Flat minima generalize for low-rank matrix recovery

Lijun Ding[*]   Dmitriy Drusvyatskiy[†]   Maryam Fazel[‡]   Zaid Harchaoui [§]

Abstract

Empirical evidence suggests that for a variety of overparameterized nonlinear models, most notably in neural network training, the growth of the loss around a minimizer strongly impacts its performance. Flat minima—those around which the loss grows slowly—appear to generalize well. This work takes a step towards understanding this phenomenon by focusing on the simplest class of overparameterized nonlinear models: those arising in low-rank matrix recovery. We analyze overparameterized matrix and bilinear sensing, robust PCA, covariance matrix estimation, and single hidden layer neural networks with quadratic activation functions. In all cases, we show that flat minima, measured by the trace of the Hessian, *exactly recover* the ground truth under standard statistical assumptions. For matrix completion, we establish weak recovery, although empirical evidence suggests exact recovery holds here as well. We conclude with synthetic experiments that illustrate our findings and discuss the effect of depth on flat solutions.

1 Introduction

Recent advances in machine learning and artificial intelligence have relied on fitting highly overparameterized models, notably deep neural networks, to observed data [28, 32, 54, 58]. In such settings, the number of parameters of the model is much greater than the number of data samples, thereby resulting in models that achieve near-zero training error. Although classical learning paradigms caution against overfitting, recent work suggests ubiquity of the "double descent" phenomenon [4], wherein significant overparameterization actually improves generalization. There is an important caveat, however, that is worth emphasizing. There is typically a continuum of models with zero training error; some of these models generalize well and some do not. Reassuringly, there is evidence that basic algorithms, such as the stochastic gradient method, are implicitly biased towards finding models that do generalize; see for example [18, 22, 23, 25, 27, 30, 31, 38, 39, 41, 50, 52]. Other seminal works [2, 3, 40] seeking to explain generalization have focused on quantifying stability, capacity, and margin bounds. Understanding generalization of overparameterized models remains an active area of research, and is the topic of our work.

Existing literature highlights two intriguing properties—**small norm** and **flat landscape**—that correlate with generalization [17, 19, 40]. Indeed, it has long been known that the magnitude of the weights plays an important role for neural network training. As a result, one typically incorporates a squared ℓ_2-penalty on the weights—called weight decay—when applying iterative methods.

One intuitive explanation is that minimizing the square Frobenius norm of the factors in matrix factorization problems is equivalent to minimizing the nuclear norm—a well-known regularizer for inducing low-rank structure [47]. Far reaching generalizations of this phenomenon for various neural network architectures have been recently pursued in [44, 45, 48]. In parallel, empirical evidence [26, 33, 42] strongly suggests that those models around which the landscape is flat—meaning the training loss grows slowly—generalize well. See Figure 1 for an illustration of flat and sharp minima. Inspired by this observation, a variety of algorithms have been proposed to explicitly bias the iterates towards flat solutions [12, 20, 29, 43], with impressive observed performance. In contrast to the magnitude of the weights, the theoretic basis for flatness is much less clear even for simple overparameterized nonlinear problems. The goal of our work is to answer the following question:

Do flat minimizers generalize for a broad family of overparameterized problems?
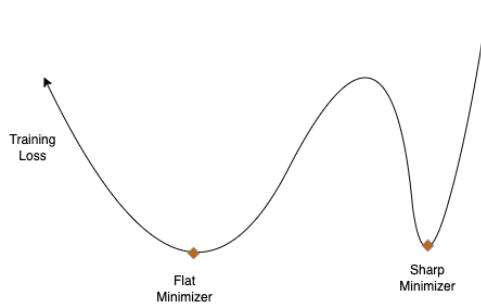

Figure 1: Flat vs. sharp minima of the training loss.

Putting generalization aside, one would hope that flat solutions are in some sense regular, occurring in a benign region where algorithms perform well. For example, numerical methods for neural network training are strongly influenced by how **balanced** the parameters appear. Namely, the set of interpolating neural networks contains models with consecutive weight matrices that are poorly scaled relative to each other [18, 49]. It has recently been shown that gradient descent in continuous time keeps the factors balanced [37, 57] for matrix factorization and for deep learning [18, 39]. Despite ubiquity of the three notions discussed so far—small norm, flatness, and balancedness— the exact relationship between them is unclear. Thus our secondary question is as follow:

Are flat minimizers nearly norm-minimal and nearly balanced
for a broad family of overparameterized problems?

1.1 Problem setting: overparameterized matrix factorization

We answer both questions in the setting of low-rank matrix factorization—a prototypical problem class often used to gain insight into more general deep learning models [18, 34, 57]. Setting the stage, consider a ground truth matrix $M_\natural \in \mathbb{R}^{d_1 \times d_2}$ with rank $r_\natural$. The goal is to recover $M_\natural$ from the observed measurements $b = \mathcal{A}(M_\natural)$ under a linear measurement map $\mathcal{A} \colon \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$. A common approach to this task is through the nonconvex optimization problem:

$$\min_{L,R} \quad f(L, R) := \left\| \mathcal{A}(LR^\top) - b \right\|_2^2 \qquad \text{with } L \in \mathbb{R}^{d_1 \times k} \text{ and } R \in \mathbb{R}^{d_2 \times k}. \tag{1}$$

2

The set of minimizers of f, which we denote by $\mathcal{S}$, consists of all solutions to the equation $\mathcal{A}(LR^\top) = b$. In order to model overparameterization, we focus on the rank-overparameterized setting $k \geq r_\natural$; indeed k can be arbitrarily large. The three notions discussed so for can be formally defined for pairs $(L, R) \in \mathcal{S}$ as follows.

- (L, R) is **norm-minimal** if it minimizes over $\mathcal{S}$ the square Frobenius norm $\|L\|_\mathrm{F}^2 + \|R\|_\mathrm{F}^2$.

- (L, R) is **balanced** if it satisfies $L^\top L = R^\top R$.

- (L, R) is **flat** if it minimizes over $\mathcal{S}$ the "scaled trace" of the Hessian, $\mathrm{str}(D^2 f(L, R))$.

Thus being norm-minimal means that (L, R) is the closest pair from $\mathcal{S}$ to the origin in Frobenius norm. Being balanced amounts to requiring L and R to have the same singular values and right-singular vectors. Flat solutions are defined in terms of the "scaled trace" of the bilinear form $D^2 f(L, R)$ defined as

$$\mathrm{str}(D^2 f(L, R)) := \tfrac{1}{d_1} \sum_{i \leq d_1, j \in [k]} D^2 f(L, R)[(e_i e_j^\top, 0_{d_2 \times k})] + \tfrac{1}{d_2} \sum_{i > d_1, j \in [k]} D^2 f(L, R)[(0_{d_1 \times k}, e_i e_j^\top)]. \quad (2)$$

where e_i and e_j are the unit coordinate vectors in $\mathbb{R}^{d_1 + d_2}$ and $\mathbb{R}^k$, respectively. In the square setting $d_1 = d_2 = d$, the scaled trace reduces to the usual trace divided by d. The scaled trace appears to have not been used previously in the literature, but is important in order to account for a possible mismatch in the dimension of the L and R factors. A number of recent papers use the trace of the Hessian to measure flatness (e.g. [17]). Other alternatives are possible, such as the maximal eigenvalue [17, 39] or the condition number [36], but we do not focus on them here. Our main contribution can be succinctly summarized as follows:

For various statistical models, flat solutions of (1) *exactly recover* $M_\natural$.
Moreover, flat solutions have nearly minimal norm and are almost balanced.

