Achieving Bayes MMSE Performance in the Sparse Signal + Gaussian White Noise Model when the Noise Level is Unknown

David Donoho and Galen Reeves
Department of Statistics
Stanford University

***Abstract*—Recent work on Approximate Message Passing algorithms in compressed sensing focuses on 'ideal' algorithms which at each iteration face a subproblem of recovering an unknown sparse signal in Gaussian white noise. The noise level in each subproblem changes from iteration to iteration in a way that depends on the underlying signal (which we don't know!). For such algorithms to be used in practice, it seems we need an estimator that achieves the MMSE when the noise level is unknown. In this paper we solve this problem using convex optimization, Stein Unbiased Risk Estimates and Huber Splines.**

I. Introduction

Recent work in Compressed Sensing (CS) [1], [2] employing the Approximate Message Passing (AMP) framework, discussed certain optimal algorithms for CS that iteratively solve a certain Sparse Signal + Gaussian White Noise problem. The algorithms in those papers repeatedly must solve subproblems with observed data

$$Y_i = X_i + Z_i, \quad i = 1, \ldots, n \tag{1}$$

where Z_i iid $N(0, \sigma^2)$ and X_i iid F_X, and the (Z_i) are independent of the (X_i). The goal is to estimate $\mathbf{X} = (X_i)$ from the data $\mathbf{Y} = (Y_i)$ with the smallest possible MSE. The minimum MSE estimator is of course the conditional expectation $E\{\mathbf{X}|\mathbf{Y}\}$. Of course, in a pure theoretical treatment there is no difficulty in simply assuming that one can compute the expectation; however, Compressed Sensing is intended for applications in which the underlying X is associated with a naturally occurring signal, is not under our control, and has a distribution F_X which cannot be assumed to be known. In fact even σ^2 depends on F_X and cannot ordinarily be assumed known in realistic applications.

In this note we propose a data-driven procedure having the ability to deliver MMSE-level performance without knowing F_X or σ^2.

To measure performance of a procedure, we say that (X, Y) are a future realization of (1) if the pair is independent of all training data $\mathbf{Y}$ and if $X \sim F_X$, while $F_{Y|X} = N(X, \sigma^2)$. The predictive Bayes Risk $\mathrm{PMSE}^*(X, Y)$ is then

$$\mathrm{PMSE}^*(X, Y) = E\big[(X - E\{X|Y\})^2\big].$$

The performance of a data-driven procedure $\hat{\eta}_n(\cdot; \mathbf{Y})$, where $\mathbf{Y}$ denotes the training data, is measured by $\mathrm{PMSE}(\hat{\eta}_n, X, Y) = E\big[(X - \hat{\eta}_n(\mathbf{Y})\})^2\big]$. The performance question we posed above becomes: can one achieve

$$\mathrm{PMSE}(\hat{\eta}_n, X, Y) \approx \mathrm{PMSE}^*(X, Y)?$$

Our proposed procedure arises from a connection between Bayesian inference and Robust statistics dating back to [3], [4]. Exploiting ideas from [5], [6] we propose a *local minimax procedure* $\hat{\eta}_n^*$, that for certain sequences (δ_n) and (α_n) that we describe explicitly, and which tend to zero with increasing n,

$$\Pr\big\{\, \mathrm{PMSE}(\eta_n^*, X, Y) \geq \mathrm{PMSE}^*(X, Y) + \delta_n \big\} < \alpha_n. \tag{2}$$

In words, we obtain theoretical guarantees on the near-Bayesian performance of a fully empirical rule.

Our proposal is based on optimization over neighborhoods of the empirical distribution, seeking the most *stable decision rule* for shrinking Y towards X. Our rule can be written in the 'Tweedie-Rule' form:

$$\eta_n^*(Y) = Y + (\sigma_n^*)^2 \frac{d}{dy} \log(f_n^*(Y)),$$

where both the density estimate f_n^* and the scale estimate σ_n^* are based on convex optimization; the density estimate is a least-informative Huber Spline. This proposal gives a conservative method of shrinkage; for example, it automatically has a limited translation property: for large values of future data Y, it translates by at most a data-derived constant. The rule has property (2).

