MINIMAX RISK OF MATRIX DENOISING BY SINGULAR VALUE THRESHOLDING

BY DAVID DONOHO[1] AND MATAN GAVISH[1,2]

Stanford University

An unknown m by n matrix X_0 is to be estimated from noisy measurements $Y = X_0 + Z$, where the noise matrix Z has i.i.d. Gaussian entries. A popular matrix denoising scheme solves the nuclear norm penalization problem $\min_X \|Y - X\|_F^2/2 + \lambda \|X\|_*$, where $\|X\|_*$ denotes the nuclear norm (sum of singular values). This is the analog, for matrices, of ℓ_1 penalization in the vector case. It has been empirically observed that if X_0 has low rank, it may be recovered quite accurately from the noisy measurement Y.

In a proportional growth framework where the rank r_n, number of rows m_n and number of columns n all tend to ∞ proportionally to each other ($r_n/m_n \to \rho$, $m_n/n \to \beta$), we evaluate the asymptotic minimax MSE $\mathcal{M}(\rho, \beta) = \lim_{m_n, n \to \infty} \inf_\lambda \sup_{\text{rank}(X) \leq r_n} \text{MSE}(X_0, \hat{X}_\lambda)$. Our formulas involve incomplete moments of the quarter- and semi-circle laws ($\beta = 1$, square case) and the Marčenko–Pastur law ($\beta < 1$, nonsquare case). For finite m and n, we show that MSE increases as the nonzero singular values of X_0 grow larger. As a result, the finite-n worst-case MSE, a quantity which can be evaluated numerically, is achieved when the signal X_0 is "infinitely strong."

The nuclear norm penalization problem is solved by applying soft thresholding to the singular values of Y. We also derive the minimax threshold, namely the value $\lambda^*(\rho)$, which is the optimal place to threshold the singular values.

All these results are obtained for general (nonsquare, nonsymmetric) real matrices. Comparable results are obtained for square symmetric nonnegative-definite matrices.

1. Introduction. Suppose we observe a single noisy matrix Y, generated by adding noise Z to an unknown matrix X_0, so that $Y = X_0 + Z$, where Z is a noise matrix. We wish to recover the matrix X_0 with some bound on the mean squared error (MSE). This is hopeless when X_0 is a completely general matrix, and the noise Z is arbitrary; but when X_0 happens to be of relatively low rank, and the noise matrix is i.i.d. standard Gaussian, one can indeed guarantee quantitatively accurate

recovery. This paper provides explicit formulas for the best possible guarantees obtainable by a popular, computationally practical procedure.

Specifically, let Y, X_0 and Z be m-by-n real matrices (a set we denote by $M_{m \times n}$), and suppose that Z has i.i.d. entries, $Z_{i,j} \sim \mathcal{N}(0, 1)$. Consider the following nuclear-norm penalization (NNP) problem:

$$(1.1) \qquad (NNP) \qquad \hat{X}_\lambda = \operatorname*{argmin}_{X \in M_{m \times n}} \frac{1}{2} \|Y - X\|_F^2 + \lambda \|X\|_*,$$

where $\|X\|_*$ denotes the sum of singular values of $X \in M_{m \times n}$, also known as the nuclear norm, $\| \cdot \|_F$ denotes square root of the sum of squared matrix entries, also known as the Frobenius norm and $\lambda > 0$ is a penalty factor. A solution to (NNP) is efficiently computable by modern convex optimization software [11]; it shrinks away from Y in the direction of smaller nuclear norm.

Measure performance (risk) by mean-squared error (MSE). When the unknown X_0 is of known rank r and belongs to a matrix class $\mathbf{X}_{m,n} \subset M_{m \times n}$, the minimax MSE of NNP is

$$(1.2) \qquad \mathcal{M}_{m,n}(r | \mathbf{X}) = \inf_\lambda \sup_{\substack{X_0 \in \mathbf{X}_{m,n} \\ \operatorname{rank}(X_0) \leq r}} \frac{1}{mn} \mathbb{E}_{X_0} \|\hat{X}_\lambda(X_0 + Z) - X_0\|_F^2,$$

namely the worst-case risk of $\hat{X}_{\lambda_*}$, where λ_* is the threshold for which this worst-case risk is the smallest possible. Here, $\mathbb{E}_{X_0}$ denotes expectation with respect to the random noise matrix Z, conditional on a given value of the signal matrix X_0, and $\hat{X}_\lambda(X_0 + Z)$ denotes the denoiser $\hat{X}_\lambda$ acting on the matrix $X_0 + Z$. Note that the symbol $\mathbf{X}$ denotes a matrix class, not a particular matrix. For square matrices, $m = n$, we write $\mathcal{M}_n(r | \mathbf{X})$ instead of $\mathcal{M}_{n,n}(r | \mathbf{X})$. In a very clear sense $\mathcal{M}_{m,n}(r | \mathbf{X})$ gives the best possible guarantee for the MSE of NNP, based solely on the rank and problem size, and not on other properties of the matrix X_0.

1.1. *Minimax MSE evaluation.*

In this paper, we calculate the minimax MSE $\mathcal{M}_{m,n}(r | \mathbf{X})$ for two matrix classes $\mathbf{X}$:

(1) *General matrices*: $\mathbf{X} = \operatorname{Mat}_{m,n}$: The signal X_0 is a real matrix $X_0 \in M_{m \times n}$ ($m \leq n$).

(2) *Symmetric matrices*: $\mathbf{X} = \operatorname{Sym}_n$: The signal X_0 is a real, symmetric positive semidefinite matrix, a set we denote by $S_+^n \subset M_{n \times n}$.

In both cases, the asymptotic MSE (AMSE) in the "large n" asymptotic setting admits considerably simpler and more accessible formulas than the minimax MSE for finite n. So in addition to the finite-n minimax MSE, we study the asymptotic setting where a sequence of problem size triplets (r_n, m_n, n) is indexed by $n \to \infty$, and where, along this sequence $m/n \to \beta \in (0, 1)$ and $r/m \to \rho \in (0, 1)$. We think

of β as the matrix shape parameter; $\beta = 1$ corresponds to a square matrix, and $\beta < 1$ to a matrix wider than it is tall. We think of ρ as the fractional rank parameter, with $\rho \approx 0$ implying low rank relative to matrix size. Using these notions we can define the asymptotic minimax MSE (AMSE)

$$\mathcal{M}(\rho, \beta | \mathbf{X}) = \lim_{n \to \infty} \mathcal{M}_{m_n, n}(r_n | \mathbf{X}).$$

We obtain explicit formulas for the asymptotic minimax MSE in terms of incomplete moments of classical probability distributions: the quarter-circle and semi-circle laws (square case $\beta = 1$) and the Marčenko–Pastur distribution (nonsquare case $\beta < 1$). Figures 1 and 2 show how the AMSE depends on the matrix class $\mathbf{X}$, the rank fraction ρ and the shape factor β. We also give explicit formulas for the optimal regularization parameter λ_*, also as a function of ρ; see Figures 3 and 4.

These minimax MSE results constitute best possible guarantees, in the sense that for the procedure in question, the MSE is actually attained at some rank r matrix, so that no better guarantee is possible for the given tuning parameter λ_*; but also, no other tuning parameter offers a better such guarantee.

1.2. *Motivations.* We see four reasons to develop these bounds.

1.2.1. *Applications.* Several important problems in modern signal and image processing, in network data analysis and in computational biology can be cast as recovery of low-rank matrices from noisy data, and nuclear norm minimization has become a popular strategy in many cases; see, for example, [2, 22] and references therein. Our results provide sharp limits on what such procedures can hope to achieve, and validate rigorously the idea that *low rank alone* is enough to provide some level of performance guarantee; in fact, they precisely quantify the best possible guarantee.

1.2.2. *Limits on possible improvements.* One might wonder whether some other procedure offers even better guarantees than NNP. Consider then the minimax risk *over all procedures*, defined by

$$(1.3) \qquad \mathcal{M}_{m,n}^*(r | \mathbf{X}) = \inf_{\hat{X}} \sup_{\substack{X_0 \in \mathbf{X}_{m,n} \\ \text{rank}(X_0) \leq r}} \frac{1}{mn} \mathbb{E}_{X_0} \| \hat{X}(X_0 + Z) - X_0 \|_F^2,$$

where $\hat{X} = \hat{X}(Y)$ is some measurable function of the observations, and its corresponding minimax AMSE

$$\mathcal{M}^*(\rho, \beta | \mathbf{X}) = \lim_{n \to \infty} \mathcal{M}_{m_n, n}^*(r_n | \mathbf{X}),$$

where the sequences m_n and r_n are as above. Here one wants to find the best possible procedure, without regard to efficient computation. We also prove a lower

bound on the minimax MSE over all procedures, and provide an asymptotic evaluation

$$\mathcal{M}^*(\rho, \beta|\mathbf{X}) \geq \mathcal{M}^-(\rho, \beta) \equiv \rho + \beta\rho - \beta\rho^2.$$

In the square case ($\beta = 1$), this simplifies to $\mathcal{M}^*(\rho|\mathbf{X}) \geq \mathcal{M}^-(\rho) \equiv \rho(2 - \rho)$. The NNP-minimax MSE is by definition larger than the minimax MSE, $\mathcal{M}(\rho, \beta|\mathbf{X}) \geq \mathcal{M}^*(\rho, \beta|\mathbf{X})$. While there may be procedures outperforming NNP, the performance improvement turns out to be limited. Indeed, our formulas show that

$$\frac{\mathcal{M}(\rho, \beta|\mathbf{X})}{\mathcal{M}^-(\rho, \beta)} \leq 2\left(1 + \frac{\sqrt{\beta}}{1 + \beta}\right),$$

while

$$(1.4) \qquad \lim_{\rho \to 0} \frac{\mathcal{M}(\rho, \beta|\mathbf{X})}{\mathcal{M}^-(\rho, \beta)} = 2\left(1 + \frac{\sqrt{\beta}}{1 + \beta}\right).$$

For square matrices ($\beta = 1$), this simplifies to

$$(1.5) \qquad \frac{\mathcal{M}(\rho|\mathbf{X})}{\mathcal{M}^-(\rho)} \leq 3, \qquad \lim_{\rho \to 0} \frac{\mathcal{M}(\rho|\mathbf{X})}{\mathcal{M}^-(\rho)} = 3.$$

In words, the potential improvement in minimax AMSE of *any* other matrix denoising procedure over NNP is at most a factor of 3; and if any such improvement were available, it would only be available in extreme low-rank situations. Actually obtaining such an improvement in performance guarantees is an interesting research challenge.

