - \frac{2}{n}\left(\frac{\sigma^2}{g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)}\right)\langle \mathbf{y} - \boldsymbol{\nu}_2, \mathbf{w}\rangle. \tag{57}$$

We also have

$$\frac{1}{n}\left\|\boldsymbol{\theta}-\boldsymbol{\nu}_2\right\|^2=\frac{1}{n}\left\|\mathbf{y}-\boldsymbol{\nu}_2-\mathbf{w}\right\|^2=\frac{1}{n}\left\|\mathbf{y}-\boldsymbol{\nu}_2\right\|^2+\frac{1}{n}\left\|\mathbf{w}\right\|^2-\frac{2}{n}\left\langle\mathbf{y}-\boldsymbol{\nu}_2,\mathbf{w}\right\rangle$$

and so,

$$-\frac{2}{n}\left\langle\mathbf{y}-\boldsymbol{\nu}_2,\mathbf{w}\right\rangle=\frac{1}{n}\left\|\boldsymbol{\theta}-\boldsymbol{\nu}_2\right\|^2-\frac{1}{n}\left\|\mathbf{y}-\boldsymbol{\nu}_2\right\|^2-\frac{1}{n}\left\|\mathbf{w}\right\|^2. \tag{58}$$

Using (58) in (57), we obtain

$$\begin{aligned}\frac{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{JS_2}\|^2}{n}&=\frac{\sigma^4\|\mathbf{y}-\boldsymbol{\nu}_2\|^2/n}{\left(g\left(\|\mathbf{y}-\boldsymbol{\nu}_2\|^2/n\right)\right)^2}+\frac{\|\mathbf{w}\|^2}{n}\\&\quad+\left(\frac{\sigma^2}{g\left(\|\mathbf{y}-\boldsymbol{\nu}_2\|^2/n\right)}\right)\left(\frac{1}{n}\left\|\boldsymbol{\theta}-\boldsymbol{\nu}_2\right\|^2-\frac{1}{n}\left\|\mathbf{y}-\boldsymbol{\nu}_2\right\|^2-\frac{1}{n}\left\|\mathbf{w}\right\|^2\right).\end{aligned} \tag{59}$$

We now use the following results whose proofs are given in Appendix B.3 and Appendix B.4.

Lemma 13.

$$\frac{1}{n}\left\|\mathbf{y}-\boldsymbol{\nu}_2\right\|^2\doteq\alpha_n+\sigma^2+\kappa_n\delta+o(\delta). \tag{60}$$

where α_n is given by (24).

Lemma 14.

$$\frac{1}{n}\left\|\boldsymbol{\theta}-\boldsymbol{\nu}_2\right\|^2\doteq\beta_n+\kappa_n\delta+o(\delta) \tag{61}$$

where β_n is given by (23).

Using Lemma 7 together with (60), we have

$$g\left(\frac{\|\mathbf{y}-\boldsymbol{\nu}_2\|^2}{n}\right)\doteq g(\alpha_n+\sigma^2)+\kappa_n\delta+o(\delta). \tag{62}$$

Using (60), (61) and (62) together with Lemmas 5, 6, and 9, we obtain

$$\begin{aligned}\frac{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{JS_2}\|^2}{n}&\doteq\frac{\sigma^4\left(\alpha_n+\sigma^2\right)}{\left(g\left(\alpha_n+\sigma^2\right)\right)^2}+\sigma^2+\left(\frac{\sigma^2}{g\left(\alpha_n+\sigma^2\right)}\right)\left(\beta^2-\left(\alpha_n+\sigma^2\right)-\sigma^2\right)+\kappa_n\delta+o(\delta) \qquad (63)\\&=\begin{cases}\beta_n+\kappa_n\delta+o(\delta), & \text{if } \alpha_n<0,\\ \frac{\beta_n\sigma^2}{\alpha_n+\sigma^2}+\kappa_n\delta+o(\delta) & \text{otherwise.}\end{cases}\end{aligned}$$

Therefore, for any $\epsilon>0$,

$$\mathbb{P}\left(\left|\frac{\|\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{JS_2}\|^2}{n}-\left[\min\left(\beta_n,\frac{\beta_n\sigma^2}{\alpha_n+\sigma^2}\right)+\kappa_n\delta+o(\delta)\right]\right|\geq\epsilon\right)\leq Ke^{-\frac{nk\min(\epsilon^2,1)}{\|\boldsymbol{\theta}\|^2/n}}.$$

This proves (21) and hence, the first part of the theorem.

To prove the second part of the theorem, we use the following definition and result.

Definition 6.1. *(Uniform Integrability [17, p. 81]) A sequence $\{X_n\}_{n=1}^{\infty}$ is said to be uniformly integrable (UI) if*

$$\lim_{K\to\infty}\left(\limsup_{n\to\infty}\mathbb{E}\left[|X_n|\mathbf{1}_{\{|X_n|\geq K\}}\right]\right)=0. \tag{64}$$

Fact 1. *[18, Sec. 13.7] Let* $\{X_n\}_{n=1}^{\infty}$ *be a sequence in* $\mathcal{L}^1$*, equivalently* $\mathbb{E}|X_n| < \infty$, $\forall n$*. Also, let* $X \in \mathcal{L}^1$*. Then* $X_n \xrightarrow{\mathcal{L}^1} X$*, i.e.,* $\mathbb{E}(|X_n - X|) \to 0$*, if and only if the following two conditions are satisfied:*

1. $X_n \xrightarrow{P} X$,
2. *The sequence* $\{X_n\}_{n=1}^{\infty}$ *is UI.*

Now, consider the individual terms of the RHS of (59). Using Lemmas 5, 6 and 7, we obtain

$$\frac{\sigma^4 \|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n}{\left(g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)\right)^2} \doteq \frac{\sigma^4\left(\alpha_n + \sigma^2\right)}{\left(g\left(\alpha_n + \sigma^2\right)\right)^2} + \kappa_n \delta + o(\delta),$$

and so, from Lemma 3,

$$S_n := \frac{\sigma^4 \|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n}{\left(g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)\right)^2} - \left[\frac{\sigma^4\left(\alpha_n + \sigma^2\right)}{\left(g\left(\alpha_n + \sigma^2\right)\right)^2} + \kappa_n \delta + o(\delta)\right] \xrightarrow{a.s.} 0. \tag{65}$$

Similarly, we obtain

$$\begin{aligned} T_n &:= \frac{\sigma^2 \|\boldsymbol{\theta} - \boldsymbol{\nu}_2\|^2/n}{g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)} - \left[\frac{\beta_n \sigma^2}{g(\alpha_n + \sigma^2)} + \kappa_n \delta + o(\delta)\right] \xrightarrow{a.s.} 0, \\ U_n &:= \frac{\sigma^2 \|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n}{g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)} - \left[\frac{\sigma^2(\alpha_n + \sigma^2)}{g(\alpha_n + \sigma^2)} + \kappa_n \delta + o(\delta)\right] \xrightarrow{a.s.} 0, \\ V_n &:= \frac{\sigma^2 \|\mathbf{w}\|^2/n}{g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)} - \left[\frac{\sigma^4}{g(\alpha_n + \sigma^2)} + \kappa_n \delta + o(\delta)\right] \xrightarrow{a.s.} 0. \end{aligned} \tag{66}$$

Now, using (59) and (63), we can write

$$\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_2}\|^2}{n} - \left(\min\left(\beta_n, \frac{\beta_n \sigma^2}{\alpha_n + \sigma^2}\right) + \kappa_n \delta + o(\delta)\right) = S_n + T_n - U_n - V_n + \left(\frac{\|\mathbf{w}\|^2}{n} - \sigma^2\right).$$

Note from Jensen's inequality that $|\mathbb{E}[X_n] - \mathbb{E}X| \leq \mathbb{E}(|X_n - X|)$. We therefore have

$$\begin{aligned} &\left|\frac{1}{n} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}) - \left[\min\left(\beta_n, \frac{\beta_n \sigma^2}{\alpha_n + \sigma^2}\right) + \kappa_n \delta + o(\delta)\right]\right| \\ &\leq \mathbb{E}\left|\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_2}\|^2}{n} - \left[\min\left(\beta_n, \frac{\beta_n \sigma^2}{\alpha_n + \sigma^2}\right) + \kappa_n \delta + o(\delta)\right]\right| \\ &= \mathbb{E}\left|S_n + T_n - U_n - V_n + \frac{\|\mathbf{w}\|^2}{n} - \sigma^2\right| \leq \mathbb{E}\left|S_n\right| + \mathbb{E}\left|T_n\right| - \mathbb{E}\left|U_n\right| - \mathbb{E}\left|V_n\right| + \mathbb{E}\left|\frac{\|\mathbf{w}\|^2}{n} - \sigma^2\right|. \end{aligned} \tag{67}$$

We first show that $\frac{\|\mathbf{w}\|^2}{n} \xrightarrow{\mathcal{L}^1} \sigma^2$, i.e.,

$$\lim_{n \to \infty} \mathbb{E}\left[\left|\frac{\|\mathbf{w}\|^2}{n} - \sigma^2\right|\right] = 0. \tag{68}$$

This holds because

$$\mathbb{E}\left[\left|\frac{\|\mathbf{w}\|^2}{n} - \sigma^2\right|\right] = \int_0^\infty \mathbb{P}\left(\left|\frac{\|\mathbf{w}\|^2}{n} - \sigma^2\right| > x\right) dx \overset{(i)}{\leq} \int_0^1 2e^{-nkx^2} dx + \int_1^\infty 2e^{-nkx} dx$$
$$\leq \int_0^\infty 2e^{-nkx^2} dx + \int_0^\infty 2e^{-nkx} dx = \frac{2}{\sqrt{nk}} \int_0^\infty e^{-t^2} dt + \frac{2}{nk} \int_0^\infty e^{-t} dt \overset{n\to\infty}{\longrightarrow} 0,$$

where inequality (i) is due to Lemma 9.