In the literature of Mathematical Statistics, several papers consider a related problem when σ is known. To discuss them, recall that in the model (1) we have Tweedie's formula ([7], [8])

$$E\{X_i|Y_i = y\} = y + \sigma^2 \frac{d}{dy} \log(f_Y(y)), \tag{3}$$

where $f_Y = \frac{d}{dy} F_Y(y)$ is the marginal density of Y. Tweedie's formula says that, from observations of Y alone, and perfect knowledge of F_Y and of σ, we can compute $E\{X|Y\}$ without using F_X explicitly. Zhang [9], [10], Brown and Greenshtein [11] combine this formula with classical Kernel Density Estimation, while Jiang and Zhang [12] combine this formula with (computationally intractable) maximum likelihood estimation. The approach we develop is based on the idea that many Tweedie rules could be consistent with the training data $\mathbf{Y}$,

but some might be very weird or unstable. In our approach, we obtain a rule which is guaranteed to be stable; in fact we seek the Tweedie rule which is optimally robust among all those rules which are consistent with the data. Our style of analysis is based on the Stein Risk Function and minimization of Fisher information using convex optimization. Our approach also solves the σ-unknown problem, and a full analysis, not presented here, shows that it works without any modification when the errors $\mathbf{Z}$ are non-Gaussian, but are asymptotically Gaussian in the sense of Stein's method for proving limit theorems. The previous statistical work explicitly needs exact Gaussianity and exact knowledge of σ^2, whereas our proposal needs neither. In this short announcement, we are only able to describe the framework and construct the estimator; a fuller analysis will be presented in a journal article.

Recent work in the compressed sensing literature has also considered adaptive estimation over certain parametric classes of distributions. Vila and Schniter [13] and Krzakala et al. [14] propose techniques combining the AMP framework with expectation maximization (EM), and Kamilov et al. [15] propose a generalization that uses maximum likelihood estimation within each EM iteration.

II. The Construction: σ^2 known

A. Stein Risk Functional

Suppose we have a proposed nonlinear estimator $\eta(y) : \mathbf{R} \mapsto \mathbf{R}$. Define the corresponding 'score' function $\psi : \mathbf{R} \mapsto \mathbf{R}$ by $\psi(y) = y - \eta(y)$; for a cdf G, define the Stein's Risk functional (SRF)

$$\mathrm{SRF}(\psi, G, \sigma) = \sigma^2 - 2\sigma^2 E_G \psi'(Y) + E_G \psi^2(Y)$$

where $E_G H(Y) = \int H(y) dG(y)$. By Stein [16] we have that, when Y and X are independent, $Y = X + Z$, $Z \sim N(0, \sigma^2)$, and $\psi(y) = y - \eta(y)$

$$E(\eta(Y) - X)^2 = \mathrm{SRF}(\psi, F_Y, \sigma^2).$$

Our approach will be to use empirical data to design a score function which is near optimal for underlying distributions which are near the empirical CDF.

We make a few standard remarks. Let $\eta_Y^*(y) = E\{X|Y = y\}$; then the *optimal MSE* aka Bayes Risk is:

$$\begin{aligned}\mathrm{PMSE}^*(X, Y) &= \inf_{\eta} E(\eta(Y) - X)^2 \\ &= E(\eta_Y^*(Y) - X)^2 = \sigma^2 \cdot (1 - \sigma^2 I(F_Y))\end{aligned}$$

where $I(F) = \int (f'/f)^2 f dy$ is the Fisher Information for location [17]. In relation to the SRF, we have that for $\psi^*(y) = y - \eta_Y^*(y)$:

$$\mathrm{PMSE}^*(X, Y) = \mathrm{SRF}(\psi^*, F_Y, \sigma^2) = \inf_{\psi} \mathrm{SRF}(\psi, F_Y, \sigma^2).$$

