The Optimal Hard Threshold for Singular Values is $4/\sqrt{3}$

Matan Gavish and David L. Donoho

arXiv:1305.5870v3 [stat.ME] 4 Jun 2014

***Abstract*—We consider recovery of low-rank matrices from noisy data by hard thresholding of singular values, in which empirical singular values below a threshold λ are set to 0. We study the asymptotic MSE (AMSE) in a framework where the matrix size is large compared to the rank of the matrix to be recovered, and the signal-to-noise ratio of the low-rank piece stays constant. The AMSE-optimal choice of hard threshold, in the case of n-by-n matrix in white noise of level σ, is simply $(4/\sqrt{3})\sqrt{n}\sigma \approx 2.309\sqrt{n}\sigma$ when σ is known, or simply $2.858 \cdot y_{med}$ when σ is unknown, where y_{med} is the median empirical singular value. For nonsquare m by n matrices with $m \neq n$ the thresholding coefficients $4/\sqrt{3}$ and 2.858 are replaced with different provided constants that depend on m/n. Asymptotically, this thresholding rule adapts to unknown rank and unknown noise level in an optimal manner: it is *always* better than hard thresholding at any other value, and is *always* better than ideal Truncated SVD (TSVD), which truncates at the true rank of the low-rank matrix we are trying to recover. Hard thresholding at the recommended value to recover an n-by-n matrix of rank r guarantees an AMSE at most $3\,nr\sigma^2$. In comparison, the guarantees provided by TSVD, optimally tuned singular value soft thresholding and the best guarantee achievable by any shrinkage of the data singular values are $5\,nr\sigma^2$, $6\,nr\sigma^2$, and $2\,nr\sigma^2$, respectively. The recommended value for hard threshold also offers, among hard thresholds, the best possible AMSE guarantees for recovering matrices with bounded nuclear norm. Empirical evidence suggests that performance improvement over TSVD and other popular shrinkage rules can be substantial, for different noise distributions, even in relatively small n.**

***Index Terms*—Singular values shrinkage, optimal threshold, low-rank matrix denoising, unique admissible, scree plot elbow truncation, quarter circle law, bulk edge.**

I. Introduction

SUPPOSE that we are interested in an unknown m-by-n matrix X, thought to be either exactly or approximately of low rank, but we only observe a single noisy m-by-n matrix Y, obeying $Y = X + \sigma Z$. The noise matrix Z has independent, identically distributed, zero-mean entries. The matrix X is a (non-random) parameter, and we wish to estimate it with some bound on the mean squared error (MSE).

The default estimation technique for our task is *Truncated SVD* (TSVD) [2]: Write

$$Y = \sum_{i=1}^{m} y_i \mathbf{u}_i \mathbf{v}_i' \tag{1}$$

for the Singular Value Decomposition of the data matrix Y, where $\mathbf{u}_i \in \mathbb{R}^m$ and $\mathbf{v}_i \in \mathbb{R}^n$, $i = 1, \ldots, m$ are the left and right singular vectors of Y corresponding to the singular value y_i. The TSVD estimator is

$$\hat{X}_r = \sum_{i=1}^{r} y_i \mathbf{u}_i \mathbf{v}_i' \,,$$

where $r = rank(X)$, assumed known, and $y_1 \geq \ldots \geq y_m$. Being the best approximation of rank r to the data in the least squares sense [3], and therefore the Maximum Likelihood estimator when Z has Gaussian entries, the TSVD is arguably as ubiquitous in science and engineering as linear regression [4]–[9].

When the true rank r of the signal X is unknown, one might try to form an estimate $\hat{r}$ and then apply the TSVD $\hat{X}_{\hat{r}}$. Extensive literature has formed on methods to estimate r: we point to the early [9], [10] (in Factor Analysis and Principal Component Analysis), the recent [11]–[13] (in our setting of Singular Value Decomposition), and reference therein. It is instructive to think about rank estimation (using any method), followed by TSVD, simply as *hard thresholding* of the data singular values, where only components $y_i \mathbf{u}_i \mathbf{v}_i'$ for which y_i passes a specified threshold, are included in $\hat{X}$. Let $\eta_H(y, \tau) = y \mathbf{1}_{\{y \geq \tau\}}$ denote the hard thresholding nonlinearity, and consider Singular Value Hard Thresholding (SVHT)

$$\hat{X}_\tau = \sum_{i=1}^{m} \eta_H(y_i; \tau)\, \mathbf{u}_i \mathbf{v}_i' \,. \tag{2}$$

In words, $\hat{X}_\tau$ sets to 0 any data singular value below τ.

Matrix denoisers explicitly or implicitly based on hard thresholding of singular values have been proposed by many authors, including [12]–[20]. As a common example of implicit SVHT denoising, consider the standard practice of estimating r by plotting the singular values of Y in decreasing order, and looking for a "large gap" or "elbow" (Figure 1, left panel). When X is exactly or approximately low-rank and the entries of Z are white noise of zero mean and unit variance, the empirical distribution of the singular values of the m-by-n matrix $Y = X + \sigma Z$ forms a quarter-circle bulk whose edge lies approximately at $(1 + \sqrt{\beta}) \cdot \sqrt{n}\sigma$, with $\beta = m/n$ [21]. Only data singular values that are larger than the bulk edge are noticeable in the empirical distribution (Figure 1, right plot). Since the singular value plot "elbow" is located at the bulk edge, the popular method of TSVD at the "elbow" is an approximation of *bulk-edge hard thresholding*, $\hat{X}_{(1+\sqrt{\beta})\sqrt{n}\sigma}$.

The authors are with the Department of Statistics, Stanford University, Stanford, CA 94305 USA (e-mail: gavish@stanford.edu; donoho@stanford.edu)

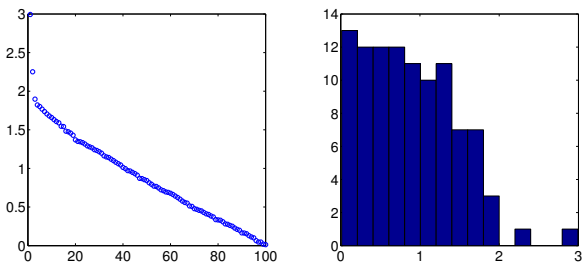

Fig. 1. Singular values of a data matrix $Y \in M_{100,100}$ sampled from the model $Y = X + Z/\sqrt{100}$, where $X_{1,1} = 1.7$, $X_{2,2} = 2.5$ and $X_{i,j} = 0$ elsewhere. In Matlab, a sample can be generated using the command `y=svd(diag([[1.7 2.5] zeros(1,98)])+randn(100)/sqrt(100))`. Left: the singular values of Y plotted in decreasing order, known in Principal Components Analysis as a Scree plot. Right: the singular values of Y in a histogram with 15 bins. Note the bulk edge approximately at 2, and note that the location of the top two singular values of Y is approximately $X_{i,i} + 1/X_{i,i}$ $(i = 1, 2)$, in agreement with Lemma 2.

A. Questions

Let us measure the denoising performance of a denoiser $\hat{X}$ at a signal matrix X using Mean Square Error (MSE),

$$\left|\left|\hat{X}(Y) - X\right|\right|_F^2 = \sum_{i,j} (\hat{X}(Y)_{i,j} - X_{i,j})^2 .$$

The TSVD is an optimal rank-r approximation of the data matrix Y, in MSE. But this does not necessarily mean that it is a good, or even reasonable, estimator to the signal matrix X, which we wish to recover. We may wonder:

- *Question 1.* Assume that $rank(X)$ is unknown but small. Is there a singular value threshold τ so that SVHT $\hat{X}_\tau$ successfully adapts to unknown rank and unknown noise level, and performs as well as TSVD would, had we known the true $rank(X)$?

As we will see, it is convenient to represent the threshold as $\tau = \lambda\sqrt{n}\sigma$, where λ is a parameter typically between 1 and 10. Recently, S. Chatterjee [17] proposed that one could have a single universal choice of λ; that in a setting more general, but similar, to our setting, any $\lambda > 2$ would give near-optimal MSE, in a qualitative sense; and he specifically proposed $\lambda = 2.02$, namely $\hat{X}_{2.02\sqrt{n}\sigma}$ as a universal choice for SVHT, regardless of the shape m/n of the matrix, and regardless of the underlying signal matrix X or its rank. While the rule of [17] was originally intended to be 'fairly good' across many situations not reducible to the low-rank matrix in i.i.d noise model considered here, $\hat{X}_{2.02\sqrt{n}\sigma}$ is a specific proposal, which prompts the following question:

- *Question 2.* Is there really a single threshold parameter λ_* that provides good performance guarantees for MSE? Is that value 2.02? Is it really independent of m and n?

Finally, note that singular value hard thresholding is just one strategy for matrix denoising. It is not a-priori clear whether the whole idea of only 'keeping' or 'killing' empirical singular values based on their size makes sense. Could there exist a shrinkage rule $\eta : [0, \infty) \to [0, \infty)$, that more smoothly transitions from 'killing' to 'keeping', which leads to a much better denoising scheme? We may wonder:

- *Question 3.* How does optimally tuned SVHT compare with the performance of the *best possible shrinkage* of singular values, at least in the worst-case MSE sense?

B. Optimal location for hard thresholding of singular values

Our main results imply that, in a certain asymptotic framework, there are simple and convincing answers to these questions. Following Perry [13] and Shabalin and Nobel [22], we adopt an asymptotic framework where the matrix grows while keeping the nonzero singular values of X fixed, and the signal-to-noise ratio of those singular values stays constant with increasing n.

In this asymptotic framework, for a low-rank n-by-n matrix observed in white noise of level σ,

$$\tau_* = \frac{4}{\sqrt{3}}\sqrt{n}\sigma \approx 2.309\sqrt{n}\sigma$$

is the optimal location for the hard thresholding of singular values. For a non-square m-by-n matrix with $m \neq n$, the optimal location is

$$\tau_* = \lambda_*(\beta) \cdot \sqrt{n}\sigma, \tag{3}$$

where $\beta = m/n$. The value $\lambda_*(\beta)$ is the *optimal hard threshold coefficient* for known σ. It is given by formula (11) below and tabulated for convenience in Table I. (Note added in proof: we found that P. Perry's PhD thesis [13] proposes a threshold which can be shown to be equivalent to (3).)

C. Answers

Our central observation is as follows.

> *When a data singular value y_i is too small, then its associated singular vectors $\mathbf{u}_i, \mathbf{v}_i$ are so noisy that the component $y_i\mathbf{u}_i\mathbf{v}_i'$ should not included in $\hat{X}$. In our asymptotic framework, which models large, low-rank matrices observed in white noise, the cutoff below which y_i is* too small *is exactly $(4/\sqrt{3})\sqrt{n}\sigma$ (for square matrices).*

- *Answer to Question 1: Optimal SVHT dominates TSVD.* Optimally tuned SVHT $\hat{X}_{\tau_*}$ is *always* at least as good as TSVD $\hat{X}_r$, in terms of AMSE (Theorem 2). Unlike $\hat{X}_r$, the optimal SVHT $\hat{X}_{\tau_*}$ does not require knowledge of $r = rank(X)$. In other words, it adapts to unknown low rank while giving uniformly equal or better performance. For square matrices, the TSVD provides a guarantee on worst-case AMSE that is $5/3$ times the guarantee provided by $\hat{X}_{\tau_*}$ (Table II).
- *Answer to Question 2: Optimal SVHT dominates every other choice of Hard Threshold.* In terms of AMSE, optimally tuned SVHT $\hat{X}_{\tau_*}$ is *always* at least as good as SVHT $\hat{X}_\tau$ at any other fixed threshold $\tau = \lambda\sqrt{n}\sigma$ (Theorem 1). It is the asymptotically minimax SVHT denoiser, over matrices of small bounded rank (Theorems 3 and 4) and over matrices of small bounded nuclear norm (Theorem 5). In particular, the parameter $\lambda = 2.02$ is noticeably worse. For square matrices, $\hat{X}_{2.02\sqrt{n}\sigma}$ provides a guarantee for worst-case AMSE that is $4.26/3 \approx 1.4$ times the guarantee provided by $\hat{X}_{\tau_*}$ (Table II).