Thus, from (67), to prove (22), it is sufficient to show that $\mathbb{E}\,|S_n|$, $\mathbb{E}\,|T_n|$, $\mathbb{E}\,|U_n|$ and $\mathbb{E}\,|V_n|$ all converge to 0 as $n \to \infty$. From Fact 1 and (65), (66), this implies that we need to show that $\{S_n\}_{n=1}^\infty, \{T_n\}_{n=1}^\infty, \{U_n\}_{n=1}^\infty, \{V_n\}_{n=1}^\infty$ are UI. Considering S_n, we have

$$\frac{\sigma^4 \|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n}{\left(g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)\right)^2} \leq \sigma^2, \qquad \frac{\sigma^4\left(\alpha_n + \sigma^2\right)}{\left(g\left(\alpha_n + \sigma^2\right)\right)^2} \leq \sigma^2, \quad \forall n,$$

and since the sum of the terms in (65) that involve δ have bounded absolute value for a chosen and fixed δ (see Note 1), there exists $M > 0$ such that $\forall n$, $|S_n| \leq 2\sigma^2 + M$. Hence, from Definition 6.1, $\{S_n\}_{n=1}^\infty$ is UI. By a similar argument, so is $\{U_n\}_{n=1}^\infty$. Next, considering V_n, we have

$$\frac{\sigma^2 \|\mathbf{w}\|^2/n}{g\left(\|\mathbf{y} - \boldsymbol{\nu}_2\|^2/n\right)} \leq \frac{\|\mathbf{w}\|^2}{n}, \qquad \frac{\sigma^4}{g(\alpha_n + \sigma^2)} \leq \sigma^2, \quad \forall n,$$

and hence, $|V_n| \leq \frac{\|\mathbf{w}\|^2}{n} + \sigma^2 + M$, $\forall n$. Note from (68) and Fact 1 that $\{\|\mathbf{w}\|^2/n\}_{n=1}^\infty$ is UI. To complete the proof, we use the following result whose proof is provided in Appendix A.3.

Lemma 15. *Let $\{Y_n\}_{n=1}^\infty$ be a UI sequence of positive-valued random variables, and let $\{X_n\}_{n=1}^\infty$ be a sequence of random variables such that $|X_n| \leq cY_n + a$, $\forall n$, where c and a are positive constants. Then, $\{X_n\}_{n=1}^\infty$ is also UI.*

Hence, $\{V_n\}_{n=1}^\infty$ is UI. Finally, considering T_n in (66), we see that

$$\frac{\sigma^2 \frac{\|\boldsymbol{\theta} - \boldsymbol{\nu}_2\|^2}{n}}{g\left(\frac{\|\mathbf{y} - \boldsymbol{\nu}_2\|^2}{n}\right)} = \frac{\sigma^2 \frac{\|\mathbf{y} - \boldsymbol{\nu}_2 - \mathbf{w}\|^2}{n}}{g\left(\frac{\|\mathbf{y} - \boldsymbol{\nu}_2\|^2}{n}\right)} \leq \frac{2\sigma^2\left(\frac{\|\mathbf{y} - \boldsymbol{\nu}_2\|^2}{n} + \frac{\|\mathbf{w}\|^2}{n}\right)}{g\left(\frac{\|\mathbf{y} - \boldsymbol{\nu}_2\|^2}{n}\right)} \leq 2\left(\sigma^2 + \frac{\|\mathbf{w}\|^2}{n}\right),$$
$$\frac{\beta_n \sigma^2}{g(\alpha_n + \sigma^2)} \leq \beta_n < \infty.$$

Note that the last inequality is due to the assumption that $\limsup_{n\to\infty} \|\boldsymbol{\theta}\|^2/n < \infty$. Therefore, $|T_n| \leq 2\|\mathbf{w}\|^2/n + 2\sigma^2 + M$, $\forall n$, where M is some finite constant. Thus, by Lemma 15, T_n is UI. Therefore, each of the terms of the RHS of (67) goes to 0 as $n \to \infty$, and this completes the proof of the theorem.

6.3 Proof of Theorem 2

Let

$$\mathcal{E}_n := \left\{\mathbf{y} \in \mathbb{R}^n \ : \ \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_1}\|^2 < \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_2}\|^2\right\},$$
$$\Delta_n := \frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_2}\|^2}{n} - \frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_1}\|^2}{n}.$$

Without loss of generality, for a given $\epsilon > 0$, we can assume that $|\Delta_n| > \epsilon$ because if not, it is clear that

$$\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_H}\|^2}{n} - \min\left(\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_1}\|^2}{n}, \frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_2}\|^2}{n}\right) \leq \epsilon.$$

From (31) and Lemma 6, we obtain the following concentration inequality for the loss estimate in (32):

$$\frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1}) \doteq \frac{\rho_n \sigma^2}{\rho_n + \sigma^2}.$$

Using this together with Corollary 2, we obtain

$$\frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1}) \doteq \frac{1}{n}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_1}\|^2. \tag{69}$$

Following steps similar to those in the proof of Lemma 13, we obtain the following for the loss estimate in (35):

$$\frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}) \doteq \frac{\beta_n \sigma^2}{g(\alpha_n + \sigma^2)} + \kappa_n \delta + o(\delta). \tag{70}$$

Combining this with Theorem 1, we have

$$\frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}) \doteq \frac{1}{n}\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_2}\|^2. \tag{71}$$

Then, from (69), (71), and Lemma 4, we have $\frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1}) - \frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}) \doteq -\Delta_n$. We therefore have, for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1}) - \frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}) - (-\Delta_n) \geq \epsilon\right) \leq K e^{-\frac{nk\min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}} \tag{72}$$

for some positive constants k and K. Let $\mathbb{P}(\gamma_{\mathbf{y}} = 0, \Delta_n > \epsilon)$ denote the probability that $\gamma_{\mathbf{y}} = 0$ and $\Delta_n > \epsilon$ for a chosen $\epsilon > 0$. Therefore,

$$\begin{aligned}\mathbb{P}(\gamma_{\mathbf{y}} = 0, \Delta_n > \epsilon) &= \mathbb{P}\left(\frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1}) > \frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2})\ ,\ \Delta_n > \epsilon\right)\\ &\leq \mathbb{P}\left(\frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1}) + \Delta_n > \frac{1}{n}\hat{L}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}) + \epsilon\right) \leq K e^{-\frac{nk\min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}},\end{aligned} \tag{73}$$

where the last inequality is obtained from (72). So for any $\epsilon > 0$, we have

$$\mathbb{P}\left(\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_H}\|^2}{n} - \frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_1}\|^2}{n} \geq \epsilon \ \middle|\ \mathbf{y} \in \mathcal{E}_n\right) \leq \mathbb{P}(\gamma_{\mathbf{y}} = 0, \Delta_n > \epsilon) \leq K e^{-\frac{nk\min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}}.$$

In a similar manner, we obtain for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_H}\|^2}{n} - \frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_2}\|^2}{n} \geq \epsilon \ \middle|\ \mathbf{y} \in \mathcal{E}_n^c\right) \leq \mathbb{P}(\gamma_{\mathbf{y}} = 1, -\Delta_n > \epsilon) \leq K e^{-\frac{nk\min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}}.$$

Therefore, we arrive at

$$\mathbb{P}\left(\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_H}\|^2}{n} - \min_{i=1,2}\left(\frac{\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_i}\|^2}{n}\right) \geq \epsilon\right) \leq K e^{-\frac{nk\min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}}.$$

This proves the first part of the theorem.