1.2.3. *Parallels in minimax decision theory.* The low-rank matrix denoising problem stands in a line of now-classical problems in minimax decision theory. Consider the sparse vector denoising problem, where an unknown vector x of interest yields noisy observations $\mathbf{y} = \mathbf{x} + \mathbf{z}$ with noise $\mathbf{z} \sim_{\text{i.i.d.}} N(0, 1)$; the vector $\mathbf{x}$ is sparsely nonzero—$\#\{i : x(i) \neq 0\} \leq \varepsilon \cdot n$—with $\mathbf{z}$ and $\mathbf{x}$ independent. In words, a vector with a fraction $\leq \varepsilon$ of nonzeros is observed with noise. In this setting, consider the following ℓ_1-norm penalization problem:

$$(1.6) \qquad (P_1) \qquad \hat{\mathbf{x}}_\lambda = \underset{x \in \mathbf{R}^n}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1.$$

The sparse vector denoising problem exhibits several striking structural resemblances to low-rank matrix denoising:

- *Thresholding representation.* For a scalar y, define the soft thresholding nonlinearity by

$$\eta_\lambda(y) = \operatorname{sign}(y) \cdot \left(|y| - \lambda\right)_+.$$

In words, values larger than λ are shifted toward zero by λ, while those smaller than λ are set to zero. The solution vector $\hat{x}_\lambda$ of (P_1) obeys $(\hat{\mathbf{x}}_\lambda)_i = \eta_\lambda(y_i)$; namely, it applies η_λ coordinate wise. Similarly, the solution of (NNP) applies η_λ coordinate wise to the singular values of the noisy matrix Y.

REMARK. By this observation, (P_1) can also be called "soft thresholding" or "soft threshold denoising," and in fact, these other terms are the labels in common use. Similarly, NNP amounts to "soft thresholding of singular values." This paper will henceforth use the term *singular value soft thresholding* (SVST).

- *Sparsity/low rank analogy.* The objects to be recovered in the sparse vector denoising problem have sparse entries; those to be recovered in the low-rank matrix denoising problem have sparse singular values. Thus the fractional sparsity parameter ε is analogous to the fractional rank parameter ρ. It is natural to ask the same questions about behavior of minimax MSE in one setting (say, asymptotics as $\rho \to 0$) as in the other setting ($\varepsilon \to 0$). In fact, such comparisons turn out to be illuminating.
- *Structure of the least-favorable estimand.* Among sparse vectors x of a given fixed sparsity fraction ε, which of these is the hardest to estimate? This should maximize the mean-squared error of soft thresholding, even under the most clever choice of λ. This least-favorable configuration is singled out in the minimax AMSE

(1.7) $$M_n(\varepsilon) = \inf_\lambda \sup_{\#\{i:x(i)\neq 0\}\leq \varepsilon \cdot n} \frac{1}{n}\mathbb{E}\|\hat{x}_\lambda - x\|_2^2.$$

In this min/max, the least favorable situation has all its nonzeros, in some sense, "at infinity"; that is, all sparse vectors which place large enough values on the nonzeros are nearly least favorable, that is, essentially make the problem maximally difficult for the estimator, even when it is optimally tuned. In complete analogy, in low-rank matrix denoising we will see that all low-rank matrices, which are in an appropriate sense "sufficiently large," are thereby almost least favorable.

- *Structure of the minimax smoothing parameter.* In the sparse vector denoising AMSE (1.7) the $\lambda = \lambda(\varepsilon)$ achieving the infimum is a type of optimal regularization parameter, or optimal threshold. It decreases as ε increases, with $\lambda(\varepsilon) \to 0$ as $\varepsilon \to 1$. Paralleling this, we show that the low-rank matrix denoising AMSE (1.2) has minimax singular value soft threshold $\lambda^*(\rho)$ decreasing as ρ increases, and $\lambda^*(\rho) \to 0$ as $\rho \to 1$.

Despite these similarities, there is one major difference between sparse vector denoising and low-rank matrix denoising. In the sparse vector denoising problem, the soft-thresholding minimax MSE was compared to the minimax MSE over all procedures by Donoho and Johnstone [8]. Let $M(\varepsilon) = \lim_{n\to\infty} M_n(\varepsilon)$ denote the soft thresholding AMSE and define the minimax AMSE over all procedures via

$$M^*(\varepsilon) = \lim_{n\to\infty} \inf_{\hat{x}} \sup_{\#\{i:x(i)\neq 0\}\leq \varepsilon \cdot n} \frac{1}{n}\mathbb{E}\|\hat{x} - x\|_2^2,$$

where here $\hat{x} = \hat{x}(y)$ denotes *any* procedure which is measurable in the observations. In the limit of extreme sparsity, soft thresholding is *asymptotically minimax* [8],

$$\frac{M(\varepsilon)}{M^*(\varepsilon)} \to 1 \qquad \text{as } \varepsilon \to 0.$$

Breaking the chain of similarities, we are not able to show a similar asymptotic minimaxity for SVST in the low rank matrix denoising problem. Although equation (1.4) says that soft thresholding of singular values is asymptotically not more than a factor of 3 suboptimal, we doubt that anything better than a factor of 3 can be true; specifically, we conjecture that SVST suffers a *minimaxity gap*. For example, for $\beta = 1$, we conjecture that

$$\frac{\mathcal{M}(\rho|\mathbf{X})}{\mathcal{M}^*(\rho|\mathbf{X})} \to 3 \qquad \text{as } \rho \to 0.$$

We believe that interesting new estimators will be found improving upon singular value soft thresholding by essentially this factor of 3. Namely, there may be substantially better guarantees to be had under extreme sparsity, than those which can be offered by SVST. Settling the minimaxity gap for SVST seems a challenging new research question.

1.2.4. *Indirect observations.* Evaluating the Minimax MSE of SVST has an intriguing new motivation [6, 7, 17], arising from the newly evolving fields of compressed sensing and matrix completion.

Consider the problem of recovering an unknown matrix X_0 from *noiseless, indirect* measurements. Let $\mathcal{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^p$ be a linear operator, and consider observations

$$y = \mathcal{A}(X_0).$$

In words, $y \in \mathbb{R}^p$ contains p linear measurements of the matrix object X_0. See the closely related *trace regression* model [21] which also includes measurement noise. Can we recover X_0? It may seem that $p \geq mn$ measurements are required, and in general this would be true; but if X_0 happens to be of low rank, and $\mathcal{A}$ has suitable properties, we may need substantially fewer measurements.

Consider reconstruction by *nuclear norm minimization*,

(1.8) (P_{nuc}) $\min \|X\|_*$ subject to $y = \mathcal{A}(X)$.

Recht and co-authors found that when the matrix representing the operator $\mathcal{A}$ has i.i.d. $\mathcal{N}(0, 1)$ entries, and the matrix is of rank r, the matrix X_0 is recoverable from $p < nm$ measurements for certain combinations of p and r [18]. The operator $\mathcal{A}$ offers so-called *Gaussian measurements* when the representation of the operator as a matrix has i.i.d. Gaussian entries. Empirical work by Recht, Xu and Hassibi [19, 20], Fazel, Parillo and Recht [18], Tanner and Wei [24] and Oymak

and Hassibi [16] documented for Gaussian measurements a *phase transition* phenomenon, that is, a fairly sharp transition from success to failure as r increases, for a given p. Putting $\rho = r/m$ and $\delta = p/(mn)$ it appears that there is a critical sampling rate $\delta^*(\rho) = \delta^*(\rho; \beta)$, such that, for $\delta > \delta^*(\rho)$, NNM is successful for large m, n, while for $\delta < \delta^*(\rho)$, NNM fails. $\delta^*(\rho)$ provides a sharp "sampling limit" for low rank matrices, that is, a clear statement of how many measurements are needed to recover a low rank matrix, by a popular and computationally tractable algorithm.

In very recent work, [6, 7, 17], it has been shown empirically that the precise location of the phase transition *coincides with the minimax MSE*

$$(1.9) \qquad \delta^*(\rho; \beta) = \mathcal{M}(\rho, \beta | \mathbf{X}), \qquad \rho \in (0, 1), \beta \in (0, 1);$$

a key requirement for discovering and verifying (1.9) empirically was to obtain an explicit formula for the right-hand side; that explicit formula is derived and proven in this paper. Relationship (1.9) connects two seemingly unrelated problems: matrix denoising from direct observations and matrix recovery from incomplete measurements. Both problems are attracting a large and growing research literature. Equation (1.9) demonstrates the importance of minimax MSE calculations even in a seemingly unrelated setting where there is no noise and no statistical decision to be made!

2. Results.

2.1. *Least-favorable matrix.* We start by identifying the least-favorable situation for matrix denoising by SVST.

THEOREM 1 (The worst-case matrix for SVST has its principal subspace "at ∞"). *Define the risk function of a denoiser* $\hat{X} : M_{m \times n} \to M_{m \times n}$ *at* $X_0 \in M_{m \times n}$ *by*

$$(2.1) \qquad R(\hat{X}, X_0) := \frac{1}{m} \mathbb{E} \left\| \hat{X} \left(X_0 + \frac{1}{\sqrt{n}} Z \right) - X_0 \right\|_F^2.$$

Let $\lambda > 0$, $m \le n \in \mathbb{N}$ *and* $1 \le r \le m$. *For the worst-case risk of* $\hat{X}_\lambda$ *on* $m \times n$ *matrices of rank at most* r, *we have*

$$(2.2) \qquad \sup_{\substack{X_0 \in M_{m \times n} \\ \mathrm{rank}(X_0) \le r}} R(\hat{X}_\lambda, X_0) = \lim_{\mu \to \infty} R(\hat{X}_\lambda, \mu C),$$

where $C \in M_{m \times n}$ *is any fixed matrix of rank exactly* r.

2.2. *Minimax MSE.*   Let $W_i(m, n)$ denote the marginal distribution of the ith largest eigenvalue of a standard central Wishart matrix $W_m(I, n)$, namely, the ith largest eigenvalue of the random matrix $\frac{1}{n}ZZ'$ where $Z \in M_{m \times n}$ has i.i.d. $\mathcal{N}(0, 1)$ entries. Define for $\Lambda > 0$ and $\alpha \in \{1/2, 1\}$

(2.3)
$$\mathbf{M}_n(\Lambda; r, m, \alpha) = \frac{r}{m} + \frac{r}{n} - \frac{r^2}{mn} + \frac{r(n-r)}{mn}\Lambda^2$$
$$+ \alpha \frac{(n-r)}{mn} \sum_{i=1}^{m-r} w_i(\Lambda; m-r; n-r),$$

where

(2.4)
$$w_i(\Lambda; m, n) = \int_{\Lambda^2}^{\infty} (\sqrt{t} - \Lambda)^2 \, dW_i(m, n)(t)$$

is a combination of the complementary incomplete moments of standard central Wishart eigenvalues

$$\int_{\Lambda^2}^{\infty} t^{k/2} \, dW_i(m, n)(t)$$

for $k = 0, 1, 2$.

