

# Optimal singular value shrinkage for operator norm loss

William Leeb\*

## Abstract

We study the denoising of low-rank matrices by singular value shrinkage. Recent work of Gavish and Donoho constructs a framework for finding optimal singular value shrinkers for a wide class of loss functions. We use this framework to derive the optimal shrinker for operator norm loss. The optimal shrinker matches the shrinker proposed by Gavish and Donoho in the special case of square matrices, but differs for all other aspect ratios. We precisely quantify the gain in accuracy from using the optimal shrinker. We also show that the optimal shrinker converges to the best linear predictor in the classical regime of aspect ratio zero.

## 1 Introduction

Low-rank matrix denoising is the task of estimating a low-rank matrix  $\mathbf{X}$  from a noisy observed matrix  $\mathbf{Y} = \mathbf{X} + \mathbf{G}$ . In the setting of this paper,  $\mathbf{G}$  is a matrix with iid Gaussian entries. We study the denoising procedure known as *singular value shrinkage*, which keeps the singular vectors of  $\mathbf{Y}$  while deflating the singular values to remove the effects of noise. Singular value shrinkage is a popular and well-studied methodology [15], [8], [7], [3], [12], [4], [13], [17], [10], [11], [2], [6].

It has previously been observed, both for the matrix denoising problem and the related problem of low-rank covariance estimation, that the optimal singular value shrinker depends crucially on the choice of loss function between  $\mathbf{X}$  and the estimated matrix  $\widehat{\mathbf{X}}$  [8], [5]. The work of Gavish and Donoho from [8] provides a general framework for deriving optimal singular value shrinkers for a wide class of loss functions. The results are applicable in a high-dimensional setting where the numbers of rows and columns grow to infinity, but the aspect ratio (the number of rows divided by the number of columns) converges to a definite limit. This is a standard setting, commonly referred to as the *spiked model* [9].

In this paper, we revisit the special case of operator norm loss, defined as  $\|\widehat{\mathbf{X}} - \mathbf{X}\|_{\text{op}}$ . We employ the the framework of Gavish and Donoho from [8] to derive the optimal singular value shrinker for this loss. We show that the optimal shrinker matches the shrinker proposed in [8] in the special case of square matrices, and provide a precise comparison of the two shrinkers for all aspect ratios. We will also show that when the columns of  $\mathbf{X}$  are iid random vectors, then the optimal shrinker converges to the best linear predictor of each column in the limiting regime of aspect ratio zero, which can be interpreted as a “classical” statistical limit.

The rest of the manuscript is structured as follows. In Section 2, we will formally state the spiked model assumptions, define singular value shrinkage, describe known results about the spiked model, and review the framework of Gavish and Donoho from [8]. In Section 3, we will present the optimal shrinker for operator norm loss, and a comparison with the shrinker of [8]. In Section 4, we prove the convergence of optimal shrinkage to the best linear predictor in the classical regime. Section 5 contains detailed proofs of the main results. Section 6 provides a brief conclusion.

## 2 Preliminaries

### 2.1 Model and estimation problem

We observe a matrix  $\mathbf{Y} = \mathbf{X} + \mathbf{G}$  of size  $p$ -by- $n$ . The matrix  $\mathbf{G}$  has entries which are iid  $N(0, 1/n)$ . The matrix  $\mathbf{X}$  is rank  $r$ , with singular value decomposition

$$\mathbf{X} = \sum_{k=1}^r t_k \mathbf{u}_k \mathbf{v}_k^T. \quad (1)$$

Here,  $\mathbf{u}_1, \dots, \mathbf{u}_r$  are orthonormal vectors in  $\mathbb{R}^p$ ,  $\mathbf{v}_1, \dots, \mathbf{v}_r$  are orthonormal vectors in  $\mathbb{R}^n$ , and  $t_1 > \dots > t_r > 0$  are the singular values of  $\mathbf{X}$ .

An observation model of this form is generally referred to as a *spiked model* [9]. We study the spiked model in the asymptotic regime where both  $n$  and  $p = p(n)$  grow to infinity, and their ratio converges to a parameter  $\gamma$ , which we will refer to as the *aspect ratio*:

$$\gamma = \lim_{n \rightarrow \infty} \frac{p(n)}{n}. \quad (2)$$

---

\*School of Mathematics, University of Minnesota, Twin Cities. Minneapolis, MN.

Since  $p$  and  $n$  grow, for each  $k = 1, \dots, r$  we really have a sequence of singular vectors  $\mathbf{u}_{k,n}$  and  $\mathbf{v}_{k,n}$ , indexed by  $n$ . However, to keep the notation to a minimum we will suppress the extra index  $n$ . Crucially, we assume that the singular values  $t_1, \dots, t_r$  remain fixed, independently of  $p$  and  $n$ .

Our goal is to estimate the low-rank matrix  $\mathbf{X}$  from the noisy observed matrix  $\mathbf{Y}$ . We consider the use of operator norm loss, where the error between our estimator  $\widehat{\mathbf{X}}$  and the true matrix  $\mathbf{X}$  is given by:

$$\mathcal{L}_{p,n}(\widehat{\mathbf{X}}, \mathbf{X}) = \|\widehat{\mathbf{X}} - \mathbf{X}\|_{\text{op}}. \quad (3)$$