For the second part, fix $\epsilon > 0$. First suppose that $\hat{\boldsymbol{\theta}}_{JS_1}$ has lower risk. For a given $\boldsymbol{\theta}$, let

$$\begin{aligned}
\mathcal{A}_{JS_1}(\boldsymbol{\theta}) &:= \{\mathbf{y} \in \mathbb{R}^n : \gamma_{\mathbf{y}} = 1\},\\
\mathcal{A}_{JS_{2a}}(\boldsymbol{\theta}) &:= \{\mathbf{y} \in \mathbb{R}^n : \gamma_{\mathbf{y}} = 0, \ \text{ and } \Delta_n \leq \epsilon\},\\
\mathcal{A}_{JS_{2b}}(\boldsymbol{\theta}) &:= \mathbb{R}^n \backslash (\mathcal{A}_{JS_1}(\boldsymbol{\theta}) \cup \mathcal{A}_{JS_{2a}}(\boldsymbol{\theta})) = \{\mathbf{y} \in \mathbb{R}^n : \gamma_{\mathbf{y}} = 0, \ \text{and } \Delta_n > \epsilon\}.
\end{aligned}$$

Denoting $\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|\mathbf{y}-\boldsymbol{\theta}\|^2}{2\sigma^2}\right)$ by $\phi(\mathbf{y};\boldsymbol{\theta})$, we have

$$\begin{aligned}
R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_H}) &= \int_{\mathcal{A}_{JS_1}(\boldsymbol{\theta})} \phi(\mathbf{y};\boldsymbol{\theta}) \|\hat{\boldsymbol{\theta}}_{JS_1} - \boldsymbol{\theta}\|^2 \, d\mathbf{y} + \int_{\mathcal{A}_{JS_{2a}}(\boldsymbol{\theta}) \cup \mathcal{A}_{JS_{2b}}(\boldsymbol{\theta})} \phi(\mathbf{y};\boldsymbol{\theta}) \|\hat{\boldsymbol{\theta}}_{JS_2} - \boldsymbol{\theta}\|^2 \, d\mathbf{y}\\
&\overset{(a)}{\leq} \int_{\mathcal{A}_{JS_1}(\boldsymbol{\theta})} \phi(\mathbf{y};\boldsymbol{\theta}) \|\hat{\boldsymbol{\theta}}_{JS_1} - \boldsymbol{\theta}\|^2 \, d\mathbf{y} + \int_{\mathcal{A}_{JS_{2a}}(\boldsymbol{\theta})} \phi(\mathbf{y};\boldsymbol{\theta}) \left(\|\hat{\boldsymbol{\theta}}_{JS_1} - \boldsymbol{\theta}\|^2 + n\epsilon\right) d\mathbf{y}\\
&\quad + \int_{\mathcal{A}_{JS_{2b}}(\boldsymbol{\theta})} \phi(\mathbf{y};\boldsymbol{\theta}) \|\hat{\boldsymbol{\theta}}_{JS_2} - \boldsymbol{\theta}\|^2 \, d\mathbf{y}\\
&\overset{(b)}{\leq} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1}) + n\epsilon + \left(\mathbb{P}\left(\gamma_{\mathbf{y}} = 0, \Delta_n > \epsilon\right) \int_{\mathcal{A}_{JS_{2b}}(\boldsymbol{\theta})} \phi(\mathbf{y};\boldsymbol{\theta}) \|\hat{\boldsymbol{\theta}}_{JS_2} - \boldsymbol{\theta}\|^4 \, d\mathbf{y}\right)^{1/2}\\
&\overset{(c)}{\leq} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_1}) + n\epsilon + Ke^{-\frac{nk\min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}} \left(\mathbb{E}\|\hat{\boldsymbol{\theta}}_{JS_2} - \boldsymbol{\theta}\|^4\right)^{1/2} \qquad (74)
\end{aligned}$$

where step (a) uses the definition of $\mathcal{A}_{JS_{2a}}$, in step (b) the last term is obtained using the Cauchy-Schwarz inequality on the product of the functions $\sqrt{\phi(\mathbf{y};\boldsymbol{\theta})}$, and $\sqrt{\phi(\mathbf{y};\boldsymbol{\theta})}\|\hat{\boldsymbol{\theta}}_{JS_2} - \boldsymbol{\theta}\|^2$. Step (c) is from (73).

Similarly, when $\hat{\boldsymbol{\theta}}_{JS_2}$ has lower risk, we get

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_H}) \leq R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_2}) + n\epsilon + Ke^{-\frac{nk\min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}} \left(\mathbb{E}\|\hat{\boldsymbol{\theta}}_{JS_1} - \boldsymbol{\theta}\|^4\right)^{1/2}. \qquad (75)$$

Hence, from (74)-(75), we obtain

$$\frac{1}{n} R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_H}) \leq \frac{1}{n}\left[\min_{i=1,2}\left(R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_i})\right) + n\epsilon + Ke^{-\frac{nk\min(\epsilon^2,1)}{\max(\|\boldsymbol{\theta}\|^2/n,1)}} \max_{i=1,2}\left(\left(\mathbb{E}\|\hat{\boldsymbol{\theta}}_{JS_i} - \boldsymbol{\theta}\|^4\right)^{1/2}\right)\right].$$

Now, noting that by assumption, $\limsup_{n\to\infty}\left(\mathbb{E}\|\hat{\boldsymbol{\theta}}_{JS_i} - \boldsymbol{\theta}\|^4\right)^{1/2}/n$ is finite, we get

$$\limsup_{n\to\infty} \frac{1}{n}\left[R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_H}) - \min_{i=1,2}\left(R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_i})\right) - \epsilon\right] \leq 0.$$

Since this is true for every $\epsilon > 0$, we therefore have

$$\limsup_{n\to\infty} \frac{1}{n}\left[R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_H}) - \min_{i=1,2}\left(R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS_i})\right)\right] \leq 0. \qquad (76)$$

This completes the proof of the theorem.

Note 3. *Note that in the best case scenario,* $\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_H}\|^2 = \min\left(\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_1}\|^2, \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{JS_2}\|^2\right)$*, which occurs when for each realization of* $\mathbf{y}$*, the hybrid estimator picks the better of the two rival estimators* $\hat{\boldsymbol{\theta}}_{JS_1}$ *and* $\hat{\boldsymbol{\theta}}_{JS_2}$*. In this case, the inequality in* (76) *is strict, provided that there are realizations of* $\mathbf{y}$ *with non-zero probability measure for which one estimator is strictly better than the other.*

6.4 Proof of Theorem 3

The proof is similar to that of Theorem 1, so we only provide a sketch. Note that for a_i, b_i, real-valued and finite, $i = 1, \cdots, n$, with $a_i < b_i$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbf{1}_{\{a_i<w_i\leq b_i\}}\right] = \frac{1}{n}\sum_{i=1}^{n}\left[Q\left(\frac{a_i}{\sigma}\right) - Q\left(\frac{b_i}{\sigma}\right)\right],$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[w_i\mathbf{1}_{\{a_i<w_i\leq b_i\}}\right] = \frac{\sigma}{n\sqrt{2\pi}}\sum_{i=1}^{n}\left(e^{-a_i^2/2\sigma^2} - e^{-b_i^2/2\sigma^2}\right).$$

Since $\mathbf{1}_{\{a_i<w_i\leq b_i\}} \in [0,1]$, it follows that $w_i\mathbf{1}_{\{a_i<w_i\leq b_i\}} \in [m_i, n_i]$ where $m_i = \min(0, a_i)$, $n_i = \max(0, b_i)$. So, from Hoeffding's inequality, we obtain

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}\mathbf{1}_{\{a_i<w_i\leq b_i\}} - \sum_{i=1}^{n}\left[Q\left(\frac{a_i}{\sigma}\right) - Q\left(\frac{b_i}{\sigma}\right)\right]\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2},$$

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}w_i\mathbf{1}_{\{a_i<w_i\leq b_i\}} - \frac{\sigma}{\sqrt{2\pi}}\sum_{i=1}^{n}\left(e^{-\frac{a_i^2}{2\sigma^2}} - e^{-\frac{b_i^2}{2\sigma^2}}\right)\right| \geq \epsilon\right) \leq 2e^{-\frac{2n\epsilon^2}{\sum_{i=1}^{n}(n_i-m_i)^2}}.$$

Subsequently, the steps of Lemma 13 are used to obtain

$$\begin{aligned}\frac{1}{n}\left\|\mathbf{y} - \boldsymbol{\nu}_{\mathbf{y}_L}\right\|^2 \doteq \frac{\|\boldsymbol{\theta}\|^2}{n} - \sum_{j=0}^{L}\frac{c_j^2}{n}\sum_{i=1}^{n}\left[Q\left(\frac{\mu_j - \theta_i}{\sigma}\right) - Q\left(\frac{\mu_{j-1} - \theta_i}{\sigma}\right)\right]\\ - \left(\frac{2}{n}\right)\left(\frac{\sigma}{\sqrt{2\pi}}\right)\left(\sum_{j=1}^{L}c_j\sum_{i=1}^{n}\left[e^{-\frac{\left(\mu_j-\theta_i\right)^2}{2\sigma^2}} - e^{-\frac{\left(\mu_{j-1}-\theta_i\right)^2}{2\sigma^2}}\right]\right) + \kappa_n\delta + o(\delta).\end{aligned} \tag{77}$$

Finally, employing the steps of Lemma 14, we get

$$\frac{1}{n}\left\|\boldsymbol{\theta} - \boldsymbol{\nu}_{\mathbf{y}_L}\right\|^2 \doteq \frac{\|\boldsymbol{\theta}\|^2}{n} - \sum_{j=0}^{L}\frac{c_j^2}{n}\sum_{i=1}^{n}\left[Q\left(\frac{\mu_j - \theta_i}{\sigma}\right) - Q\left(\frac{\mu_{j-1} - \theta_i}{\sigma}\right)\right] + \kappa_n\delta + o(\delta).$$

The subsequent steps of the proof are along the lines of that of Theorem 1.