THEOREM 2 (An implicit formula for the finite-n minimax MSE).   *The minimax MSE of SVST over m-by-n matrices of rank at most r is given by*

$$\mathcal{M}_n(r, m | \text{Mat}) = \min_{\Lambda \geq 0} \mathbf{M}_n(\Lambda; r, m, 1) \qquad and$$
$$\mathcal{M}_n(r | \text{Sym}) = \min_{\Lambda \geq 0} \mathbf{M}_n(\Lambda; r, n, 1/2),$$

where the minimum on the right-hand sides is unique.

In fact, we will see that $\mathbf{M}_n(\Lambda; r, m, \alpha)$ is convex in Λ. As the densities of the standard central Wishart eigenvalues $W_i(m, n)$ are known [25], this makes it possible, in principle, to tabulate the finite-n minimax risk.

2.3. *Asymptotic minimax MSE.*   A more accessible formula is obtained by calculating the large-n asymptotic minimax MSE, where $r = r(n)$ and $m = m(n)$ both grow proportionally to n. Let us write *minimax AMSE* for asymptotic minimax MSE. For the case $\mathbf{X}_{m,n} = \text{Mat}_{m,n}$ we assume a limiting rank fraction $\rho = \lim_{n \to \infty} r/m$ and limiting aspect ratio $\beta = \lim_{n \to \infty} m/n$ and consider

(2.5)
$$\mathcal{M}(\rho, \beta | \text{Mat}) = \lim_{n \to \infty} \mathcal{M}_n(r, m | \text{Mat})$$
$$= \lim_{n \to \infty} \inf_{\lambda} \sup_{\substack{X_0 \in M_{\lceil \beta n \rceil \times n} \\ \text{rank}(X_0) \leq \rho \beta n}} \frac{1}{mn} \mathbb{E} \|\hat{X}_\lambda - X_0\|_F^2.$$

Similarly, for the case $\mathbf{X}_{m,n} = \mathrm{Sym}_n$, we assume a limiting rank fraction $\rho = \lim_{n\to\infty} r/n$ and consider

$$
\mathcal{M}(\rho|\mathrm{Sym}) = \lim_{n\to\infty} \mathcal{M}_n(r|\mathrm{Sym})
$$

(2.6)

$$
= \lim_{n\to\infty} \inf_{\lambda} \sup_{\substack{X_0 \in S_+^n \\ \mathrm{rank}(X_0) \le \rho n}} \frac{1}{n^2} \mathbb{E}\|\hat{X}_\lambda - X_0\|_F^2.
$$

The Marčenko–Pastur distribution [15] gives the asymptotic empirical distribution of Wishart eigenvalues. It has density

(2.7)
$$
p_\gamma(t) = \frac{1}{2\pi\gamma t}\sqrt{(\gamma_+ - t)(t - \gamma_-)} \cdot \mathbf{1}_{[\gamma_-,\gamma_+]}(t),
$$

where $\gamma_\pm = (1 \pm \sqrt{\gamma})^2$. Define the complementary incomplete moments of the Marčenko–Pastur distribution

(2.8)
$$
P_\gamma(x; k) = \int_x^{\gamma_+} t^k p_\gamma(t)\, dt.
$$

Finally, let

$$
\mathbf{M}(\Lambda; \rho, \tilde{\rho}, \alpha)
$$

(2.9)
$$
= \rho + \tilde{\rho} - \rho\tilde{\rho} + (1 - \tilde{\rho})
$$

$$
\times \Big[\rho\Lambda^2
$$

$$
+ \alpha(1-\rho)\Big(P_\gamma(\Lambda^2; 1) - 2\Lambda P_\gamma\Big(\Lambda^2; \frac{1}{2}\Big) + \Lambda^2 P_\gamma(\Lambda^2; 0) \Big) \Big],
$$

with $\gamma = \gamma(\rho, \tilde{\rho}) = (\tilde{\rho} - \rho\tilde{\rho})/(\rho - \rho\tilde{\rho})$.

THEOREM 3 (An explicit formula for the minimax AMSE). *For the minimax AMSE of SVST we have*

(2.10)
$$
\mathcal{M}(\rho, \beta|\mathrm{Mat}) = \min_{0 \le \Lambda \le \gamma_+} \mathbf{M}(\Lambda; \rho, \beta\rho, 1),
$$

(2.11)
$$
\mathcal{M}(\rho|\mathrm{Sym}) = \min_{0 \le \Lambda \le \gamma_+} \mathbf{M}(\Lambda; \rho, \rho, 1/2),
$$

with $\gamma_+ = (1 + \sqrt{(\beta - \beta\rho)/(1 - \beta\rho)})^2$, where the minimum on the right-hand sides is unique. Moreover, for any $0 < \beta \le 1$, the function $\rho \mapsto \mathcal{M}(\rho, \beta|\mathrm{Mat})$ is continuous and increasing on $\rho \in [0, 1]$, with $\mathcal{M}(0, \beta|\mathrm{Mat}) = 0$ and $\mathcal{M}(1, \beta|\mathrm{Mat}) = 1$. The same is true for $\mathcal{M}(\rho|\mathrm{Sym})$.

The curves $\rho \mapsto \mathcal{M}(\rho, \beta|\mathrm{Mat})$, for different values of β, are shown in Figure 1. The curves $\rho \mapsto \mathcal{M}(\rho, \beta|\mathrm{Mat})$ and $\rho \mapsto \mathcal{M}(\rho, \beta|\mathrm{Mat})$ are shown in Figure 2.
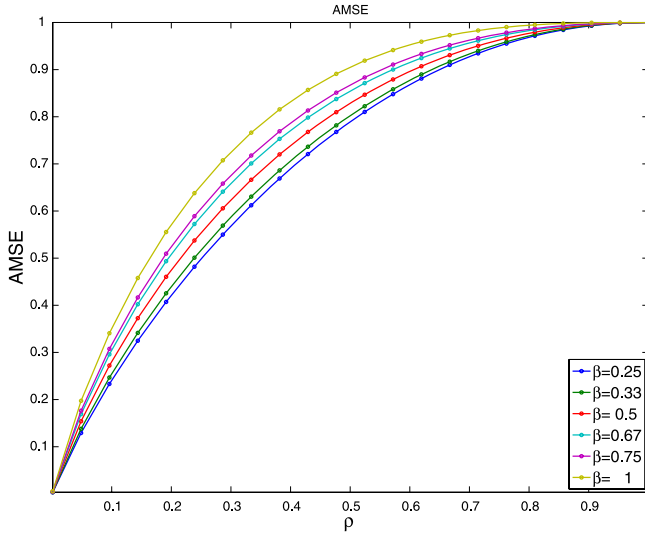
FIG. 1.   *The minimax AMSE curves for case* Mat, *defined in* (2.10), *for a few values of* β.

2.4. *Computing the minimax AMSE.* To compute $\mathcal{M}(\rho, \beta|\text{Mat})$ and $\mathcal{M}(\rho|\text{Sym})$ we need to minimize (2.9). Define

$$(2.12) \qquad \Lambda_*(\rho, \beta, \alpha) = \underset{\Lambda}{\operatorname{argmin}}\, \mathbf{M}(\Lambda; \rho, \tilde{\rho}, \alpha).$$
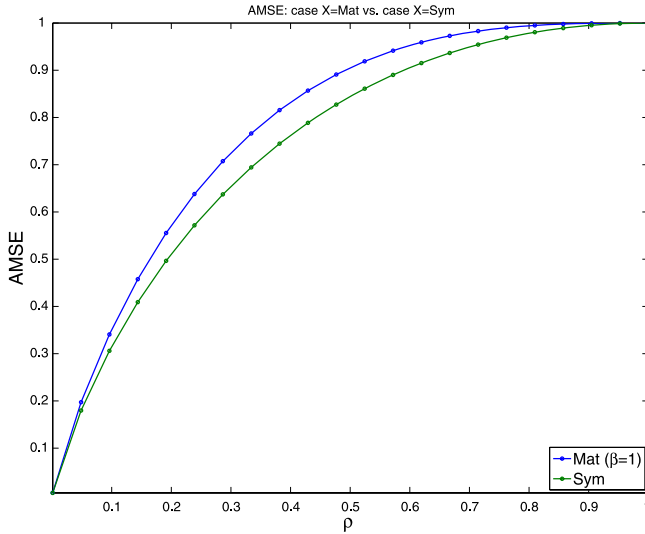

FIG. 2.   *The minimax AMSE curves for case* Mat *with* $\beta = 1$ *and case* Sym.

THEOREM 4 (A characterization of the minimax AMSE for general β).   *For any $\alpha \in \{1/2, 1\}$ and $\beta \in (0, 1]$, the function $\rho \mapsto \Lambda_*(\rho, \beta, \alpha)$ is decreasing on $\rho \in [0, 1]$ with*

(2.13)
$$\lim_{\rho \to 0} \Lambda_*(\rho, \beta, \alpha) = \Lambda_*(0, \beta, \alpha) = 1 + \sqrt{\beta} \qquad and$$

(2.14)
$$\lim_{\rho \to 1} \Lambda_*(\rho, \beta, \alpha) = \Lambda_*(1, \beta, \alpha) = 0.$$

For $\rho \in (0, 1)$, the minimizer $\Lambda_*(\rho, \beta, \alpha)$ is the unique root of the equation in Λ

(2.15)
$$P_\gamma\left(\Lambda^2; \frac{1}{2}\right) - \Lambda \cdot P_\gamma(\Lambda^2; 0) = \frac{\Lambda \rho}{\alpha(1 - \rho)},$$

where the left-hand side of (2.15) is a decreasing function of Λ.

The minimizer $\Lambda_*(\rho, \beta, \alpha)$ can therefore be determined numerically by binary search. [In fact, we will see that Λ_* is the unique minimizer of the convex function $\Lambda \mapsto \mathbf{M}(\Lambda; \rho, \tilde{\rho}, \alpha)$.] Evaluating $\mathcal{M}(\rho, \beta | \mathrm{Mat})$ and $\mathcal{M}(\rho | \mathrm{Sym})$ to precision ϵ thus requires $O(\log(1/\epsilon))$ evaluations of the complementary incomplete Marčenko–Pastur moments (2.8).

For square matrices ($\beta = 1$), this computation turns out to be even simpler, and only requires evaluation of elementary trigonometric functions.