We will consider the class of singular value shrinkage estimators, which keep the top  $r$  singular vectors of  $\mathbf{Y}$  while changing the singular values. More precisely, we consider an estimator of the form

$$\widehat{\mathbf{X}}^{\mathbf{q}} = \sum_{k=1}^r q_k \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T, \quad (4)$$

where  $\hat{\mathbf{u}}_1, \dots, \hat{\mathbf{u}}_r$  and  $\hat{\mathbf{v}}_1, \dots, \hat{\mathbf{v}}_r$  are the top  $r$  singular vectors of  $\mathbf{Y}$ , and  $\mathbf{q} = (q_1, \dots, q_r)$  is the vector of singular values of  $\widehat{\mathbf{X}}^{\mathbf{q}}$ . As it turns out, for any specified choice of  $\mathbf{q} = (q_1, \dots, q_r)$ , the asymptotic loss

$$\mathcal{L}_{\infty}(\mathbf{q}) = \lim_{p,n \rightarrow \infty} \mathcal{L}_{p,n}(\widehat{\mathbf{X}}^{\mathbf{q}}, \mathbf{X}) \quad (5)$$

is well-defined almost surely. The task is to find the values of  $q_1, \dots, q_r$  that minimize the asymptotic loss; that is, we will find:

$$\mathbf{q}^* = \underset{\mathbf{q}}{\text{argmin}} \mathcal{L}_{\infty}(\mathbf{q}). \quad (6)$$

We also wish to evaluate the asymptotic loss  $\mathcal{L}_{\infty}(\mathbf{q}^*)$  itself.

## 2.2 Asymptotics of the spiked model

The high-dimensional spiked model has been well-studied in the statistics and random matrix literature. It is known that there are precise relationships between the SVD of the observed matrix  $\mathbf{Y}$  and the SVD of the low-rank matrix  $\mathbf{X}$ . If we write the SVD of  $\mathbf{Y}$  as

$$\mathbf{Y} = \sum_{k=1}^{\min(p,n)} \sigma_k \hat{\mathbf{u}}_k \hat{\mathbf{v}}_k^T, \quad (7)$$

then we can summarize the relevant results as follows:

**Proposition 2.1.** *For  $1 \leq k \leq r$ , the  $k^{\text{th}}$  squared singular value of  $\mathbf{Y}$  converges almost surely to the following deterministic limit:*

$$\sigma_k^2 = \begin{cases} (t_k^2 + 1) \left(1 + \frac{\gamma}{t_k^2}\right), & \text{if } t_k > \gamma^{1/4}, \\ (1 + \sqrt{\gamma})^2, & \text{if } t_k \leq \gamma^{1/4}, \end{cases} \quad (8)$$

*For  $1 \leq j, k \leq r$ , the squared cosines between the  $j^{\text{th}}$  and  $k^{\text{th}}$  singular vectors of  $\mathbf{X}$  and  $\mathbf{Y}$  converge almost surely to the following limits:*

$$\lim_{p \rightarrow \infty} \langle \hat{\mathbf{u}}_j, \mathbf{u}_k \rangle^2 = c_{jk}^2 = \begin{cases} \frac{1 - \gamma/t_k^4}{1 + \gamma/t_k^2}, & \text{if } j = k \text{ and } t_k > \gamma^{1/4}, \\ 0, & \text{if } j \neq k \text{ or } t_k \leq \gamma^{1/4}, \end{cases} \quad (9)$$

and

$$\lim_{n \rightarrow \infty} \langle \hat{\mathbf{v}}_j, \mathbf{v}_k \rangle^2 = \tilde{c}_{jk}^2 = \begin{cases} \frac{1 - \gamma/t_k^4}{1 + 1/t_k^2}, & \text{if } j = k \text{ and } t_k > \gamma^{1/4}, \\ 0, & \text{if } j \neq k \text{ or } t_k \leq \gamma^{1/4}. \end{cases} \quad (10)$$

A proof of this result may be found in [16], [1].

**Remark 1.** The signs of  $c_k$  and  $\tilde{c}_k$  are arbitrary, since the sign of a singular vector may be flipped. However, their product satisfies  $c_k \tilde{c}_k \geq 0$ , and we may therefore assume without loss of generality that  $c_k \geq 0$  and  $\tilde{c}_k \geq 0$  (see, e.g., [15]).

Proposition 2.1 describes the behavior of the top  $r$  singular components of  $\mathbf{Y}$ . For each singular value  $\sigma_k$  of  $\mathbf{Y}$  with  $\sigma_k > 1 + \sqrt{\gamma}$ , information about the corresponding component of  $\mathbf{X}$  may be recovered. In particular, we may estimate  $t_k$  by inverting formula (8):

$$t_k = \sqrt{\frac{\sigma_k^2 - 1 - \gamma + \sqrt{(\sigma_k^2 - 1 - \gamma)^2 - 4\gamma}}{2}}. \quad (11)$$

From  $t_k$ , the cosines  $c_k$  and  $\tilde{c}_k$  are estimable by directly applying formulas (9) and (10).

## 2.3 The shrinkage framework of [8]

The work of Gavish and Donoho from [8] uses the behavior of the spiked model described in Proposition 2.1 to construct a framework for finding asymptotically optimal singular value shrinkers. The key observation is that there are orthonormal bases of  $\mathbb{R}^p$  and  $\mathbb{R}^n$  in which the matrices  $\widehat{\mathbf{X}}^{\mathbf{q}}$  and  $\mathbf{X}$  may be expressed as follows:

$$\mathbf{X} = \bigoplus_{k=1}^r \begin{pmatrix} t_k & 0 \\ 0 & 0 \end{pmatrix}, \quad (12)$$

and

$$\widehat{\mathbf{X}}^{\mathbf{q}} = \bigoplus_{k=1}^r q_k \begin{pmatrix} c_k \tilde{c}_k & c_k \tilde{s}_k \\ s_k \tilde{c}_k & s_k \tilde{s}_k \end{pmatrix}, \quad (13)$$

where  $s = \sqrt{1 - c^2}$  and  $\tilde{s} = \sqrt{1 - \tilde{c}^2}$ . In other words, both  $\mathbf{X}$  and  $\widehat{\mathbf{X}}^{\mathbf{q}}$  are block-diagonal, with  $r$  2-by-2 blocks of the prescribed form. The entries of these blocks depend only on the estimable quantities  $t_k$ ,  $c_k$ , and  $\tilde{c}_k$ , and the values  $q_k$  for which we solve. Because the operator norm is orthogonally-invariant and max-decomposable over block matrices, it follows that we may express the asymptotic operator norm loss as follows:

$$\mathcal{L}_{\infty}(\mathbf{q}) = \max_{1 \leq k \leq r} \left\| \begin{pmatrix} t_k & 0 \\ 0 & 0 \end{pmatrix} - q_k \begin{pmatrix} c_k \tilde{c}_k & c_k \tilde{s}_k \\ s_k \tilde{c}_k & s_k \tilde{s}_k \end{pmatrix} \right\|_{\text{op}}. \quad (14)$$