- *Answer to Question 3. Optimal SVHT compares adequately to the optimal shrinker.* Optimally tuned SVHT $\hat{X}_{\tau_*}$ provides a guarantee on worst-case asymptotic MSE that is $3/2$ times (for square matrices) the best possible guarantee achievable by *any* shrinkage of data singular values (Table II).

These are all rigorous results, within a specific asymptotic framework, which prescribes a certain scaling of the noise level, the matrix size, and the signal-to-noise ratio as n grows. But does AMSE predict actual MSE in finite-sized problems? In Section VII we show finite-n simulations demonstrating the effectiveness of these results even at rather small problem sizes. In high signal-to-noise, all denoisers considered here perform roughly the same, and in particular the classical TSVD is a valid choice in that regime. However, in low and moderate SNR, the performance gain of optimally tuned SVHT is substantial, and can offer $30\% - 80\%$ decrease in AMSE.

D. Optimal singular value hard thresholding – In practice

For a low-rank n-by-n matrix observed in white noise of *unknown* level, one can use the data to obtain an approximation of the optimal location τ_*. Define

$$\hat{\tau}_* \approx 2.858 \cdot y_{med}\,,$$

where y_{med} is the median singular value of the data matrix Y. The notation $\hat{\tau}_*$ is meant to emphasize that this is not a fixed threshold chosen a-priori, but rather a data-dependent threshold. For a non-square m-by-n matrix with $m \neq n$, the approximate optimal location when σ is unknown is

$$\hat{\tau}_* = \omega(\beta) \cdot y_{med}\,. \tag{4}$$

The optimal hard threshold coefficient for unknown σ, denoted by $\omega(\beta)$, is not available as an analytic formula, but can easily be evaluated numerically. We provide a Matlab script for this purpose [1]; the underlying derivation appears in Section III-E below. Some values of $\omega(\beta)$ are provided in Table IV. When a high-precision value of $\omega(\beta)$ cannot be computed, one can use the approximation

$$\omega(\beta) \approx 0.56\beta^3 - 0.95\beta^2 + 1.82\beta + 1.43\,. \tag{5}$$

The optimal SVHT for unknown noise level, $\hat{X}_{\hat{\tau}_*}$, is very simple to implement and does not require any tuning parameters. The denoised matrix $\hat{X}_{\hat{\tau}_*}(Y)$ can be computed using just a few code lines in a high-level scripting language. For example, in Matlab:

```
beta = size(Y,1) / size(Y,2);
omega = 0.56*beta^3 - 0.95*beta^2 + ...
     1.82*beta + 1.43;
[U D V] = svd(Y);
y = diag(Y);
y( y < (omega * median(y) ) ) = 0;
Xhat = U * diag(y) * V';
```

Here we have used the approximation (5). We recommend, whenever possible, to use a function `omega(beta)`, such as the one we provide in the code supplement [1], to compute the coefficient $\omega(\beta)$ to high precision.

In our asymptotic framework, τ_* and $\hat{\tau}_*$ enjoy exactly the same optimality properties. This means that $\hat{X}_{\hat{\tau}_*}$ adapts to unknown low rank *and* to unknown noise level. Empirical evidence suggest that their performance for finite n is similar. As a result, the answers we provide above hold for the threshold $\hat{\tau}_*$ when the noise level is unknown, just as they hold for the threshold τ_* when the noise level is known.

II. Preliminaries and setting

Column vectors are denoted by boldface lowercase letters, such as $\mathbf{v}$, their transpose is $\mathbf{v}'$ and their i-th coordinate is v_i. The Euclidean inner product and norm on vectors are denoted by $\langle \mathbf{u}\,,\, \mathbf{v} \rangle$ and $||\mathbf{u}||_2$, respectively. Matrices are denotes by uppercase letters, such as X, its transpose is X' and their i,j-th entry is $A_{i,j}$. $M_{m\times n}$ denotes the space of real m-by-n matrices, $\langle X\,,\, Y \rangle = \sum_{i,j} X_{i,j} Y_{i,j}$ denotes the Hilbert-Schmidt inner product, and $||X||_F$ denotes the corresponding Frobenius norm on $M_{m\times n}$. For simplicity we only consider $m \leq n$. We denote matrix denoisers, or estimators, by $\hat{X} : M_{m\times n} \to M_{m\times n}$. The symbols $\xrightarrow{a.s.}$ and $\overset{a.s.}{=}$ denote almost sure convergence and equality of a.s. limits, respectively.

A. Scaling considerations in singular value thresholding

With the exception of TSVD, when σ is known, all the denoisers we discuss operate by shrinkage of data singular values, namely are of the form

$$\hat{X} : \sum_{i=1}^{m} y_i \mathbf{u}_i \mathbf{v}_i' \mapsto \sum_{i=1}^{m} \eta(y_i;\lambda) \mathbf{u}_i \mathbf{v}_i' \tag{6}$$

where Y is given by (1) and $\eta : [0,\infty) \to [0,\infty)$ is some univariate shrinkage rule. As we will see, in the general model $Y = X + \sigma Z$, the noise level in the singular values of Y is $\sqrt{n}\sigma$. Instead of specifying a different shrinkage rule that depends on the matrix size n, we calibrate our shrinkage rules to the "natural" model $Y = X + Z/\sqrt{n}$. In this convention, shrinkage rules stay the same for every value of n, and we conveniently abuse notation by writing $\hat{X}$ as in (6) for any $\hat{X} : M_{m\times n} \to M_{m\times n}$, keeping m and n implicit. To apply any denoiser $\hat{X}$ below to data from the general model $Y = X + \sigma Z$, use the denoiser

$$\hat{X}^{(n,\sigma)}(Y) = \sqrt{n}\sigma \cdot \hat{X}(Y/\sqrt{n}\sigma)\,. \tag{7}$$

For example, to apply the SVHT

$$\hat{X}_\lambda : \sum_{i=1}^{m} y_i \mathbf{u}_i \mathbf{v}_i' \mapsto \sum_{i=1}^{m} \eta_H(y_i;\lambda) \mathbf{u}_i \mathbf{v}_i'$$

to data sampled from the model $Y = X + \sigma Z$, use $\hat{X}_\tau$, with

$$\tau = \lambda \cdot \sqrt{n}\sigma\,.$$

Throughout the text, we use $\hat{X}_\lambda$ to denote SVHT calibrated for noise level $1/\sqrt{n}$ and $\hat{X}_\tau$ to denote SVHT calibrated for a specific general model $Y = X + \sigma Z$.

To translate the AMSE of any denoiser $\hat{X}$, calibrated for noise level $1/\sqrt{n}$, to an approximate MSE of the corresponding denoiser $\hat{X}^{(n,\sigma)}$, calibrated for a model $Y = X + \sigma Z$, we use the identity

$$\left|\left|\hat{X}^{(n,\sigma)}(Y) - X\right|\right|_F^2 = n \cdot \sigma^2 \cdot \left|\left|\hat{X}(X/(\sqrt{n}\sigma)) + Z/\sqrt{n}) - X/(\sqrt{n}\sigma)\right|\right|_F^2 .$$

Below, we spell out this translation of AMSE where appropriate.

B. Asymptotic framework and problem statement

In this paper, we consider a sequence of increasingly larger denoising problems $Y_n = X_n + Z_n/\sqrt{n}$, with $X_n, Z_n \in M_{m_n,n}$, satisfying the following assumptions:

1) *Invariant white noise:* The entries of Z_n are i.i.d samples from a distribution with zero mean, unit variance and finite fourth moment. To simplify the formal statement of our results, we assume that this distribution is *orthogonally invariant* in the sense that Z_n follows the same distribution as AZ_nB, for every orthogonal $A \in M_{m_n,m_n}$ and $B \in M_{n,n}$. This is the case, for example, when the entries of Z_n are Gaussian. In Section VI we revisit this restriction and discuss general (not necessarily invariant) white noise.
2) *Fixed signal column span* $(x_1, \ldots, x_r)$*:* Let the rank $r > 0$ be fixed and choose a vector $\mathbf{x} \in \mathbb{R}^r$ with coordinates $\mathbf{x} = (x_1, \ldots, x_r)$ such that $x_1 \geq \ldots \geq x_r > 0$. Assume that for all n,

$$X_n = U_n \, diag(x_1, \ldots, x_r, 0, \ldots, 0) \, V_n' \tag{8}$$

is an arbitrary[1] singular value decomposition of X_n, where $U_n \in M_{m_n,m_n}$ and $V_n \in M_{n,n}$.
3) *Asymptotic aspect ratio* β*:* The sequence m_n is such that $m_n/n \to \beta$. To simplify our formulas, we assume that $0 < \beta \leq 1$.

Let $\hat{X}$ be any singular value shrinkage denoiser calibrated, as discussed above, for noise level $1/\sqrt{n}$. Define the Asymptotic MSE (AMSE) of an $\hat{X}$ at a signal $\mathbf{x}$ by the (almost sure) limit[2]

$$\mathbf{M}(\hat{X}, \mathbf{x}) \stackrel{a.s.}{=} \lim_{n\to\infty} \left|\left|\hat{X}(Y_n) - X_n\right|\right|_F^2 . \tag{9}$$

Adopting the asymptotic framework above, we seek singular value thresholding rules $\hat{X}_\lambda$ that minimize the AMSE $\mathbf{M}(\hat{X}_\lambda, \mathbf{x})$. As we will see, in this framework there are simple, satisfying answers to the questions posed in the introduction.

[1]While the signal rank r and nonzero signal singular values $x_1, \ldots, x_r$ are shared by all matrices X_n, the signal left and right singular vectors U_n and V_n are unknown and arbitrary.

[2]Our results imply that the AMSE is well-defined as a function of the signal singular values $\mathbf{x}$.

β	$\lambda_*(\beta)$	β	$\lambda_*(\beta)$
0.05	1.5066	0.55	2.0167
0.10	1.5816	0.60	2.0533
0.15	1.6466	0.65	2.0887
0.20	1.7048	0.70	2.1229
0.25	1.7580	0.75	2.1561
0.30	1.8074	0.80	2.1883
0.35	1.8537	0.85	2.2197
0.40	1.8974	0.90	2.2503
0.45	1.9389	0.95	2.2802
0.50	1.9786	1.00	2.3094

TABLE I
SOME OPTIMAL HARD THRESHOLD COEFFICIENTS $\lambda_*(\beta)$ FROM (3). FOR m-BY-n MATRIX IN KNOWN NOISE LEVEL σ (WITH $m/n = \beta$), THE OPTIMAL SVHT DENOISER $\hat{X}_{\tau_*}$ SETS TO ZERO ALL DATA SINGULAR VALUES BELOW THE THRESHOLD $\tau_* = \lambda_*(\beta)\sqrt{n}\sigma$.

III. Results

Define the *optimal hard threshold for singular values* for n-by-n square matrices by

$$\lambda_* = \frac{4}{\sqrt{3}} . \tag{10}$$

More generally, define the optimal threshold for m-by-n matrices with $m/n = \beta$ by

$$\lambda_*(\beta) \stackrel{\text{def}}{=} \sqrt{2(\beta+1) + \frac{8\beta}{(\beta+1) + \sqrt{\beta^2 + 14\beta + 1}}} . \tag{11}$$

Some values of $\lambda_*(\beta)$ are provided in Table I.