THEOREM 5 (A characterization of the minimax AMSE for $\beta = 1$).   *We have*

$$\mathbf{M}(\Lambda; \rho, \rho, \alpha) = \rho(2 - \rho)$$

(2.16)
$$+ (1 - \rho)\big[\rho\Lambda^2 + \alpha(1 - \rho)\big(Q_2(\Lambda) - 2\lambda Q_1(\Lambda) + \Lambda^2 Q_0(\Lambda)\big)\big],$$

where

(2.17)
$$Q_0(x) = \frac{1}{\pi} \int_x^2 \sqrt{4 - t^2}\, dt$$
$$= 1 - \frac{x}{2\pi}\sqrt{4 - x^2} - \frac{2}{\pi} a\tan\left(\frac{x}{\sqrt{4 - x^2}}\right),$$

(2.18)
$$Q_1(x) = \frac{1}{\pi} \int_x^2 t\sqrt{4 - t^2}\, dt = \frac{1}{3\pi}(4 - x^2)^{3/2},$$

(2.19)
$$Q_2(x) = \frac{1}{\pi} \int_x^2 t^2\sqrt{4 - t^2}\, dt$$
$$= 1 - \frac{1}{4\pi}x\sqrt{4 - x^2}(x^2 - 2) - \frac{2}{\pi} a\sin\left(\frac{x}{2}\right)$$

are the complementary incomplete moments of the quarter circle law. Moreover, for $\alpha \in \{1/2, 1\}$

$$(2.20) \qquad \Lambda_*(\rho, \rho, \alpha) = 2 \cdot \sin(\theta_\alpha(\rho)),$$

where $\theta_\alpha(\rho) \in [0, \pi/2]$ is the unique solution to the transcendental equation

$$(2.21) \qquad \theta + \cot(\theta) \cdot \left(1 - \frac{1}{3}\cos^2(\theta)\right) = \frac{\pi(1 + \alpha^{-1}\rho - \rho)}{2(1 - \rho)}.$$

The left-hand side of (2.21) *is a decreasing function of θ.*

In [4] we make available a Matlab script, and a web-based calculator for evaluating $\mathcal{M}(\rho, \beta|\text{Mat})$ and $\mathcal{M}(\rho|\text{Sym})$. The implementation provided employs binary search to solve (2.15) [or (2.21) for $\beta = 1$] and then feeds the minimizer Λ_* into (2.9) [or into (2.16) for $\beta = 1$].

2.5. *Asymptotically optimal tuning for the SVST threshold λ.* The crucial functional Λ_*, defined in (2.12), can now be explained as the optimal (minimax) threshold of SVST in a special system of units. Let $\lambda_*(m, n, r|\mathbf{X})$ denote the minimax tuning threshold, namely

$$\lambda_*(m, n, r|\mathbf{X}) = \underset{\lambda}{\operatorname{argmin}} \underset{\substack{X_0 \in \mathbf{X}_{m,n} \\ \operatorname{rank}(X_0) \leq r}}{\sup} \frac{1}{mn} \mathbb{E}_{X_0} \|\hat{X}_\lambda(X_0 + Z) - X_0\|_F^2.$$

THEOREM 6 (Asymptotic minimax tuning of SVST). *Consider again a sequence $n \mapsto (m(n), r(n))$ with a limiting rank fraction $\rho = \lim_{n \to \infty} r/m$ and a limiting aspect ratio $\beta = \lim_{n \to \infty} m/n$. For the asymptotic minimax tuning threshold we have*

$$\lim_{n \to \infty} \frac{1}{\sqrt{n}} \lambda_*(m, n, r|\text{Mat}) = \sqrt{(1 - \beta\rho)} \cdot \Lambda_*(\rho, \beta, 1) \quad \text{and}$$

$$\lim_{n \to \infty} \frac{1}{\sqrt{n}} \lambda_*(n, r|\text{Sym}) = \sqrt{(1 - \rho)} \cdot \Lambda_*(\rho, 1, 1/2).$$

The curves $\rho \mapsto \lim_{n \to \infty} \lambda_*(m, n, r|\text{Mat})/\sqrt{n}$, namely the scaled asymptotic minimax tuning threshold for SVST, are shown in Figure 3 for different values of β. The curves $\rho \mapsto \lim_{n \to \infty} \lambda_*(n, n, r|\text{Mat})/\sqrt{n}$ and $\rho \mapsto \lim_{n \to \infty} \lambda_*(n, r|\text{Sym})/\sqrt{n}$ are shown in Figure 4.

2.6. *Parametric representation of the minimax AMSE for square matrices.* For square matrices ($\rho = \tilde{\rho}$, $\beta = 1$) the minimax curves $\mathcal{M}(\rho, 1|\text{Mat})$ and $\mathcal{M}(\rho|\text{Sym})$ admit a parametric representation in the $(\rho, \mathcal{M})$ plane using elementary trigonometric functions.
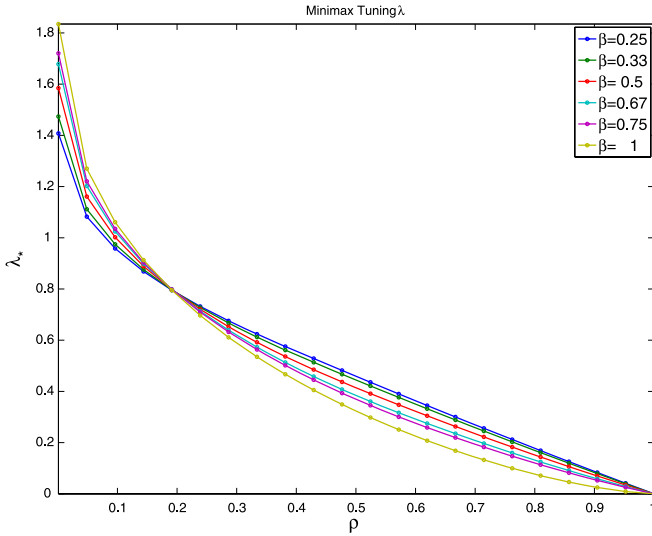
FIG. 3. (*Nonsquare cases.*) *The scaled asymptotic minimax tuning threshold for SVST,* $\rho \mapsto \lim_{n \to \infty} \lambda_*(m, n, r | \mathrm{Mat}) / \sqrt{n}$, *when* $m/n \to \beta$ *and* $r/m \to \rho$, *for a few values of* β.
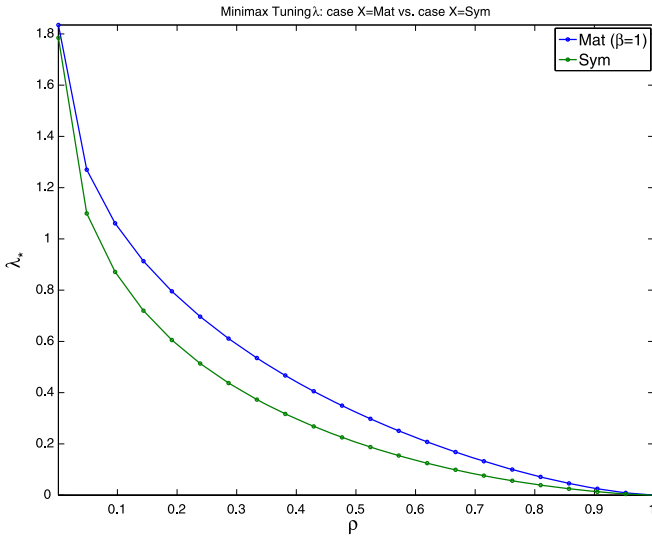

FIG. 4. (*Square case.*) *The scaled asymptotic minimax tuning threshold for SVST,* $\rho \mapsto \lim_{n \to \infty} \lambda_*(n, n, r | \mathrm{Mat}) / \sqrt{n}$ *and* $\rho \mapsto \lim_{n \to \infty} \lambda_*(n, r | \mathrm{Sym}) / \sqrt{n}$, *when* $r/m \to \rho$.

THEOREM 7 (Parametric representation of the minimax AMSE curve for $\beta = 1$). *As θ ranges over $(0, \pi/2)$,*

$$\rho(\theta) = 1 - \frac{\pi/2}{\theta + (\cot(\theta) \cdot (1 - (1/3)\cos^2(\theta)))},$$

$$\mathcal{M}(\theta) = 2\rho(\theta) - \rho^2(\theta) + 4\rho(\theta)(1 - \rho(\theta))\sin^2(\theta)$$
$$+ \frac{4}{\pi}(1 - \rho)^2 \left[(\pi - 2\theta)\left(\frac{5}{4} - \cos(\theta)^2\right) + \frac{\sin(2\theta)}{12}(\cos(2\theta) - 14)\right]$$

is a parametric representation of $\rho \mapsto \mathcal{M}(\rho, \rho|\mathrm{Mat})$, and similarly

$$\rho(\theta) = 1 - \frac{\theta + (\cot(\theta) \cdot (1 - (1/3)\cos^2(\theta))) - \pi/2}{\theta + (\cot(\theta) \cdot (1 - (1/3)\cos^2(\theta))) + \pi/2},$$

$$\mathcal{M}(\theta) = 2\rho(\theta) - \rho^2(\theta) + 4\rho(\theta)(1 - \rho(\theta))\sin^2(\theta)$$
$$+ \frac{2}{\pi}(1 - \rho)^2 \left[(\pi - 2\theta)\left(\frac{5}{4} - \cos(\theta)^2\right) + \frac{\sin(2\theta)}{12}(\cos(2\theta) - 14)\right]$$

is a parametric representation of $\rho \mapsto \mathcal{M}(\rho|\mathrm{Sym})$.

2.7. *Minimax AMSE in the low-rank limit $\rho \approx 0$.*

THEOREM 8 (Minimax AMSE to first order in ρ near $\rho = 0$). *For the behavior of the minimax curves near $\rho = 0$, we have*

$$\mathcal{M}(\rho, \beta|\mathrm{Mat}) = 2(1 + \sqrt{\beta} + \beta) \cdot \rho + o(\rho)$$

and in particular

$$\mathcal{M}(\rho, 1|\mathrm{Mat}) = 6\rho + o(\rho).$$

Moreover,

$$\mathcal{M}(\rho|\mathrm{Sym}) = 6\rho + o(\rho).$$

The minimax AMSE curves $\rho \mapsto \mathcal{M}(\rho, \beta|\mathrm{Mat})$ for small values of ρ, and the corresponding approximation slopes $2(1 + \sqrt{\beta} + \beta)$ are shown in Figure 5 for several values of β. We find it surprising that asymptotically, *symmetric positive definite matrices are no easier to recover than general square matrices*. This phenomenon is also seen in the case of sparse vector denoising, where in the limit of extreme sparsity, the nonnegativity of the nonzeros does not allow one to reduce the minimax MSE.[3] We note that this first-order AMSE near $\rho = 0$ agrees with a different asymptotic model for minimax MSE of SVST over large low-rank matrices [10]. There, the asymptotic prediction for AMSE near $\rho = 0$ is found to be in agreement with the empirical finite-n MSE.