Consequently, each optimal  $q_k^*$  may in principle be solved for independently:

$$q_k^* = \underset{q_k}{\operatorname{argmin}} \left\| \begin{pmatrix} t_k & 0 \\ 0 & 0 \end{pmatrix} - q_k \begin{pmatrix} c_k \tilde{c}_k & c_k \tilde{s}_k \\ s_k \tilde{c}_k & s_k \tilde{s}_k \end{pmatrix} \right\|_{\text{op}}. \quad (15)$$

The remaining piece is to solve the minimization (15). In Section 3, we will present a simple, estimable formula for the optimal  $q_k^*$ .

## 3 Optimal shrinkage

### 3.1 The optimal singular values

In this section, we derive the optimal singular values  $q_k^*$  and the resulting asymptotic operator norm loss  $\mathcal{L}_{\infty}(\mathbf{q}^*)$ . The main result is the following:

**Theorem 3.1.** *The optimal singular value shrinkage estimator  $\widehat{\mathbf{X}}^{\mathbf{q}^*}$  of  $\mathbf{X}$  from  $\mathbf{Y}$  has singular values*

$$q_k^* = \begin{cases} t_k \sqrt{\frac{t_k^2 + \min\{1, \gamma\}}{t_k^2 + \max\{1, \gamma\}}}, & \text{if } \sigma_k > 1 + \sqrt{\gamma}, \\ 0, & \text{otherwise} \end{cases}, \quad (16)$$

where  $1 \leq k \leq r$ . The loss  $\|\widehat{\mathbf{X}}^{\mathbf{q}^*} - \mathbf{X}\|_{\text{op}}$  converges almost surely to

$$\mathcal{L}_{\infty}(\mathbf{q}^*) = \max_{1 \leq k \leq r} t_k \sqrt{1 - \min\{c_k^2, \tilde{c}_k^2\}} = t_1 \sqrt{1 - \min\{c_1^2, \tilde{c}_1^2\}}. \quad (17)$$

We defer the proof of Theorem 3.1 to Section 5.1.

**Remark 2.** The optimal  $q_k^*$  and loss  $\mathcal{L}_{\infty}(\mathbf{q}^*)$  may be consistently estimated from the observed singular values  $\sigma_k$  of  $\mathbf{Y}$  using (8), (9), and (10).

**Remark 3.** The paper [8] proposes the shrinker  $\widehat{\mathbf{X}}^{\mathbf{q}}$  with singular values  $q_k = t_k$ . The optimal singular values  $q_k^*$  we derive in Theorem 3.1 are equal to  $t_k$  when  $\gamma = 1$  (the case of square matrices). In all other cases, however, the singular values  $q_k = t_k$  are strictly suboptimal. In Figure 1, we plot both shrinkers  $q = q^*$  and  $q = t$  and their asymptotic losses as functions as functions of the observed singular value  $\sigma$ .

### 3.2 Comparison with the shrinker from [8]

The paper [8] proposes use of the singular values  $q_k = t_k$  for operator norm loss. This is a natural procedure, as it replace the noisy singular value  $\sigma_k$  of  $\mathbf{Y}$  with the “true” singular value  $t_k$  of  $\mathbf{X}$ . However, according to Theorem 3.1, this choice is only optimal when  $\gamma = 1$ . We can quantify the gap in performance by comparing the relative errors. To simplify notation, we will consider only the rank 1 setting, and subsequently drop subscripts; the same results hold when  $\mathbf{X}$  is rank  $r$ .

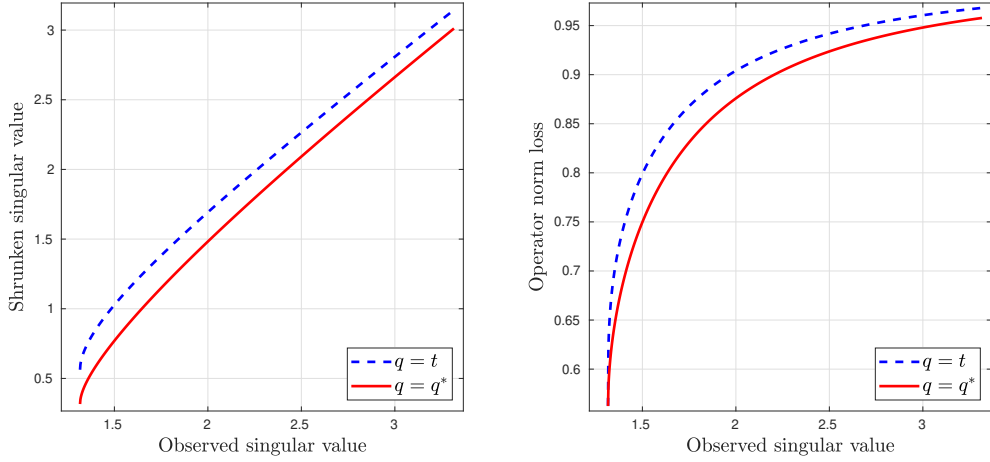


Figure 1: Left: The singular values  $q = q^*$  and  $q = t$  as functions of the observed singular value  $\sigma$ . Right: The asymptotic operator norm losses as functions of  $\sigma$ .