B. Local Minimax Theorem

We now robustify the notion of PMSE using tools from Robust statistics [5], [6]. Given a *central CDF* F_0, define the Kolmogorov-Smirnov neighborhood of F_0 of radius $\kappa > 0$ by

$$\mathcal{F}(\kappa; F_0) = \{G : |G - F_0|_{KS} \le \kappa\},$$

where $|\cdot|_{KS}$ denotes the KS distance:

$$|F - G|_{KS} = \sup_t |F(t) - G(t)|.$$

Consider the *local minimax problem*

$$(P_{F_0,\kappa,\sigma}) \qquad \inf_{\psi} \sup_{G \in \mathcal{F}(\kappa; F_0)} \mathrm{SRF}(\psi, G, \sigma^2)$$

Lemma II.1. *$\mathrm{SRF}(\psi, G, \sigma^2)$ is convex quadratic in ψ and affine in G. The collection of CDF's in the neighborhood $\mathcal{F}(\kappa; F_0)$ is convex and vaguely compact. Hence for fixed σ, the Huber minimax theorem applies. The minimax problem has a saddlepoint in pure strategies. There is a pair $(\psi_\kappa^*, F_\kappa^*)$ such that*

$$\mathrm{SRF}(\psi_\kappa^*, G, \sigma^2) \le \mathrm{SRF}(\psi_\kappa^*, F_\kappa^*, \sigma^2) \le \mathrm{SRF}(\psi, F_\kappa^*, \sigma^2).$$

The pair can be characterized as follows: F_κ^* is the least-informative distribution in the KS-neighborhood

$$F_\kappa^* = \arg \min_{\mathcal{F}(\kappa; F_0)} I(G), \tag{4}$$

and $\psi_{\kappa,\sigma}^*$ is proportional to its Fisher score function for location:

$$\psi_{\kappa^*,\sigma} = -\sigma^2 \cdot \frac{d}{dy} \log(f_\kappa^*(y)). \tag{5}$$

Remarks

- The problem of minimizing Fisher Information over a KS neighborhood has been studied before, in robust statistics; see [5], [18]. The case where $F_0 = N(0, 1)$ is detailed in [19].
- While σ appears in the problem statement, and indeed the solution $\psi_{\kappa^*,\sigma}$ depends on σ, this dependence is very simple: as the final display, (5) reveals, all the solutions $\psi_{\kappa^*,\sigma}$ are proportional to a single fixed function $\frac{d}{dy} \log(f_\kappa^*(y))$.
- The same type of lemma can be shown for the sub neighborhood

  $$\mathcal{G}(\kappa; F_0, \sigma) = \{G : G = \Phi_\sigma \star H \ \& \ |G - F_0|_{KS} \le \kappa\}.$$

  Using this subneighborhood would give tighter bounds in everything that follows, and be more 'intrinsic' to the structure of our problem, because it is only for distributions $F_Y \in \mathcal{G}(\kappa; F_0, \sigma)$ that SRF has a valid interpretation as giving the Bayes MSE. Outside this subneighborhood, SRF can simply be viewed as a regularization of the Bayes MSE, that extends the Bayes MSE to CDF's where the standard assumptions (1) fail.
  We don't focus on this subneighborhood because it renders the optimization problem more complicated. For example, the solution of the minimax problem over the neighborhood $G(\kappa; F_0, \sigma)$ would not depend so simply on sigma as in the problem we do study.

Denote the value of the minimax problem $(P_{F_0,\kappa,\sigma})$ by

$$M(F_0, \kappa) = M(F_0, \kappa; \sigma) = Val(P_{F_0,\kappa,\sigma})$$

The triangle inequality for $|\cdot|_{KS}$ implies the basic monotonicity

$$\mathcal{F}(\kappa_1; F_1) \subset \mathcal{F}(\kappa_0; F_0) \Longrightarrow M(F_1, \kappa_1, \sigma) \le M(F_0, \kappa_0, \sigma).$$