The exact recovery guarantee may be striking at first because flat solutions are distinct from minimal norm solutions, and thus do not correspond to nuclear norm minimization over $\mathcal{S}$. Yet, our main result shows that flat solutions do exactly recover the ground truth $M_\natural$ under standard statistical assumptions. The precise statistical models for which this is the case are matrix and bilinear sensing, robust PCA (or PCA with outliers), covariance matrix estimation, and single hidden layer neural networks with quadratic activation functions. Moreover, we prove weak recovery for the matrix completion problem, though our numerical experiments suggest that exact recovery holds here as well.

1.2 Main results and outline of the paper

We next outline our main results and the arguments that underpin them. We begin in Section 2 with the idealized "population level" setting where $\mathcal{A}$ is the identity map. In this case, we show that there is no distinction between flat, norm-minimal, and balanced solutions. As soon as $\mathcal{A}$ deviates from the identity, however, all three notions become distinct in general.

An immediate difficulty with analyzing flat solutions of the problem (1) with a general measurement map $\mathcal{A}$ is that flat solutions are defined as minimizers of a highly nonconvex optimization problem corresponding to minimizing the scaled trace over the solution set. In Section 3,
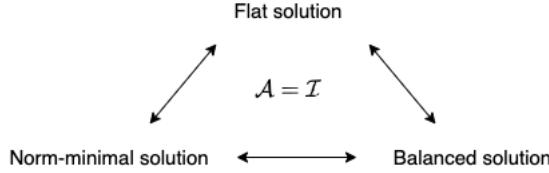
Figure 2: Equivalence between balanced, minimal norm, and flat solutions when $\mathcal{A} = \mathcal{I}$.

we derive a simple convex relaxation of flat minimizers. Setting the notation, let us write $\mathcal{A}$ as $\mathcal{A}(X) = (\langle A_i, X\rangle, \ldots, \langle A_m, X\rangle)$ for some matrices $A_i \in \mathbb{R}^{d_1 \times d_2}$ and define the "rescaling" matrices

$$D_1 := \left(\frac{1}{md_2}\sum_{i=1}^{m} A_i A_i^\top\right)^{\frac{1}{2}} \quad \text{and} \quad D_2 := \left(\frac{1}{md_1}\sum_{i=1}^{m} A_i^\top A_i\right)^{\frac{1}{2}}. \tag{3}$$

We will show in Theorem 3.2 that flat solutions can be identified with minimizers of the problem

$$\min_{X \in \mathbb{R}^{d_1 \times d_2}:\ \mathrm{rank}(X) \leq k} \|D_1 X D_2\|_* \qquad \text{subject to} \qquad \mathcal{A}(X) = b. \tag{4}$$

It is worthwhile to note that without the D_1 and D_2 matrices and without the rank constraint, the problem (4) is classically known to characterize norm-minimal solutions and is known as nuclear norm minimization. Herein, we already see the distinction between the two solution concepts. A natural convex relaxation for flat solutions simply drops the rank constraint:

$$\min_{X \in \mathbb{R}^{d_1 \times d_2}} \|D_1 X D_2\|_* \qquad \text{subject to} \qquad \mathcal{A}(X) = b. \tag{5}$$

Summarizing, verifying that flat solutions exactly recover $M_\flat$ is reduced to showing that $M_\flat$ (which has rank $r_\flat$) is the unique solution of the convex problem (5).

In Section 4, we will show that if the linear map $\mathcal{A}$ satisfies ℓ_2/ℓ_2 or ℓ_1/ℓ_2 restricted isometry properties (RIP) and the rescaling matrices D_1 and D_2 are sufficiently close to the identity, then $M_\flat$ is the unique solution of (5). As a consequence, we deduce that flat solutions exactly recover $M_\flat$ for matrix sensing [8, 47] and bilinear sensing [1, 35] problems with Gaussian design. The former corresponds to the setting where the entries of A_i are independent standard Gaussian random variables, while the latter corresponds to the setting $A_i = a_i b_i^\top$ where $a_i \in \mathbb{R}^{d_1}$ and $b_i \in \mathbb{R}^{d_2}$ are independent standard Gaussian vectors. The end result is the following theorem. Simplifying notation, we set $d_{\max} = \max\{d_1, d_2\}$ and $d_{\min} = \min\{d_1, d_2\}$.

Theorem 1.1 (Matrix and bilinear sensing (Informal))**.** *Suppose that $\mathcal{A}$ is generated according to a Gaussian matrix sensing or bilinear sensing model. Then as long as we are in the regime $m \gtrsim r_\flat d_{\max}$ and $d_{\min} \gtrsim \log m$, with high probability, any flat solution (L_f, R_f) satisfies $L_f R_f^\top = M_\flat$, and is nearly norm-minimal and nearly balanced.*

Note that our requirement on the sample size $m \gtrsim r_\flat d_{\max}$ matches the known regime for exact recovery with nuclear norm minimization [6, 8]. Since we are interested in the high dimensional regime, the extra condition $d_{\min} \gtrsim \log(m)$ can be assumed without harm. Appendix A presents a generalization of this result when the measurements b are corrupted by noise.

We next move on to analyzing the matrix completion problem in Section 5. We focus on the Bernoulli model, wherein each matrix A_i takes the form $A_i = \xi_{ij} e_i e_j^\top$, where e_i and e_j denote the i'th and j'th coordinate vectors in $\mathbb{R}^d$ and ξ_{ij} are independent Bernoulli random variables with

4

success probability $p \in (0,1)$. The main difficulty with analyzing the matrix completion problem is that the linear map $\mathcal{A}$ does not have good restricted isometry properties. Moreover, the existing techniques for analyzing the nuclear norm relaxation of the matrix completion problem [9, 46] do not directly apply to the problem (5) because of the dependence between the rescaling matrices D_1, D_2, and the observation map $\mathcal{A}$. Consequently, we settle for an approximate recovery guarantee.

Theorem 1.2 (Matrix completion (Informal)). *Suppose that $\mathcal{A}$ is generated from the Bernoulli matrix completion model with success probability $p > 0$ and let $\mu > 0$ be the incoherence parameter of $M_\natural$.[1] Then provided we are in the regime $p \gtrsim \frac{1}{\gamma}\sqrt{\frac{r_\natural \log(d_{\max})}{d_{\min}}}$, with high probability, any flat solution (L_f, R_f) satisfies $\left\| L_f R_f^\top - M_\natural \right\|_* \leq \gamma \|M_\natural\|_*$ and is nearly norm-minimal and nearly balanced.*

Hence according to this theorem, in order to conclude that flat solutions achieve a constant relative error, we must be in the regime $p \gtrsim \sqrt{\frac{r_\natural \log d_{\max}}{d_{\min}}}$. This is a stronger requirement than is needed for exact recovery of the ground truth matrix by nuclear norm minimization [13], which is $p \gtrsim \mu r_\natural \log(\mu r_\natural)\frac{\log(d_{\max})}{d_{\min}}$. We stress, however, that our numerical results suggest that flat solutions exactly recovery the ground truth matrix in this wider parameter regime.