**Proposition 3.2.** Suppose  $\mathbf{X}$  is a rank 1 matrix with singular value  $t > \gamma^{1/4}$ . Let  $\widehat{\mathbf{X}}^t$  denote the singular value shrinkage denoiser with  $q = t$ , and let  $\widehat{\mathbf{X}}^{q^*}$  denote the optimal singular value shrinker from Theorem 3.1. Then

$$\frac{\mathcal{L}_\infty(q^*)}{\mathcal{L}_\infty(t)} = \lim_{n \rightarrow \infty} \frac{\|\mathbf{X} - \widehat{\mathbf{X}}^{q^*}\|_{\text{op}}}{\|\mathbf{X} - \widehat{\mathbf{X}}^t\|_{\text{op}}} = \sqrt{\frac{1 + \min\{c, \bar{c}\}}{1 + \max\{c, \bar{c}\}}}, \quad (18)$$

where the limit holds almost surely as  $p/n \rightarrow \gamma$ .

In the next result, we derive a limiting expression for the error ratio as  $\gamma \rightarrow 0$ .

**Proposition 3.3.** Define the asymptotic error ratio  $R$ :

$$R = \sqrt{\frac{1 + \min\{c, \bar{c}\}}{1 + \max\{c, \bar{c}\}}}. \quad (19)$$

For fixed  $t > 0$ ,  $R = R(\gamma, t)$  is an increasing function of  $\gamma < \min\{t^4, 1\}$ , and its minimum value is

$$\lim_{\gamma \rightarrow 0} R(\gamma, t) = \sqrt{\frac{1}{2} \left( 1 + \sqrt{\frac{t^2}{t^2 + 1}} \right)}. \quad (20)$$

**Remark 4.** If the columns of  $\mathbf{X}$  are  $n$  iid random vectors from a rank  $r$  distribution in  $\mathbb{R}^p$ , then the  $\gamma = 0$  limit considered in Proposition 3.3 can be thought of informally as the “classical” setting, where the number of observations  $n$  grows faster than the number of features  $p$ .

Proposition 3.3 shows that the performance of the optimal shrinker  $q = q^*$  over  $q = t$  should be most evident when  $\gamma$  and  $t$  are both small. To illustrate this, in Figure 2 we compare error curves of  $q = t$  and  $q = q^*$  as a function of  $\gamma \in [0, 1]$ , where the signal strength  $t$  is chosen to decrease with  $\gamma$  as  $t = \gamma^{1/4} + 1/20$ . The left panel shows the relative errors (the error divided by the signal norm  $t$ ) for the two shrinkers as a function of  $\gamma$ , and the right panel shows the ratio of the errors. As  $\gamma$  and  $t$  both decrease, the relative performance of the optimal shrinker  $q = q^*$  increases over the shrinker  $q = t$ .

## 4 Convergence to the best linear predictor

In this section, we consider the setting where the columns of  $\mathbf{X}$  are iid random vectors from a distribution in  $\mathbb{R}^p$  with mean zero. We will write each column of  $\mathbf{X}$  as  $X_j/\sqrt{n}$ , where  $X_j$  is a random vector of the following form:

$$X_j = \sum_{k=1}^r t_k z_{jk} \mathbf{u}_k, \quad (21)$$

where the  $z_{jk}$  are mean zero, unit variance sub-Gaussian random variables, and the  $\mathbf{u}_k$  are the orthonormal principal components of  $X_j$ .

**Remark 5.** In the new setting, each  $t_k$  is the standard deviation of  $X_j$  along the principal component  $\mathbf{u}_k$ , not the singular value of  $\mathbf{X}$ ; and  $\mathbf{u}_1, \dots, \mathbf{u}_r$  are not the left singular vectors of  $\mathbf{X}$ . However, in the large  $n$  limit, the singular values of  $\mathbf{X}$  converge almost surely to  $t_1, \dots, t_r$ , and the left singular vectors of  $\mathbf{X}$  almost surely make zero angle with, respectively,  $\mathbf{u}_1, \dots, \mathbf{u}_r$ . In this sense, the new notation is consistent with the old notation.

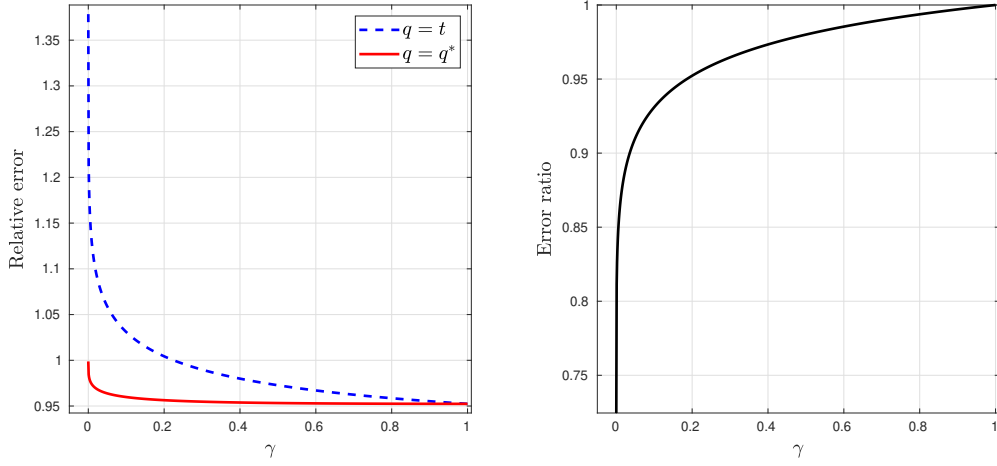


Figure 2: Left: Relative errors  $\|\widehat{\mathbf{X}}^q - \mathbf{X}\|_{\text{op}}/t$  for  $q = q^*$  and  $q = t$  as a function of  $\gamma$ , where  $t = \gamma^{1/4} + 1/20$ . Right: The ratio of the errors  $\|\widehat{\mathbf{X}}^{q^*} - \mathbf{X}\|_{\text{op}}/\|\widehat{\mathbf{X}}^t - \mathbf{X}\|_{\text{op}}$  as a function of  $\gamma$ .