---

[3]Compare results in [8] with [9]. To be clear, in both matrix denoising and vector denoising, there is an MSE advantage for each fixed positive rank fraction/sparsity fraction. It is just that the benefit goes away as either fraction tends to 0.
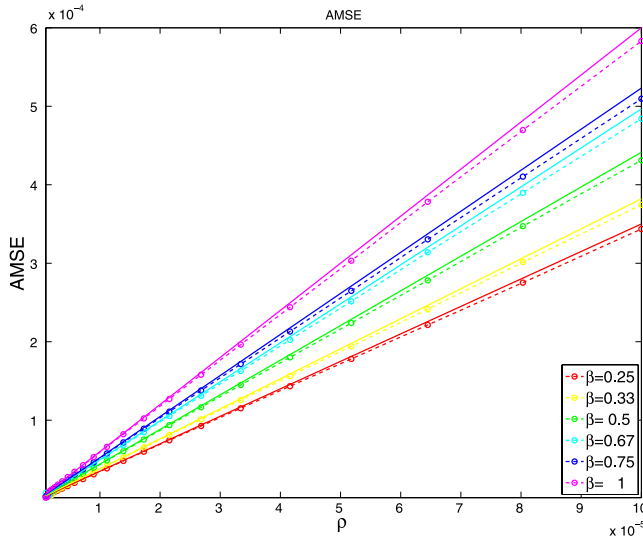
FIG. 5. *The minimax AMSE curves* $\rho \mapsto \mathcal{M}(\rho, \beta | \mathrm{Mat})$ *for small values of* ρ (*dashed lines*) *and the corresponding approximation slopes* $2(1 + \sqrt{\beta} + \beta)$ (*solid lines*).

2.8. *AMSE vs. the asymptotic global minimax MSE.* In (1.3) we have introduced global minimax MSE $\mathcal{M}_{m,n}^*(r | \mathbf{X})$, namely the minimax risk over *all* measurable denoisers $\hat{X} : M_{m \times n} \to M_{m \times n}$. To define the large-$n$ asymptotic global minimax MSE analogous to (2.5), consider sequences where $r = r(n)$ and $m = m(n)$ both grow proportionally to n, such that both limits $\rho = \lim_{n \to \infty} r/m$ and $\beta = \lim_{n \to \infty} m/n$ exist. Define the asymptotic global minimax MSE

$$(2.22) \qquad \mathcal{M}^*(\rho, \beta | \mathbf{X}) = \lim_{n \to \infty} \mathcal{M}_{m,n}^*(r | \mathbf{X}).$$

THEOREM 9. (1) *For the global minimax MSE we have*

$$(2.23) \qquad \mathcal{M}_{m,n}^*(r | \mathbf{X}) \geq \frac{r}{m} + \frac{r}{n} - \frac{r^2 + r}{mn}$$

for case Mat, *and if* $m = n$, *for case* Sym.

(2) *For the asymptotic global minimax MSE we have*

$$(2.24) \qquad \mathcal{M}^*(\rho, \beta | \mathbf{X}) \geq \rho + \tilde{\rho} - \rho\tilde{\rho}$$

for case Mat, *and if* $\beta = 1$, *for case* Sym. *Here* $\tilde{\rho} = \beta\rho$.

(3) *Let*

$$(2.25) \qquad \mathcal{M}^-(\rho, \beta) = \rho + \tilde{\rho} - \rho\tilde{\rho}$$

denote our lower bound on asymptotic global minimax MSE. Then

$$(2.26) \qquad \frac{\mathcal{M}(\rho, \beta | \mathbf{X})}{\mathcal{M}^-(\rho, \beta)} \leq 2\left(1 + \frac{\sqrt{\beta}}{1 + \beta}\right)$$

and

$$(2.27) \qquad \lim_{\rho \to 0} \frac{\mathcal{M}(\rho, \beta | \mathbf{X})}{\mathcal{M}^-(\rho, \beta)} = 2\left(1 + \frac{\sqrt{\beta}}{1 + \beta}\right).$$

2.9. *Outline of this paper.* The body of the paper proves the above results. Section 3 introduces notation, and proves auxiliary lemmas. In Section 4 we characterize the worst-case MSE of SVST for matrices of a fixed size (Theorem 1). In Section 5 we derive formula (2.3) for the worst-case MSE, and prove Theorem 2. In Section 6 we pass to the large-n limit and derive formula (2.9), which provides the worst-case asymptotic MSE in the large-n limit (Theorem 3). In Section 7 we investigate the minimizer of the asymptotic worst-case MSE function, and its minimum, namely the minimax AMSE, and prove Theorem 4. In Section 8 we extend our scope from SVST denoisers to all denoisers, investigate the global minimax MSE and prove Theorem 9. In the interest of space, Theorems 5, 6 7 and 8 are proved in the supplemental article [5]. The supplemental article also contains a derivation of the Stein unbiased risk estimate for SVST, which is instrumental in the proof of Theorem 1, and other technical auxiliary lemmas.

3. Preliminaries.

3.1. *Scaling.* Our main object of interest, the worst-case MSE of SVST,

$$(3.1) \qquad \sup_{\substack{X_0 \in M_{m \times n} \\ \mathrm{rank}(X_0) \le \rho m}} \frac{1}{mn} \mathbb{E} \|\hat{X} - X_0\|_F^2,$$

is more conveniently expressed using a specially calibrated risk function. Since the SVST denoisers are scale-invariant, namely

$$\mathbb{E}_X \|X - \hat{X}(X + \sigma Z)\|_F^2 = \sigma^2 \mathbb{E}_X \left\| \frac{X}{\sigma} - \hat{X}\left(\frac{X}{\sigma} + Z\right) \right\|_F^2,$$

we are free to introduce the scaling $\sigma = n^{-1/2}$ and define the risk function of a denoiser $\hat{X} : M_{m \times n} \to M_{m \times n}$ at $X_0 \in M_{m \times n}$ by

$$(3.2) \qquad R(\hat{X}, X_0) := \frac{1}{m} \mathbb{E} \left\| \hat{X}\left(X_0 + \frac{1}{\sqrt{n}} Z\right) - X_0 \right\|_F^2.$$

Then, the worst-case MSE of $\hat{X}$ at X_0 is given by

$$(3.3) \qquad \sup_{\substack{X_0 \in M_{m \times n} \\ \mathrm{rank}(X_0) \le \rho m}} \frac{1}{mn} \mathbb{E} \|\hat{X} - X_0\|_F^2 = \sup_{\substack{X_0 \in M_{m \times n} \\ \mathrm{rank}(X_0) \le \rho m}} R(\hat{X}, X_0).$$

To vary the SNR in the problem, it will be convenient to vary the norm of the signal matrix X_0 instead, namely, to consider $Y = \mu X_0 + \frac{1}{\sqrt{n}} Z$ with $\frac{1}{m} \|X_0\|_F^2 = 1$.

3.2. *Notation*. Vectors are denoted by boldface lowercase letters, such as $\mathbf{v}$, and their entries by v_i. Matrices are denoted by uppercase letters, such as A, and their entries by $A_{i,j}$. Throughout this text, Y will denote the data matrix $Y = \mu X_0 + \frac{1}{\sqrt{n}} Z$. We use $M_{m \times n}$ and O_m to denote the set of real-valued m-by-n matrices, and group of m-by-m orthogonal matrices, respectively. $\| \cdot \|_F$ denotes the Frobenius matrix norm on $M_{m \times n}$, namely the Euclidean norm of a matrix considered as a vector in $\mathbb{R}^{mn}$. We denote matrix multiplication by either AB or $A \cdot B$. We use the following convenient notation for matrix diagonals: for a matrix $X \in M_{m \times n}$, we denote by $X_\Delta \in \mathbb{R}^m$ its main diagonal,

$$(3.4) \qquad (X_\Delta)_i = X_{i,i}, \qquad 1 \le i \le m.$$

Similarly, for a vector $\mathbf{x} \in \mathbb{R}^m$, and $n \ge m$ that we suppress in our notation, we denote by $\mathbf{x}_\Delta \in M_{m \times n}$ the "diagonal" matrix

$$(3.5) \qquad (\mathbf{x}_\Delta)_{i,j} = \begin{cases} x_i, & 1 \le i = j \le m, \\ 0, & \text{otherwise.} \end{cases}$$

We use a "fat" singular value decomposition (SVD) of $X \in M_{m \times n}$ $X = U_X \cdot \mathbf{x}_\Delta \cdot V_X'$, with $U_X \in M_{m \times m}$ and $V_X \in M_{n \times n}$. Note that the SVD is not uniquely determined, and in particular $\mathbf{x}$ can contain the singular values of X in any order. Unless otherwise noted, we will assume that the entries of $\mathbf{x}$ are nonnegative and sorted in nonincreasing order, $x_1 \ge \cdots \ge x_m \ge 0$. When $m < n$, the last $n - m$ columns of V_Y are not uniquely determined; we will see that our various results do not depend on this choice. Note that with the "fat" SVD, the matrices Y and $U_Y' \cdot Y \cdot V_Y$ have the same dimensionality, which simplifies the notation we will need.

When appropriate, we let univariate functions act on vectors entry-wise, namely, for $\mathbf{x} \in \mathbb{R}^n$ and $f : \mathbb{R} \to \mathbb{R}$, we write $f(\mathbf{x}) \in \mathbb{R}^n$ for the vector with entries $f(\mathbf{x})_i = f(x_i)$.

3.3. $\hat{X}_\lambda$ *acts by soft thresholding of the data singular values.* By orthogonal invariance of the Frobenius norm, (1.1) is equivalent to

$$(3.6) \qquad \hat{\mathbf{x}}_\lambda = \operatorname*{argmin}_{\mathbf{x} \in \mathbb{R}^n} \tfrac{1}{2} \|\mathbf{y} - \mathbf{x}\|_2^2 + \lambda \|\mathbf{x}\|_1,$$

through the relation $\hat{X}_\lambda(Y) = U_Y \cdot (\hat{\mathbf{x}}_\lambda)_\Delta \cdot V_Y'$. It is well known that the solution to (3.6) is given by $\hat{\mathbf{x}}_\lambda = \mathbf{y}_\lambda$, where $\mathbf{y}_\lambda = (\mathbf{y} - \lambda)_+$ denotes coordinate-wise soft thresholding of $\mathbf{y}$ with threshold λ. The SVST estimator (1.1) is therefore given by [12]

$$(3.7) \qquad \hat{X}_\lambda : Y \mapsto U_Y \cdot (\mathbf{y}_\lambda)_\Delta \cdot V_Y'.$$

Note that (3.7) is well defined, that is, $\hat{X}_\lambda(Y)$ does not depend on the particular SVD $Y = U_Y \cdot (\mathbf{y})_\Delta \cdot V_Y'$ chosen.

In case Sym, observe that the solution to (1.1) is constrained to lie in the linear subspace of symmetric matrices. The solution is the same whether the noise matrix $Z \in M_{n \times n}$ has i.i.d. standard normal entries, or whether Z is a symmetric Wigner matrix $\frac{1}{2}(Z_1 + Z_1')$ where $Z_1 \in M_{n \times n}$ has i.i.d. standard normal entries. Below, we assume that the data in case Sym is of the form $Y = \mu X_0 + Z/\sqrt{n}$ where $X_0 \in S_+^n$ and Z has this Wigner form, namely, the singular values $\mathbf{y}$ are the absolute values of eigenvalues of the symmetric matrix Y.