7 Concluding remarks

In this paper, we presented a class of shrinkage estimators that take advantage of the large dimensionality to infer the clustering structure of the parameter values from the data. This structure is then used to construct an attracting vector for the shrinkage estimator. A good cluster-based attracting vector enables significant risk reduction over the ML-estimator even when $\boldsymbol{\theta}$ is composed of several inhomogeneous quantities.

We obtained concentration bounds for the squared-error loss of the constructed estimators and convergence results for the risk. The estimators have significantly smaller risks than the regular JS-estimator for a wide range of $\boldsymbol{\theta} \in \mathbb{R}^n$, even though they do not dominate the regular (positive-part) JS-estimator for finite n.

An important next step is to test the performance of the proposed estimators on real data sets. It would be interesting to adapt these estimators and analyze their risks when the sample values are bounded by a known value, i.e., when $|\theta_i| \leq \tau$, $\forall i = 1, \cdots, n$, with τ known. Another open question is how one should decide the maximum number of clusters to be considered for the hybrid estimator.

An interesting direction for future research is to study confidence sets centered on the estimators in this paper, and compare them to confidence sets centered on the positive-part JS-estimator, which were studied in [19, 20].

The James-Stein estimator for colored Gaussian noise, i.e., for $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ known, has been studied in [21], and variants have been proposed in [22], [23]. It would be interesting to extend the ideas in this paper to the case of colored Gaussian noise, and to noise that has a general sub-Gaussian distribution. Yet another research direction is to construct multi-dimensional target subspaces from the data that are more general than the cluster-based subspaces proposed here. The goal is to obtain greater risk savings for a wider range of $\boldsymbol{\theta} \in \mathbb{R}^n$, at the cost of having a more complex attractor.

Appendices

A Proofs of General Lemmas

A.1 Proof of Lemma 10

Note that $\mathbb{E}\left[w_i 1_{\{w_i > a_i\}}\right] = \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{a_i^2}{2\sigma^2}}$. So, with

$$X := \sum_{i=1}^{n} w_i 1_{\{w_i > a_i\}} - \frac{\sigma}{\sqrt{2\pi}} \sum_{i=1}^{n} e^{-\frac{a_i^2}{2\sigma^2}},$$

we have $\mathbb{E}X = 0$. Let $m_i := \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{a_i^2}{2\sigma^2}}$, and consider the moment generating function (MGF) of X. We have

$$\begin{aligned}\mathbb{E}\left[e^{\lambda X}\right] &= \prod_{i=1}^{n} \frac{e^{-\lambda m_i}}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{\left(\lambda w_i 1_{\{w_i > a_i\}}\right)} e^{-\frac{w_i^2}{2\sigma^2}} dw_i = \prod_{i=1}^{n} \frac{e^{-\lambda m_i}}{\sqrt{2\pi\sigma^2}} \left[\int_{a_i}^{\infty} e^{\lambda w_i} e^{-\frac{w_i^2}{2\sigma^2}} dw_i + \int_{-\infty}^{a_i} e^{-\frac{w_i^2}{2\sigma^2}} dw_i\right] \\ &= \prod_{i=1}^{n} e^{-\lambda m_i} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \int_{a_i}^{\infty} e^{\lambda w_i} e^{-\frac{w_i^2}{2\sigma^2}} dw_i + 1 - Q\left(\frac{a_i}{\sigma}\right)\right] \\ &= \prod_{i=1}^{n} e^{-\lambda m_i} \left[e^{\frac{\lambda^2\sigma^2}{2}} Q\left(\frac{a_i}{\sigma} - \lambda\sigma\right) + 1 - Q\left(\frac{a_i}{\sigma}\right)\right].\end{aligned} \tag{78}$$

Now, for any positive real number b, consider the function

$$f(x; b) = e^{-\frac{b}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}} \left[e^{\frac{b^2}{2}} Q(x - b) + 1 - Q(x)\right].$$

Note that the RHS of (78) can be written as $\prod_{i=1}^{n} f(\frac{a_i}{\sigma}; \lambda\sigma)$. We will bound the MGF in (78) by bounding $f(x; b)$.

Clearly, $f(-\infty;b) = e^{\frac{b^2}{2}}$, and since $b > 0$, we have for $x \leq 0$,

$$\begin{aligned}
&f(x;b) < e^{-\frac{b}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}} \left[e^{\frac{b^2}{2}} Q\left(x-b\right) + e^{\frac{b^2}{2}} \left(1 - Q\left(x\right)\right) \right] \\
&= \left(e^{-\frac{b}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}} \right) \left(e^{\frac{b^2}{2}} \right) \left[Q\left(x-b\right) + 1 - Q\left(x\right) \right] = \left(e^{-\frac{b}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}} \right) \left(e^{\frac{b^2}{2}} \right) \left(1 + \int_{x-b}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du \right) \\
&\overset{(i)}{=} \left(e^{-\frac{b}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}} \right) \left(e^{\frac{b^2}{2}} \right) \left(1 + \frac{b}{\sqrt{2\pi}} e^{-\frac{c^2}{2}} \right), \quad \text{for some } c \in (x-b, x) \\
&\overset{(j)}{\leq} \left(e^{-\frac{b}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}} \right) \left(e^{\frac{b^2}{2}} \right) \left(e^{\frac{b}{\sqrt{2\pi}}e^{-\frac{c^2}{2}}} \right) \overset{(k)}{<} e^{\frac{b^2}{2}}
\end{aligned}$$

where (i) is from the first mean value theorem for integrals, (j) is because $e^x \geq 1 + x$ for $x \geq 0$, and (k) is because for $x \leq 0$, $e^{-x^2} > e^{-(x-b)^2}$ for $b > 0$. Therefore,

$$\sup_{x \in (-\infty, 0]} f(x;b) = e^{\frac{b^2}{2}}. \tag{79}$$

Now, for $x \geq 0$, consider

$$h(x) := \frac{f(-x;b) - f(x;b)}{e^{-\frac{b}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}}} = e^{\frac{b^2}{2}} \left[Q\left(-x-b\right) - Q\left(x-b\right) \right] + Q\left(x\right) - Q\left(-x\right).$$

We have $h(0) = 0$ and

$$\begin{aligned}
\sqrt{2\pi}\frac{dh(x)}{dx} &= e^{\frac{b^2}{2}} \left[e^{-\frac{(x+b)^2}{2}} + e^{-\frac{(x-b)^2}{2}} \right] - 2e^{-\frac{x^2}{2}} \\
&= e^{\frac{-x^2}{2}} \left[e^{-bx} + e^{bx} \right] - 2e^{-\frac{x^2}{2}} = e^{\frac{-x^2}{2}} \left[e^{-bx} + e^{bx} - 2 \right] = 2e^{\frac{-x^2}{2}} \left[\cosh(bx) - 1 \right] \geq 0
\end{aligned}$$

because $\cosh(bx) \geq 1$. This establishes that $h(x)$ is monotone non-decreasing in $[0, \infty)$ with $h(0) = 0$, and hence, for $x \in [0, \infty)$,

$$h(x) \geq 0 \Rightarrow f(-x;b) \geq f(x;b). \tag{80}$$

Finally, from (79) and (80), it follows that

$$\sup_{x \in (-\infty, \infty)} f(x;b) = e^{\frac{b^2}{2}}. \tag{81}$$

Using (81) in (78), we obtain $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{n\lambda^2\sigma^2}{2}}$. Hence, applying the Chernoff trick, we have for $\lambda > 0$:

$$\mathbb{P}\left(X \geq \epsilon\right) = \mathbb{P}\left(e^{\lambda X} \geq e^{\lambda\epsilon}\right) \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda\epsilon}} \leq e^{-\left(\lambda\epsilon - \frac{n\lambda^2\sigma^2}{2}\right)}.$$