We next focus on the problem of Robust Principal Component Analysis (PCA) in Section 6. Though this problem is not of the form (1), we will see that flat solutions (appropriately defined) exactly recover the ground truth under reasonable assumptions. Specifically, following [7, 11], the robust PCA problem asks to find a low-rank matrix $M_\natural \in \mathbb{R}^{d_1 \times d_2}$ that has been corrupted by sparse noise $S_\natural$. Thus, we observe a matrix $Y \in \mathbb{R}^{d_1 \times d_2}$ of the form

$$Y = M_\natural + S_\natural, \tag{6}$$

where the matrix $S_\natural$ is assumed to have at most $l_\natural$ nonzero entries in any column and in any row. A popular formulation of the problem (see [24, Eqn. (19)], [21, Eqn. (6)]) takes the form

$$\min_{L,R}\ f(L,R) := \text{dist}_\Omega^2(Y - LR^\top), \tag{7}$$

where dist_Ω^2 is the square Frobenius distance to the sparsity-inducing set $\Omega := \{S \mid \|S\|_{1,1} \leq \|S_\natural\|_{1,1}\}$. The objective function f is C^1-smooth but not C^2-smooth. Therefore, in order to measure flatness, we approximate f near a basepoint $(\tilde{L}, \tilde{S})$ by a certain C^2-smooth local model $f_{\tilde{L},\tilde{R}}(L,R)$, introduced in [24, Section 4.2], [21, Section 4.3]. See Section 6 for a precise definition of $f_{\tilde{L},\tilde{R}}(L,R)$. We then define a minimizer of (7) to be *flat* if it minimizes the scaled trace $\text{str}(D^2 f_{L,R}(L,R))$ over all $(L,R) \in \mathcal{S}$. We will prove the following theorem, which largely follows from the results of [14].

Theorem 1.3 (Robust PCA (Informal)). *Let μ be the strong incoherence parameter of $M_\natural$.[2] Then, in the regime $l_\natural \lesssim \frac{d_{\min}}{\mu r_\natural}$, any flat minimizer (L_f, R_f) satisfies $L_f R_f^\top = M_\natural$.*

Section 7 analyzes the last problem class of the paper, motivated by the problems of covariance matrix estimation and training of shallow neural networks. Setting the stage, consider a ground truth matrix $M_\natural$ satisfying

$$M_\natural = U_{1,\natural}U_{1,\natural}^\top - U_{2,\natural}U_{2,\natural}^\top,$$

for some matrices $U_{1,\natural} \in \mathbb{R}^{d \times r_1}$ and $U_{2,\natural} \in \mathbb{R}^{d \times r_2}$. The goal is to recover $M_\natural$ from the observations

$$b_i = x_i^\top M_\natural x_i, \tag{8}$$

---

[1]See (34) for the definition of the incoherence parameter μ.
[2]See (48) for the definition of the strong incoherence parameter μ.

where $x_1, \ldots, x_m \overset{\text{iid}}{\sim} N(0, I_d)$. Note that in the special case $r_2 = 0$, this problem reduces to covariance matrix estimation [15] and further reduces to phase retrieval when $r_1 = 1$ [10]. The added generality allows to also model shallow neural networks with quadratic activation functions; see details below. A natural optimization formulation of the problem takes the form

$$\min_{U_1 \in \mathbb{R}^{d \times k_1}, \, U_2 \in \mathbb{R}^{d \times k_2}} \quad f(U_1, U_2) = \frac{1}{m} \left\| \mathcal{A}(U_1 U_1^\top - U_2 U_2^\top - M_\natural) \right\|_2^2, \tag{9}$$

where the sensing matrices are $A_i = x_i x_i^\top$ and $k_i \geq r_i$ for $i = 1, 2$. Since $\text{str}(D^2 f(U_1, U_2)) = d \, \text{tr}(D^2 f(U_1, U_2))$, we declare a minimizer $(U_{1,f}, U_{2,f})$ to be *flat* if it has minimal trace $\text{tr}(D^2 f(U_1, U_2))$ among all minimizers of (9). We prove the following.

Theorem 1.4 (Exact recovery). *In the regime $m \gtrsim C(r_1 + r_2)d$ and $d \gtrsim C \log m$, with high probability, any flat solution $(U_{f,1}, U_{f,2})$ of (9) satisfies $U_{f,1} U_{f,1}^\top - U_{f,2} U_{f,2}^\top = M_\natural$.*

Here, our requirement on the sample size $m \gtrsim C(r_1 + r_2)d$ coincides with the known requirement for exact recovery by nuclear norm minimization [15] in terms of r and d. An interesting example of (9) arises from a model of shallow neural networks, analyzed in [34, 51] for the purpose of studying energy landscape around saddle points. Namely, suppose that given an input vector $x \in \mathbb{R}^d$ a response vector $y(x)$ is generated by the "teacher neural network"

$$y(U_\natural, x) = v^\top q(U_\natural^\top x),$$

where the output weight vector $v \in \mathbb{R}^r$ has r_1 positive entries and r_2 negative entries, the hidden layer weight matrix $U_\natural$ has dimensions $d \times r_\natural$, and we use a quadratic activation $q(s) = s^2$ applied coordinate-wise. We get to observe a set of m pairs $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, where the features x_i are drawn as $x_i \overset{\text{iid}}{\sim} N(0, I_d)$ and the output values are $y_i = y(x_i)$. The goal is to fit the data with an overparameterized "student neural network"

$$\hat{y}(U, x) = u^\top q(U^\top x),$$

with hidden weights $U \in \mathbb{R}^{d \times k}$ and output layer weights $u = (\mathbf{1}_{k_1}, -\mathbf{1}_{k_2})$, where $k_1 \geq r_1$, and $k_2 \geq r_2$. It is straightforward to see that by partitioning the matrix $U = [U_1, U_2]$, this problem is exactly equivalent to recovering the matrix $M_\natural = U_\natural \, \text{diag}(v) U_\natural^\top$ from the observations (8).

Section 8 numerically validates our theoretical results. Section 9 summarizes our findings and speculates about the role of depth on generalization properties of flat solutions.