**Remark 6.** The assumption that  $X_j$  has mean zero is easily removed by subtracting the sample mean from each  $X_j$ .

It is known [14] that the best linear predictor of  $X_j$  from  $Y_j$  has the following form:

$$\widehat{X}_j^{\text{BLP}} = \sum_{k=1}^r \frac{t_k^2}{t_k^2 + 1} \langle Y_j, \mathbf{u}_k \rangle \mathbf{u}_k. \quad (22)$$

The next result shows that optimal singular value shrinkage with operator norm loss converges to the best linear predictor when  $\gamma = 0$ . We will let  $\widehat{\mathbf{X}}^{q^*} = [\widehat{X}_j^{q^*}, \dots, \widehat{X}_j^{q^*}]/\sqrt{n}$  and  $\widehat{\mathbf{X}}^{\text{BLP}} = [\widehat{X}_j^{\text{BLP}}, \dots, \widehat{X}_j^{\text{BLP}}]/\sqrt{n}$ .

**Theorem 4.1.** *In the limit  $n \rightarrow \infty$  and  $p/n \rightarrow 0$ ,*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \|\widehat{X}_j^{q^*} - \widehat{X}_j^{\text{BLP}}\|^2 = \lim_{n \rightarrow \infty} \|\widehat{\mathbf{X}}^{q^*} - \widehat{\mathbf{X}}^{\text{BLP}}\|_{\text{F}}^2 = 0, \quad (23)$$

where the limit holds almost surely.

Theorem 4.1 is proven in Section 5.3. It is a consequence of the following result, whose proof follows from [13]:

**Lemma 4.2.** *Let  $\widehat{\mathbf{X}}^q$  be any singular value shrinker, with singular values  $q_1, \dots, q_r$ . Define the linear predictor  $\widetilde{X}_j^q$  by:*

$$\widetilde{X}_j^q = \sum_{k=1}^r \frac{q_k}{\sigma_k} \langle Y_j, \mathbf{u}_k \rangle \mathbf{u}_k, \quad (24)$$

where  $\sigma_1, \dots, \sigma_r$  are the top  $r$  singular values of  $\mathbf{Y}$ , and  $\hat{\mathbf{u}}_1, \dots, \hat{\mathbf{u}}_r$  are the top  $r$  left singular vectors of  $\mathbf{Y}$ . Then

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=1}^n \|\widehat{X}_j^q - \widetilde{X}_j^q\|^2 = \lim_{n \rightarrow \infty} \|\widehat{\mathbf{X}}^q - \widetilde{\mathbf{X}}^q\|_{\text{F}}^2 = 0, \quad (25)$$

where the limit holds almost surely in the limit  $n \rightarrow \infty$ ,  $p/n \rightarrow 0$ .

**Remark 7.** Lemma 4.2 states that singular value shrinkage converges to a linear predictor of the form (24) when  $\gamma = 0$  (the “classical” regime of zero aspect ratio). Comparing (24) to the form of the best linear predictor (22), we see that the shrinker  $\widehat{\mathbf{X}}^q$  will converge to the best linear predictor if and only if

$$q_k = \frac{\sigma_k t_k^2}{t_k^2 + 1} = t_k \sqrt{\frac{t_k^2}{t_k^2 + 1}}, \quad 1 \leq k \leq r, \quad (26)$$

when  $\gamma = 0$ . Any other choice of shrinker, including  $q_k = t_k$ , will result in convergence to a suboptimal linear filter in the  $\gamma = 0$  regime.

To illustrate Theorem 4.1 numerically, we draw  $n$  iid observations from a spiked model in  $\mathbb{R}^p$ , for increasing values of  $n \geq p$ . We take the rank  $r = 1$  and  $t = t_1 = 1.1$  (to ensure the signal is detectable for all  $n \geq p$ ). We apply the best linear predictor  $\widehat{X}_j^{\text{BLP}}$  (which assumes the principal component  $\mathbf{u} = \mathbf{u}_1$  is known), optimal shrinkage  $\widehat{X}_j^{q^*}$ , and the suboptimal shrinker  $\widehat{X}_j^t$ . In Figure 3, we plot the average operator norm error over 4000 runs of the experiment, as a function of  $n$ .

The error for the BLP is approximately constant, since the BLP does not vary with the sample size  $n$ . As  $n$  grows, the error for optimal shrinkage approaches that of the BLP, because  $\widehat{X}_j^{q^*}$  converges to the BLP  $\widehat{X}_j^{\text{BLP}}$ . By contrast, the error for the suboptimal shrinker converges to a strictly larger value, since  $\widehat{X}_j^t$  converges to the suboptimal linear predictor  $\widetilde{X}_j^t$ .

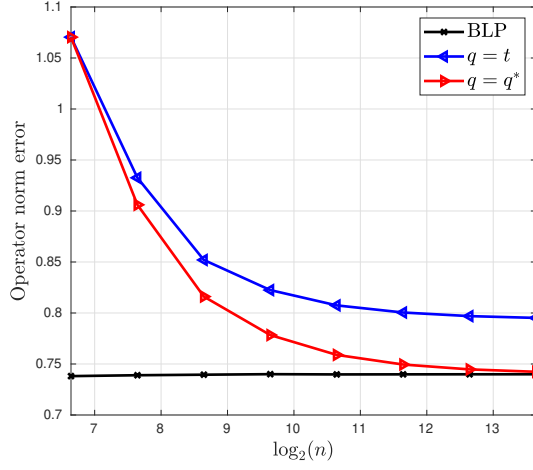


Figure 3: Operator norm errors for  $p = 100$  and increasing  $n$  for the BLP (22) and the shrinkers  $\widehat{\mathbf{X}}^q$ ,  $q = q^*$  and  $q = t$ . The signal is rank 1 with  $t = 1.1$ . Errors are averaged over 4000 runs.