Choosing $\lambda = \frac{\epsilon}{n\sigma^2}$ which minimizes $e^{-\left(\lambda\epsilon - \frac{n\lambda^2\sigma^2}{2}\right)}$, we get $\mathbb{P}\left(X \geq \epsilon\right) \leq e^{-\frac{\epsilon^2}{2n\sigma^2}}$ and so,

$$\mathbb{P}\left(\frac{X}{n} \geq \epsilon\right) = \mathbb{P}\left(\frac{1}{n}\left(\sum_{i=1}^{n} w_i 1_{\{w_i > a_i\}} - \frac{\sigma}{\sqrt{2\pi}} \sum_{i=1}^{n} e^{-\frac{a_i^2}{2\sigma^2}}\right) \geq \epsilon\right) \leq e^{-\frac{n\epsilon^2}{2\sigma^2}}. \tag{82}$$

To obtain the lower tail inequality, we use the following result:

Fact 2. *[24, Thm. 3.7]. For independent random variables X_i satisfying $X_i \geq -M$, for $1 \leq i \leq n$, we have for any $\epsilon > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}[X_i] \leq -\epsilon\right) \leq e^{-\frac{\epsilon^2}{2\left(\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] + \frac{M\epsilon}{3}\right)}}.$$

So, for $X_i = w_i 1_{\{w_i > a_i\}}$, we have $X_i \geq \min\{0, a_i, i = 1, \cdots, n\}$, and $\mathbb{E}\left[X_i^2\right] \leq \sigma^2$, $\forall i = 1, \cdots, n$. Clearly, we can take $M = -\min\{0, a_i, i = 1, \cdots, n\} < \infty$. Therefore, for any $\epsilon > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}[X_i] \leq -\epsilon\right) \leq e^{-\frac{\epsilon^2}{2\left(\sum_{i=1}^{n} \mathbb{E}\left[X_i^2\right] + \frac{M\epsilon}{3}\right)}} \leq e^{-\frac{\epsilon^2}{2\left(n\sigma^2 + \frac{M\epsilon}{3}\right)}}$$

and hence,

$$\mathbb{P}\left(\frac{1}{n}\left(\sum_{i=1}^{n} w_i 1_{\{w_i > a_i\}} - \frac{\sigma}{\sqrt{2\pi}} \sum_{i=1}^{n} e^{-\frac{a_i^2}{2\sigma^2}}\right) \leq -\epsilon\right) \leq e^{-\frac{n\epsilon^2}{2\left(\sigma^2 + \frac{M\epsilon}{3}\right)}}. \tag{83}$$

Using the upper and lower tail inequalities obtained in (82) and (83), respectively, we get

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} w_i 1_{\{w_i > a_i\}} - \frac{\sigma}{\sqrt{2\pi}} \sum_{i=1}^{n} e^{-\frac{a_i^2}{2\sigma^2}}\right| \geq \epsilon\right) \leq 2e^{-\frac{n\epsilon^2}{2\left(\sigma^2 + \frac{M\epsilon}{3}\right)}} \leq 2e^{-nk\min(\epsilon, \epsilon^2)}$$

where k is a positive constant (this is due to M being finite). This proves (52). The concentration inequality in (53) can be similarly proven, and will not be detailed here.

A.2 Proof of Lemma 11

We first prove (55). Then (54) immediately follows by setting $\theta_i = 1$, $\forall i$.

Let us denote the event whose probability we want to bound by $\mathcal{E}$. In our case,

$$\mathcal{E} = \left\{\frac{1}{n}\left|\sum_{i=1}^{n} \theta_i 1_{\{y_i > f(\mathbf{y})\}} - \sum_{i=1}^{n} \theta_i 1_{\{y_i > a\}}\right| \geq \epsilon\right\}.$$

Then, for any $t > 0$, we have

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &\leq \mathbb{P}(\mathcal{E},\, \{a < f(\mathbf{y}) \leq a + t\}) + \mathbb{P}(\mathcal{E},\, \{a - t \leq f(\mathbf{y}) \leq a\}) + \mathbb{P}(|f(\mathbf{y}) - a| > t) \\
&= \mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} \theta_i 1_{\{a < y_i \leq f(\mathbf{y})\}}\right| \geq \epsilon,\, \{a < f(\mathbf{y}) \leq a + t\}\right) \\
&\quad + \mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} \theta_i 1_{\{f(\mathbf{y}) < y_i \leq a\}}\right| \geq \epsilon,\, \{a - t < f(\mathbf{y}) \leq a\}\right) + \mathbb{P}(|f(\mathbf{y}) - a| > t) \\
&\leq \mathbb{P}\left(\left[\frac{1}{n}\sum_{i=1}^{n} |\theta_i| 1_{\{a < y_i \leq a + t\}}\right] \geq \epsilon\right) + \mathbb{P}\left(\left[\frac{1}{n}\sum_{i=1}^{n} |\theta_i| 1_{\{a - t < y_i \leq a\}}\right] \geq \epsilon\right) + 2e^{-nkt^2}.
\end{aligned} \tag{84}$$

Now,

$$\mathbb{P}(1_{\{a < y_i \leq a + t\}} = 1) = \int_a^{a+t} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \theta_i)^2}{2\sigma^2}} dy_i \leq \frac{t}{\sqrt{2\pi\sigma^2}} \tag{85}$$

where we have used $e^{-\frac{(y_i-\theta_i)^2}{2\sigma^2}} \leq 1$. Let $Y := \frac{1}{n}\sum_{i=1}^n Y_i$ where $Y_i := |\theta_i| 1_{\{a<y_i \leq a+t\}}$. Then, from (85), we have

$$0 \leq \mathbb{E}Y = \frac{1}{n}\sum_{i=1}^n |\theta_i| \, \mathbb{P}(1_{\{a<y_i\leq a+t\}} = 1) \leq \frac{t}{n\sqrt{2\pi\sigma^2}} \sum_{i=1}^n |\theta_i|.$$

Since $Y_i \in [0, |\theta_i|]$, from Hoeffding's inequality, for any $\epsilon_1 > 0$, we have

$$\mathbb{P}\left(Y - \mathbb{E}Y \geq \epsilon_1\right) \leq e^{-\frac{2n\epsilon_1^2}{\|\boldsymbol{\theta}\|^2/n}} \Rightarrow \mathbb{P}\left(Y \geq \epsilon_1 + \frac{t\|\boldsymbol{\theta}\|_1}{n\sqrt{2\pi\sigma^2}}\right) \leq e^{-\frac{2n\epsilon_1^2}{\|\boldsymbol{\theta}\|^2/n}},$$

where $\|\boldsymbol{\theta}\|_1 := \sum_{i=1}^n |\theta_i|$. Now, set $\epsilon_1 = \epsilon/2$ and

$$t = \frac{\epsilon\sqrt{\pi\sigma^2/2}}{\|\boldsymbol{\theta}\|_1/n} \tag{86}$$

to obtain

$$\mathbb{P}\left(Y \geq \epsilon\right) \leq e^{-\frac{n\epsilon^2}{2\|\boldsymbol{\theta}\|^2/n}} \Rightarrow \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |\theta_i| 1_{\{a<y_i\leq a+t\}} \geq \epsilon\right) \leq e^{-\frac{n\epsilon^2}{2\|\boldsymbol{\theta}\|^2/n}}. \tag{87}$$

A similar analysis yields

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |\theta_i| 1_{\{a-t<y_i\leq a\}} \geq \epsilon\right) \leq e^{-\frac{n\epsilon^2}{2\|\boldsymbol{\theta}\|^2/n}}. \tag{88}$$

Using (87) and (88) in (84) and recalling that t is given by (86), we obtain

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n \theta_i 1_{\{y_i > f(\mathbf{y})\}} - \sum_{i=1}^n \theta_i 1_{\{y_i > a\}}\right| \geq \epsilon\right) \leq 2\left(e^{-\frac{n\epsilon^2 k\pi\sigma^2}{2\|\boldsymbol{\theta}\|_1^2/n^2}} + e^{-\frac{n\epsilon^2}{2\|\boldsymbol{\theta}\|^2/n}}\right) \leq 4e^{-\frac{nk\epsilon^2}{\|\boldsymbol{\theta}\|^2/n}}$$

where k is a positive constant. The last inequality holds because $\|\boldsymbol{\theta}\|_1^2/n^2 < \|\boldsymbol{\theta}\|^2/n$ (by the Cauchy-Schwarz inequality), and $\limsup_{n\to\infty} \|\boldsymbol{\theta}\|^2/n < \infty$ (by assumption). This proves (55).