A. Optimally tuned SVHT asymptotically dominates TSVD and any SVHT

Our primary result is simply that $\hat{X}_{\lambda_*}$ always has equal or better AMSE compared to SVHT with any other choice of threshold, and compared to TSVD. In other words, from the ideal perspective of our asymptotic framework, the decision-theoretic picture is very straightforward: TSVD is asymptotically inadmissible, and so is any SVHT with $\lambda \neq \lambda_*$. We note that since AMSE of SVHT with $\lambda < 1+\sqrt{\beta}$ in our framework turns out to be infinite, here and below we need only consider SVHT with $\lambda > 1+\sqrt{\beta}$. As discussed in Section VIII, AMSE calculation in the case where the threshold λ is placed exactly at the bulk edge $1+\sqrt{\beta}$ is a little more subtle and lies beyond our current scope.

Theorem 1. The threshold λ_* is asymptotically optimal for SVHT. *Let $0 < \beta \leq 1$. For any $\lambda > 1+\sqrt{\beta}$, any $r \in \mathbb{N}$ and any $\mathbf{x} \in \mathbb{R}^r$, the AMSE* (9) *of the SVHT denoiser $\hat{X}_\lambda$ is well defined and*

$$\mathbf{M}(\hat{X}_{\lambda_*}, \mathbf{x}) \leq \mathbf{M}(\hat{X}_\lambda, \mathbf{x}) , \tag{12}$$

where $\lambda* = \lambda_*(\beta)$ is the optimal threshold (11)*. Moreover, if $\lambda \neq \lambda_*(\beta)$, strict inequality holds at least at one point $\mathbf{x}_*(\lambda) \in \mathbb{R}^r$.*

We can therefore say that $\lambda_*(\beta)$ is *asymptotically unique admissible* for SVHT. In particular, the popular practice of hard thresholding close to the bulk edge is asymptotically inadmissible. The popular Truncated SVD $\hat{X}_r$ is asymptotically inadmissible, too:

Theorem 2. Asymptotic inadmissibility of TSVD. *Let $0 < \beta \le 1$. For any $r \in \mathbb{N}$ and any $\mathbf{x} \in \mathbb{R}^r$, the AMSE of the TSVD estimator $\hat{X}_r$ is well defined, and*

$$\mathbf{M}(\hat{X}_{\lambda_*}, \mathbf{x}) \le \mathbf{M}(\hat{X}_r, \mathbf{x}) \,. \tag{13}$$

Moreover, strict inequality holds at least at one point $\mathbf{x}_*(\lambda) \in \mathbb{R}^r$.

Figure 2 shows the uniform ordering of the AMSE curves, stated in Theorems 1 and 2, for a few values of β.

To apply the optimal hard threshold to m-by-n matrices sampled from the general model $Y = X + \sigma Z$, by translating $\hat{X}_{\lambda_*}$ using Eq. (7), we find the optimal threshold

$$\tau_* = \lambda_* \cdot \sqrt{n}\sigma \,.$$

Note that Theorem 1 obviously does not imply that for any finite matrix X and $\tau \neq \tau_*$ we have $\left|\left|\hat{X}_{\tau_*}(X) - X\right|\right|_F^2 \le \left|\left|\hat{X}_{\tau}(X) - X\right|\right|_F^2$. However, empirical evidence discussed in Section VII suggests that even for relatively small matrices, e.g $n \sim 20$, the performance gain from using $\hat{X}_{\tau_*}$ is noticeable, and becomes substantial in low SNR.

B. Minimaxity over matrices of bounded rank

Theorem 1 implies that $\hat{X}_{\lambda_*}$ is asymptotically minimax among SVHT denoisers, over the class of matrices of a given low rank. Our next result explicitly characterizes the least favorable signal and the asymptotic minimax MSE.

Theorem 3. *In the asymptotic square case $\beta = 1$, the following holds.*

1) **Asymptotically Least Favorable signal for SVHT.** *Let $\lambda > 2$. Then*

$$\mathit{argmax}_{\mathbf{x}\in\mathbb{R}^r} \mathbf{M}(\hat{X}_\lambda, \mathbf{x}) = x_*(\lambda) \cdot (1, \ldots, 1) \in \mathbb{R}^r \,, \tag{14}$$

where

$$x_*(\lambda) = \frac{\lambda + \sqrt{\lambda^2 - 4}}{2} \,.$$

2) **Minimax AMSE of SVHT.** *For the AMSE of the SVHT denoiser* (2) *we have*

$$\min_{\lambda>2} \max_{\mathbf{x}\in\mathbb{R}^r} \mathbf{M}(\hat{X}_\lambda, \mathbf{x}) \;\; = \;\; 3\,r \,. \tag{15}$$

3) **Asymptotically minimax tuning of SVHT threshold.** *For the AMSE of the SVHT denoiser* (2) *we have*

$$\mathit{argmin}_{\lambda>2} \max_{\mathbf{x}\in\mathbb{R}^r} \mathbf{M}(\hat{X}_\lambda, \mathbf{x}) \;\; = \;\; \frac{4}{\sqrt{3}} \,. \tag{16}$$

In words, in our asymptotic framework, the least favorable signal for SVHT is fully degenerate. We will see in Lemma 2 below that the least favorable location for signal singular values, $x_*(\lambda)$, is such that the top r observed data singular values fall exactly on the chosen threshold λ.

Theorem 4. *For a general asymptotic aspect ratio $0 < \beta \le 1$, the following holds. Let $\lambda > 1 + \sqrt{\beta}$, then*

$$\mathit{argmax}_{\mathbf{x}\in\mathbb{R}^r} \mathbf{M}(\hat{X}_\lambda, \mathbf{x}) = x_*(\lambda) \cdot (1, \ldots, 1) \in \mathbb{R}^r \,, \tag{17}$$

where

$$x_*(\lambda) = \sqrt{\frac{\lambda^2 - \beta - 1 + \sqrt{(\lambda^2 - \beta - 1)^2 - 4\beta}}{2}} \,. \tag{18}$$

Moreover,

$$\min_{\lambda>1+\sqrt{\beta}} \max_{\mathbf{x}\in\mathbb{R}^r} \mathbf{M}(\hat{X}_\lambda, \mathbf{x}) = \frac{r}{2} \cdot \left[(\beta+1) + \sqrt{\beta^2 + 14\beta + 1}\right] \tag{19}$$

and

$$\mathit{argmin}_{\lambda>1+\sqrt{\beta}} \max_{\mathbf{x}\in\mathbb{R}^r} \mathbf{M}(\hat{X}_\lambda, \mathbf{x}) = \sqrt{2(\beta+1) + \frac{8\beta}{(\beta+1) + \sqrt{\beta^2 + 14\beta + 1}}} \,. \tag{20}$$

C. Comparison of worst-case AMSE

By Theorem 1, the AMSE of optimally tuned SVHT $\hat{X}_{\lambda_*}$ is always lower than the AMSE of other choices for the hard threshold location. One way to measure how much worse the other choices are, and to compare $\hat{X}_{\lambda_*}$ with other popular matrix denoisers, is to evaluate their worst-case AMSE.

Table II compares the guarantees provided on AMSE by shrinkage rules mentioned, for the square matrix case $m = n$ in the model $Y = X + Z/\sqrt{n}$. For the general noise $Y = X + \sigma Z$ multiply each guarantee by $n\sigma^2$.

1) TSVD: The AMSE of the TSVD $\hat{X}_r$ is calculated in Lemma 5 below. A simple calculation shows that, in the square matrix case ($\beta = 1$)

$$\max_{\mathbf{x}\in\mathbb{R}^r} \mathbf{M}(\hat{X}_r, \mathbf{x}) = 5r \,.$$

This is $5/3$ times the corresponding worst-case AMSE of $\hat{X}_{\lambda_*}$.

2) Hard Thresholding near the bulk edge: Lemma 4 provides the AMSE of the SVHT denoiser $\hat{X}_\lambda$, for any $\lambda > 1 + \sqrt{\beta}$. A simple calculation shows that

$$\max_{\mathbf{x}\in\mathbb{R}^r} \mathbf{M}(\hat{X}_{2.02}, \mathbf{x}) = 4.26r \,,$$

providing the worst-case AMSE of the Universal Singular Value Threshold (USVT) of [17]. When thresholding near the bulk edge 2, the change in worse-case AMSE for just a small increase in the threshold λ is drastic (see Figure 2). The reason for this phenomenon is discussed in section IV.

3) Soft Thresholding: Many authors have considered matrix denoising by applying the soft thresholding nonlinearity $\eta_S(y; s) = (|y| - s)_+ \cdot \mathrm{sign}(y)$, instead of hard thresholding, to the data singular values. The denoiser

$$\hat{X}_s = \sum_{i=1}^{n} \eta_S(y_i; s) \mathbf{u}_i \mathbf{v}_i'$$

is known as Singular Value Soft Thresholding (SVST) or SVT; See [29]–[31] and references therein. In our asymptotic framework, following reasoning similar to the proof of Theorem 1, one finds that the AMSE of SVST is well defined, and that the optimal (namely, asymptotically unique admissible) tuning s_* for the soft threshold is exactly at the bulk edge $1 + \sqrt{\beta}$. In the square case, the AMSE guarantee of optimally-tuned SVST $\hat{X}_{s_*}$ turns out to be $6r$. This is twice as large as that for the optimally tuned SVHT $\hat{X}_{\lambda_*}$. It is interesting to now that both

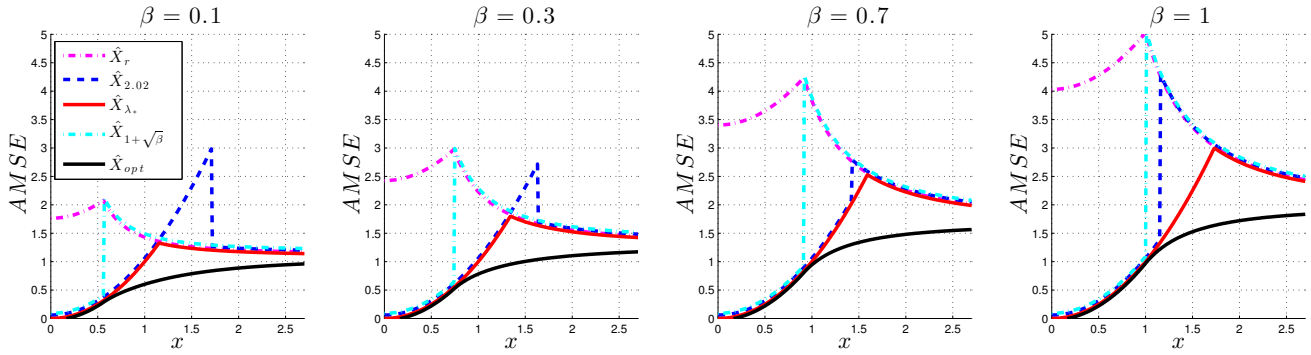

Fig. 2. AMSE against signal amplitude x for denoisers discussed: TSVD $\hat{X}_r$, universal hard threshold $\hat{X}_{2.02}$ from [17], and optimally tuned SVHT proposed here $\hat{X}_{\lambda_*}$. Also shown: the limiting AMSE of $\hat{X}_\lambda$ as $\lambda \to 1+\sqrt{\beta}$ (denoted $\hat{X}_{1+\sqrt{\beta}}$), and optimal singular value shrinkage $\hat{X}_{opt}$ from [28]. Different aspect ratios β are shown; $r = 1$ everywhere; curves jittered in vertical axis to avoid overlap.

optimal tuning for the soft threshold λ and the corresponding best-possible AMSE guarantee agree with calculations done in an altogether different asymptotic model, in which one first takes $n \to \infty$ with rank $r/n \to \rho$, and only then takes $\rho \to 0$ [31, sec. 8]. We also note that the worst-case AMSE of SVST is obtained in the limit of very high SNR, where SVHT does very well in comparison. When both are optimally tuned, SVHT does not dominate SVST across all matrices; In fact, soft thresholding does better than hard thresholding in low SNR (Figure 3). For example, in the square case, when the signal is near $\sqrt{3}$ (the least favorable location for $\hat{X}_{\lambda_*}$), the AMSE of $\hat{X}_{s_*}$ is $(7 - 8/\sqrt{3})r \approx 2.38r$, compared to $3r$, the worse-case AMSE of $\hat{X}_{\lambda_*}$.