Notation. Throughout, we let $\mathbb{R}^d$ denote the d-dimensional Euclidean space, equipped with the usual dot-product $\langle x, y \rangle = x^\top y$ and the induced Euclidean norm $\| \cdot \|_2$. More generally, the symbol $\| \cdot \|_p$ will denote the ℓ_p norm on $\mathbb{R}^d$. Given two numbers d_1 and d_2, which will be clear from context, we set $d_{\max} := \max\{d_1, d_2\}$ and $d_{\min} := \min\{d_1, d_2\}$. The Euclidean space of $d_1 \times d_2$ real matrices $\mathbb{R}^{d_1 \times d_2}$ will always be equipped with the trace inner product $\langle X, Y \rangle = \text{tr}(X^\top Y)$ and the induced Frobenius norm $\|X\|_{\text{F}} = \sqrt{\langle X, X \rangle}$. The nuclear norm $\|X\|_*$ of any matrix $X \in \mathbb{R}^{d_1 \times d_2}$ is the sum of its singular values. We will often use the characterization of the nuclear norm [53, Lemma 1]:

$$\|X\|_* = \min_{X = LR^\top} \|L\|_{\text{F}} \|R\|_{\text{F}} = \min_{X = LR^\top} \tfrac{1}{2} (\|L\|_{\text{F}}^2 + \|R\|_{\text{F}}^2). \tag{10}$$

2 Norm-minimal, flat, and balanced solutions with an identity measurement map

In this section, we focus on the idealized objective (1) where the measurement map $\mathcal{A}$ is the identity:

$$\min_{L \in \mathbb{R}^{d_1 \times k}, \ R \in \mathbb{R}^{d_2 \times k}} f(L, R) = \left\| LR^\top - M_\natural \right\|_F^2, \tag{11}$$

Recall that $M_\natural \in \mathbb{R}^{d_1 \times d_2}$ is a rank $r_\natural$ matrix, $k \geq r_\natural$ is arbitrary, and the set of minimizers $\mathcal{S}$ of (11) coincides with the solution set of the equation $LR^\top = M_\natural$. We will show in this section that in this setting there is no distinction between norm-minimal, flat, and balanced solutions. As soon as the measurement map $\mathcal{A}$ is not the identity, the three notions become distinct; this remains true even under standard statistical models as our numerical experiments show. Nonetheless, the simplified setting $\mathcal{A} = \mathcal{I}$ explored in this section will serve as motivation for the rest of the paper.

We begin with the following lemma that provides a convenient expression for $\mathrm{str}(D^2 f(L, R))$.

Lemma 2.1 (Scaled trace). *The second-order derivative of the function f at any $(L, R) \in \mathcal{S}$ is the quadratic form:*

$$D^2 f(L, R)[U, V] = 2 \left\| LV^\top + UR^\top \right\|_F^2. \tag{12}$$

Consequently, the scaled trace is simply

$$\mathrm{str}(D^2 f(L, R)) = 2(\|L\|_F^2 + \|R\|_F^2). \tag{13}$$

Proof. A straightforward computation shows for any pair (L, R) the expression

$$D^2 f(L, R)[U, V] = 4\langle LR^\top - M_\natural, UV^\top \rangle + 2 \left\| LV^\top + UR^\top \right\|_F^2.$$

For pairs $(L, R) \in \mathcal{S}$, the first term on the right is zero yielding the claimed expression (12). To see the expression for the scaled trace, let $e_i \in \mathbb{R}^{d_1 + d_2}$ and $e_j \in \mathbb{R}^k$ be the i'th and j'th coordinate vectors. A quick computation shows $D^2 f(L, R)[e_i e_j^\top] = 2\|R_j\|_F^2$ for $i \leq d_1$ and $D^2 f(L, R))[e_i e_j^\top] = 2\|L_j\|_F^2$ for $i > d_1$. Therefore, from the definition (2), the scaled trace becomes

$$\mathrm{str}(D^2 f(L, R)) := \frac{1}{d_1} \sum_{i \leq d_1} \sum_{j \in [k]} D^2 f(L, R)[e_i e_j^\top] + \frac{1}{d_2} \sum_{i > d_1} \sum_{j \in [k]} D^2 f(L, R)[e_i e_j^\top] = 2(\|L\|_F^2 + \|R\|_F^2),$$

as claimed. $\qquad\square$

We are now ready to prove the claimed equivalence between the three properties.

Lemma 2.2 (Equivalence). *Norm-minimal, flat, and balanced solutions of (11) all coincide.*

Proof. First, the equivalence of flat and norm-minimal solutions follows directly from the expression (13) in Lemma 2.1. Next, we prove the equivalence between minimal norm and balanced solutions. Suppose $(L, R) \in S$ is balanced. The equality $L^\top L = R^\top R$ implies that L and R have the same nonzero singular values and the same set of right singular vectors. Therefore, we may form compact singular value decompositions $L = U_1 \Sigma V^\top$ and $R = U_2 \Sigma V^\top$. Since equality $LR^\top = M_\natural$ holds, we see that $U_1 \Sigma^2 U_2^\top = M_\natural$. Hence, the nuclear norm of $M_\natural$ is simply $\|M_\natural\|_* = \mathrm{tr}(\Sigma^2)$. Noting the equality $\frac{1}{2}\left(\|L\|_F^2 + \|R\|_F^2\right) = \mathrm{tr}(\Sigma^2)$ along with (10), we deduce that (L, R) is a minimal norm

solution, as claimed. Conversely, suppose that (L, R) is a minimal norm solution. Define the function

$$\varphi(B) = \frac{1}{2}\|LB\|_{\mathrm{F}}^2 + \frac{1}{2}\left\|RB^{-\top}\right\|_{\mathrm{F}}^2,$$

over the open set of $k \times k$ invertible matrices B. Clearly $B = I_k$ is a local minimizer of φ and therefore $\nabla\varphi(I_k)$ must be the zero matrix. A quick computation yields the expression $\nabla\varphi(I_k) = L^\top L - R^\top R$, and therefore (L, R) is balanced, as claimed. $\qquad\square$

3   Convex relaxation and regularity of flat solutions

In this section, we begin investigating flat minimizers of the problem (1) with general linear measurement maps $\mathcal{A}$. It will be convenient to write the linear map $\mathcal{A}(X)$ in coordinates as

$$\mathcal{A}(X) = (\langle A_1, X\rangle, \langle A_2, X\rangle \ldots, \langle A_m, X\rangle),$$

where $A_i \in \mathbb{R}^{d_1 \times d_2}$ are some matrices. As always, $\mathcal{S}$ denotes the set of solutions to the equation $\mathcal{A}(LR^\top) = b$. We will make use of the following two "rescaling" matrices:

$$D_1 = \left(\frac{1}{md_2}\sum_{i=1}^{m} A_i A_i^\top\right)^{\frac{1}{2}} \quad \text{and} \quad D_2 = \left(\frac{1}{md_1}\sum_{i=1}^{m} A_i^\top A_i\right)^{\frac{1}{2}}. \tag{14}$$

The section presents two main results: Theorems 3.2 and 3.3. The former presents a convex relaxation for verifying that a solution is flat, while the latter shows that flat solutions are nearly balanced and nearly norm-minimal, whenever the matrices D_1 and D_2 are well-conditioned.

3.1   A convex relaxation for flat solutions

Flat solutions are by definition minimizers of the highly nonconvex problem $\min_{(L,R)\in\mathcal{S}} \mathrm{str}(D^2 f(L, R))$. The main result of this section is to present an appealing convex relaxation of this problem. We begin with a convenient expression for the scaled trace $\mathrm{str}(D^2 f(L, R))$. Namely, recall that Lemma 2.1 showed the equality $\mathrm{str}(D^2 f(L, R)) = 2\|L\|_{\mathrm{F}}^2 + 2\|R\|_{\mathrm{F}}^2$ in the simplified setting $\mathcal{A} = \mathcal{I}$. Lemma 3.1 provides an analogous statement for general maps $\mathcal{A}$ up to rescaling the factors by D_1 and D_2.