Next, we prove (56). Using steps very similar to (84), we have, for $t > 0$, $\epsilon > 0$,

$$\begin{aligned} &\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^n w_i 1_{\{y_i > f(\mathbf{y})\}} - \sum_{i=1}^n w_i 1_{\{y_i > a\}}\right| \geq \epsilon\right) \\ &\leq 2e^{-nkt^2} + \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |w_i| 1_{\{a<y_i\leq a+t\}} \geq \epsilon\right) + \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |w_i| 1_{\{a-t<y_i\leq a\}} \geq \epsilon\right). \end{aligned} \tag{89}$$

Now, let $Y := \frac{1}{n}\sum_{i=1}^n Y_i$ where

$$Y_i := |w_i| 1_{\{a<y_i\leq a+t\}} = |w_i| 1_{\{a-\theta_i<w_i\leq a-\theta_i+t\}}.$$

Noting that $|w_i| \leq t + |a-\theta_i|$ when $w_i \in [a-\theta_i, a-\theta_i+t]$, we have

$$\mathbb{E}[Y_i] = \int_{a-\theta_i}^{a-\theta_i+t} \frac{|w|}{\sqrt{2\pi\sigma^2}} e^{-w^2/2\sigma^2} dw \overset{(i)}{=} t\left(\frac{|c|}{\sqrt{2\pi\sigma^2}} e^{-c^2/2\sigma^2}\right) \overset{(j)}{\leq} \frac{t}{\sqrt{2\pi e}}.$$

Note that (i) is from the mean value theorem for integrals with $c \in (a-\theta_i, a-\theta_i+t)$, and (j) is because $xe^{-x^2} \leq 1/\sqrt{2e}$ for $x \geq 0$. Hence

$$0 \leq \mathbb{E}[Y] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y_i] \leq \frac{t}{\sqrt{2\pi e}}.$$

As each Y_i takes values in an interval of length at most t, by Hoeffding's inequality we have for any $\epsilon_1 > 0$

$$\mathbb{P}(Y \geq \epsilon_1 + \mathbb{E}[Y]) \leq 2e^{-2n\epsilon_1^2/t^2} \Rightarrow \mathbb{P}\left(Y \geq \epsilon_1 + \frac{t}{\sqrt{2\pi e}}\right) \leq 2e^{-2n\epsilon_1^2/t^2}. \tag{90}$$

Now, set $\frac{t}{\sqrt{2\pi e}} = \sqrt{\epsilon_1}$. Using this value of t in the RHS of (90), we obtain

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}|w_i|\mathbf{1}_{\{a<y_i\leq a+t\}} \geq \epsilon_1 + \sqrt{\epsilon_1}\right) \leq 2e^{-nk_1\epsilon_1}$$

where $k_1 = 1/(\pi e)$. Setting $\epsilon_1 + \sqrt{\epsilon_1} = \epsilon$, we get $\sqrt{\epsilon_1} = \frac{\sqrt{4\epsilon+1}-1}{2}$. Using the following inequality for $x > 0$:

$$\left(\sqrt{1+x}-1\right)^2 \geq \begin{cases} x^2/32, & 0 \leq x \leq 3 \\ 3x/4, & x > 3, \end{cases}$$

we obtain,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}|w_i|\mathbf{1}_{\{a<y_i\leq a+t\}} \geq \epsilon\right) \leq 2e^{-nk\min(\epsilon^2,\epsilon)} \tag{91}$$

where k is a positive constant. Using similar steps, it can be shown that the third term on the RHS of (89) can also be bounded as

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}|w_i|\mathbf{1}_{\{a-t<y_i\leq a\}} \geq \epsilon\right) \leq 2e^{-nk\min(\epsilon^2,\epsilon)}. \tag{92}$$

This completes the proof of (56).

A.3 Proof of Lemma 15

Since $\{Y_n\}_{n=1}^{\infty}$ is UI, from Definition 6.1, we have $\lim_{K\to\infty}\left(\limsup_{n\to\infty}\mathbb{E}\left[Y_n\mathbf{1}_{\{Y_n\geq K\}}\right]\right) = 0$. Therefore,

$$\mathbb{E}\left[|X_n|\mathbf{1}_{\{|X_n|\geq K\}}\right] \leq \mathbb{E}\left[c|Y_n|\mathbf{1}_{\{|X_n|\geq K\}}\right] + \mathbb{E}\left[a\mathbf{1}_{\{|X_n|\geq K\}}\right] \leq c\mathbb{E}\left[Y_n\mathbf{1}_{\{cY_n+a\geq K\}}\right] + a\mathbb{E}\left[\mathbf{1}_{\{cY_n+a\geq K\}}\right]$$
$$= c\mathbb{E}\left[Y_n\mathbf{1}_{\left\{Y_n\geq\frac{K-a}{c}\right\}}\right] + a\mathbb{E}\left[\mathbf{1}_{\left\{Y_n\geq\frac{K-a}{c}\right\}}\right] = c\mathbb{E}\left[Y_n\mathbf{1}_{\left\{Y_n\geq\frac{K-a}{c}\right\}}\right] + a\mathbb{P}\left(Y_n \geq \frac{K-a}{c}\right).$$

So,

$$\lim_{K\to\infty}\left(\limsup_{n\to\infty}\mathbb{E}\left[|X_n|\mathbf{1}_{\{|X_n|\geq K\}}\right]\right) \leq c\lim_{K\to\infty}\left(\limsup_{n\to\infty}\mathbb{E}\left[Y_n\mathbf{1}_{\left\{Y_n\geq\frac{K-a}{c}\right\}}\right]\right)$$
$$+ a\lim_{K\to\infty}\left(\limsup_{n\to\infty}\mathbb{P}\left(Y_n \geq \frac{K-a}{c}\right)\right) = 0.$$

B Proofs of Lemmas related to JS-estimators

B.1 Proof of Lemma 1

We first prove (16). Then, (17) and (18) immediately follow by setting $\theta_i = 1$, for $1 \leq i \leq n$.

From Lemma 11, for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n}\theta_i 1_{\{y_i > \bar{y}\}} - \sum_{i=1}^{n}\theta_i 1_{\{y_i > \bar{\theta}\}}\right|\right) \leq 4e^{-\frac{nk\epsilon^2}{\|\boldsymbol{\theta}\|^2/n}}. \tag{93}$$

Since $\theta_i 1_{\{y_i > \bar{\theta}\}} \in \{0, \theta_i\}$ are independent for $1 \leq i \leq n$, from Hoeffding's inequality, we have, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\theta_i 1_{\{y_i > \bar{\theta}\}} - \frac{1}{n}\sum_{i=1}^{n}\theta_i \mathbb{E}\left[1_{\{y_i > \bar{\theta}\}}\right]\right| > \epsilon\right) \leq 2e^{-\frac{2n\epsilon^2}{\|\boldsymbol{\theta}\|^2/n}}. \tag{94}$$

Also for each i,

$$\mathbb{E}\left[1_{\{y_i > \bar{\theta}\}}\right] = \mathbb{P}\left(y_i > \bar{\theta}\right) = \mathbb{P}\left(w_i > \bar{\theta} - \theta_i\right) = Q\left(\frac{\bar{\theta} - \theta_i}{\sigma}\right).$$

Therefore, from (93) and (94), we obtain

$$\frac{1}{n}\sum_{i=1}^{n}\theta_i 1_{\{y_i > \bar{y}\}} \doteq \frac{1}{n}\sum_{i=1}^{n}\theta_i Q\left(\frac{\bar{\theta} - \theta_i}{\sigma}\right). \tag{95}$$

The second concentration result in (16) immediately follows by writing $1_{\{y_i \leq \bar{y}\}} = 1 - 1_{\{y_i > \bar{y}\}}$.

To prove (14), we write

$$\frac{1}{n}\sum_{i=1}^{n} y_i 1_{\{y_i > \bar{y}\}} = \frac{1}{n}\sum_{i=1}^{n}\theta_i 1_{\{y_i > \bar{y}\}} + \frac{1}{n}\sum_{i=1}^{n} w_i 1_{\{y_i > \bar{y}\}}.$$

Hence, we have to show that

$$\frac{1}{n}\sum_{i=1}^{n} w_i 1_{\{y_i > \bar{y}\}} \doteq \frac{\sigma}{n\sqrt{2\pi}}\sum_{i=1}^{n} e^{-\frac{(\bar{\theta} - \theta_i)^2}{2\sigma^2}}. \tag{96}$$

From Lemma 11, for any $\epsilon > 0$, we have

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} w_i 1_{\{y_i > \bar{y}\}} - \sum_{i=1}^{n} w_i 1_{\{y_i > \bar{\theta}\}}\right|\right) \leq 4e^{-\frac{nk\epsilon^2}{\|\boldsymbol{\theta}\|^2/n}}. \tag{97}$$

Now,

$$\mathbb{E}\left[w_i 1_{\{y_i > \bar{\theta}\}}\right] = \int_{-\infty}^{\infty} w_i 1_{\{y_i > \bar{\theta}\}}\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{w_i^2}{2\sigma^2}}\right)dw_i = \int_{\bar{\theta} - \theta_i}^{\infty}\frac{w_i}{\sqrt{2\pi\sigma^2}}e^{-\frac{w_i^2}{2\sigma^2}}dw_i = \frac{\sigma}{\sqrt{2\pi}}e^{-\frac{(\bar{\theta} - \theta_i)^2}{2\sigma^2}}.$$

Using Lemma 10, we get, for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\left|\sum_{i=1}^{n} w_i 1_{\{w_i > \bar{\theta}\}} - \frac{\sigma}{\sqrt{2\pi}}\sum_{i=1}^{n} e^{-\frac{(\bar{\theta} - \theta_i)^2}{2\sigma^2}}\right| \geq \epsilon\right) \leq 2e^{-nk\min(\epsilon, \epsilon^2)}. \tag{98}$$

We obtain (96) by combining (97) and (98).