4) Optimal Singular Value Shrinker: Our focus in this paper is denoising by singular value hard thresholding (SVHT), where $\hat{X}_\lambda$ acts applying a hard thresholding nonlinearity to each of the data singular values. As mentioned in the introduction, one may ask how SVHT compares to other singular value shrinkage denoisers, which use a different nonlinearity that may be more suitable to the problem at hand. In a special case of our asymptotic framework, Perry [13] and Shabalin and Nobel [22] have derived an optimal singular value shrinker $\hat{X}_{opt}$. Proceeding along this line, in [28] we explore optimal shrinkage of singular values under various loss functions and develop a simple expression for the optimal shrinkers. Calibrated for the model $X + Z/\sqrt{n}$, in the square setting $m = n$, this shrinker takes the form

$$\hat{X}_{opt} : \sum_{i=1}^{n} y_i \mathbf{u}_i \mathbf{v}_i' \mapsto \sum_{i=1}^{n} \eta_{opt}(y_i) \mathbf{u}_i \mathbf{v}_i'$$

where

$$\eta_{opt}(x) = \sqrt{(x^2 - 4)_+}\,.$$

In our asymptotic framework, this rule dominates in AMSE essentially *any other* estimator based on singular value shrinkage, at any configuration of the non-zero signal singular values $\mathbf{x}$. The AMSE of the optimal shrinker (in the square matrix case) at $\mathbf{x} \in \mathbb{R}^r$ is [28]

$$\mathbf{M}(\hat{X}_{opt}, \mathbf{x}) = \sum_{i=1}^{r} \begin{cases} 2 - \frac{1}{x_i^2} & x_i \geq 1 \\ x_i^2 & 0 \leq x_i \leq 1 \end{cases}. \qquad (21)$$

(See Figure 2.) It follows that the worst-case AMSE of $\hat{X}_{opt}$ is

$$\max_{\mathbf{x} \in \mathbb{R}^r} \mathbf{M}(\hat{X}_{opt}, \mathbf{x}) = 2r$$

in the square case. We conclude that, for square matrices, in worst-case AMSE, singular value hard thresholding at the optimal location is 50% worse than the best possible singular value shrinker, Truncated SVD or SVHT just above the bulk-edge (which roughly equals the widely used Scree-plot elbow truncation) is 250% worse, and singular value soft thresholding is 300% worse.

D. Minimaxity over matrices of bounded nuclear norm

So far we have considered minimaxity over the class of matrices of at most rank r, where r is given. In [17], the author considered minimax estimation over a different class of matrices, namely nuclear norm balls. For a given constant ξ, this is the class of all matrices for which the nuclear norm is at most ξ. Recall that the nuclear norm of a matrix $X \in M_{m \times n}$, whose vector of singular values is $\mathbf{x} \in \mathbb{R}^m$, is given by $||\mathbf{x}||_1$. Our next result shows that $\hat{X}_{\lambda_*}$ is minimax optimal over this class as well. Specifically, it is the minimax estimator, in AMSE, among all SVHT rules, over a given Nuclear Norm ball. We note that unlike Theorems 3 and 4, this result does not follow directly from Theorem 1. We restrict our discussion to square matrices ($\beta = 1$); the general nonsquare case is handled similarly.

Theorem 5. *Let $\lambda > 2$ and let $\xi = r \cdot (\lambda + \sqrt{\lambda^2 - 4})/2$ for some $r \in \mathbb{N}$.*

1) *The least favorable singular value configuration obeys*

$$argmax_{||\mathbf{x}||_1 \leq \xi} \mathbf{M}(\hat{X}_\lambda, \mathbf{x}) = x_*(\lambda) \cdot (1, \ldots, 1) \in \mathbb{R}^r\,, \qquad (22)$$

where

$$x_*(\lambda) = \frac{\lambda + \sqrt{\lambda^2 - 4}}{2}\,.$$

2) *The best achievable inequality between nuclear norm ξ and AMSE of a hard threshold rule is:*

$$\min_{\lambda > 2} \max_{||\mathbf{x}||_1 \leq \xi} \mathbf{M}(\hat{X}_\lambda, \mathbf{x}) \;=\; \sqrt{3} \cdot \xi\,. \qquad (23)$$

Shrinker	Standing notation	Guarantee on AMSE
Optimal singular value shrinker	$\hat{X}_{opt}$	$2r$
Optimally tuned SVHT	$\hat{X}_{\lambda_*}$	$3r$
Universal Singular Value Threshold [17]	$\hat{X}_{2.02}$	$\approx 4.26r$
TSVD	$\hat{X}_r$	$5r$
Optimally tuned SVST	$\hat{X}_{s_*}$	$6r$

TABLE II

A COMPARISON OF GUARANTEES ON AMSE PROVIDED BY SINGULAR VALUE SHRINKAGE RULES DISCUSSED, FOR THE SQUARE MATRIX CASE $m = n$ IN THE MODEL $Y = X + Z/\sqrt{n}$. FOR THE GENERAL MODEL $Y = X + \sigma Z$ MULTIPLY EACH GUARANTEE BY $n\sigma^2$.

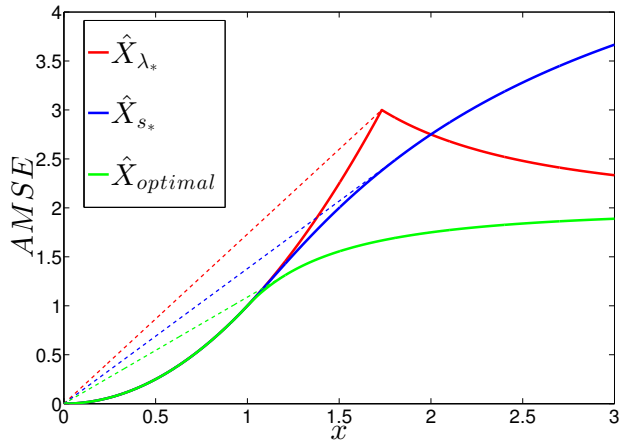

Fig. 3. AMSE of optimally tuned SVHT (red), optimally tuned SVST (blue) and the optimal singular value shrinker [28] (green), for square case ($\beta = 1$) and $r = 1$. The best available constant from Table III is the slope of the convex envelope (dashed) of the AMSE curve (solid), namely, the slope of the secant running from the origin to the inflection point of the AMSE curve. Although the maximum (worst-case AMSE) of optimally tuned SVST $\hat{X}_{s_*}$ is higher than that of optimally tuned SVHT $\hat{X}_{\lambda_*}$, the slope of its convex envelope is lower.

3) *The threshold achieving this inequality is*

$$argmin_{\lambda>2} \max_{||\mathbf{x}||_1 \le \xi} \mathbf{M}(\hat{X}_\lambda, \mathbf{x}) \quad = \quad \frac{4}{\sqrt{3}}\,. \qquad (24)$$

As an alternative to comparing denoisers by comparing their guarantees on AMSE over a prescribed rank r, one can compare denoisers based on the best available constant C in the inequality

$$\min_{\lambda>2} \max_{||\mathbf{x}||_1 \le \xi} \mathbf{M}(\hat{X}_\lambda, \mathbf{x}) \quad = \quad C \cdot \xi\,. \qquad (25)$$

The results in the square matrix case are summarized in Table III. Each constant is derived from the AMSE formula for the respective denoiser, as cited above. To understand why the best available constant for optimally tuned SVST is smaller than than of optimally tuned SVHT, consider Figure 3.

E. When the noise level σ is unknown

When the noise level in which Y is observed is unknown, it no longer makes sense to use $\hat{X}_{\lambda_*}$, which is calibrated for a specific noise level. We now describe a method to estimate the optimal hard threshold from the data matrix Y. To emphasize that the resulting denoiser is ready for use on data from the general model $Y = X + \sigma Z$, we denote this estimated threshold by $\hat{\tau}_*$, and the SVHT denoiser by $\hat{X}_{\hat{\tau}_*}$. To this end, we are required to estimate the unknown noise level σ. In the closely related Spiked Covariance Model, there are existing methods for estimation of an unknown noise level; see for example [32] and references therein.

Consider the following robust estimator for the parameter σ in the model $Y = X + \sigma Z$:

$$\hat{\sigma}(Y) \stackrel{\text{def}}{=} \frac{y_{med}}{\sqrt{n \cdot \mu_\beta}}\,, \qquad (26)$$

where y_{med} is a median singular value of Y and μ_β is the median of the the Marčenko-Pastur distribution, namely, the unique solution in $\beta_- \le x \le \beta_+$ to the equation

$$\int_{\beta_-}^{x} \frac{\sqrt{(\beta_+ - t)(t - \beta_-)}}{2\pi t} dt = \frac{1}{2}\,,$$

where $\beta_\pm = (1 \pm \sqrt{\beta})^2$. Define the optimal hard threshold for a data matrix $Y \in M_{m\times n}$ observed in unknown noise level, with $m/n = \beta$, by plugging in $\hat{\sigma}(Y)$ instead of σ in Eq. (3):

$$\hat{\tau}_*(\beta, Y) \stackrel{\text{def}}{=} \lambda_*(\beta) \cdot \sqrt{n}\,\hat{\sigma}(Y) = \frac{\lambda_*(\beta)}{\sqrt{\mu_\beta}} y_{med}\,.$$

Writing $\omega(\beta) = \lambda_*(\beta)/\sqrt{\mu_\beta}$, the threshold is

$$\hat{\tau}_*(\beta, Y) = \omega(\beta) \cdot y_{med}\,.$$

The median μ_β and hence the coefficient $\omega(\beta)$ are not available analytically; in [1] we make available a Matlab script to evaluate the coefficient $\omega(\beta)$. Some values are tabulated in Table IV for convenience. A useful approximation to ω is given as a cubic polynomial in Eq. (5) above. Empirically,

$$\max_{0.001<\beta\le 1} |\omega(\beta) - \left(0.56\beta^3 - 0.95\beta^2 + 1.82\beta + 1.43\right)| \le 0.02$$

which may be sufficient for some practical purposes if one does not have access to a more exact value of ω.

Lemma 1. For the sequence Y_n in our asymptotic framework,

$$\lim_{n\to\infty} \frac{\hat{\sigma}(Y_n)}{1/\sqrt{n}} \stackrel{a.s.}{=} 1.$$

Correlary 1. For Y_n as above and any $0 < \beta \le 1$,

$$\lim_{n\to\infty} \hat{\tau}_*(\beta, Y_n) \stackrel{a.s.}{=} \lambda_*(\beta) \cdot \lim_{n\to\infty} \sqrt{n} \cdot \hat{\sigma}(Y_n) \stackrel{a.s.}{=} \lambda_*(\beta)\,,$$

Correlary 2. For Y_n as above, any $0 < \beta \le 1$, any r and any $\mathbf{x} \in \mathbb{R}^r$, almost surely

$$\mathbf{M}(\hat{X}_{\hat{\tau}_*}, \mathbf{x}) = \mathbf{M}(\hat{X}_{\lambda_*}, \mathbf{x})\,.$$

Correlary 3. Theorem 1, Theorem 2, Theorem 3, Theorem 4 and Theorem 5 all hold if we replace the optimally-tuned

TABLE III

A COMPARISON OF BEST AVAILABLE CONSTANT IN MINIMAX AMSE, OVER NUCLEAR NORM BALLS, FOR THE SHRINKAGE RULES DISCUSSED, FOR THE SQUARE MATRIX CASE $m = n$ IN THE MODEL $Y = X + Z/\sqrt{n}$. THESE CONSTANTS ARE THE SAME FOR THE GENERAL MODEL $Y = X + \sigma Z$.