Lemma 3.1 (Scaled trace and the Frobenius norm)**.** *The second-order derivative of the function f at any $(L, R) \in \mathcal{S}$ is the quadratic form:*

$$D^2 f(L, R)[U, V] = 2\|\mathcal{A}(LV^\top + UR^\top)\|_2^2. \tag{15}$$

Moreover, the scaled trace can be written as

$$\mathrm{str}(D^2 f(L, R)) = 2m(\|D_1 L\|_{\mathrm{F}}^2 + \|D_2 R\|_{\mathrm{F}}^2). \tag{16}$$

Proof. An elementary computation yields for any (L, R) the expression

$$D^2 f(L, R)[U, V] = 4\langle\mathcal{A}(LR^\top) - b, \mathcal{A}(LV^\top + UR^\top)\rangle + 2\|\mathcal{A}(LV^\top + UR^\top)\|_2^2.$$

Noting that for any $(L, R) \in \mathcal{S}$ the first term on the right is zero yields the claimed expression (15). Next, we verify (16) by a direct calculation. To this end, the definition of the scaled trace (2) yields the expression

$$\text{str}(\mathcal{H}(L,R)) = \frac{2}{d_1}\sum_{i=1}^{d_1}\sum_{j=1}^{k}\left\|\mathcal{A}(e_i e_j^\top R^\top)\right\|_2^2 + \frac{2}{d_2}\sum_{i=1}^{d_2}\sum_{j=1}^{k}\left\|\mathcal{A}(L e_j e_i^\top)\right\|_2^2. \tag{17}$$

Let us analyze the second term on the right. Letting $A_{l,i}$ denote the i'th column of A_l, we compute

$$
\begin{aligned}
\sum_{i=1}^{d_2}\sum_{j=1}^{k}\left\|\mathcal{A}(L e_j e_i^\top)\right\|_2^2 &= \sum_{i=1}^{d_2}\sum_{j=1}^{k}\sum_{l=1}^{m}\langle A_l, L e_j e_i^\top\rangle^2 \\
&= \sum_{i=1}^{d_2}\sum_{j=1}^{k}\sum_{l=1}^{m}\langle A_{l,i}, L_j\rangle^2 \\
&= \sum_{i=1}^{d_2}\sum_{j=1}^{k}\sum_{l=1}^{m}\langle A_{l,i} A_{l,i}^\top, L_j L_j^\top\rangle \\
&= \sum_{l=1}^{m}\langle\sum_{i=1}^{d_2} A_{l,i} A_{l,i}^\top, \sum_{j=1}^{k} L_j L_j^\top\rangle = \sum_{l=1}^{m}\langle A_l A_l^\top, LL^\top\rangle = m d_2\|D_1 L\|_{\mathrm{F}}^2.
\end{aligned}
\tag{18}
$$

A similar argument shows $\|D_2 R\|_{\mathrm{F}}^2 = \frac{1}{m d_1}\sum_{i=1}^{d_1}\sum_{j=1}^{k}\left\|\mathcal{A}(e_i e_j^\top R^\top)\right\|_2^2$, completing the proof. $\qquad\square$

In particular, Lemma 3.1 implies that flat solutions are exactly the minimizers of the problem

$$\min_{L,R}\ \frac{1}{2}\Big(\|D_1 L\|_{\mathrm{F}}^2 + \|D_2 R\|_{\mathrm{F}}^2\Big) \qquad \text{subject to} \qquad \mathcal{A}(L R^\top) = b. \tag{19}$$

In turn, it follows directly from (10) that so long as D_1, D_2 are invertible, the problem (19) is equivalent to minimizing the nuclear norm over rank constrained matrices:

$$\min_{X\in\mathbb{R}^{d_1\times d_2}:\ \mathrm{rank}(X)\le k}\|X\|_* \qquad \text{subject to} \qquad \mathcal{A}(D_1^{-1} X D_2^{-1}) = b. \tag{20}$$

Therefore, a natural convex relaxation for finding the flattest solution drops the rank constraint:

$$\min_{X\in\mathbb{R}^{d_1\times d_2}}\|X\|_* \qquad \text{subject to} \qquad \mathcal{A}(D_1^{-1} X D_2^{-1}) = b. \tag{21}$$

The following theorem summarizes these observations.

Theorem 3.2 (Convex relaxation). *Suppose the matrices D_1 and D_2 are invertible. Then the problems (19) are (20) are equivalent in the following sense. Let $l = \min(k, d_{\min})$.*

1. *the optimal values of (19) and (20) are equal,*

2. *if L, R solves (19), then $X = D_1 L R^\top D_2$ is a minimizer of (20).*

3. *if a solution X of (20) has a singular value decomposition $X = U\Sigma V^\top$ for some diagonal matrix $\Sigma \in \mathbb{R}^{l\times l}$ with nonnegative entries, then the matrices $L = D_1^{-1}U\sqrt{\Sigma}$ and $R = D_2^{-1}V\sqrt{\Sigma}$ are minimizers of (19) when $l \ge k$, and the matrices $L = [D_1^{-1}U\sqrt{\Sigma}, 0_{d_1,(k-l)}]$ and $R = [D_2^{-1}V\sqrt{\Sigma}, 0_{d_2,(k-l)}]$ are minimizers of (19) when $l < k$.*

Moreover, if $X = D_1 M_\natural D_2$ is the unique minimizer of the problem (21), then any flat solution (L, R) satisfies $L R^\top = M_\natural$.

Proof. The three claims follow directly from making a variable substitution $L' = D_1 L$ and $R' = D_2 R$ and using (10). The "moreover" part follows from (21) being a convex relaxation of (20). $\quad\square$

Section 4 will verify that the convex relaxation (21) indeed recovers $M_\natural$ under restricted isometry properties on the measurement map $\mathcal{A}$, and therefore flat solutions exactly recover $M_\natural$.

3.2 Regularity of flat solutions

In this section, we show that the condition numbers of the rescaling matrices D_1 and D_2 determine balancedness and norm minimality of flat solutions. The main result is the following theorem.

Theorem 3.3 (Regularity of flat solutions). *Suppose that there exist constants $\alpha_1, \alpha_2 > 0$ satisfying $\alpha_1 I \preceq D_i \preceq \alpha_2 I$ for each $i \in \{1, 2\}$. Define the constant $\kappa := \frac{\alpha_2}{\alpha_1}$. Then any flat solution (L_f, R_f) of* (1) *satisfies the following properties.*

1. **Norm-minimal***: the pair (L_f, R_f) is approximately norm-minimal:*

$$\|L_f\|_F^2 + \|R_f\|_F^2 \leq \kappa^2 \cdot \left(\min_{\mathcal{A}(LR^\top)=b} \|L\|_F^2 + \|R\|_F^2 \right). \tag{22}$$

2. **Balanced***: The pair (L_f, R_f) is approximately balanced:*

$$\left\| L_f^\top L_f - R_f^\top R_f \right\|_* \leq 2(\kappa^2 - 1)\|M_\natural\|_*. \tag{23}$$

The proof of Theorem 3.3 relies on the following simple linear algebraic lemma.