$M(F_0, \kappa; \sigma)$ provides *an upper bound on the* PMSE^* over the neighborhood[1] $\mathcal{G}(\kappa; F_0, \sigma)$ More importantly, it provides an upper bound on the PMSE of the shrinkage rule $\eta^*_{\kappa,\sigma}(y) = y - \psi^2_{\kappa,\sigma}(y)$. Indeed, let $F_Y \in \mathcal{G}(\kappa; F_0, \sigma)$, and let $(F^*_{\kappa,\sigma}, \psi^*_{\kappa,\sigma}, \eta^*_{\kappa,\sigma})$ result from the minimax result above. Applying the Saddlepoint relation:

$$\begin{aligned}
\mathrm{PMSE}^*(X, Y) &= \mathrm{SRF}(\psi^*_Y, F_Y, \sigma) \\
&\le \mathrm{SRF}(\psi^*_{\kappa,\sigma}, F_Y, \sigma) \\
&= \mathrm{PMSE}(\eta^*_{\kappa,\sigma}, X, Y) \\
&\le \mathrm{SRF}(\psi^*_{\kappa}, F_{\kappa^*,\sigma}, \sigma) \\
&= M(F_0, \kappa; \sigma).
\end{aligned}$$

Hence the Bayes Risk at F_Y is smaller than the minimax PMSE over any neighborhood of Y; but it is not much smaller. Arguments from Section 3 will prove:

Lemma II.2. *Let F_Y be the distribution function of a random variable Y formed according to the standard model (1). The Bayes Risk and the Bayes Rule are the limiting solutions of the local minimax problem as the neighborhood size κ shrinks to zero:*

$$\lim_{\kappa \to 0} M(F_Y, \kappa) = \mathrm{PMSE}^*(X, Y),$$

and

$$\lim_{\kappa \to 0} \eta^*_\kappa(y) = \eta^*_Y(y) \qquad \text{in } L^2(F_Y).$$

In short, *our local minimax problem gives a regularization of the notion of Bayes risk.* Unlike Bayes risk, it is defined on all CDF's in a suitable neighborhood of a CDF where the Bayes risk is defined.

C. Least-Informative Tweedie Rule

Let F_n denote the empirical distribution of the entries in $\mathbf{Y}$ and consider the optimization problem

$$(\mathrm{FI}_{F_n,\kappa}) \qquad \min\{I(F) : |F - F_n|_{KS} \le \kappa\}. \tag{6}$$

The case $\kappa = 0$ of $(\mathrm{FI}_{F_n,\kappa})$ asks for the distribution with minimal Fisher information interpolating the empirical CDF, and has been studied by Huber [20]; the solution CDF is a so-called Huber Spline, having a density obeying a certain differential equation and interpolating F_n at $(Y_{(i)}, i/n)$. The general case $\kappa > 0$ inherits these features, i.e. it is again a Huber spline.

Lemma II.3. *For each $\kappa > 0$, the solution of $(\mathrm{FI}_{F_n,\kappa})$ exists, is unique, and is obtainable by solving an n-dimensional convex optimization problem.*

[1]The least upper bound would be obtained from the analogous minimax problem defined in terms of $\mathcal{G}(\kappa; F_0, \sigma)$; in contrast $M(F_0, \kappa; \sigma)$ is the least upper bound on the optimal SRF over $\mathcal{F}(\kappa; F_0, \sigma)$; but the SRF does not have the PMSE interpretation over $\mathcal{F}(\kappa; F_0, \sigma) \backslash \mathcal{G}(\kappa; F_0, \sigma)$.

Choose a 'failure probability' $\alpha_n > 2^{-n}$; we can find a value $\kappa_n = \kappa_{n,\alpha_n}$ so that

$$\Pr\{|F_n - F_Y|_{KS} \ge \kappa_n\} \le \alpha_n.$$

By the distribution-free character of the KS distance, κ_n does not depend on F_Y.

Let $F^*_n = F^*_{n,\kappa_n}$ denote the solution of (6) with $\kappa = \kappa_n$, let $\psi^*_n(y) = \psi^*(y; F_n, \kappa_n) = -\sigma^2 \cdot (\frac{d}{dy} \log f^*_n(y))$ denote the corresponding score function and $\eta^*_n(y) = y - \psi^*_n(y)$ the corresponding shrinker. These can be described more explicitly. The density f^*_n has the property that $\frac{d^2}{dy^2}\sqrt{f^*_n}$ is piecewise constant, with knots at the points of the sample CDF; and the displacement function ψ^*_n is a continuous function, made by splining together some simple rational functions. One can show that in the extreme tails the displacement function is constant, i.e. it does not oscillate wildly.