4. The least-favorable matrix for SVST is at $\|X\| = \infty$. We now prove Theorem 1, which characterizes the worst-case MSE of the SVST denoiser $\hat{X}_\lambda$ for a given λ. The theorem follows from a combination of two classical gems of the statistical literature. The first is Stein's unbiased risk estimate (SURE) from 1981, which we specialize to the SVST estimator; see also [2]. The second is Anderson's celebrated monotonicity property for the integral of a symmetric unimodal probability distribution over a symmetric convex set [1], from 1955, and more specifically its implications for monotonicity of the power function of certain tests in multivariate hypothesis testing [3]. To simplify the proof, we introduce the following definitions, which will be used in this section only.

DEFINITION 1 (A weak notion of matrix majorization based on singular values). Let $A, B \in M_{m \times n}$ have singular value vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$, respectively, which as usual we assume are sorted in nonincreasing order: $0 \leq a_m \leq \cdots \leq a_1$ and $0 \leq b_m \leq \cdots \leq b_1$. If $a_i \leq b_i$ for $i = 1, \ldots, m$, we write $A \preceq B$.

We note that by rescaling an arbitrary rank-r matrix, it is always possible to majorize any fixed matrix of rank at most r (in the sense of Definition 1).

LEMMA 1. *Let $C \in M_{m \times n}$ be a matrix of rank r, and let $X \in M_{m \times n}$ be a matrix of rank at most r. Then there exists $\mu > 0$ for which $X \preceq \mu C$.*

PROOF. Let $\mathbf{c}, \mathbf{x}$ be the vectors of singular values of C, X, respectively, each sorted in nonincreasing order. Then $c_r > 0$. Take $\mu = x_1/c_r$. For $1 \leq i \leq r$ we have $x_i \leq x_1 = \mu c_r \leq \mu c_i$, and for $r + 1 \leq i \leq m$ we have $\mu c_i = x_i = 0$. □

The above weak notion of majorization gives rise to a weak notion of monotonicity:

DEFINITION 2 (Orthogonally invariant function of a matrix argument). We say that $f : M_{m \times n} \to \mathbb{R}$ is an orthogonally invariant function if $f(U \cdot A \cdot V') = f(A)$ for all $A \in M_{m \times n}$ and all orthogonal $U \in O_m$ and $V \in O_n$.

DEFINITION 3 (SV-monotone increasing function of a matrix argument). Let $f : M_{m \times n} \to \mathbb{R}$ be orthogonally invariant. If, whenever $A \preceq B$ and $\sigma > 0$, f satisfies

$$(4.1) \qquad \mathbb{E} f(A + Z) \leq \mathbb{E} f(B + Z),$$

for $Z \in M_{m \times n}$ and $Z_{i,j} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, we say that f is singular-value-monotone increasing, or *SV-monotone increasing*.

We now provide a sufficient condition for SV-monotonicity, which follows from Anderson's seminal monotonicity result [1]. The following lemma is proved in the supplemental article [5].

LEMMA 2. *Assume that $f : M_{m \times n} \to \mathbb{R}$ can be decomposed as $f = \sum_{k=1}^{s} f_k$, where for each $1 \leq k \leq s$, $f_k : M_{m \times n} \to \mathbb{R}$ is a bounded, orthogonally invariant function. Further assume that for each $1 \leq k \leq s$, f_k is quasi-convex, in the sense that for all $c \in \mathbb{R}$, the set $f_k^{-1}((-\infty, c])$ is convex in $M_{m \times n}$. Then f is SV-monotone increasing.*

The second key ingredient in the proof of Theorem 1 is the Stein unbiased risk estimate for SVST. Let $\hat{X}$ be a weakly differentiable estimator of X_0 from data $Y = X_0 + \sigma Z$, where Z has i.i.d. standard normal entries. The Stein unbiased risk estimate [23] is a function of the data, $Y \mapsto \text{SURE}(Y)$, for which $\mathbb{E} \, \text{SURE}(Y) = \mathbb{E} \| \hat{X} - X_0 \|_2^2$. In our case, X_0, Z and Y are matrices in $M_{m \times n}$, and Stein's theorem ([23], Theorem 1) implies that for

$$\text{SURE}(Y) = mn\sigma^2 + \| \hat{X}(Y) - Y \|_F^2 + 2\sigma^2 \sum_{i,j} \frac{\partial (\hat{X}(Y) - Y)_{i,j}}{\partial Y_{i,j}},$$

we have

$$\| \hat{X} - X_0 \|_F^2 = \mathbb{E}_{X_0} \, \text{SURE}(Y).$$

In the supplemental article [5], we derive SURE for a large class of invariant matrix denoisers. As a result, we prove:

LEMMA 3 (The Stein unbiased risk estimate for SVST). *For each $\lambda > 0$, there exists an event $\mathcal{S} \subset M_{m \times n}$ and a function, $\text{SURE}_\lambda : \mathcal{S} \to \mathbb{R}$ which maps a matrix Y with singular values $\mathbf{y}$ to*

$$\text{SURE}_\lambda(Y) = m + \sum_{i=1}^{m} \left[(\min\{y_i, \lambda\})^2 - \mathbf{1}_{\{y_i < \lambda\}} - \frac{(n-m) \cdot \min\{y_i, \lambda\}}{y_i} \right]$$

$$- \frac{2}{n} \sum_{1 \leq i \neq j \leq m} \frac{\min\{y_j, \lambda\} y_j - \min\{y_i, \lambda\} y_j}{y_j^2 - y_i^2},$$

enjoying the following properties:

(1) $\mathbb{P}(\mathcal{S}) = 1$, *where $\mathbb{P}$ is the distribution of the matrix Z with $Z_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.*

(2) SURE$_\lambda$ *is a finite sum of bounded, orthogonally invariant, quasi-convex functions.*

(3) *Denoting as usual $Y = X_0 + Z/\sqrt{n} \in M_{m \times n}$, where $X_0, Z \in M_{m \times n}$ and $Z_{i,j} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, we have*

$$R(\hat{X}_\lambda, X_0) = \frac{1}{m} \mathbb{E}_{X_0} \text{SURE}_\lambda(Y).$$

Putting together Lemmas 2 and 3, we come to a crucial property of SVST.

LEMMA 4 (The risk of SVST is monotone nondecreasing in the signal singular values). *For each $\lambda > 0$, the map $X \mapsto R(\hat{X}_\lambda, X)$ is a bounded, SV-monotone increasing function. In particular, let $A, B \in M_{m \times n}$ with $A \preceq B$. Then*

(4.2) $$R(\hat{X}_\lambda, A) \leq R(\hat{X}_\lambda, B).$$

PROOF. By Lemma 3, the function $\text{SURE}_\lambda : M_{m \times n} \to \mathbb{R}$ satisfies the conditions of Lemma 2 and is therefore SV-monotone increasing. It follows that

$$R(\hat{X}_\lambda, A) = \frac{1}{m} \mathbb{E}_A \text{SURE}_\lambda(A + Z/\sqrt{n})$$

$$\leq \frac{1}{m} \mathbb{E}_B \text{SURE}_\lambda(B + Z/\sqrt{n}) = R(\hat{X}_\lambda, B).$$

To see that the risk is bounded, note that for any $X \in M_{m \times n}$, we have by Lemma 3

$$\infty < \inf_{Y \in M_{m \times n}} \mathbb{E}\,\text{SURE}_\lambda(Y) \leq R(\hat{X}_\lambda, X) \leq \sup_{Y \in M_{m \times n}} \mathbb{E}\,\text{SURE}_\lambda(Y) < \infty. \qquad \square$$

PROOF OF THEOREM 1. By Lemma 4, the map $\mu \to R(\hat{X}_\lambda, \mu C)$ is bounded and monotone nondecreasing in μ. Hence $\lim_{\mu \to \infty} R(\hat{X}_\lambda, \mu C)$ exists and is finite, and

(4.3) $$R(\hat{X}_\lambda, \mu_0 C) \leq \lim_{\mu \to \infty} R(\hat{X}_\lambda, \mu C)$$

for all $\mu_0 > 0$. Since $\text{rank}(C) = r$, obviously

$$\sup_{\text{rank}(X_0) \leq r} R(\hat{X}_\lambda, X_0) \geq \lim_{\mu \to \infty} R(\hat{X}_\lambda, \mu C),$$

and we only need to show the reverse inequality. Let $X_0 \in M_{m \times n}$ be an arbitrary matrix of rank at most r. By Lemma 1 there exists μ_0 such that $X_0 \preceq \mu_0 C$. It now follows from Lemma 4 and (4.3) that

$$R(\hat{X}_\lambda, X_0) \leq R(\hat{X}_\lambda, \mu_0 C) \leq \lim_{\mu \to \infty} R(\hat{X}_\lambda, \mu C). \qquad \square$$

5. Worst-case MSE. Let λ and $r \leq m \leq n$, and consider them fixed for the remainder of this section. Our second main result, Theorem 2, follows immediately from Theorem 1, combined with the following lemma, which is proved in the supplemental article [5].

LEMMA 5. *Let $X_0 \in M_{m \times n}$ be of rank r. Then*

$$\lim_{\mu \to \infty} R(\hat{X}_\lambda, \mu X_0) = \mathbf{M}_n \left(\frac{\lambda}{\sqrt{1 - r/n}}; r, m, \alpha \right),$$

as defined in (2.3), *with $\alpha = 1$ for case* Mat *and $\alpha = 1/2$ for case* Sym.

In the supplemental article [5] we prove the following lemma:

LEMMA 6. *The function $\Lambda \mapsto \mathbf{M}_n(\Lambda; r, m, \alpha)$, defined in* (2.3) *on $\Lambda \in [0, \infty)$, is convex and obtains a unique minimum.*

Our second main result is an immediate consequence:

PROOF OF THEOREM 2. Let $C \in M_{m \times n}$ be an arbitrary fixed matrix of rank r. For case Mat, by Theorem 1 and Lemma 5,

$$\mathcal{M}_n(r, m | \text{Mat}) = \inf_{\lambda} \sup_{\substack{X_0 \in M_{m \times n} \\ \text{rank}(X_0) \leq r}} R(\hat{X}_\lambda, X_0) = \inf_{\lambda > 0} \lim_{\mu \to \infty} R(\hat{X}_\lambda, \mu C)$$

$$= \inf_{\lambda > 0} \mathbf{M}_n \left(\frac{\lambda}{\sqrt{1 - r/n}}; r, m, 1 \right)$$

$$= \min_{\Lambda > 0} \mathbf{M}_n(\Lambda; r, m, 1),$$

where we have used Lemma 6, which also asserts that the minimum is unique.