Lemma 3.4. *Consider two symmetric matrices $Q_1 \in \mathbb{R}^{d_1 \times d_1}$ and $Q_2 \in \mathbb{R}^{d_2 \times d_2}$. Suppose that there exist constants $\alpha_1, \alpha_2 > 0$ satisfying $\alpha_1 I \preceq Q_i \preceq \alpha_2 I$ for each $i \in \{1, 2\}$. Define the constant $\kappa = \frac{\alpha_2}{\alpha_1}$. Then given any matrix $X \in \mathbb{R}^{d_1 \times d_2}$, any minimizer (L, R) of the problem*

$$\min_{\tilde{L}, \tilde{R}: \, Q_1 \tilde{L} \tilde{R}^\top Q_2 = X} \frac{1}{2} \left(\left\| Q_1 \tilde{L} \right\|_F^2 + \left\| Q_2 \tilde{R} \right\|_F^2 \right), \tag{24}$$

satisfies the inequality:

$$\left\| L^\top L - R^\top R \right\|_* \leq \left(1 - \kappa^{-2} \right) \left(\|L\|_F^2 + \|R\|_F^2 \right). \tag{25}$$

Proof. Lemma 2.2 implies that the pair $(Q_1 L, Q_2 R)$ is balanced, meaning $L^\top Q_1^2 L = R^\top Q_2^2 R$. Hence, we may decompose $L^\top L - R^\top R$ in the following way:

$$L^\top L - R^\top R = \left(L^\top L - \frac{L^\top Q_1^2 L}{\alpha_2^2} \right) + \left(\frac{R^\top Q_2^2 R}{\alpha_2^2} - R^\top R \right). \tag{26}$$

We bound the first term on the right as follows,

$$\begin{aligned}
\left\| L^\top L - \frac{L^\top Q_1^2 L}{\alpha_2^2} \right\|_* = \left\| L^\top \left(I - \frac{1}{\alpha_2^2} Q_1^2 \right) L \right\|_* &\overset{(a)}{\leq} \left\| L^\top \left(I - \frac{1}{\alpha_2^2} Q_1^2 \right) \right\|_F \|L\|_F \\
&\overset{(b)}{\leq} \left\| I - \frac{1}{\alpha_2^2} Q_1^2 \right\|_{op} \|L\|_F^2 \\
&\leq \left(1 - \kappa^{-2} \right) \|L\|_F^2.
\end{aligned} \tag{27}$$

Here, (a) and (b) follow, respectively, from the basic inequalities: $\|FG\|_* \leq \|F\|_F \|G\|_F$ and $\|FG\|_F \leq \|F\|_{op} \|G\|_F$, which hold for all matrices F and G with compatible dimensions. A similar argument yields the inequality

$$\left\| R^\top R - \frac{R^\top Q_2^2 R}{\alpha_2^2} \right\|_* \leq \left(1 - \kappa^{-2} \right) \|R\|_F^2. \tag{28}$$

The claimed estimate (25) follows immediately. $\qquad\square$

We are now ready to prove Theorem 3.3.

Proof of Theorem 3.3. We first prove inequality (22). To this end, for any $(L, R) \in \mathcal{S}$, we successively estimate:

$$
\begin{aligned}
\alpha_1^2 \Big(\|L_f\|_{\mathrm{F}}^2 + \|R_f\|_{\mathrm{F}}^2 \Big) &\leq \|D_1 L_f\|_{\mathrm{F}}^2 + \|D_2 R_f\|_{\mathrm{F}}^2 \\
&\leq \|D_1 L\|_{\mathrm{F}}^2 + \|D_2 R\|_{\mathrm{F}}^2 \\
&\leq \alpha_2^2 \Big(\|L\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2 \Big),
\end{aligned}
\tag{29}
$$

where the second inequality follows from the characterization (19) of flat solutions. Taking the infimum over pairs $(L, R) \in \mathcal{S}$ completes the proof of (22).

We next verify (23). To this end, define the matrix $X := D_1 L_f R_f^\top D_2$. Then clearly (L_f, R_f) is a minimizer of the problem

$$
\min_{\tilde{L}, \tilde{R}: \ D_1 \tilde{L} \tilde{R}^\top D_2 = X} \frac{1}{2} \Big(\big\| D_1 \tilde{L} \big\|_{\mathrm{F}}^2 + \big\| D_2 \tilde{R} \big\|_{\mathrm{F}}^2 \Big).
\tag{30}
$$

Lemma 3.4 therefore guarantees the estimate

$$
\Big\| L_f^\top L_f - R_f^\top R_f \Big\|_* \leq \big(1 - \kappa^{-2} \big) \Big(\|L_f\|_{\mathrm{F}}^2 + \|R_f\|_{\mathrm{F}}^2 \Big).
$$

The already established estimate (22) ensures

$$
\|L_f\|_{\mathrm{F}}^2 + \|R_f\|_{\mathrm{F}}^2 \leq \kappa^2 \Big(\|L\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2 \Big) \qquad \forall (L, R) \in \mathcal{S}.
$$

In particular, minimizing the right hand-side over L, R satisfying $M_\natural = L R^\top$ yields an upper bound of $2\kappa^2 \|M_\natural\|_*$. The proof is complete. $\qquad \square$

4 Flat minima under RIP conditions: matrix and bilinear sensing

The previous section motivates a two-part strategy for showing that flat minima exactly recover the ground truth (Theorem 3.2) and are automatically nearly balanced and nearly norm-minimal (Theorem 3.3). The first step is to argue that the convex relaxation (21) admits $D_1 M_\natural D_2$ as its unique minimizer. The second step is to argue that the condition numbers of the matrices D_1 and D_2 are close to one. In this section, we follow this recipe for problems satisfying ℓ_2/ℓ_2 and ℓ_1/ℓ_2 restricted isometry properties (defined below). The main two examples will be the following random ensembles.

Definition 4.1 (Matrix and bilinear sensing)**.** We introduce the following definitions.

1. We say that $\mathcal{A}$ is a *Gaussian ensemble* if the entries of A_i are i.i.d standard normal random variables $N(0, 1)$.

2. We say that $\mathcal{A}$ is a *Gaussian bilinear ensemble* if the matrices A_i take the form $A_i = a_i b_i^\top$ where the entries of a_i and b_i are i.i.d. standard normal random variables $N(0, 1)$

The main results of the section is the following theorem, stated here informally.

Theorem 4.2 (Matrix and bilinear sensing (Informal))**.** *Suppose we are in one of the settings:*

1. $\mathcal{A}$ is a Gaussian ensemble and $m \gtrsim r_\natural d_{\max}$,

2. $\mathcal{A}$ is a Gaussian bilinear ensemble, $m \gtrsim r_\natural d_{\max}$, and $d_{\min} \gtrsim \log m$.

Then with high probability, any flat solution (L_f, R_f) of (1) satisfies $L_f R_f^\top = M_\natural$, and is nearly norm-minimal and nearly balanced:

$$\|L_f\|_F^2 + \|R_f\|_F^2 \le (1 + \delta)\left(\min_{\mathcal{A}(LR^\top)=b} \|L\|_F^2 + \|R\|_F^2\right)$$

$$\left\|L_f^\top L_f - R_f^\top R_f\right\|_* \lesssim \delta \|M_\natural\|_*.$$

We begin by formally defining the restricted isometry property of a measurement map $\mathcal{A}(\cdot)$.