D. Consistency of the Local Minimax procedure

We propose the random procedure $\eta^*_n(y)$ be used as an empirical Tweedie's formula. How does its PMSE behave?

In effect, η^*_n is the result of applying Lemma II.1 to the problem $(P_{F_n,\kappa_n,\sigma})$. By the minimax theorem, Lemma II.1, we have the random upper bound

$$\mathrm{PMSE}(\eta^*_n, X, Y) \le M(F_n, |F_n - F_Y|_{KS}).$$

On the other hand the triangle inequality for KS distance gives the random inclusion

$$\mathcal{F}(F_n, |F_n - F_Y|_{KS}) \subset \mathcal{F}(F_Y, 2|F_n - F_Y|_{KS}).$$

Hence

$$M(F_n, |F_n - F_Y|_{KS}) \le M(F_Y, 2|F_n - F_Y|_{KS}).$$

Summarizing:

Lemma II.4. *Let (Y_i) and (X_i) obey the standard assumptions. Then the MSE of the random procedure η^*_n applied to a future realization (X, Y) obeys:*

$$\{|F_Y - F_n|_{KS} \le \kappa_n\} \Longrightarrow \{\mathrm{PMSE}(\eta^*_n, X, Y) \le M(F_Y, 2\kappa_n)\}$$

In short, if F_n is not too far from its sampling parent F_Y the random procedure we proposed has *MSE controlled by the regularized Bayes risk* $M(F_Y, 2\kappa_n)$. In view of Lemma II.2 this regularized Bayes risk is close to the Bayes risk.

We now choose $\alpha_n = n^{-M}$ for some $M \gg 0$. This corresponds to $\kappa_n \sim C\sqrt{\frac{\log(n)}{n}}$. It follows that $M(F_Y, 2\kappa_n) = \mathrm{PMSE}^*(X, Y)(1 + o(1))$. We have

Corollary II.1. (Consistency). *Choose $\kappa_n = C\sqrt{\frac{\log(n)}{n}}$,* independently *of F_Y.*

$$P\left\{1 \le \frac{\mathrm{PMSE}(\eta^*_n, X, Y)}{\mathrm{PMSE}^*(X, Y)} \le (1 + o(1))\right\} \to 1, \qquad n \to \infty.$$

This proves our initial goal (2).

III. Construction: Case σ^2 unknown, but X sparse

In case σ^2 is unknown, we have the problem of identifiability. If F_X contains a normal component, $F_X = F_0 \star \Phi_\tau$, then we have $F_Y = F_X \star \Phi_\sigma = F_0 \star \Phi_{\sqrt{\tau^2+\sigma^2}}$. Not knowing the true noise level, we can't decide empirically between σ and $\sqrt{\sigma^2+\tau^2}$.

A. The largest normal component

We *can* identify the *largest* normal component of F_Y; define the functional

$$\sigma^+(F) = \sup\{s > 0 : \exists(s, H) : F = H \star \Phi_s\}.$$

In the previous example, the functional σ^+ is (at least) $\sigma^2+\tau^2$; it could be larger still, if F_0 has a normal component.

σ^+ is an example of a discontinuous functional about which we can in general make only one-sided inference. For an in-depth discussion of one-sided inference, see [21].

Lemma III.1. *σ^+ is upper semicontinuous for weak convergence of CDF's.*

Define the nonnegative quantity

$$\sigma^+(\kappa; F_0) = \sup\{s > 0 : |G - F_0| \le \kappa; G = \Phi_s \star H\};$$

this gives the largest value of σ consistent with being nearby F_0, i.e. within KS distance κ.