Shrinker	Standing notation	Best possible constant C in Eq. (25)
Optimal singular value shrinker	$\hat{X}_{opt}$	1
Optimally tuned SVHT	$\hat{X}_{\lambda_*}$	$\sqrt{3} \approx 1.73$
USVT of [17]	$\hat{X}_{2.02}$	≈ 3.70
TSVD	$\hat{X}_r$	5
Optimally tuned SVST	$\hat{X}_{s_*}$	≈ 1.38

β	$\omega(\beta)$	β	$\omega(\beta)$
0.05	1.5194	0.55	2.2365
0.10	1.6089	0.60	2.3021
0.15	1.6896	0.65	2.3679
0.20	1.7650	0.70	2.4339
0.25	1.8371	0.75	2.5011
0.30	1.9061	0.80	2.5697
0.35	1.9741	0.85	2.6399
0.40	2.0403	0.90	2.7099
0.45	2.106	0.95	2.7832
0.50	2.1711	1.00	2.8582

TABLE IV

SOME VALUES OF THE OPTIMAL HARD THRESHOLD COEFFICIENT FOR UNKNOWN NOISE LEVEL, $\omega(\beta)$ OF EQ. (4) FOR AN m-BY-n MATRIX IN UNKNOWN NOISE LEVEL (WITH $m/n = \beta$), THE OPTIMAL SVHT DENOISER $\hat{X}_{\hat{\tau}_*}$ SETS TO ZERO ALL DATA SINGULAR VALUES BELOW THE THRESHOLD $\tau_* = \omega(\beta) y_{med}$, WHERE y_{med} IS THE MEDIAN SINGULAR VALUE OF THE DATA MATRIX Y. CALCULATED USING FUNCTION PROVIDED IN THE CODE SUPPLEMENT [1].

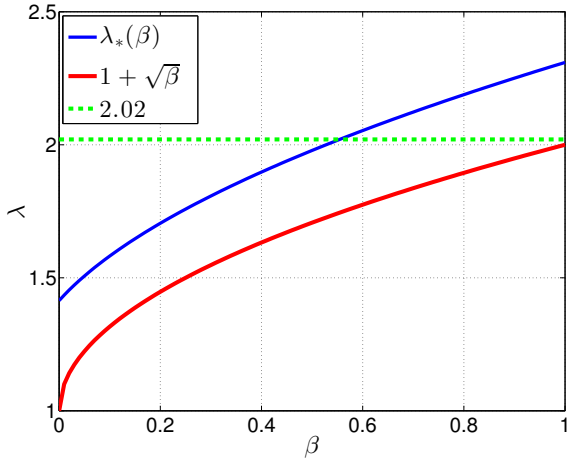

Fig. 4. The optimal threshold $\lambda_*(\beta)$ from (11) against β. Also shown are bulk-edge $1 + \sqrt{\beta}$, which is the hard threshold corresponding to TSVD in our setting, and the USVT threshold 2.02 from [17].

SVHT for known σ, $\hat{X}_{\lambda_*(\beta)}$, by optimally-tuned SVHT for unknown σ, $\hat{X}_{\hat{\tau}_*}$.

IV. DISCUSSION

A. The optimal threshold $\lambda_*(\beta)$ and the bulk edge $1 + \sqrt{\beta}$

Figure 4 shows the optimal threshold $\lambda_*(\beta)$ over β. The edge of the quarter circle bulk $1 + \sqrt{\beta}$, the hard threshold that best emulates TSVD in our setting, is shown for comparison. In the null case $X = 0$, the largest data singular value is located asymptotically exactly at the bulk edge, $1 + \sqrt{\beta}$. It might seem that just above the bulk edge is a natural place to set a threshold, since anything smaller could be the product of a pure noise situation. However, for $\beta > 0.2$, the optimal hard threshold $\lambda^*(\beta)$ is 15-20% larger than the bulk edge; as $\beta \to 0$, it grows about 40% larger. Inspecting the proof of Theorem 1 and particularly the expression for AMSE of SVHT (Lemma 4), one finds the reason: one component of the AMSE is due to the angle between the signal singular vectors and the data singular vectors. This angle converges to a nonzero value as $n \to \infty$ (given explicitly in Lemma 3) which grows as SNR decreases. When some data singular value y_i is too close to the bulk, its corresponding singular vectors are too badly rotated, and the rank-one matrix $y_i \mathbf{u}_i \mathbf{v}_i'$ it contributes to the denoiser hurts the AMSE more than it helps. For example, for square matrices $\beta = 1$, this situation is most acute when the signal singular value is just barely larger than $x_i = 1$, causing the corresponding data singular value y_i to be just barely larger than the bulk edge, which for square matrices is located at 2. A SVHT denoiser thresholding just above the bulk edge would include the component $y_i \mathbf{u}_i \mathbf{v}_i'$, incurring an AMSE about 5 times larger than the AMSE incurred by excluding y_i from the reconstruction. The optimal threshold $\lambda_*(\beta)$ keeps such singular values out of the picture; this is why it is necessarily larger than the bulk edge. The precise value of $\lambda_*(\beta)$ is the precise point at which it becomes advantageous to include the rank-one contribution of a singular value y_i in the reconstruction.

B. The optimal threshold $\lambda_*(\beta)$ relative to the USVT $\hat{X}_{2.02}$

As mentioned in the introduction, S. Chatterjee has recently discussed SVHT in a broad class of situations [17]. Translating his much broader discussion to the confines of the present context, he observed that *any* $\lambda > 2$ can serve as a *universal hard threshold for singular values* (USVT), offering fairly good performance regardless of the matrix shape m/n and the underlying signal matrix X. The author makes the specific recommendation $\lambda = 2.02$ and writes:

> *"The algorithm manages to cut off the singular values at the 'correct' level, depending on the structure of the unknown parameter matrix. The adaptiveness of the USVT threshold is somewhat similar in spirit to that of the SureShrink algorithm of Donoho and Johnstone."*

Keeping in mind that the scope of [17] is much broader than the one considered here, we would like to evaluate this proposal, in the setting of low rank matrix in white noise, and specifically in our asymptotic framework. Figure

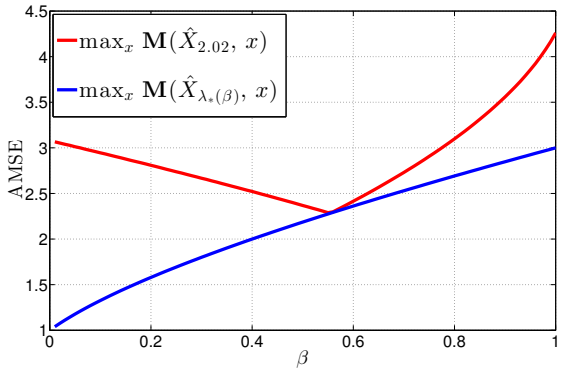

Fig. 5. Worst-case AMSE against the shape parameter β, for two choices of hard threshold: $\lambda = 2.02$ from [17], and optimal threshold λ_* from (11).

4 includes the value 2.02: indeed, this threshold is larger than the bulk edge, for any $0 < \beta \leq 1$, so Chatterjee's $\hat{X}_{2.02}$ rule asymptotically set to zero all singular values which *could* arise due to an underlying noise-only situation. When $\lambda_*(\beta) < 2.02$, the $\hat{X}_{2.02}$ rule sometimes "kills" singular values that the optimal threshold deems good enough for keeping, and when $\lambda_*(\beta) > 2.02$, the $\hat{X}_{2.02}$ rule sometimes "keeps" singular values that did in fact arise from signal, but are so close to the bulk that the optimal threshold declares them unusable.

For $\beta = 1$, the guarantee on worst-case AMSE obtained by using $\lambda = 2.02$ over matrices of rank r is about $4.26r$, roughly 140% larger than the guarantee obtained by using the minimax threshold $\lambda = 4/\sqrt{3}$ (See Figure 2). For square matrices, the regret for preferring USVT to optimally-tuned SVHT can be substantial: in low SNR ($x \approx 1$), using the threshold $\lambda = 2.02$ incurs roughly twice the AMSE of the minimax threshold $4/\sqrt{3}$.

We note that unlike the optimally tuned SVHT $\hat{X}_{\lambda_*}$, the USVT $\hat{X}_{2.02}$ does not take into account the shape factor β, namely the ratio of number of rows to number of columns of the matrix in question. A comparison of worst-case AMSE between the fixed threshold choice $\lambda = 2.02$ and the optimal hard threshold $\lambda = \lambda_*(\beta)$ is shown in Figure 5. The two curves intersect at $\lambda \approx 0.55$, where the optimal threshold (11) is approximately 2.02.

One might argue that [17] proposed 2.02 based on its MSE performance over classes of matrices bounded in nuclear norm. But also for that purpose, 2.02 is noticeably outperformed by $\lambda^*(\beta)$. Arguing as in Theorem 5 we obtain, in the square case:

$$\max_{||\mathbf{x}||_1 \leq \xi} \mathbf{M}(\hat{X}_{2.02}, \mathbf{x}) \quad \approx \quad 3.70 \cdot \xi \,. \tag{27}$$

The coefficient 3.70 is about 110% larger than the best coefficient achievable by SVHT, namely $C = \sqrt{3}$ in (25).

One should keep in mind that USVT is applicable for a wide range of noise models, e.g. in stochastic block models. [17] is the first, to the best of out knowledge, to suggest that a matrix denoising procedure as simple as SVHT could have universal optimality properties. In our asymptotic framework of low-rank matrices in white noise, the 2.02 threshold performs fairly well in AMSE, except for very small values of β (Figure 2); but one often gets a substantial AMSE improvement by switching to the rule we recommend. Since our recommendation dominates in AMSE, there is no downside to making this switch – i.e. there is no configuration of signal singular values x which could make one regret this switch.

V. Proofs

Setting additional notation required in the proofs, let

$$X_n = \sum_{i=1}^{r} x_i \, \mathbf{a}_{n,i} \, \mathbf{b}'_{n,i}$$

be a sequence of signal matrices in our asymptotic framework, so that $\mathbf{a}_{n,i} \in \mathbb{R}^{m_n}$ (resp. $\mathbf{b}_{n,i} \in \mathbb{R}^n$) is the left (resp. right) singular vector corresponding to the singular value x_i, namely, i-th column of U_n (resp. V_n) in (39). Similarly, let Y_n be a corresponding sequence of observed matrices in our framework, and write

$$Y_n = \sum_{i=1}^{m_n} y_{n,i} \, \mathbf{u}_{n,i} \, \mathbf{v}'_{n,i}$$

so that $\mathbf{u}_{n,i} \in \mathbb{R}^m$ (resp. $\mathbf{v}_{n,i} \in \mathbb{R}^n$) is the left (resp. right) singular vector corresponding to the singular value $y_{n,i}$. (Note that $\{\mathbf{a}_{n,i}\}$ and $\{\mathbf{b}_{n,i}\}$ are unknown, arbitrary, non-random vectors.)

Our main results depend on Lemma 4, a formula for the AMSE of SVHT. This formula in turn depends on Lemma 2 and Lemma 3. Both follow from recent key results due to [25].