Now let $C \in S_+^n$ be an arbitrary, fixed symmetric positive semidefinite matrix of rank r. For case Sym, by the same lemmas,

$$\mathcal{M}_n(r | \text{Sym}) = \inf_{\lambda} \sup_{\substack{X_0 \in M_{m \times n} \\ \text{rank}(X_0) \leq r}} R(\hat{X}_\lambda, X_0) = \inf_{\lambda} \lim_{\mu \to \infty} R(\hat{X}_\lambda, \mu C)$$

$$= \inf_{\lambda} \mathbf{M}_n \left(\frac{\lambda}{\sqrt{1 - r/n}}; r, 1/2 \right) = \min_{\Lambda} \mathbf{M}_n(\Lambda; r, 1/2). \qquad \square$$

6. Worst-case AMSE. Toward the proof of our third main result, Theorem 3, let λ be fixed. We first show that in the proportional growth framework, where the rank $r(n)$, number of rows $m(n)$ and number of columns n all tend to ∞ proportionally to each other, the key quantity in our formulas can be evaluated by complementary incomplete moments of a Marčenko–Pastur distribution, instead of a sum of complementary incomplete moments of Wishart eigenvalues.

DEFINITION 4.   For a pair of matrices $X_0, Z \in M_{m \times n}$, we denote by $\zeta(X_0, Z | \mathrm{Mat}) = (\zeta_1, \ldots, \zeta_{m-r})$ the singular values, in nonincreasing order, of

$$(6.1) \qquad \Pi_m \cdot Z \cdot \Pi_n' \in M_{(m-r) \times (n-r)},$$

where $\Pi_m : \mathbb{R}^m \to \mathbb{R}^{m-r}$ is the projection of $\mathbb{R}^m$ on $\mathrm{null}(X_0') = \mathrm{Im}(X_0)^\perp$ and $\Pi_n : \mathbb{R}^n \to \mathbb{R}^{n-r}$ is the projection on $\mathrm{null}(X_0)$. Similarly, for a pair of matrices $X_0, Z \in M_{n \times n}$, denote by $\zeta(X_0, Z | \mathrm{Sym}) = (\zeta_1, \ldots, \zeta_{m-r})$ the eigenvalues, in nonincreasing order, of

$$(6.2) \qquad \Pi_m \cdot \tfrac{1}{2}(Z + Z') \cdot \Pi_n' \in M_{(n-r) \times (n-r)}.$$

LEMMA 7.   *Consider sequences* $n \mapsto r(n)$ *and* $n \mapsto m(n)$ *and numbers* $0 < \beta \leq 1$ *and* $0 \leq \rho \leq 1$ *such that* $\lim_{n \to \infty} r(n)/m(n) = \rho$ *and* $\lim_{n \to \infty} m(n)/n = \beta$. *Let* $(\zeta_1(n), \ldots, \zeta_{m-r}(n)) = \zeta(X_0, Z | \mathbf{X})$, *as in Definition* 4, *where* $Z \in M_{m \times n}$ *has i.i.d.* $\mathcal{N}(0, 1)$ *entries. Define* $\gamma = (\beta - \rho\beta)/(1 - \rho\beta)$ *and* $\gamma_\pm = (1 \pm \sqrt{\gamma})^2$, *and let* $0 \leq \Lambda \leq \sqrt{\gamma_+}$. *Then*

$$\lim_{n \to \infty} \frac{1}{m} \sum_{i=1}^{m-r} \mathbb{E}\left(\frac{\zeta_i}{\sqrt{n-r}} - \Lambda \right)_+^2 = (1 - \rho) \int_{\Lambda^2}^{\gamma_+} (\sqrt{t} - \Lambda)^2 \frac{\sqrt{(\gamma_+ - t)(t - \gamma_-)}}{2\pi t \gamma} \, dt.$$

PROOF.   Write $\xi_i = \zeta_i^2/(n - r)$, and recall that by the Marčenko–Pastur law [15],

$$\lim_{n \to \infty} \frac{1}{m-r} \sum_{i=1}^{m-r} \delta_{\xi_i} \overset{w}{=} P_\gamma,$$

in the sense of weak convergence of probability measures, where P_γ is the Marčenko–Pastur probability distribution with density $p_\gamma = dP_\gamma/dt$ given by (2.7). Now,

$$\lim_{n \to \infty} \frac{1}{m} \sum_{i=1}^{m-r} (\sqrt{\xi_i} - \Lambda)_+^2 = \lim_{n \to \infty} \frac{1}{m} \sum_{i=1}^{m-r} \int_0^\infty (\sqrt{t} - \Lambda)_+^2 \delta_{\xi_i}(t) \, dt$$

$$= \lim_{n \to \infty} \left(1 - \frac{r}{m} \right) \int_0^\infty (\sqrt{t} - \Lambda)_+^2 \frac{1}{m-r} \sum_{i=1}^{m-r} \delta_{\xi_i}(t) \, dt$$

$$= (1 - \rho) \int_0^{\gamma_+} (\sqrt{t} - \Lambda)_+^2 p_\gamma(t) \, dt$$

as required.   □

LEMMA 8.   *Let* $m(n)$ *and* $r(n)$ *such that* $\lim_{n \to \infty} m/n = \beta$ *and* $\lim_{n \to \infty} r/m = \rho$, *and set* $\tilde{\rho} = \beta\rho$. *Then*

$$\lim_{n \to \infty} \sup_{\substack{X_0 \in M_{m \times n} \\ \mathrm{rank}(X_0) \leq r}} R(\hat{X}_\lambda, X_0) = \mathbf{M}\left(\frac{\lambda}{\sqrt{1 - \tilde{\rho}}}; \rho, \tilde{\rho}, \alpha \right),$$

where the right-hand side is defined in (2.9), *with* $\alpha = 1$ *for case* Mat *and* $\alpha = 1/2$ *for case* Sym.

PROOF. For case Mat, let $C(n) \in M_{m \times n}$ be an arbitrary fixed matrix of rank r. For case Sym, $C(n) \in S_+^n$ an arbitrary, fixed symmetric positive semidefinite matrix of rank r. By Theorem 1 and Lemma 5,

$$\lim_{n \to \infty} \sup_{\substack{X_0 \in M_{m \times n} \\ \operatorname{rank}(X_0) \leq r}} R(\hat{X}_\lambda, X_0)$$

$$= \lim_{n \to \infty} \lim_{\mu \to \infty} R(\hat{X}_\lambda, \mu C(n))$$

$$= \lim_{n \to \infty} \left[\frac{r}{m} + \frac{r}{n} - \frac{r^2}{mn} + \frac{r}{m} \lambda^2 \right.$$

$$\left. + \alpha \frac{n-r}{mn} \sum_{i=1}^{m-r} \mathbb{E} \left(\frac{\zeta_i}{\sqrt{n-r}} - \frac{\lambda}{\sqrt{1-r/n}} \right)_+^2 \right]$$

$$= \rho + \tilde{\rho} - \rho\tilde{\rho} + (1-\tilde{\rho})\rho\Lambda^2$$

$$+ \alpha(1-\rho)(1-\tilde{\rho}) \int_{\Lambda^2}^{\gamma_+} (\sqrt{t} - \Lambda)^2 M P_\gamma(t) \, dt$$

$$= \mathbf{M} \left(\frac{\lambda}{\sqrt{1-\tilde{\rho}}}; \rho, \tilde{\rho}, \alpha \right),$$

where we have used Lemma 7 and set $\Lambda = \lambda / \sqrt{1-\tilde{\rho}}$. □

In the supplemental article we prove a variation of Lemma 6 for the asymptotic setting:

LEMMA 9. *The function* $\Lambda \mapsto \mathbf{M}(\Lambda; \rho, \tilde{\rho}, \alpha)$, *defined in* (2.9) *on* $\Lambda \in [0, \gamma_+]$, *where* $\gamma_+ = (1 + \sqrt{(\tilde{\rho} - \rho\tilde{\rho})/(\rho - \rho\tilde{\rho})})^2$, *is convex and obtains a unique minimum.*

This allows us to the prove our third main result.

PROOF OF THEOREM 3. By Lemma 8,

$$\mathcal{M}(\rho, \beta | \mathbf{X}) = \lim_{n \to \infty} \inf_\lambda \sup_{\substack{X_0 \in M_{m \times n} \\ \operatorname{rank}(X_0) \leq r}} R(\hat{X}_\lambda, X_0)$$

$$= \inf_\lambda \lim_{n \to \infty} \sup_{\substack{X_0 \in M_{m \times n} \\ \operatorname{rank}(X_0) \leq r}} R(\hat{X}_\lambda, X_0)$$

$$= \inf_\lambda \mathbf{M} \left(\frac{\lambda}{\sqrt{1-\tilde{\rho}}}; \rho, \tilde{\rho}, \alpha \right) = \min_\Lambda \mathbf{M}(\Lambda; \rho, \tilde{\rho}, \alpha),$$

with $\alpha = 1$ for case Mat and $\alpha = 1/2$ for case Sym, where we have used Lemma 9, which also asserts that the minimum is unique. $\square$

7. Minimax AMSE. Having established that the asymptotic worst-case MSE (2.9) satisfies (2.10) and (2.11), we turn to its minimizer Λ_*. The notation follows (2.12).

PROOF OF THEOREM 4. By equation (4.2) in the supplemental article [5], the condition

$$\frac{d\mathbf{M}(\Lambda; \rho, \tilde{\rho}, \alpha)}{d\Lambda} = 0$$

is thus equivalent, for any $\rho \in [0, 1]$, to

$$(7.1) \qquad f(\Lambda, \rho) := \rho\Lambda - \alpha(1 - \rho)\int_{\Lambda^2}^{\gamma_+} (\sqrt{t} - \Lambda)p_\gamma(t)\,dt = 0,$$

establishing (2.15) in particular for $0 < \rho < 1$. By Lemma 9, the minimum exists and is unique; namely this equation has a unique root in Λ. One directly verifies that $f(1 + \sqrt{\beta}, 0) = f(0, 1) = 0$. The limits (2.13) and (2.14) follow from the fact that $\rho \mapsto \Lambda_*(\rho, \cdot)$ is decreasing. To establish this, it is enough to observe that $\partial f/\partial\rho > 0$ for all (Λ, ρ), which can be verified directly. $\square$

Theorem 5, which provides more a explicit formula for the minimax AMSE in square matrix case ($\beta = 1$), is proved in the supplemental article [5].

8. Global minimax MSE and AMSE. In this section we prove Theorem 9, which provides a lower bound on the minimax risk of the family of all measurable matrix denoisers (as opposed to the family of SVST denoisers considered so far) over m-by-n matrices of rank at most r. Consider the class of singular-value matrix denoisers, namely all mappings $Y \mapsto \hat{X}(Y)$ that act on the data Y only through their singular values. More specifically, consider all denoisers $\hat{X} : M_{m\times n} \to M_{m\times n}$ of the form

$$(8.1) \qquad \hat{X}(Y) = U_Y \cdot \hat{\mathbf{x}}(\mathbf{y})_\Delta \cdot V_Y',$$

where $Y = U_Y \cdot \mathbf{y}_\Delta \cdot V_Y'$ and $\hat{\mathbf{x}} : [0, \infty)^m \to [0, \infty)^m$. (Note that this class contains SVST denoisers but does not exhaust all measurable denoisers.) The mapping in (8.1) is not well defined in general, since the SVD of Y, and in particular the order of the singular values in the vector $\mathbf{y}$, is not uniquely determined. However, (8.1) is well defined when each function $\hat{x}_i : [0, \infty) \to [0, \infty)$ is invariant under permutations of its coordinates. Since the equality $Y = U_Y \cdot \mathbf{y}_\Delta \cdot V_Y'$ may hold for vectors $\mathbf{y}$ with negative entries, we are led to the following definition.