## 5 Proofs

### 5.1 Proof of Theorem 3.1

Theorem 3.1 is a consequence of the following result:

**Lemma 5.1.** *Let  $t > 0$ , and  $0 \leq c \leq 1$ ,  $0 \leq \tilde{c} \leq 1$ . For any  $q \in \mathbb{R}$ , define the 2-by-2 matrix  $\mathbf{D}(q)$  by*

$$\mathbf{D}(q) = \begin{pmatrix} t & 0 \\ 0 & 0 \end{pmatrix} - q \begin{pmatrix} c\tilde{c} & c\tilde{s} \\ s\tilde{c} & s\tilde{s} \end{pmatrix}, \quad (27)$$

where  $s = \sqrt{1 - c^2}$  and  $\tilde{s} = \sqrt{1 - \tilde{c}^2}$ . Define the function  $F$  by:

$$F(q) = \|\mathbf{D}(q)\|_{\text{op}}^2, \quad (28)$$

the squared operator norm of  $\mathbf{D}(q)$ . Then the value  $q^*$  that minimizes  $F$  defined in (28) is given by

$$q^* = t \cdot \frac{\min\{c, \tilde{c}\}}{\max\{c, \tilde{c}\}}, \quad (29)$$

when  $c$  and  $\tilde{c}$  are not both 0; and  $q^*$  may be taken as any value with  $|q^*| \leq t$  if  $c = \tilde{c} = 0$ . Furthermore, the value of  $F$  at  $q^*$  is given by:

$$F(q^*) = t^2 \cdot \max\{s^2, \tilde{s}^2\}. \quad (30)$$

**Remark 8.** Lemma 5.1 is applicable for general parameters  $t > 0$  and  $0 \leq c \leq 1$ ,  $0 \leq \tilde{c} \leq 1$ , even when they do not satisfy the relationships (8), (9) and (10).

We will first prove Theorem 3.1 assuming Lemma 5.1, and then return to the proof of Lemma 5.1 itself in Section 5.1.1. Without loss of generality, we assume  $\gamma \leq 1$ . If  $\sigma_k > 1 + \sqrt{\gamma}$ , or equivalently  $t_k > \gamma^{1/4}$ , Lemma 5.1 says that the optimal singular value is given by  $q_k^* = t_k \tilde{c}_k / c_k$ . Using the expressions (9) and (10) for  $c_k$  and  $\tilde{c}_k$ , formula (16) follows immediately. When  $\sigma \leq 1 + \sqrt{\gamma}$ , or equivalently  $t_k \leq \gamma^{1/4}$ , both  $c_k$  and  $\tilde{c}_k$  are zero. Consequently,  $q_k^* = 0$  is optimal.

All the remains is to show the error formula:

$$\max_{1 \leq k \leq r} t_k \sqrt{1 - \min\{c_k^2, \tilde{c}_k^2\}} = t_1 \sqrt{1 - \min\{c_1^2, \tilde{c}_1^2\}}. \quad (31)$$

Without loss of generality, assume  $\gamma \leq 1$ , so  $\tilde{c} \leq c$ . Then

$$t^2(1 - \tilde{c}^2) = t^2 \frac{t^2 + \gamma}{t^4 + t^2} = \frac{t^2 + \gamma}{t^2 + 1} \quad (32)$$

is an increasing function of  $t$ , and is equal to  $\sqrt{\gamma}$  (the largest possible error for any component  $t_k \leq \gamma^{1/4}$ ) when  $t = \gamma^{1/4}$ . Consequently, the maximum error is achieved at  $t = t_1$ .

### 5.1.1 Proof of Lemma 5.1

First, suppose  $c = \tilde{c} = 0$ . Then

$$\mathbf{D}(q) = \begin{pmatrix} t & 0 \\ 0 & -q \end{pmatrix}, \quad (33)$$

and so  $F(q) = \|\mathbf{D}(q)\|_{\text{op}}^2 = \max\{t^2, q^2\}$ . Consequently, any  $q$  with  $|q| \leq t$  minimizes  $F(q)$ , and since  $s = \tilde{s} = 1$ ,  $F(q) = t^2 \cdot \max\{s^2, \tilde{s}^2\}$  for such  $q$ .

Next, assume that  $c$  and  $\tilde{c}$  are not both 0. The proof when  $c = \tilde{c}$  is identical to the proof for optimal shrinkage of eigenvalues for covariance estimation contained in [5]; so we will assume that  $c \neq \tilde{c}$ . We will use the expression for the operator norm as a function of  $q$  derived in [8]. We have:

$$\begin{aligned} F(q) &= \frac{q^2 + t^2 - 2qtc\tilde{c} + \sqrt{(q^2 + t^2 - 2qtc\tilde{c})^2 - 4(tqs\tilde{s})^2}}{2} \\ &= \frac{A(q) + \sqrt{A(q)^2 - 4B(q)^2}}{2}, \end{aligned} \quad (34)$$

where

$$A(q) = q^2 + t^2 - 2qtc\tilde{c}, \quad (35)$$

and

$$B(q) = -tqs\tilde{s}. \quad (36)$$

The function  $F(q)$  is differentiable whenever  $A(q)^2 - 4B(q)^2 > 0$ . Furthermore, at points where  $F$  is differentiable, its derivative is given by

$$F'(q) = \frac{1}{2} \left( A'(q) + \frac{A(q)A'(q) - 4B(q)B'(q)}{\sqrt{A(q)^2 - 4B(q)^2}} \right). \quad (37)$$

Suppose, without loss of generality, that  $\tilde{c} < c$ . Then  $q^* = t\tilde{c}/c$ . First, we have

$$A'(q) = 2q - 2tc\tilde{c}. \quad (38)$$

Consequently

$$A'(q^*) = 2t\frac{\tilde{c}}{c} - 2tc\tilde{c} = 2t\frac{\tilde{c}}{c}(1 - c^2) = 2t\frac{\tilde{c}}{c}s^2. \quad (39)$$