Lemma 2. Asymptotic data singular values. For $1 \leq i \leq r$,

$$\lim_{n\to\infty} y_{n,i} \overset{a.s.}{=} \begin{cases} \sqrt{\left(x_i + \frac{1}{x_i}\right)\left(x_i + \frac{\beta}{x_i}\right)} & x_i > \beta^{1/4} \\ 1 + \sqrt{\beta} & x_i \leq \beta^{1/4} \end{cases} \tag{28}$$

Lemma 3. Asymptotic angle between signal and data singular vectors. Let $1 \leq i \neq j \leq r$ and assume that x_i has degeneracy d, namely, there are exactly d entries of $\mathbf{x}$ equal to x_i. If $x_i > \beta^{1/4}$, we have

$$d \cdot \lim_{n\to\infty} \left|\langle \mathbf{a}_{n,i} \, , \, \mathbf{u}_{n,j} \rangle\right|^2 \overset{a.s.}{=} \begin{cases} \frac{x_i^4 - \beta}{x_i^4 + \beta x_i^2} & x_i = x_j \\ 0 & x_i \neq x_j \end{cases} , \tag{29}$$

and, a slightly different formula,

$$d \cdot \lim_{n\to\infty} \left|\langle \mathbf{b}_{n,i} \, , \, \mathbf{v}_{n,j} \rangle\right|^2 \overset{a.s.}{=} \begin{cases} \frac{x_i^4 - \beta}{x_i^4 + x_i^2} & x_i = x_j \\ 0 & x_i \neq x_j \end{cases} . \tag{30}$$

If however $x_i \leq \beta^{1/4}$, then we have

$$\lim_{n\to\infty} \left|\langle \mathbf{a}_{n,i} \, , \, \mathbf{u}_{n,j} \rangle\right| \overset{a.s.}{=} \lim_{n\to\infty} \left|\langle \mathbf{b}_{n,i} \, , \, \mathbf{v}_{n,j} \rangle\right| \overset{a.s.}{=} 0 \,.$$

To appeal to these results, we need to show that our asymptotic framework satisfies the assumptions of [25]. By [21] the limiting law of the singular values of $Z_n/\sqrt{n}$ is the quarter-circle density

$$f(x) = \frac{\sqrt{4\beta - (x^2 - 1 - \beta)^2}}{\pi \beta x} \mathbf{1}_{[1-\sqrt{\beta}, 1+\sqrt{\beta}]}(x) \,; \tag{31}$$

by [26], $y_{n,1} \xrightarrow{a.s.} 1+\sqrt{\beta}$; by [27], $y_{n,m_n} \xrightarrow{a.s.} 1-\sqrt{\beta}$. This satisfies assumptions 2.1, 2.2 and 2.3 of [25], respectively. Formulas (28), (29) and (30), as seen in [25, example 3.1], depend only on the shape of the limiting distribution (31) and not on any Gaussian assumptions.

Using Lemma 2 and Lemma 3, we can calculate the AMSE (9) of the hard thresholding estimator $\hat{X}_\lambda$, for given threshold λ, at a matrix of specific aspect ratio β and signal singular values $\mathbf{x}$:

Lemma 4. AMSE of singular value hard thresholding. Fix $r > 0$ and $\mathbf{x} \in \mathbb{R}^r$. Let $\{X_n(\mathbf{x})\}_{n=1}^\infty$ and $\{Z_n\}_{n=1}^\infty$ be matrix sequences in our asymptotic framework, and let $\lambda \geq 1+\sqrt{\beta}$. Then

$$\mathbf{M}(\hat{X}_\lambda, \mathbf{x}) = \sum_{i=1}^r M(\hat{X}_\lambda, x_i) \tag{32}$$

where

$$M(\hat{X}_\lambda, x) = \begin{cases} (x+\frac{1}{x})(x+\frac{\beta}{x}) - (x^2 - \frac{2\beta}{x^2}) & x \geq x_*(\lambda) \\ x^2 & x < x_*(\lambda) \end{cases} \tag{33}$$

and $x_*(\lambda)$ is given by Eq. (18).

Figure 2 shows the AMSE of Lemma 4, in square case $\beta = 1$ and nonsquare cases $\beta = 0.1$, $\beta = 0.3$ and $\beta = 0.7$.

Proof: By definition,

$$\hat{X}_\lambda(Y_n) = \sum_{i=1}^{m_n} \eta_H(y_{n,i};\lambda)\, \mathbf{u}_{n,i}\, \mathbf{v}'_{n,i}\,,$$

where $\eta_H(y,\tau) = y\, \mathbf{1}_{\{y \geq \tau\}}$. Observe that

$$\begin{aligned} &\left|\left|\hat{X}_\lambda(Y_n) - X_n\right|\right|_F^2 = \\ &\quad \langle \hat{X}_\lambda(Y_n) - X_n\,,\, \hat{X}_\lambda(Y_n) - X_n\rangle = \\ &\quad \langle \hat{X}_\lambda(Y_n)\,,\, \hat{X}_\lambda(Y_n)\rangle + \\ &\qquad \langle X_n\,,\, X_n\rangle - 2\langle \hat{X}_\lambda(Y_n)\,,\, X_n\rangle = \\ &\quad \sum_{i=1}^{m_n} \eta_H(y_{n,i};\lambda)^2 + \sum_{i=1}^r x_i^2 - \\ &\qquad 2\sum_{i,j=1}^r x_i\, \eta_H(y_{n,j};\lambda)\, \langle \mathbf{a}_i \mathbf{b}'_i\,,\, \mathbf{u}_{n,j}\mathbf{v}'_{n,j}\rangle = \\ &\quad \sum_{i=r+1}^{m_n} \eta_H(y_{n,i};\lambda)^2 + \sum_{i=1}^r \left[\eta_H(y_{n,i};\lambda)^2 + x_i^2 - \right. \\ &\qquad \left. 2x_i \sum_{j=1}^r \eta_H(y_{n,j};\lambda)\, \langle \mathbf{a}_i \mathbf{b}'_i\,,\, \mathbf{u}_{n,j}\mathbf{v}'_{n,j}\rangle \right]. \end{aligned} \tag{34}$$

Since $y_{n,r+1} \xrightarrow{a.s.} 1+\sqrt{\beta} < \lambda$, the leftmost term above converges almost surely to zero. When $0 \leq x_i \leq \beta^{1/4}$, by Lemma 2 and Lemma 3, only the term x_i survives and Eq. (33) holds. Assume now that $x_i > \beta^{1/4}$. We now consider the a.s. limiting value of each of the remaining terms in (34). For the term $\sum_{i=1}^r \eta_H(y_{n,i};\lambda)^2$, by Lemma 2, for $i = 1,\ldots,r$ we have

$$\lim_{n\to\infty} \eta_H(y_{n,i};\lambda)^2 \overset{a.s.}{=} \begin{cases} (x_i+\frac{1}{x_i})(x_i+\frac{\beta}{x_i}) & (x_i+\frac{1}{x_i})(x_i+\frac{\beta}{x_i}) \geq \lambda^2 \\ 0 & (x_i+\frac{1}{x_i})(x_i+\frac{\beta}{x_i}) < \lambda^2 \end{cases}.$$

Turning to the rightmost term of (34), by Lemma 3, for $i,j = 1,\ldots,r$ we find that it equals

$$\lim_{n\to\infty} \langle \mathbf{a}_i\,,\, \mathbf{u}_{n,j}\rangle \langle \mathbf{b}_i\,,\, \mathbf{v}'_{n,j}\rangle \overset{a.s.}{=} \begin{cases} \frac{1}{d_i} \frac{x_i^4 - \beta}{x_i^3 \sqrt{(x_i+\beta/x_i)(x_i+1/x_i)}} & x_i = x_j \\ 0 & x_i \neq x_j \end{cases} \tag{35}$$

where $d_i = \#\{j \,|\, x_j = x_i\}$. Furthermore, since for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^m$ and $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^n$ we have $\langle \mathbf{x}_1\mathbf{y}'_1\,,\, \mathbf{x}_2\mathbf{y}'_2\rangle = \langle \mathbf{x}_1\,,\, \mathbf{x}_2\rangle\langle \mathbf{y}_1\,,\, \mathbf{y}_2\rangle$, we find that for $i = 1,\ldots,r$,

$$\sum_{j=1}^r \eta(y_{n,j};\lambda)\, \langle \mathbf{a}_i\mathbf{b}'_i\,,\, \mathbf{u}_{n,j}\mathbf{v}'_{n,j}\rangle = \sum_{j=1}^r \eta(y_{n,j};\lambda)\, \langle \mathbf{a}_i\,,\, \mathbf{u}_{n,j}\rangle \langle \mathbf{b}_i\,,\, \mathbf{v}'_{n,j}\rangle\,.$$

For the rightmost term of (34) we conclude that

$$\begin{aligned} &\lim_{n\to\infty} x_i \sum_{j=1}^r \eta(y_{n,j};\lambda)\, \langle \mathbf{a}_{i i}'\,,\, \mathbf{u}_{n,j}\mathbf{v}'_{n,j}\rangle \overset{a.s.}{=} \\ &\quad \sum_{1\leq j\leq r\,:\, x_j = x_i} \lim_{n\to\infty} \cdot \frac{\eta(y_{n,j};\lambda)\cdot(x_i^4-\beta)}{d_i x_i^2 \sqrt{(x_i+\beta/x_i)(x_i+1/x_i)}} = \\ &\quad \begin{cases} \frac{x_i^4-\beta}{x_i^2} & (x_i+\frac{1}{x_i})(x_i+\frac{\beta}{x_i}) \geq \lambda^2 \\ 0 & (x_i+\frac{1}{x_i})(x_i+\frac{\beta}{x_i}) < \lambda^2 \end{cases}, \end{aligned}$$

where we have used Lemma 2 again. Collecting the terms, we find for the limiting value of (34) that

$$\lim_{n\to\infty} \left|\left|\hat{X}_\lambda(Y_n) - X_n\right|\right|_F^2 \overset{a.s.}{=} \sum_{i=1}^r M(\hat{X}_\lambda, x_i)\,, \tag{36}$$

where $M(\hat{X}_\lambda, x)$ is given by (33) as required. ∎

For the TSVD, the same argument gives:

Lemma 5. AMSE of TSVD. Fix $r > 0$ and $\mathbf{x} \in \mathbb{R}^r$. Let $\{X_n(\mathbf{x})\}_{n=1}^\infty$ and $\{Z_n\}_{n=1}^\infty$ be matrix sequences in our asymptotic framework, and let $\lambda \geq 1+\sqrt{\beta}$. Then

$$\mathbf{M}(\hat{X}_r, \mathbf{x}) = \sum_{i=1}^r M(\hat{X}_r, x_i) \tag{37}$$

where

$$M(\hat{X}_r, x) = \begin{cases} (x+\frac{1}{x})(x+\frac{\beta}{x}) - (x^2 - \frac{2\beta}{x^2}) & x \geq \beta^{1/4} \\ (1+\sqrt{\beta})^2 + x^2 & x \leq \beta^{1/4} \end{cases}. \tag{38}$$

We now to turn to prove our main results.

Proof of Theorem 1: Let $x_* = x_*(\lambda_*(\beta))$ where $\lambda_*(\beta)$ is defined in (11) and $x_*(\lambda)$ is defined in (18). Then

$$x_*^2 = \left(x_* + \frac{1}{x_*}\right)\left(x_* + \frac{\beta}{x_*}\right) - \left(x_*^2 - \frac{2\beta}{x_*^2}\right).$$

It follows that for all $x > 0$ and $\lambda \geq 1 + \sqrt{\beta}$,

$$\begin{aligned} &M(\hat{X}_{\lambda_*}, x) \\ &\quad = \min\left\{x^2\,,\, \left(x + \frac{1}{x}\right)\left(x + \frac{\beta}{x}\right) - \left(x^2 - \frac{2\beta}{x^2}\right)\right\} \\ &\quad \leq M(\hat{X}_\lambda, x) \end{aligned}$$

and the theorem follows from Eq. (32). □

Figure 6 provides a visual explanation of this proof for the square ($\beta = 1$) case.

Proof of Theorem 2: For $x < \beta^{1/4}$, by Lemma 4 and Lemma 5 we have

$$M(\hat{X}_{\lambda_*}, x) = x^2 \leq x^2 + (1 + \sqrt{\beta})^2 = M(\hat{X}_r, x)\,.$$

For $x \geq \beta^{1/4}$, by Lemma 5 and Theorem 1 we have

$$M(\hat{X}_{\lambda_*}, x) \leq M(\hat{X}_r, x)\,.$$

□

Proof of Theorems 3 and 4: Theorem 3 is a special case of Theorem 4. By (32), it is enough to consider the univariate function $x \mapsto M(\hat{X}_\lambda, x)$ defined in (33). The theorem follows from Lemma 4 using the following simple observation.