DEFINITION 5. By *singular-value denoiser* we mean any measurable mapping $\hat{X}: M_{m \times n} \to M_{m \times n}$ which takes the form (8.1), where each entry of $\hat{\mathbf{x}}$ is a function $\hat{x}_i : \mathbb{R}^m \to \mathbb{R}$ that is invariant under permutation and sign changes of its coordinates. We let $\mathcal{D}$ denote the class of such mappings.

For a detailed introduction to real-valued or matrix-valued functions which depend on a matrix argument only through its singular values, see [13, 14]. The following lemma is proved in the supplemental article [5].

LEMMA 10 (Singular-value denoisers can only improve in worst-case). *Let $\hat{X}_1 : M_{m \times n} \to M_{m \times n}$ be an arbitrary measurable matrix denoiser. There exists a singular-value denoiser $\hat{X}$ such that*

$$\sup_{\substack{X_0 \in M_{m \times n} \\ \operatorname{rank}(X_0) \leq r}} R(\hat{X}, X_0) \leq \sup_{\substack{X_0 \in M_{m \times n} \\ \operatorname{rank}(X_0) \leq r}} R(\hat{X}_1, X_0).$$

PROOF OF THEOREM 9. We consider the case $\mathbf{X} = \operatorname{Mat}_{m,n}$. By Lemma 10, it is enough to show that

$$\frac{r}{m} + \frac{r}{n} - \frac{r^2 + r}{mn} \leq \sup_{\substack{X_0 \in \mathbf{X}_{m,n} \\ \operatorname{rank}(X_0) \leq r}} R(\hat{X}, X_0),$$

where $\hat{X} \in \mathcal{D}$ is an arbitrary singular-value denoiser. Indeed, let $X_0 \in M_{m \times n}$ be a fixed arbitrary matrix of rank r. The calculation leading to equation (3.9) in the supplemental article [5] is valid for any rule in $\mathcal{D}$, and implies that $R(\hat{X}(Y), X_0) \geq 1 - \frac{1}{m}\mathbb{E}\|\mathbf{z}\|_2^2$, where $Y = U_Y \cdot \mathbf{y}_\Delta \cdot V_Y'$ and

(8.2) $$\mathbf{z} = \frac{1}{\sqrt{n}}(U_Y' \cdot Z \cdot V)_\Delta.$$

Write $Y_\mu = \mu X_0 + Z/\sqrt{n} = U_\mu \cdot (\mathbf{y}_\mu)_\Delta \cdot V_\mu'$, and let $\mathbf{z}_\mu = \frac{1}{\sqrt{n}}(U_\mu' \cdot Z \cdot V_\mu)_\Delta$. We therefore have

$$\sup_{\substack{X_0 \in \mathbf{X}_{m,n} \\ \operatorname{rank}(X_0) \leq r}} R(\hat{X}, X_0) \geq \lim_{\mu \to \infty} R(\hat{X}, \mu X_0) \geq 1 - \frac{1}{m} \lim_{\mu \to \infty} \mathbb{E}\|z_\mu\|_2^2.$$

Combining equations (3.17) and (3.15) in the supplemental article [5], we have

$$\frac{1}{m} \sum_{i=r+1}^{m} \lim_{\mu \to \infty} \mathbb{E}(z_{\mu,i})^2 = 1 - \frac{r}{m} - \frac{r}{n} + \frac{r^2}{mn}.$$

A similar argument yields $\frac{1}{m} \sum_{i=1}^{r} \lim_{\mu \to \infty} \mathbb{E}(z_{\mu,i})^2 = \frac{r}{mn}$, and the first part of the theorem follows. The second part of the theorem follows since, taking the limit

$n \to \infty$ as prescribed, we have $r/m \to \rho$, $r/n \to \tilde{\rho}$ and $r/mn \to 0$. For the third part of the theorem, we have by Theorem 8,

$$\lim_{\rho \to 0} \frac{\mathcal{M}(\rho, \beta | \mathbf{X})}{\mathcal{M}^-(\rho, \beta)} = \lim_{\rho \to 0} \frac{\mathcal{M}(\rho, \beta | \mathbf{X})}{\rho + \beta\rho + \beta\rho^2} = \frac{2(1 + \sqrt{\beta} + \beta)}{1 + \beta}$$
$$= 2\left(1 + \frac{\sqrt{\beta}}{1 + \beta}\right). \qquad \square$$

9. Discussion. In the Introduction, we pointed out several ways that these matrix denoising results for SVST estimation of low-rank matrices parallel results for soft thresholding of sparse vectors. Our derivation of the minimax MSE formulas exposed two more parallels:

- *Common structure of minimax MSE formulas.* The minimax MSE formula vector denoising problem involves certain incomplete moments of the standard Gaussian distribution [7]. The matrix denoising problem involves completely analogous incomplete moments, only replacing the Gaussian by the Marčenko–Pastur distribution or (in the square case $\beta = 1$) the quarter-circle law.
- *Monotonicity of SURE.* In both settings, the least-favorable estimand places the signal "at ∞," which yields a convenient formula for Minimax MSE [7]. In each setting, validation of the least-favorable estimation flows from monotonicity, in an appropriate sense, of Stein's unbiased risk estimate within that specific setting.

SUPPLEMENTARY MATERIAL

Proofs and additional discussion (DOI: 10.1214/14-AOS1257SUPP; .pdf). In this supplementary material we prove Theorems 5, 6, 7, 8 and other lemmas. We also derive the Stein unbiased risk Estimate (SURE) for SVST, which is instrumental in the proof of Theorem 1. Finally, we discuss similarities between singular value thresholding and sparse vector thresholding.

REFERENCES

[1] ANDERSON, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proc. Amer. Math. Soc.* **6** 170–176. MR0069229

[2] CANDÈS, E. J., SING-LONG, C. A. and TRZASKO, J. D. (2013). Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Trans. Signal Process.* **61** 4643–4657. MR3105401

[3] DAS GUPTA, S., ANDERSON, T. W. and MUDHOLKAR, G. S. (1964). Monotonicity of the power functions of some tests of the multivariate linear hypothesis. *Ann. Math. Statist.* **35** 200–205. MR0158474

[4] DONOHO, D. L. and GAVISH, M. (2013). Companion website for the article the phase transition of matrix recovery from Gaussian measurements matches the minimax MSE of matrix denoising. Available at http://www.runmycode.org/CompanionSite/Site265.

[5] DONOHO, D. and GAVISH, M. (2014). Supplement to "Minimax risk of matrix denoising by singular value thresholding." DOI:10.1214/14-AOS1257SUPP.

[6] DONOHO, D. L., GAVISH, M. and MONTANARI, A. (2013). The phase transition of matrix recovery from Gaussian measurements matches the minimax MSE of matrix denoising. *Proc. Natl. Acad. Sci. USA* **110** 8405–8410. MR3082268

[7] DONOHO, D. L., JOHNSTONE, I. and MONTANARI, A. (2013). Accurate prediction of phase transitions in compressed sensing via a connection to minimax denoising. *IEEE Trans. Inform. Theory* **59** 3396–3433. MR3061255

[8] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Minimax risk overl ℓ_p-balls forl ℓ_q-error. *Probab. Theory Related Fields* **303** 277–303.

[9] DONOHO, D. L., JOHNSTONE, I. M., HOCH, J. C. and STERN, A. S. (1992). Maximum entropy and the nearly black object. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **54** 41–81. MR1157714

[10] GAVISH, M. and DONOHO, D. L. (2014). The optimal hard threshold for singular values is $4/\sqrt{3}$. *IEEE Trans. Inform. Theory* **60** 5040–5053.

[11] GRANT, M. and BOYD, S. P. (2010). CVX: Matlab software for disciplined convex programming, version 2.0 beta. Available at http://cvxr.com/cvx, September 2013.

[12] KOLTCHINSKII, V., LOUNICI, K. and TSYBAKOV, A. B. (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Statist.* **39** 2302–2329. MR2906869

[13] LEWIS, A. S. (1995). The convex analysis of unitarily invariant matrix functions. *J. Convex Anal.* **2** 173–183. MR1363368

[14] LEWIS, A. S. and SENDOV, H. S. (2005). Nonsmooth analysis of singular values. I. Theory. *Set-Valued Var. Anal.* **13** 213–241. MR2162512

[15] MARCENKO, V. and PASTUR, L. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics USSR Sbornik* **1** 457–483.

[16] OYMAK, S. and HASSIBI, B. (2010). New null space results and recovery thresholds for matrix rank minimization. Preprint. Available at http://arxiv.org/pdf/1011.6326v1.pdf.

[17] OYMAK, S. and HASSIBI, B. (2012). On a relation between the minimax risk and the phase transitions of compressed recovery. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing* 1018–1025. IEEE, Piscataway, NJ. Available at http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6483330.

[18] RECHT, B., FAZEL, M. and PARRILO, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* **52** 471–501. MR2680543

[19] RECHT, B., XU, W. and HASSIBI, B. (2008). Necessary and sufficient conditions for success of the nuclear norm heuristic for rank minimization. In *Proceedings of the* 47*th IEEE Conference on Decision and Control Cancun*, *Mexico*.

[20] RECHT, B., XU, W. and HASSIBI, B. (2011). Null space conditions and thresholds for rank minimization. *Math. Program.* **127** 175–202. MR2776714

[21] ROHDE, A. and TSYBAKOV, A. B. (2011). Estimation of high-dimensional low-rank matrices. *Ann. Statist.* **39** 887–930. MR2816342

[22] SHABALIN, A. and NOBEL, A. (2010). Reconstruction of a low-rank matrix in the presence of Gaussian noise. Preprint. Available at arXiv:1007.4148.

[23] STEIN, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151. MR0630098

[24] TANNER, J. and WEI, K. (2013). Normalized iterative hard thresholding for matrix completion. *SIAM J. Sci. Comput.* **35** S104–S125. MR3120763

[25] ZANELLA, A., CHIANI, M. and WIN, M. Z. (2009). On the marginal distribution of the eigenvalues of Wishart matrices. *IEEE Transactions on Communications* **57** 1050–1060.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL, 390 SERRA MALL
STANFORD, CALIFORNIA 94305-4065
USA
E-MAIL: donoho@stanford.edu
            gavish@stanford.edu