Definition 4.3 (Restricted isometry property). A linear map $\mathcal{A} \colon \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ satisfies an ℓ_p/ℓ_2 *restricted isometry property (RIP) with parameters* (r, δ_1, δ_2) *if the estimate*

$$\delta_1 \|X\|_{\mathrm{F}} \le \frac{\|\mathcal{A}(X)\|_p}{m^{1/p}} \le \delta_2 \|X\|_{\mathrm{F}}, \tag{31}$$

holds for all matrices $X \in \mathbb{R}^{d_1 \times d_2}$ with rank at most r.

In this paper we will be primarily interested in ℓ_2/ℓ_2 and ℓ_1/ℓ_2 restricted isometry properties. In particular, the two random measurement models in Theorem 4.2 satisfy these two properties. The following two lemmas are from [8, Theorem 2.3], [47, Theorem 4.2], and [6, Theorem 2.2].

Lemma 4.4 (ℓ_2/ℓ_2 RIP in matrix sensing). *Let $\mathcal{A}$ be a Gaussian ensemble. Then for any $\delta \in (0, 1)$, there exist constants $c, C > 0$ depending only on δ such that as long as $m \ge cr(d_1 + d_2)$, with probability at least $1 - \exp(-Cm)$, the linear map $\mathcal{A}$ satisfies ℓ_2/ℓ_2 RIP with parameters $(2r, 1 - \delta, 1 + \delta)$.*

Lemma 4.5 (ℓ_1/ℓ_2 RIP in bilinear sensing). *Let $\mathcal{A}$ be a Gaussian bilinear ensemble. For any positive integer $k \ge 2$, there exist constants $c, C > 0$ depending only on k and numerical constants $\delta_1, \delta_2 > 0$ and such that in the regime $m \ge cr(d_1 + d_2)$, with probability at least $1 - \exp(-Cm)$, the measurement map $\mathcal{A}$ satisfies ℓ_1/ℓ_2 RIP with parameters (kr, δ_1, δ_2).*

Our goal is to show that under RIP conditions, with reasonable parameters, flat solutions exactly recover the ground truth $M_\natural$. We will need the following lemma, whose proof is immediate from definitions.

Lemma 4.6. *Let $\mathcal{A}(\cdot)$ be a linear map satisfying an ℓ_p/ℓ_2 RIP with parameters (r, δ_1, δ_2). Let Q_1, Q_2 be two positive definite matrices satisfying $\alpha_1 I \preceq Q_i \preceq \alpha_2 I$ for all $i \in \{1, 2\}$ for some $\alpha_1, \alpha_2 > 0$. Then the linear map $\mathcal{B}(\cdot) := \mathcal{A}(Q_1^{-1} \cdot Q_2^{-1})$ satisfies an ℓ_p/ℓ_2 RIP with parameters $(r, \alpha_2^{-2}\delta_1, \alpha_1^{-2}\delta_2)$.*

The following lemma will be our main technical tool; it establishes that if $\mathcal{A}(\cdot)$ satisfies RIP, then so does the perturbed map $\mathcal{B}(\cdot) = \mathcal{A}(Q_1^{-1} \cdot Q_2^{-1})$, provided that the condition numbers of the positive definite matrices Q_1 and Q_2 are sufficiently close to one.

Lemma 4.7. *Consider two positive definite matrices Q_1, Q_2 and constants $\alpha_1, \alpha_2 > 0$ satisfying $\alpha_1 I \preceq Q_i \preceq \alpha_2 I$ for each $i \in \{1, 2\}$. Define $\kappa = \alpha_2/\alpha_1$ and let $\mathcal{A}(\cdot)$ be a linear map satisfying one of the following conditions.*

1. *The map $\mathcal{A}$ satisfies ℓ_2/ℓ_2 RIP with parameters $(5r_\natural, \delta_1, \delta_2)$, where $\delta_1 \ge \frac{9}{10}$ and $\delta_2 \kappa^2 \le \frac{11}{10}$.*

2. *The map $\mathcal{A}$ satisfies ℓ_1/ℓ_2 RIP with parameters $(lr_\natural, \delta_1, \delta_2)$, where $\frac{\delta_2 \kappa^2}{\delta_1} < \sqrt{l}$.*

12

Then $Q_1 M_\natural Q_2$ is the unique solution of the following convex program

$$\min_{X \in \mathbb{R}^{d_1 \times d_2}} \|X\|_* \qquad subject\ to \qquad \mathcal{A}(Q_1^{-1} X Q_2^{-1}) = b. \tag{32}$$

Proof. Define the map $\mathcal{B}(Z) = \alpha_2^2 \mathcal{A}(Q_1^{-1} Z Q_2^{-1})$. Then Lemma 4.6 implies that $\mathcal{B}$ in the first case satisfies ℓ_2/ℓ_2 RIP with parameters $(5r_\natural, \delta_1, \delta_2 \kappa^2)$ and in the second case satisfies ℓ_1/ℓ_2 RIP with parameters $(lr_\natural, \delta_1, \delta_2 \kappa^2)$. An application of [47, Theorem 3.3] in the first case and [6, Theorem 2.1] in the second guarantees that $Q_1 M_\natural Q_2$ is the unique solution of (32), as claimed. $\square$

It remains to estimate the condition number κ of the matrices D_1 and D_2 defined in (14) under RIP (or statistical assumptions). The following lemma shows that D_1 and D_2 are automatically well conditioned under ℓ_2/ℓ_2 RIP.

Lemma 4.8 (Conditioning of D_i under ℓ_2/ℓ_2 RIP). *Suppose that the linear map $\mathcal{A}$ satisfies ℓ_2/ℓ_2 RIP with parameters $(1, \delta_1, \delta_2)$. Then the relation $\delta_1 I_{d_i} \preceq D_i \preceq \delta_2 I_{d_i}$ holds for all $i \in \{1, 2\}$*

Proof. In the proof of Lemma 3.1 (equation (18)), we actually showed the expression:

$$\sum_{i=1}^{d_2} \sum_{j=1}^{k} \left\| \mathcal{A}(L e_j e_i^\top) \right\|_2^2 = m d_2 \|D_1 L\|_{\mathrm{F}}^2 \qquad \forall L \in \mathbb{R}^{d_1 \times k}.$$

Now for any vector $v \in \mathbb{R}^{d_1}$, we can take the matrix $L = [v, 0_{d_1, k-1}]$ in the above equation. Appealing to the ℓ_2/ℓ_2 RIP condition on $\mathcal{A}$, we deduce

$$\delta_1^2 \|v\|_2^2 \leq \|D_1 v\|_2^2 \leq \delta_2^2 \|v\|_2^2.$$

A similar argument shows that D_2 satisfies the analogous inequality with $v \in \mathbb{R}^{d_2}$. $\square$

The ℓ_1/ℓ_2 RIP property does not in general imply a good bound on the condition numbers of D_1 and D_2. Instead we will directly show that under the Gaussian design for bilinear sensing, the matrices D_1 and D_2 are well-conditioned. This is the content of the following lemma.