Let $0 < \beta \leq 1$ and $\lambda > 1 + \sqrt{\beta}$. Denote by $x_*(\lambda)$ the unique positive solution to the equation $(x + 1/x)(x + \beta/x) = \lambda^2$. Let λ_* be the unique solution to the equation in λ

$$x_*^4(\lambda) - (\beta + 1)x_*^2 - 3\beta = 0\,.$$

Then for the function $M(\hat{X}_\lambda, x)$ defined in (33), we have

$$\begin{aligned} \text{argmax}_{x>0} M(\hat{X}_\lambda, x) &= x_*(\lambda) \\ \text{argmin}_{\lambda > 1+\sqrt{\beta}} \max_{x>0} M(\hat{X}_\lambda, x) &= \lambda_* \\ \text{min}_{\lambda > 1+\sqrt{\beta}} \max_{x>0} M(\hat{X}_\lambda, x) &= x_*(\lambda_*)^2\,. \end{aligned}$$

□

Note that the least favorable situation occurs when $x_1 = \ldots = x_r = x_*(\lambda)$, and that $x_*(\lambda)$ is precisely the value of x for which the corresponding limiting data singular value satisfies $y_{n,i} \xrightarrow{a.s.} \lambda$. In other words, the least favorable situation occurs when the data singular values all coincide with each other and with the chosen hard threshold.

Proof of Lemma 1: Let F_n denote the empirical cumulative distribution function (CDF) of the squared singular values of Y_n. Write $y_{med,n}^2 = Median(F_n)$ where $Median(\cdot)$ is a functional which takes as argument the CDF and delivers the median of that CDF. Under our asymptotic framework, almost surely, F_n converges weakly to a limiting distribution, F_{MP}, the CDF of the Marčenko-Pastur distribution with shape parameter β [21]. This distribution has a positive density throughout its support, in particular at its median. The median functional is continuous for weak convergence at F_0, and hence, almost surely,

$$y_{med,n}^2 = Median(F_n) \xrightarrow{a.s.} Median(F_0) = \mu_\beta, \qquad n \to \infty\,.$$

It follows that,

$$\lim_{n\to\infty} \frac{\hat{\sigma}(Y_n)}{1/\sqrt{n}} \overset{a.s.}{=} \lim_{n\to\infty} \frac{y_{med,n}}{\sqrt{\mu_\beta}} \overset{a.s.}{=} 1\,.$$

□

VI. General white noise

Our results were formally stated for a sequence of models of the form $Y = X + \sigma Z$, where X is a non-random matrix to be estimated, and the entries of Z are i.i.d samples from a distribution that is orthogonally invariant (in the sense that the matrix Z follows the same distribution as AZB, for any orthogonal $A \in M_{m,m}$ and $B \in M_{n,n}$). While Gaussian noise is orthogonally invariant, many common distributions, which one could consider to model white observation noise, are not.

One attractive feature of the discussion on optimal choice of singular value hard threshold, presented above, is that the AMSE $\mathbf{M}(\hat{X}, \mathbf{x})$ only depends on the signal matrix X through its rank, or more specifically, through its nonzero singular values $\mathbf{x}$. If the distribution of Z is not orthogonally invariant, MSE (or AMSE) losses this property and depends on properties of X other than its rank. This point is discussed extensively in [22].

In general white noise, which is not necessarily orthogonally invariant, one can still allow MSE to depend on X only through its singular values by placing a prior distribution on X and shifting to a model where it is a random, instead of a fixed, matrix. Specifically, consider an alternative asymptotic framework to the one in Section II-B, in which the sequence denoising problems $Y_n = X_n + Z_n/\sqrt{n}$ satisfies the following assumptions:

1) *General white noise:* The entries of Z_n are i.i.d samples from a distribution with zero mean, unit variance and finite fourth moment.
2) *Fixed signal column span and uniformly distributed signal singular vectors:* Let the rank $r > 0$ be fixed and choose a vector $\mathbf{x} \in \mathbb{R}^r$ with coordinates $\mathbf{x} = (x_1, \ldots, x_r)$. Assume that for all n,
$$X_n = U_n\, diag(x_1, \ldots, x_r, 0, \ldots, 0)\, V_n' \tag{39}$$
is a singular value decomposition of X_n, where U_n and V_n are uniformly distributed random orthogonal matrices. Formally, U_n and V_n are sampled from the Haar distribution on the m-by-m and n-by-n orthogonal group, respectively.
3) *Asymptotic aspect ratio β:* The sequence m_n is such that $m_n/n \to \beta$.

The second assumption above implies that X_n a "generic" choice of matrix with nonzero singular values $\mathbf{x}$, or equivalently, a generic choice of coordinate systems in which the linear operator corresponding to X is expressed.

The results of [25], which we have used, hold in this case as well. It follows that Lemma 4 and Lemma 5, and consequently all our main results, hold under this alternative framework. In short, in general white noise, all our results hold if one is willing to only specify the signal singular values, rather than the signal matrix, and consider a "generic" signal matrix with these singular values.

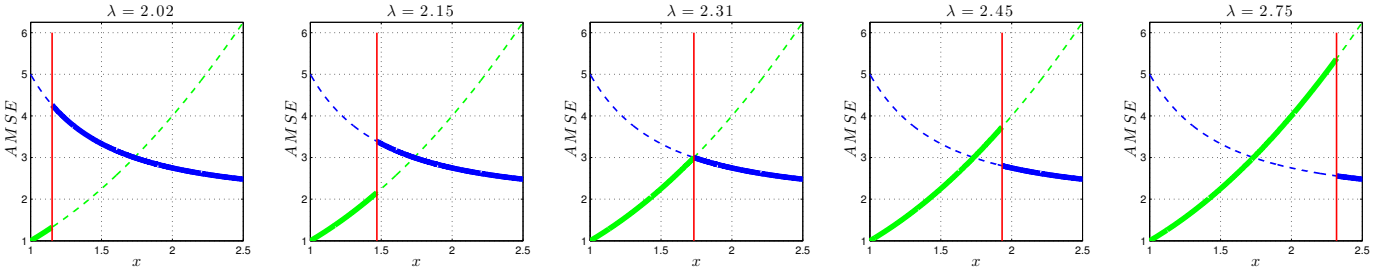

Fig. 6. The AMSE profiles of (32) for $\beta = 1$, $r = 1$ and several threshold values λ. Green: $x \mapsto x^2$. Blue: $x \mapsto 2 + 3/x^2$. Horizontal line: location of the cutoff $x_*(\lambda)$ solving $x_* + 1/x_* = \lambda$. Solid line: AMSE curve.

VII. Empirical comparison of MSE with AMSE

We have calculated the exact optimal threshold τ_* in a certain asymptotic framework. The practical significance of our results hinges on the validity of the AMSE as an approximation to MSE, for values of (m, n, r) and error distributions encountered in practice. This in turn depends on the simultaneous convergence of three terms:

- Convergence of the top data singular values $y_{n,i}$ $(1 \leq i \leq r)$ to the limit in Lemma 2,
- Convergence of the angle between the top data singular vectors $\mathbf{u}_{n,i}, \mathbf{v}_{n,i}$ and their respective signal singular vectors to the limit in Lemma 3, and
- Convergence of the rest of the data singular values $y_{n,i}$ $(r + 1 \leq i \leq m)$ to the interval $[0, 1 + \sqrt{\beta}]$.

Analysis of each of these terms for specific error distributions is beyond our current scope. Figure 7 contains a few sample comparisons of AMSE and empirical MSE we have performed. The matrix sizes and number of Monte Carlo draws are small enough to demonstrate that AMSE is a reasonable approximation even for relatively small low-rank matrices. As convergence of the empirical spectrum to its limit is known to depend on moments of the underlying distributions, we include results for different error distributions. AMSE is found to be a useful proxy to MSE even in small matrix sizes. AMSE of SVHT was found to be inaccurate when: (i) the rank fraction is nontrivial (e.g $n = 50$, $r = 4$ shown at the bottom of Figure 7); (ii) the threshold λ is very close to the approximate bulk edge $1 + \sqrt{m/n}$. In case (i), interaction effects between singular values, which are ignored in our asymptotic framework, start to have non-negligible effect. In case (ii), where the discontinuity of the SVHT nonlinearity is placed close to the bulk edge, the distribution of the largest "non-signal" singular value $y_{n,r+1}$, which is known in some cases to be asymptotically a Tracy-Widom distribution [23], becomes important. Indeed, some data singular values from the bulk manage to pass the threshold λ and cause their singular vectors to be included in the estimator $\hat{X}_\lambda$. Our derivation of AMSE assumed however that no such singular vectors are included in $\hat{X}_\lambda$, since $y_{n,r+1} \xrightarrow{a.s.} 1 + \sqrt{\beta}$. Note however that the main recommendation of this paper is that one should *not* threshold at or near he bulk edge, as explained in detail above. Therefore, from a practical perspective, the inaccuracy of AMSE for SVHT with λ near the bulk edge is slightly irrelevant.

VIII. Conclusion

The asymptotic framework considered here is perhaps the simplest nontrivial model for matrix denoising. It allows one to calculate, in AMSE, basically any quantity of interest, for any denoiser of interest. The fundamental elements of matrix denoising in white noise, which underly more complicated models, are present yet understandable and quantifiable. For example, the AMSE of any denoiser based on singular value shrinkage contains a component due to noise contamination in the data singular vectors, and this component determines a fundamental lower bound on AMSE.

We conjecture that results calculated in this model, which are not attached to a specific assumption on rank (e.g, the constants in Table III, which determine the minimax AMSE over nuclear norm balls) remain essentially correct in more complicated models.

The decision-theoretic landscape as it appears through the naive prism of our asymptotic framework is extremely simple: there is a unique admissible hard thresholding rule, and moreover a unique admissible shrinkage rule, for singular values. This is of course quite different from the situation encountered, for example, in estimating normal means. The reason is the extreme simplicity of our model. For example, we have replaced the data singular values, which are random for finite matrix size, with their almost sure limits, and in effect neglected their random fluctuations around these limits. These fluctuations are now well understood (see for example [33], [34]). We have ignored this structure. However, including these second-order terms in the asymptotic distributions is only likely to achieve second-order improvements in MSE over our suggested optimal truncation threshold.

Reproducible Research

In the code supplement [1] we offer a Matlab software library that includes:

1) A function that calculates the optimal shrinkage coefficient in known or unknown noise level.
2) Scripts that generate each of the figures in this paper.
3) A script that generates figures similar to Figure 7, which compare AMSE to MSE in various situations.