Similarly, (15) can be shown using Lemma 11 and Lemma 10 to establish that

$$\frac{1}{n}\sum_{i=1}^{n} w_i 1_{\{w_i \leq \bar{y}\}} \doteq -\frac{\sigma}{n\sqrt{2\pi}}\sum_{i=1}^{n} e^{-\frac{(\bar{\theta} - \theta_i)^2}{2\sigma^2}}.$$

B.2 Proof of Lemma 2

From Lemma 12, we have, for any $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{n}\frac{\sigma^2}{2\delta}\left|\sum_{i=0}^{n} 1_{\{|y_i-\bar{y}|\leq\delta\}} - \sum_{i=0}^{n} 1_{\left\{\left|y_i-\bar{\theta}\right|\leq\delta\right\}}\right| \geq \epsilon\right) \leq 8e^{-\frac{nk\epsilon^2\delta^2}{\sigma^4}}. \tag{99}$$

Further, from Hoeffding's inequality,

$$\mathbb{P}\left(\frac{1}{n}\left|\frac{\sigma^2}{2\delta}\sum_{i=0}^{n} 1_{\left\{\left|y_i-\bar{\theta}\right|\leq\delta\right\}} - \frac{\sigma^2}{2\delta}\sum_{i=0}^{n}\mathbb{E}\left[1_{\left\{\left|y_i-\bar{\theta}\right|\leq\delta\right\}}\right]\right| \geq \epsilon\right) \leq 2e^{-\frac{8n\delta^2\epsilon^2}{\sigma^4}}. \tag{100}$$

Also,

$$\frac{\sigma^2}{2\delta}\sum_{i=0}^{n}\mathbb{E}\left[1_{\left\{\left|y_i-\bar{\theta}\right|\leq\delta\right\}}\right] = \frac{\sigma^2}{2\delta}\sum_{i=0}^{n}\mathbb{P}\left(\left|y_i-\bar{\theta}\right|\leq\delta\right) = \frac{\sigma^2}{2\delta}\sum_{i=0}^{n}\int_{\bar{\theta}-\delta}^{\bar{\theta}+\delta}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-\theta_i)^2}{2\sigma^2}}dy_i. \tag{101}$$

From the first mean value theorem for integrals, $\exists\varepsilon_i \in (-\delta,\delta)$ such that

$$\int_{\bar{\theta}-\delta}^{\bar{\theta}+\delta}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-\theta_i)^2}{2\sigma^2}}dy_i = 2\delta\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(\bar{\theta}+\varepsilon_i-\theta_i)^2}{2\sigma^2}}\right)$$

and so the RHS of (101) can be written as

$$\frac{\sigma^2}{2n\delta}\sum_{i=0}^{n}\int_{\bar{\theta}-\delta}^{\bar{\theta}+\delta}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-\theta_i)^2}{2\sigma^2}}dy_i = \frac{\sigma}{n\sqrt{2\pi}}\sum_{i=0}^{n}e^{-\frac{(\bar{\theta}+\varepsilon_i-\theta_i)^2}{2\sigma^2}}.$$

Now, let $x_i := \bar{\theta}-\theta_i$. Then, since $|\epsilon_i| \leq \delta$, we have

$$\frac{1}{n\sqrt{2\pi\sigma^2}}\sum_{i=0}^{n}\left|e^{-\frac{x_i^2}{2\sigma^2}} - e^{-\frac{(x_i+\varepsilon_i)^2}{2\sigma^2}}\right| \leq \delta\max\left(\frac{d\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}\right)}{dx}\right) = \frac{\delta}{\sigma^2\sqrt{2\pi e}}.$$

Therefore,

$$\frac{\sigma^2}{2n\delta}\sum_{i=0}^{n}\int_{\bar{\theta}-\delta}^{\bar{\theta}+\delta}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y_i-\theta_i)^2}{2\sigma^2}}dy_i = \frac{\sigma}{n\sqrt{2\pi}}\sum_{i=0}^{n}e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}} + \kappa_n\delta \tag{102}$$

where $|\kappa_n| \leq \frac{1}{\sqrt{2\pi e}}$. Using (102) in (101), and then the obtained result in (100) and (99), the proof of the lemma is complete.

B.3 Proof of Lemma 13

We have

$$\frac{1}{n}\|\mathbf{y}-\boldsymbol{\nu}_2\|^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(y_i-a_1)^2 1_{\{y_i>\bar{y}\}} + \sum_{i=1}^{n}(y_i-a_2)^2 1_{\{y_i\leq\bar{y}\}}\right]. \tag{103}$$

Now,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(y_i-a_1)^2 1_{\{y_i>\bar{y}\}} &= \frac{1}{n}\left[\sum_{i=1}^{n}y_i^2 1_{\{y_i>\bar{y}\}} + \sum_{i=1}^{n}a_1^2 1_{\{y_i>\bar{y}\}} - 2\sum_{i=1}^{n}a_1 y_i 1_{\{y_i>\bar{y}\}}\right] \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}w_i^2 1_{\{w_i>\bar{y}-\theta_i\}} + \sum_{i=1}^{n}\theta_i^2 1_{\{y_i>\bar{y}\}} + 2\sum_{i=1}^{n}\theta_i w_i 1_{\{w_i>\bar{y}-\theta_i\}} + \sum_{i=1}^{n}a_1^2 1_{\{y_i>\bar{y}\}} - 2\sum_{i=1}^{n}a_1 y_i 1_{\{y_i>\bar{y}\}}\right]\end{aligned}$$

and similarly,

$$\frac{1}{n}\sum_{i=1}^{n}(y_i-a_2)^2\,1_{\{y_i\le\bar{y}\}} =$$
$$\frac{1}{n}\left[\sum_{i=1}^{n} w_i^2 1_{\{w_i\le\bar{y}-\theta_i\}}+\sum_{i=1}^{n}\theta_i^2 1_{\{y_i\le\bar{y}\}}+2\sum_{i=1}^{n}\theta_i w_i 1_{\{w_i\le\bar{y}-\theta_i\}}+\sum_{i=1}^{n}a_2^2 1_{\{y_i\le\bar{y}\}}-2\sum_{i=1}^{n}a_2 y_i 1_{\{y_i\le\bar{y}\}}\right].$$

Therefore, from (103)

$$\begin{aligned}\frac{1}{n}\left\|\mathbf{y}-\boldsymbol{\nu}_2\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}w_i^2+\frac{\|\boldsymbol{\theta}\|^2}{n}+\frac{2}{n}\sum_{i=1}^{n}\theta_i w_i \\ +\frac{1}{n}\left(\sum_{i=1}^{n}a_1^2 1_{\{y_i>\bar{y}\}}-2\sum_{i=1}^{n}a_1 y_i 1_{\{y_i>\bar{y}\}}+\sum_{i=1}^{n}a_2^2 1_{\{y_i\le\bar{y}\}}-2\sum_{i=1}^{n}a_2 y_i 1_{\{y_i\le\bar{y}\}}\right).\end{aligned} \tag{104}$$

Since $\frac{1}{n}\sum_{i=1}^{n}\theta_i w_i\sim\mathcal{N}\left(0,\frac{\|\boldsymbol{\theta}\|^2}{n^2}\right)$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\theta_i w_i\right|\ge\epsilon\right)\le e^{-\frac{n\epsilon^2}{2\|\boldsymbol{\theta}\|^2/n}}. \tag{105}$$

From Lemma 9, we have, for any $\epsilon>0$,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}w_i^2-\sigma^2\right|\ge\epsilon\right)\le 2e^{-nk\min(\epsilon,\epsilon^2)}$$

where k is a positive constant. Next, we claim that

$$a_1\doteq c_1+\kappa_n\delta+o(\delta),\quad a_2\doteq c_2+\kappa_n\delta+o(\delta), \tag{106}$$

where c_1,c_2 are defined in (25). The concentration in (106) follows from Lemmas 1 and 2, together with the results on concentration of products and reciprocals in Lemmas 5 and 6, respectively. Further, using (106) and Lemma 5 again, we obtain $a_1^2\doteq c_1^2+\kappa_n\delta+o(\delta)$ and

$$a_1^2\left(\frac{1}{n}\sum_{i=1}^{n}1_{\{y_i>\bar{y}\}}\right)\doteq\frac{c_1^2}{n}\sum_{i=1}^{n}Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)+\kappa_n\delta+o(\delta). \tag{107}$$