We also have

$$A(q^*) = t^2 \left( 1 + \frac{\tilde{c}^2}{c^2} - 2\tilde{c}^2 \right), \quad (40)$$

and so

$$A(q^*)A'(q^*) = 2t^3 s^2 \frac{\tilde{c}}{c} \left( 1 + \frac{\tilde{c}^2}{c^2} - 2\tilde{c}^2 \right). \quad (41)$$

Next, observe that for all  $q$ ,

$$B'(q) = -ts\tilde{s}. \quad (42)$$

We also have

$$B(q^*) = -t^2 \frac{\tilde{c}}{c} s\tilde{s}, \quad (43)$$

and so

$$B(q^*)B'(q^*) = t^3 s^2 \frac{\tilde{c}}{c} (1 - \tilde{c}^2). \quad (44)$$

Combining (41) and (44), we obtain:

$$\begin{aligned} A(q^*)A'(q^*) - 4B(q^*)B'(q^*) &= 2t^3 s^2 \frac{\tilde{c}}{c} \left( 1 + \frac{\tilde{c}^2}{c^2} - 2\tilde{c}^2 - 2 + 2\tilde{c}^2 \right) \\ &= 2t^3 s^2 \frac{\tilde{c}}{c} \left( \frac{\tilde{c}^2}{c^2} - 1 \right) \\ &= 2t^3 s^2 \frac{\tilde{c}}{c^3} (\tilde{c}^2 - c^2). \end{aligned} \quad (45)$$

Next, we observe that we may write

$$A(q^*)^2 - 4B(q^*)^2 = t^4 \left(1 - \frac{\tilde{c}^2}{c^2}\right)^2, \quad (46)$$

and consequently,

$$\sqrt{A(q^*)^2 - 4B(q^*)^2} = \frac{t^2}{c^2}(c^2 - \tilde{c}^2), \quad (47)$$

where we have used the fact that  $c > \tilde{c}$ . Note that (47) implies that  $F$  is differentiable at  $q^*$  whenever  $c \neq \tilde{c}$ .

Combining (45) and (47), we get:

$$\frac{A(q^*)A'(q^*) - 4B(q^*)B'(q^*)}{\sqrt{A(q^*)^2 - 4B(q^*)^2}} = -2t \frac{\tilde{c}}{c} s^2. \quad (48)$$

Consequently,

$$\begin{aligned} F'(q^*) &= \frac{1}{2} \left( A'(q^*) + \frac{A(q^*)A'(q^*) - 4B(q^*)B'(q^*)}{\sqrt{A(q^*)^2 - 4B(q^*)^2}} \right) \\ &= \frac{1}{2} \left( 2t \frac{\tilde{c}}{c} s^2 - 2t \frac{\tilde{c}}{c} s^2 \right) \\ &= 0. \end{aligned} \quad (49)$$

Next, we evaluate  $F(q^*)$ . From (40) and (47), we get

$$\begin{aligned} F(q^*) &= \frac{1}{2} \left( t^2 \left( 1 + \frac{\tilde{c}^2}{c^2} - 2\tilde{c}^2 \right) + \frac{t^2}{c^2} (c^2 - \tilde{c}^2) \right) \\ &= \frac{1}{2} \left( t^2 + t^2 \frac{\tilde{c}^2}{c^2} - 2t^2 \tilde{c}^2 + t^2 - t^2 \frac{\tilde{c}^2}{c^2} \right) \\ &= t^2(1 - \tilde{c}^2), \end{aligned} \quad (50)$$

which is the desired result.

## 5.2 Proof of Propositions 3.2 and 3.3

### 5.2.1 Proof of Proposition 3.2

The error formula for  $q = t$  found in [8]:

$$\mathcal{L}_\infty(t) = t \cdot \sqrt{1 - c\tilde{c} + |c - \tilde{c}|}. \quad (51)$$

Assuming, without loss of generality, that  $\gamma \leq 1$  and hence  $\tilde{c} \leq c$ , then

$$\mathcal{L}_\infty(t) = t \cdot \sqrt{1 - c\tilde{c} + c - \tilde{c}} = t \cdot \sqrt{(1 - \tilde{c})(1 + c)}. \quad (52)$$

Since  $\mathcal{L}_\infty(q^*) = t\sqrt{1 - \tilde{c}^2} = t\sqrt{(1 - \tilde{c})(1 + \tilde{c})}$  from Theorem 3.1, dividing  $\mathcal{L}_\infty(t)$  by  $\mathcal{L}_\infty(q^*)$  gives the result.

### 5.2.2 Proof of Proposition 3.3

The value  $t > 0$  is fixed, and we consider the range  $\gamma < \min\{t^4, 1\}$ . Define the function

$$E(\gamma) = R(\gamma, t)^2 = \frac{1 + \tilde{c}(\gamma)}{1 + c(\gamma)}, \quad (53)$$

where we treat the cosines  $c = \sqrt{(t^4 - \gamma)/(t^4 + \gamma t^2)}$  and  $\tilde{c} = \sqrt{(t^4 - \gamma)/(t^4 + t^2)}$  as functions of  $\gamma$ . Then the derivative of  $E(\gamma)$  is:

$$E'(\gamma) = \frac{1}{2c(\gamma)\tilde{c}(\gamma)(1 + c(\gamma))^2} \left( \frac{\tilde{c}'(\gamma)(1 + \tilde{c}(\gamma))(t^2 + 1)}{(t^2 + \gamma)^2} - \frac{c(\gamma)(1 + c(\gamma))}{t^4 + t^2} \right). \quad (54)$$