Acknowledgements

The authors would like to thank Andrea Montanari for pointing to the work of Shabalin and Nobel, Drew Nobel

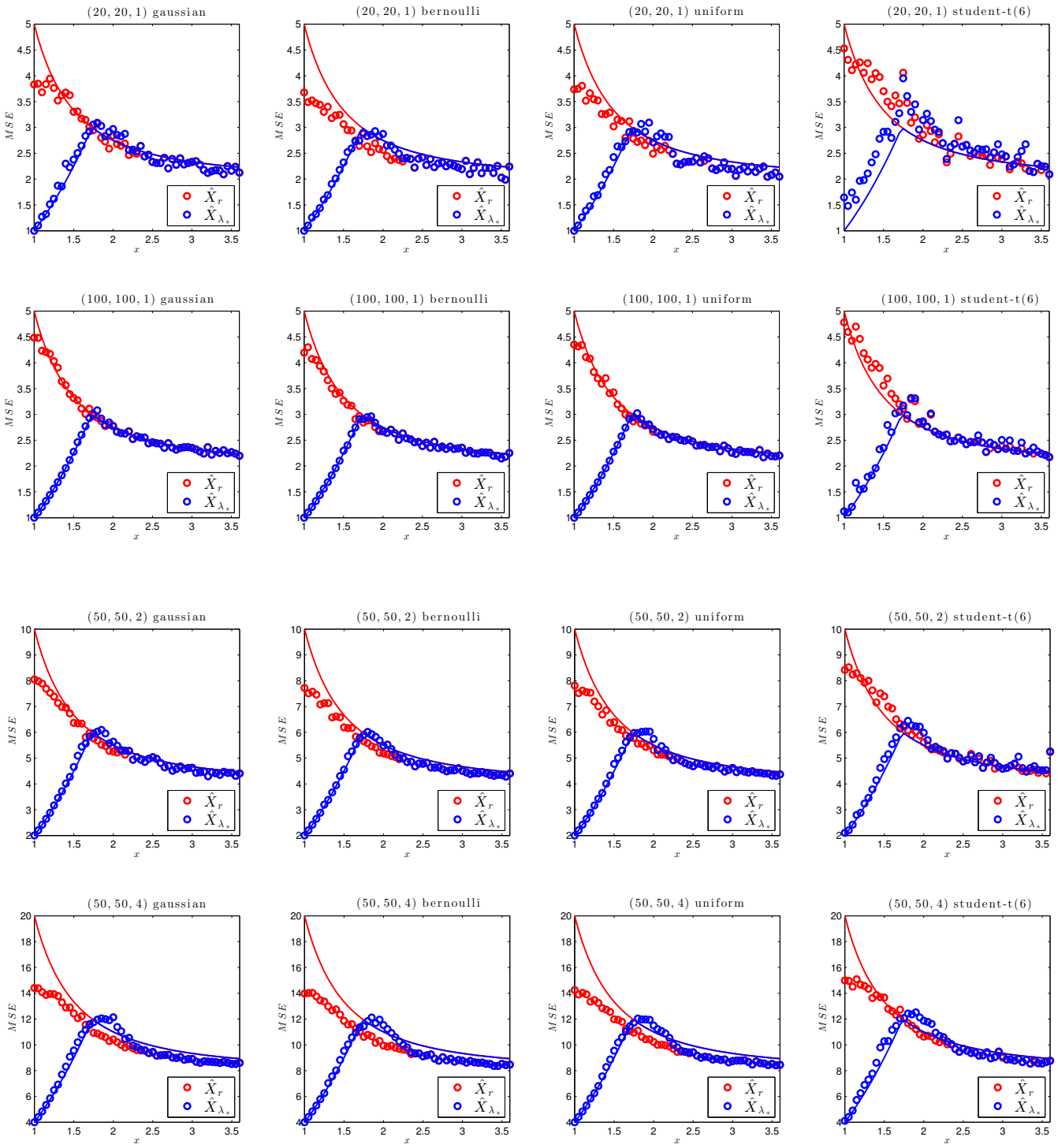

Fig. 7. The AMSE (solid line) and empirical MSE (circles) of TSVD $\hat{X}_r$ and optimal SVHT $\hat{X}_{\lambda_*}$ for $\beta = 1$ and signal singular value $x \geq 1$ which correspond to leading data singular values that fall beyond the bulk egde. In a given panel, for a given value of x, the blue and the red dots were generated by first generating a signal matrix X, and then averaging each of the losses $\left|\left|\hat{X}(X + Z/\sqrt{n}) - X\right|\right|_F^2$ over the same 50 Monte Carlo draws of the noise matrix Z. Each column of panels represents a different noise with zero mean and unit variance: Gaussian, Bernoulli on ± 1, uniform on $[-0.5, 0.5]$ and Student's t-distribution with 6 degreees of freedom; Panel titles indicate (m, n, r) and the noise distribution. Top rows: $r = 1$, different values of $m = n$. Bottom rows: $m = n = 50$, different values of r (for $r > 1$, signal singular values are all equal). *Reproducibility advisory:* script to generate figure, and to perform similar experiments, is included in code supplement [1].

and Sourav Chatterjee for helpful discussions, Art Owen for pointing to the work of Perry, and the anonymous referees for their useful suggestions. This work was partially supported by NSF DMS 0906812 (ARRA). MG was partially supported

by a William R. and Sara Hart Kimball Stanford Graduate Fellowship.

REFERENCES

[1] D. L. Donoho and M. Gavish, "Code supplement to 'The Optimal Hard Threshold for Singular Values is $4/\sqrt{3}$'," http://purl.stanford.edu/vg705qn9070, 2014, accessed 27 March 2014. [Online]. Available: http://purl.stanford.edu/vg705qn9070

[2] G. Golub and W. Kahan, "Calculating the Singular Values and Pseudo-Inverse of a Matrix," *Journal of the Society for Industrial & Applied Mathematics: Series B*, vol. 2, no. 2, pp. 205–224, 1965.

[3] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, 1936.

[4] O. Alter, P. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, pp. 10 101–10 106, Aug. 2000.

[5] D. Jackson, "Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches," *Ecology*, 1993. Available: http://www.jstor.org/stable/10.2307/1939574

[6] T. D. Lagerlund, F. W. Sharbrough, and N. E. Busacker, "Spatial filtering of multichannel electroencephalographic recordings through principal component analysis by singular value decomposition." *Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society*, vol. 14, no. 1, pp. 73–82, Jan. 1997.

[7] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. a. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies." *Nature genetics*, vol. 38, no. 8, pp. 904–9, Aug. 2006.

[8] O. Edfors and M. Sandell, "OFDM channel estimation by singular value decomposition," *IEEE Transactions on Communications*, vol. 46, no. 7, pp. 931-939, 1998.

[9] R. Cattell, "The scree test for the number of factors," *Multivariate behavioral research*, 1966.

[10] S. Wold, "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Components Models," *Technometrics*, vol. 20, no. 4, pp. 397–405, 1978. [Online]. Available: http://www.tandfonline.com/doi/pdf/10.1080/00401706.1978.10489693

[11] P. D. Hoff, "Model averaging and dimension selection for the singular value decomposition," Sep. 2006. [Online]. Available: http://arxiv.org/abs/math/0609042

[12] A. B. Owen and P. O. Perry, "Bi-cross-validation of the SVD and the nonnegative matrix factorization," *The Annals of Applied Statistics*, vol. 3, no. 2, pp. 564–594, Jun. 2009. [Online]. Available: http://projecteuclid.org/euclid.aoas/1245676186

[13] P. O. Perry, "Cross validation for unsupervised learning," *PhD Thesis, Department of Statistics, Stanford University*, 2009. [Online]. Available: http://http://arxiv.org/abs/0909.3052

[14] D. Achlioptas and F. McSherry, "Fast Computation of Low Rank Matrix Approximations," in *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, 2001, pp. 611–618. [Online]. Available: http://dl.acm.org/citation.cfm?id=380858

[15] Y. Azar, A. Fiat, A. R. Karlin, F. McSherry, and J. Saia, "Spectral Analysis of Data," in *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, 2001, pp. 619–626. [Online]. Available: http://dl.acm.org/citation.cfm?id=380859

[16] P. J. Bickel and E. Levina, "Covariance regularization by thresholding," *The Annals of Statistics*, vol. 36, no. 6, pp. 2577–2604, Dec. 2008. [Online]. Available: http://projecteuclid.org/euclid.aos/1231165180

[17] S. Chatterjee, "Matrix estimation by universal singular value thresholding,", 2010. [Online]. Available: arxiv.org/abs/1212.1247

[18] R. H. Keshavan and S. Oh, "OptSpace : A Gradient Descent Algorithm on the Grassman Manifold for Matrix Completion," 2009. [Online]. Available: arxiv.org/abs/0910.5260

[19] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transcations on Information Theory*, vol. 56, no. 6, pp.2980-2998, 2010. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5466511

[20] J. Tanner and K. Wei, "Normalized iterative hard thresholding for matrix completion," 2012. [Online]. Available: http://people.maths.ox.ac.uk/tanner/papers/TaWei_NIHT.pdf

[21] Z. Bai, and J. W. Silverstein, "Spectral Analysis of Large Dimensional Random Matrices (2nd Edition)," 2010. Springer New York. doi:10.1007/978-1-4419-0661-8

[22] A. Shabalin and A. Nobel, "Reconstruction of a Low-rank Matrix in the Presence of Gaussian Noise," *Journal of Multivariate Analysis*, vol. 118, pp. 67–76, 2013.

[23] I. M. Johnstone, "On the distribution of the largest eigenvalue in principal components analysis," *Annals of Statistics*, vol. 29, no. 2, pp. 295–327, 2001.

[24] D. L. Donoho, M. Gavish and I. M. Johnstone, "Optimal Shrinkage of Eigenvalues in the Spiked Covariance Model," *Stanford University Statistics Department technical report 2013-10*, 2013. [Online]. Available: http://arxiv.org/abs/1311.0851

[25] F. Benaych-Georges and R. R. Nadakuditi, "The singular values and vectors of low rank perturbations of large rectangular random matrices," *Journal of Multivariate Analysis*, vol. 111, pp. 120–135, Oct. 2012.

[26] Y. Q. Yin, Z. D. Bai, and P. R. Krishnaiah, "On the limit of the largest eigenvalue of the large dimensional sample covariance matrix," *Probability Theory and Related Fields*, vol. 78, pp. 509-521, 1988.

[27] Y. Q. Yin, Z. D. Bai, "Limit of the smallest eigenvalue of a large dimensional sample covariance matrix," *The annals of Probability*, vol. 21, no. 3, pp.1275-1294, 1993.

[28] M. Gavish and D. L. Donoho, "Optimal Shrinkage of Singular Values," *Stanford University Statistics Department technical report 2014-08*, 2014. [Online]. Available: http://arxiv.org/abs/1405.7511

[29] J.-F. Cai, E. J. Candès, and Z. Shen, "A Singular Value Thresholding Algorithm for Matrix Completion," *SIAM Journal on Optimization*, vol. 20, no. 4, p. 1956, 2008. [Online]. Available: http://arxiv.org/abs/0810.3286

[30] E. J. Candès, C. A. Sing-long, and J. D. Trzasko, "Unbiased Risk Estimates for Singular Value Thresholding and Spectral Estimators," *IEEE Transactions on Signal Processing*, vol. 61, no. 19, pp. 4643–4657, 2012.

[31] D. L. Donoho and M. Gavish, "Minimax Risk of Matrix Denoising by Singular Value Thresholding," *Stanford University Statistics Department technical report 2013-03*, 2013. [Online]. Available: http://arxiv.org/abs/1304.2085

[32] S. Kritchman and B. Nadler, "Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory," *Signal Processing, IEEE Transactions*, vol. 57, no. 10, pp. 3930–3941, 2009.

[33] Z. Bai and J.-f. Yao, "Central limit theorems for eigenvalues in a spiked population model," *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, vol. 44, no. 3, pp. 447–474, Jun. 2008.

[34] D. Shi, "Asymptotic Joint Distribution of Extreme Sample Eigenvalues and Eigenvectors in the Spiked Population Model," Apr. 2013. [Online]. Available: http://arxiv.org/abs/1304.6113

Matan Gavish received the dual B.Sc. degree in Mathematics and Physics from Tel Aviv University (TAU) in 2006 and the M.Sc. degree in Mathematics from the Hebrew University of Jerusalem in 2008. He is currently a doctoral student in Statistics at Stanford University, in collaboration with the Yale University program in Applied Mathematics. His research interests include applied harmonic analysis, high-dimensional statistics and computing. He was in the Adi Lautman Interdisciplinary Program for outstanding students at TAU from 2002 to 2006 and held a William R. and Sara Hart Kimball Stanford Graduate Fellowship from 2009 to 2012.

David L. Donoho is a professor at Stanford University. His research interests include computational harmonic analysis, high-dimensional geometry, and mathematical statistics. Dr. Donoho received the Ph.D. degree in Statistics from Harvard University, and holds honorary degrees from University of Chicago and Ecole Polytechnique Federale de Lausanne. He is a member of the American Academy of Arts and Sciences and the US National Academy of Sciences, and a foreign associate of the French Académie des sciences.