Similarly,

$$\begin{aligned}a_1\left(\frac{2}{n}\sum_{i=1}^{n}y_i 1_{\{y_i>\bar{y}\}}\right)&\doteq\frac{2c_1}{n}\left(\sum_{i=1}^{n}\theta_i Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)+\frac{\sigma}{\sqrt{2\pi}}\sum_{i=1}^{n}e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}}\right)+\kappa_n\delta+o(\delta)\\&\doteq\frac{2c_1}{n}\left(c_1\sum_{i=1}^{n}Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)+\frac{\sigma}{\sqrt{2\pi}}\sum_{i=1}^{n}e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}}\right)+\kappa_n\delta+o(\delta)\\&\doteq\frac{2c_1^2}{n}\sum_{i=1}^{n}Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)+\frac{2c_1\sigma}{n\sqrt{2\pi}}\sum_{i=1}^{n}e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}}+\kappa_n\delta+o(\delta).\end{aligned} \tag{108}$$

Employing the same steps as above, we get

$$a_2^2\left(\frac{1}{n}\sum_{i=1}^{n}1_{\{y_i\leq\bar{y}\}}\right)\doteq\frac{c_2^2}{n}\sum_{i=1}^{n}Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)+\kappa_n\delta+o(\delta), \tag{109}$$

$$a_2\left(\frac{2}{n}\sum_{i=1}^{n}y_i1_{\{y_i\leq\bar{y}\}}\right)\doteq\frac{2c_2^2}{n}\sum_{i=1}^{n}Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)-\frac{2c_2\sigma}{n\sqrt{2\pi}}\sum_{i=1}^{n}e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}}+\kappa_n\delta+o(\delta). \tag{110}$$

Therefore, using (105)-(110) in (104), we finally obtain

$$\begin{aligned}\frac{1}{n}\left\|\mathbf{y}-\boldsymbol{\nu}_2\right\|^2\doteq\frac{\|\boldsymbol{\theta}\|^2}{n}+\sigma^2-\frac{c_1^2}{n}\sum_{i=1}^{n}Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)-\frac{c_2^2}{n}\sum_{i=1}^{n}Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)\\ -\left(\frac{2}{n}\right)\left(\frac{\sigma}{\sqrt{2\pi}}\right)\left(\sum_{i=1}^{n}e^{-\frac{(\bar{\theta}-\theta_i)^2}{2\sigma^2}}\right)(c_1-c_2)+\kappa_n\delta+o(\delta),\end{aligned}$$

which completes the proof of the lemma.

B.4 Proof of Lemma 14

The proof is along the same lines as that of Lemma 13. We have

$$\begin{aligned}\frac{1}{n}\left\|\boldsymbol{\theta}-\boldsymbol{\nu}_2\right\|^2&=\frac{1}{n}\left[\sum_{i=1}^{n}(\theta_i-a_1)^2\,\mathbf{1}_{\{y_i>\bar{y}\}}+\sum_{i=1}^{n}(\theta_i-a_2)^2\,\mathbf{1}_{\{y_i\leq\bar{y}\}}\right]\\ &=\frac{\|\boldsymbol{\theta}\|^2}{n}+\frac{1}{n}\left(\sum_{i=1}^{n}a_1^2\mathbf{1}_{\{y_i>\bar{y}\}}-2\sum_{i=1}^{n}a_1\theta_i\mathbf{1}_{\{y_i>\bar{y}\}}+\sum_{i=1}^{n}a_2^2\mathbf{1}_{\{y_i\leq\bar{y}\}}-2\sum_{i=1}^{n}a_2\theta_i\mathbf{1}_{\{y_i\leq\bar{y}\}}\right)\\ &\doteq\frac{\|\boldsymbol{\theta}\|^2}{n}+\frac{1}{n}\left(c_1^2\sum_{i=1}^{n}Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)-2c_1\sum_{i=1}^{n}\theta_iQ\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)\right)\\ &\quad+\frac{1}{n}\left(c_2^2\sum_{i=1}^{n}Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)-2c_2\sum_{i=1}^{n}\theta_iQ^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)\right)+\kappa_n\delta+o(\delta)\\ &\doteq\frac{\|\boldsymbol{\theta}\|^2}{n}-\frac{c_1^2}{n}\sum_{i=1}^{n}Q\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)-\frac{c_2^2}{n}\sum_{i=1}^{n}Q^c\left(\frac{\bar{\theta}-\theta_i}{\sigma}\right)+\kappa_n\delta+o(\delta).\end{aligned}$$

Acknowledgement

This work was supported in part by a Marie Curie Career Integration Grant (Grant Agreement No. 631489) and an Early Career Grant from the Isaac Newton Trust. The authors thank Richard Samworth for useful discussions on James-Stein estimators, and Andrew Barron and an anonymous referee for their comments which led to a much improved manuscript.

References

[1] W. James and C. M. Stein, "Estimation with Quadratic Loss," in *Proc. Fourth Berkeley Symp. Math. Stat. Probab.*, pp. 361–380, 1961.

[2] E. L. Lehmann and G. Casella, *Theory of Point Estimation.* Springer, New York, NY, 1998.

[3] B. Efron and C. Morris, "Data Analysis Using Stein's Estimator and Its Generalizations," *J. Amer. Statist. Assoc.*, vol. 70, pp. 311–319, 1975.

[4] D. V. Lindley, "Discussion on Professor Stein's Paper," *J. R. Stat. Soc.*, vol. 24, pp. 285–287, 1962.

[5] C. Stein, "Estimation of the mean of a multivariate normal distribution," *Ann. Stat.*, vol. 9, pp. 1135–1151, 1981.

[6] B. Efron and C. Morris, "Stein's estimation rule and its competitors—an empirical Bayes approach," *J. Amer. Statist. Assoc.*, vol. 68, pp. 117–130, 1973.

[7] E. George, "Minimax Multiple Shrinkage Estimation," *Ann. Stat.*, vol. 14, pp. 188–205, 1986.

[8] E. George, "Combining Minimax Shrinkage Estimators," *J. Amer. Statist. Assoc.*, vol. 81, pp. 437–445, 1986.

[9] G. Leung and A. R. Barron, "Information theory and mixing least-squares regressions," *IEEE Trans. Inf. Theory*, vol. 52, no. 8, pp. 3396–3410, 2006.

[10] G. Leung, *Improving Regression through Model Mixing.* PhD thesis, Yale University, 2004.

[11] A. J. Baranchik, "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution," *Tech. Report, 51, Stanford University*, 1964.

[12] P. Shao and W. E. Strawderman, "Improving on the James-Stein Positive Part Estimator," *Ann. Stat.*, vol. 22, no. 3, pp. 1517–1538, 1994.

[13] Y. Maruyama and W. E. Strawderman, "Necessary conditions for dominating the James-Stein estimator," *Ann. Inst. Stat. Math.*, vol. 57, pp. 157–165, 2005.

[14] R. Beran, *The unbearable transparency of Stein estimation. Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*, pp. 25–34. Institute of Mathematical Statistics, 2010.

[15] I. M. Johnstone, *Gaussian estimation: Sequence and wavelet models.* [Online]: `http://statweb.stanford.edu/~imj/GE09-08-15.pdf`, 2015.

[16] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, 2013.

[17] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference.* Springer, New York, NY, 2nd ed., 2005.

[18] D. Williams, *Probability with Martingales.* Cambridge University Press, 1991.

[19] J. T. Hwang and G. Casella, "Minimax confidence sets for the mean of a multivariate normal distribution," *The Annals of Statistics*, vol. 10, no. 3, pp. 868–881, 1982.

[20] R. Samworth, "Small confidence sets for the mean of a spherically symmetric distribution," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 3, pp. 343–361, 2005.

[21] M. E. Bock, "Minimax estimators of the mean of a multivariate normal distribution," *Ann. Stat.*, vol. 3, no. 1, pp. 209–218, 1975.

[22] J. H. Manton, V. Krishnamurthy, and H. V. Poor, "James-Stein State Filtering Algorithms," *IEEE Trans. Sig. Process.*, vol. 46, pp. 2431–2447, Sep. 1998.

[23] Z. Ben-Haim and Y. C. Eldar, "Blind Minimax Estimation," *IEEE Trans. Inf. Theory*, vol. 53, pp. 3145–3157, Sep. 2007.

[24] F. Chung and L. Lu, "Concentration inequalities and martingale inequalities: a survey," *Internet Mathematics*, vol. 3, no. 1, pp. 79–127, 2006.