Lemma III.2. *The largest normal component σ^+ is estimable from observations of Y. Indeed, the estimator*

$$\hat\sigma_n^+ = \sigma^+(\kappa_n; F_n),$$

where $\kappa_n = C\sqrt{\frac{\log(n)}{n}}$, is consistent for $\sigma^+(F_Y)$:

$$\hat\sigma_n^+ \to_{a.s.} \sigma^+(F_Y), \qquad n \to \infty.$$

Proof. We have

$$P\{|F_n - F_Y|_{KS} \ge \kappa_n\} \to 0, \qquad n \to \infty,$$

and the bracketing relationship

$$\{|F_n - F_Y|_{KS} \le \kappa_n\} \\ \Longrightarrow \{\sigma^+_{F_Y} \le \sigma^+(\kappa_n; F_n) \le \sigma^+(2\kappa_n; F_Y)\}.$$

Uppersemicontinuity gives

$$\sigma^+(2\kappa_n; F_Y) \to \sigma^+_{F_Y}, \quad \kappa_n \to 0;$$

combining these displays proves the claim. □

B. Identifiability of σ^2 under Sparsity

In the main application we consider, the distribution of X obeys a sparsity constraint:

$$P\{X = 0\} \ge (1 - \varepsilon). \tag{7}$$

This allows σ, and not only σ^+, to be identifiable.

Lemma III.3. *Under assumption (7), then F_Y has a unique normal component; i.e. there is only one value σ for which there exists a representation $F_Y = \Phi_\sigma \star F_X$, with F_X a CDF having positive mass at 0. In particular $\sigma = \sigma^+_{F_Y}$.*

Lemma III.4. *Suppose that $H_0(\{0\}) \ge 1 - \varepsilon$ with $\varepsilon < 1$ and $F_0 = \Phi_{\sigma_0} \star H_0$, then*

$$\sigma^+(\kappa; F_0) \to \sigma_0,$$

as $\kappa \to 0$.

Note that *it is not necessary to know* $\varepsilon \in (0, 1)$ in order to get the benefits of identifiability and consistency; they accrue to the estimator $\hat\sigma_n^+$ regardless.

IV. The General Method

The proposed procedure is outlined as follows:

- *Neighborhood Size*. For $C > 0$, let $\kappa_n = C\sqrt{\frac{\log(n)}{n}}$.
- *Local Minimax Shape.* Let f_n^* be the density of the distribution F_n^* solving the problem of minimizing Fisher Information over a KS neighborhood of F_n of radius κ_n. Specifically, define the problem

  $$(Shape_{\kappa, F_n}) \quad \min\{I(F) : |F - F_n|_{KS} \le \kappa\},$$

  where F_n is the empirical distribution, and let f_n^* be the solution of $(Shape_{\kappa_n, F_n})$.
- *Largest Normal Component.* Let $\hat\sigma_n$ denote the estimated largest normal component, i.e. the σ solving

  $$\sup\{\sigma : \exists(H, \sigma) \& |F_n - \Phi_\sigma \star H| \le \kappa_n\}.$$

- *Quantitatively Robust Tweedie's Rule.*

  $$\widehat{\eta_n^*}(y) = y - (\hat\sigma_n)^2 \frac{d}{dy} \log(f_n^*(y)).$$

Note that, in the proposed procedure:

- It is not necessary to specify σ.
- It is not necessary to specify the sparsity ε.
- The minimax shape is a Huber spline [20].
- Both steps, as explained below, can be implemented using standard convex optimization software. The minimax shape problem is a finite-dimensional convex optimization problem. The Largest Normal Component problem is approximable by a finite-dimensional convex optimization problem.

Theorem IV.1. *Suppose that X obeys the sparsity constraint (7); but we do not know the value of ε. We have, for explicitly describable sequences δ_n and α_n both tending to zero,*

$$P\{PMSE(\hat\eta_n^*, X, Y) = PMSE^*(X, Y) + \delta_n\} \le \alpha_n,$$

where $\alpha_n \to 0$.

- Here α_n is universal and does not depend on the distribution of (X, Y). We can make α_n go to zero at any desired polynomial rate, by enlarging the constant C in the definition of κ_n.