Consequently, to show  $E'(\gamma) > 0$ , it is enough to show that

$$\frac{\tilde{c}'(\gamma)(1 + \tilde{c}(\gamma))(t^2 + 1)}{(t^2 + \gamma)^2} - \frac{c(\gamma)(1 + c(\gamma))}{t^4 + t^2} > 0, \quad (55)$$



or equivalently that

$$\frac{\tilde{c}(\gamma)}{c(\gamma)} \frac{1 + \tilde{c}(\gamma)}{1 + c(\gamma)} \left( \frac{t^2 + 1}{t^2 + \gamma} \right)^2 t^2 - 1 > 0. \quad (56)$$

Since  $\gamma < 1$ ,  $\tilde{c}(\gamma) < c(\gamma)$ ; consequently,

$$\frac{\tilde{c}(\gamma)}{c(\gamma)} \frac{1 + \tilde{c}(\gamma)}{1 + c(\gamma)} > \frac{\tilde{c}(\gamma)^2}{c(\gamma)^2} = \frac{t^2 + \gamma}{t^2 + 1}. \quad (57)$$

Hence it is enough to show

$$\left( \frac{t^2 + 1}{t^2 + \gamma} \right) t^2 - 1 > 0; \quad (58)$$

but this follows immediately, since  $\gamma < t^4$ .

The limit of  $R(\gamma, t)$  as  $\gamma \rightarrow 0$  follows immediately from the limits:

$$\lim_{\gamma \rightarrow 0} c(\gamma) = \lim_{\gamma \rightarrow 0} \sqrt{\frac{t^4 - \gamma}{t^4 + \gamma t^2}} = 1 \quad (59)$$

and

$$\lim_{\gamma \rightarrow 0} \tilde{c}(\gamma) = \lim_{\gamma \rightarrow 0} \sqrt{\frac{t^4 - \gamma}{t^4 + t^2}} = \sqrt{\frac{t^2}{t^2 + 1}}. \quad (60)$$

### 5.3 Proof of Theorem 4.1

Since each component is treated separately, we drop the subscripts. Following Remark 7, Theorem 4.1 will be proven if we show that:

$$q^* = \frac{\sigma t^2}{t^2 + 1} = t \sqrt{\frac{t^2}{t^2 + 1}} \quad (61)$$

when  $\gamma = 0$ . To prove this identity, observe that when  $\gamma = 0$ ,  $c = 1$ , and

$$\tilde{c} = \sqrt{\frac{t^2}{t^2 + 1}}. \quad (62)$$

Since  $q^* = t \cdot \min\{c, \tilde{c}\} = t\tilde{c}$ , the result follows immediately.

## 6 Conclusion

We have considered the problem of estimating a low-rank matrix  $\mathbf{X}$  from noisy observations  $\mathbf{Y} = \mathbf{X} + \mathbf{G}$ , where we measure the error by operator norm loss  $\|\mathbf{X} - \hat{\mathbf{X}}\|_{\text{op}}$ . We have proven (Theorem 3.1) that the optimal singular value shrinker has singular values of the form (16). For square matrices ( $\gamma = 1$ ), the optimal singular values agree with those proposed in [8], though the two methods and the resulting errors differ increasingly as  $\gamma$  differs from 1, or equivalently as the matrix becomes more rectangular (Proposition 3.3). We have also shown (Theorem 4.1) that when the columns of  $\mathbf{X}$  are iid vectors in  $\mathbb{R}^p$ , then in the classical regime ( $\gamma = 0$ ) the optimal shrinker converges to the best linear predictor.

## Acknowledgements

I acknowledge support from the NSF BIGDATA program IIS 1837992 and BSF award 2018230.

## References

- [1] Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *Journal of Multivariate Analysis*, 111:120–135, 2012.
- [2] Jérémie Bigot, Charles Deledalle, and Delphine Féral. Generalized SURE for optimal shrinkage of singular values in low-rank matrix denoising. *Journal of Machine Learning Research*, 18:1–50, 2017.
- [3] Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1):177–214, 2015.

- [4] Edgar Dobriban, William Leeb, and Amit Singer. Optimal prediction in the linearly transformed spiked model. *Annals of Statistics*, 48(1):491–513, 2020.
- [5] David L. Donoho, Matan Gavish, and Iain M. Johnstone. Optimal shrinkage of eigenvalues in the spiked covariance model. *Annals of Statistics*, 46(4):1742–1778, 2018.
- [6] Matan Gavish and David L. Donoho. Minimax risk of matrix denoising by singular value thresholding. *The Annals of Statistics*, 42(6):2413–2440, 2014.
- [7] Matan Gavish and David L. Donoho. The optimal hard threshold for singular values is  $4/\sqrt{3}$ . *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.
- [8] Matan Gavish and David L. Donoho. Optimal shrinkage of singular values. *IEEE Transactions on Information Theory*, 63(4):2137–2152, 2017.
- [9] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327, 2001.
- [10] Julie Josse and Sylvain Sardy. Adaptive shrinkage of singular values. *Statistics and Computing*, 26:715–724, 2016.
- [11] Julie Josse and Stefan Wager. Bootstrap-based regularization for low-rank matrix estimation. *Journal of Machine Learning Research*, 17:1–29, 2016.
- [12] William Leeb. Matrix denoising for weighted loss functions and heterogeneous signals. *arXiv:1902.09474*, 2020.
- [13] William Leeb and Elad Romanov. Optimal spectral shrinkage and PCA with heteroscedastic noise. *arXiv:1811.02201*, 2019.
- [14] D. J. C. MacKay. Deconvolution. In *Information Theory, Inference and Learning Algorithms*, pages 550–551. Cambridge University Press, Cambridge, UK, 2004.
- [15] Raj Rao Nadakuditi. OptShrink: An algorithm for improved low-rank signal matrix denoising by optimal, data-driven singular value shrinkage. *IEEE Transactions on Information Theory*, 60(5):3002–3018, 2014.
- [16] Debashis Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617–1642, 2007.
- [17] Andrey A. Shabalin and Andrew B. Nobel. Reconstruction of a low-rank matrix in the presence of Gaussian noise. *Journal of Multivariate Analysis*, 118:67–76, 2013.