- Here we may take

$$\delta_n = M(2\kappa_n, F_Y, \sigma^2) - \mathrm{PMSE}^*(X,Y) + \Big[2|\sigma^+(2\kappa_n, F_Y)^2 - \sigma^2| \times (1 + (\sigma^2 + \sigma^+(2\kappa_n, F_Y)^2) I(F_Y))\Big].$$

Proof of Theorem IV.1. Let F_n^* denote the least-favorable distribution achieving the minimum Fisher information over the neighborhood $\{G : |F_n - G|_{KS} \leq \kappa_n\}$. On the event $\{|F - F_n|_{KS} \leq \kappa_n\}$, we have

$$\begin{aligned}
\mathrm{PMSE}(\widehat{\psi_n^*}, X, Y) &= \mathrm{SRF}(\widehat{\psi_n^*}, F_Y, \widehat{\sigma_n}^2) \\
&\leq \mathrm{SRF}(\widehat{\psi_n^*}, F_n^*, \widehat{\sigma_n}^2) \\
&\leq \mathrm{SRF}(\psi_n^*, F_n^*, \sigma^2) \\
&\quad + |\widehat{\sigma_n}^2 - \sigma^2| \cdot (1 + (\sigma^2 + \widehat{\sigma_n}^2) I(F_n^*)) \\
&\leq M(2\kappa_n, F_Y, \sigma^2) \\
&\quad + \Big[|\sigma^+(2\kappa_n, F_Y)^2 - \sigma^2| \times \\
&\qquad (1 + (\sigma^2 + \sigma^+(2\kappa_n, F_Y)^2) I(F_Y))\Big],
\end{aligned}$$

where we used Lemma IV.1 below, taking $G = F_n^*$, and we make a distinction between $\widehat{\psi_n^*}$, the score function with estimated scale and ψ_n^* the score function with scale perfectly known.

On the same event, we have

$$\begin{aligned}
\mathrm{PMSE}(\widehat{\psi_n^*}, X, Y) &= \mathrm{SRF}(\widehat{\psi_n^*}, F_Y, \widehat{\sigma_n}^2) \\
&\geq \mathrm{SRF}(\psi_{\widehat{\sigma_n}, Y}, F_Y, \widehat{\sigma_n}^2) \\
&\geq \mathrm{SRF}(\psi_{\sigma, Y}, F_Y, \sigma^2) \\
&\quad - |\widehat{\sigma_n}^2 - \sigma^2| \cdot (1 + (\sigma^2 + \widehat{\sigma_n}^2) I(F_Y)) \\
&\geq \mathrm{PMSE}^*(X,Y) \\
&\quad - \Big[|\sigma^+(2\kappa_n, F_Y)^2 - \sigma^2| \times \\
&\qquad (1 + (\sigma^2 + \sigma^+(2\kappa_n, F_Y)^2) I(F_Y))\Big],
\end{aligned}$$

where we used Lemma 4.1 below, with $G = F_Y$, and we make a distinction between $\psi_{Y,\widehat{\sigma}}$, the score function based on the true underlying F_Y but estimated scale $\widehat{\sigma}$, and $\psi_{Y,\sigma}$, the score function with both F_Y and scale σ perfectly known.

Combining these displays

$$|\mathrm{PMSE}(\widehat{\psi_n^*}, X, Y) - \mathrm{PMSE}^*(X,Y)| \leq \delta_n.$$

Finally, the probability of the event's complement is bounded by

$$P\{|F_n - F_Y|_{KS} \geq \kappa_n\} \leq \alpha_n. \quad \square$$

Lemma IV.1. *Let $\sigma_1 \neq \sigma_0$. Then with G a CDF and $\psi_{G,\sigma}$ the score function $\psi_{G,\sigma} = -\sigma^2 \frac{d}{dy} \log(g(y))$, we have*

$$\begin{aligned}
&|\mathrm{SRF}(\psi_{G,\sigma_1}, G, \sigma_1^2) - \mathrm{SRF}(\psi_{G,\sigma_0}, G, \sigma_0^2)| \\
&\leq |\sigma_1^2 - \sigma_0^2| \cdot (1 + (\sigma_0^2 + \sigma_1^2) I(G)).
\end{aligned}$$

References

[1] G. Reeves, "Sparsity pattern recovery in compressed sensing," Ph.D. dissertation, University of California, Berkeley, 2011.

[2] D. L. Donoho, A. Javanmard, and A. Montanari, "Information-theoretically optimal compressed sensing via spatial coupling and approximate message passing," in *Proceedings of the IEEE International Symposium on Information Theory (ISIT 2012)*, Cambridge, MA, July 2012.

[3] A. Marazzi, *Algorithm, Routines and S functions for robust statistics*. Pacific Grove, CA: Wadsworth and Brooks/Cole, 1993.

[4] P. J. Bickel and J. R. Collins, "Minimizing Fisher information over mixtures of distributions," *Sankhya A*, vol. 45, pp. 1–19, 1983.

[5] P. J. Huber, "Robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[6] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics - The Approach Based on Influence Functions*. Wiley, 1986.

[7] H. Robbins, "An empirical Bayes approach to statistics," in *Proc. 3-rd Berkeley Symposium on Mathematical Statistics and Probability*, U. of California Press, Ed., Berkeley and Los Angeles, 1956, pp. 157–163.

[8] B. Efron, "Tweedie's formula and selection bias," *Journal of the American Statistical Association*, vol. 106, no. 492, pp. 1602–1614, 2011.

[9] C.-H. Zhang, "Empirical Bayes and compound estimation of normal means," *Statistica Sinica*, vol. 7, pp. 181–193, 1997.

[10] ——, "General empirical Bayes wavelet methods and exactly adaptive minimax estimation," *The Annals of Statistics*, vol. 22, no. 1, pp. 54–100, 2005.

[11] L. D. Brown and E. Greenshtein, "Nonparametric empirical Bayes and compound decision approaches to estimation of a high-dimensional vector of normal means," *The Annals of Statistics*, vol. 37, no. 4, pp. 1685–1704, 2009.

[12] W. Jiang and C.-H. Zhang, "General maximum likelihood empirical Bayes estimation of normal means," *The Annals of Statistics*, vol. 37, no. 4, pp. 1647–1684, 2009.

[13] J. P. Vila and P. Schniter, "Expectation-maximization gaussian-mixture approximate message passing," in *Proc. 46th Annual Conference on Information Sciences and Systems*, Princeton, NJ, March 2012, (See also http://arxiv.org/abs/1207.3107).

[14] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborova, "Probabilistic reconstruction in compressed sensing: Algorithms, phase diagrams, and threshold achieving matrices," June 2012, arXiv: http://arxiv.org/abs/1206.3953.

[15] U. S. Kamilov, S. Rangan, A. K. Fletcher, and M. Unser, "Approximate message passing with consistent parameter estimation and applications to sparse learning," in *Proc. 23rd Annual Conference on Neural Information Processing Systems*, Lake Taho, NV, December 2012, pp. 2447–2455, (See also http://arxiv.org/abs/1207.3859).

[16] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, 1981.

[17] P. J. Bickel, "Minimax estimation of the mean of a normal distribution whenthe parameter space is restricted," *Annals of Statistics*, vol. 9, pp. 1301–1309, 1982.

[18] J. Sacks and D. Ylvisaker, "A note on Huber's robust estimation of a location parameter," *The Annals of Mathematical Statistics*, vol. 43, no. 4, pp. 1068–1075, 1972.

[19] P. J. Huber and E. M. Ronchetti, *Robust Statistics*. New York: Wiley, 2009.

[20] P. J. Huber, "Fisher information and spline interpolation," *The Annals of Statistics*, vol. 2, no. 5, pp. 1029–1033, 1974.

[21] D. L. Donoho, "One-sided inference about functionals of a density," *The Annals of Statistics*, vol. 16, no. 4, pp. 1390–1420, 1988.

[22] R. Carroll and P. Hall, "Optimal rates of convergence for deconvolving a density," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1184–1186, Dec 1988.