Lemma 4.9 (Conditioning of D_i for bilinear sensing). *Let $\mathcal{A}$ be a Gaussian bilinear ensemble. Then there exist constant $c_1, c_2, c_3, c_4 > 0$ such that for any $\delta \in (0, 1)$ as long as we are in the regime $m \geq \frac{c_3 d_{\max}}{\delta^2}$ and $\log(m) \leq c_4 \delta^2 d_{\min}$, the estimate holds:*

$$\mathbb{P}\{(1-\delta) I_{d_i} \preceq D_i \preceq (1+\delta) I_{d_i} \ \ \forall i \in \{1, 2\}\} \geq 1 - c_2 e^{-c_1 d_{\min} \delta^2}.$$

Proof. First observe $A_i A_i^\top = \|b_i\|_2^2 a_i a_i^\top$ for each index i. Bernstein's inequality [55, Theorem 2.8.3] implies

$$\mathbb{P}\left\{ \left| \frac{1}{d_2} \|b_i\|_2^2 - 1 \right| \geq \delta \right\} \leq 2 \exp(-c_1 d_2 \delta^2) \qquad \forall \delta > 0, \ \forall i \in [m].$$

Taking a union bound, we can therefore be sure that with probability at least $1 - m \exp(-c_1 d_2 \delta^2)$ the estimate

$$1 - \delta \leq \frac{\|b_i\|_2^2}{d_2} \leq 1 + \delta$$

holds simultaneously for all $i = 1, \ldots, m$. In this event, we estimate

$$(1-\delta) a_i a_i^\top \preceq \frac{1}{d_2} A_i A_i^\top \preceq (1+\delta) a_i a_i^\top$$

13

Therefore, after summing for $i = 1, \ldots, m$ we deduce

$$(1 - \delta)\frac{1}{m}\sum_{i=1}^{m}a_i a_i^\top \preceq D_1^2 \preceq (1 + \delta_2)\frac{1}{m}\sum_{i=1}^{m}a_i a_i^\top.$$

Concentration of covariance matrices [55, Exercise 4.7.3] in turn implies that the estimate

$$\left\|\frac{1}{m}A^\top A - I_{d_1}\right\|_{\mathrm{op}} \le c_2\left(\sqrt{\frac{d_1 + u}{m}} + \frac{d_1 + u}{m}\right),$$

holds with probability at least $1 - 2\exp(-u)$. Taking a union bound, we therefore deduce

$$(1 - \delta)\left(1 - c_2\left(\sqrt{\frac{d_1 + u}{m}} + \frac{d_1 + u}{m}\right)\right)I_{d_1} \preceq D_1^2 \preceq (1 + \delta_2)\left(1 + c_2\left(\sqrt{\frac{d_1 + u}{m}} + \frac{d_1 + u}{m}\right)\right)I_{d_1}.$$

holds with probability at least $1 - m\exp(-c_1 d_2 \delta^2) - 2\exp(-u)$. Setting $u = d_1$, we see that there is a constant c_3 such that as long as $m \ge c_3\frac{\max\{d_1, d_2\}}{\delta^2}$, we have

$$(1 - \delta)^2 I_{d_1} \preceq D_1^2 \preceq (1 + \delta)^2 I_{d_1}$$

with probability at least $1 - m\exp(-c_1 d_2 \delta^2) - 2\exp(-d_1)$. The result follows. $\qquad\square$

The following are the two main results of the section.

Theorem 4.10 (Exact recovery in matrix sensing). *Suppose that $\mathcal{A}$ is a Gaussian ensemble. Then there exists a constant c_0 such that the following hold for any $\delta \in (0, c_0)$. There exist constants $c, C > 0$ depending only on δ such that in the regime $m \ge cr_\natural(d_1 + d_2)$, with probability at least $1 - \exp(-Cm)$, any flat solution (L_f, R_f) of (1) satisfies $L_f R_f^\top = M_\natural$ and is automatically nearly norm-minimal and nearly balanced:*

$$\|L_f\|_F^2 + \|R_f\|_F^2 \le \left(\frac{1 + \delta}{1 - \delta}\right)^2 \cdot \left(\min_{\mathcal{A}(LR^\top)=b}\|L\|_F^2 + \|R\|_F^2\right)$$

$$\left\|L_f^\top L_f - R_f^\top R_f\right\|_* \le 2\left(\left(\frac{1 + \delta}{1 - \delta}\right)^2 - 1\right)\|M_\natural\|_*.$$

Proof. Lemma 4.4 shows that for any $\delta \in (0, 1)$, there exist constants $c_1, C_1 > 0$ depending only on δ such that as long as $m \ge c_1 r(d_1 + d_2)$, with probability at least $1 - \exp(-C_1 m)$, the linear map $\mathcal{A}$ satisfies ℓ_2/ℓ_2 RIP with parameters $(r, 1 - \delta, 1 + \delta)$. In this event, Lemma 4.8 ensures that the condition number κ of D_1 and D_2 is bounded by $\frac{1+\delta}{1-\delta}$. Set now $r = 5r_\natural$ and choose any $\delta \le 0.1$ satisfying $(1 + \delta)\left(\frac{1+\delta}{1-\delta}\right)^2 \le \frac{11}{10}$. An application of Lemma 4.7 and Theorem 3.3 completes the proof. $\qquad\square$

Theorem 4.11 (Exact recovery in bilinear sensing). *Suppose that $\mathcal{A}$ is a Gaussian bilinear ensemble. Then for any $\delta \in (0, 1)$ there exist numerical constants $c, C, c_1, c_2, c_3, c_4 > 0$ depending only on δ such that in the regime $m \ge cr_\natural(d_1 + d_2)$ and $\log(m) \le c_4 d_{\min}$, with probability at least $1 - c_3\exp(-C d_{\min})$ any flat solution (L_f, R_f) of (1) satisfies $L_f R_f^\top = M_\natural$ and is automatically nearly norm-minimal and nearly balanced:*

$$\|L_f\|_F^2 + \|R_f\|_F^2 \le \left(\frac{1 + \delta}{1 - \delta}\right)^2 \cdot \left(\min_{\mathcal{A}(LR^\top)=b}\|L\|_F^2 + \|R\|_F^2\right)$$

$$\left\|L_f^\top L_f - R_f^\top R_f\right\|_* \le 2\left(\left(\frac{1 + \delta}{1 - \delta}\right)^2 - 1\right)\|M_\natural\|_*.$$

Proof. For any integer $k \in \mathbb{N}$, Lemma 4.5 ensures that there exist numerical constants $\delta_1, \delta_2 > 0$ and constants $c_0, C_0 > 0$ depending only on l such that in the regime $m \geq c_0 r(d_1 + d_2)$, with probability at least $1 - \exp(-C_0 m)$, the measurement map $\mathcal{A}$ satisfies ℓ_1/ℓ_2 RIP with parameters $(lr_\natural, \delta_1, \delta_2)$. Lemma 4.9 in turn ensures there exist numerical constant $c_5, c_6, c_7, c_8 > 0$ such that for any $\delta \in (0, 1)$ as along as we are in the regime, $m \geq \frac{c_7 d_{\max}}{\delta^2}$ and $\log(m) \leq c_8 \delta^2 d_{\min}$, the estimate holds:

$$\mathbb{P}\{(1 - \delta)I_{d_i} \preceq D_i \leq (1 + \delta)I_{d_i} \quad \forall i \in \{1, 2\}\} \geq 1 - c_5 e^{-c_6 d_{\min} \delta^2}.$$

Therefore in this regime, we may upper bound the condition number κ of D_1 and D_2 by $\frac{1+\delta}{1-\delta}$. In light of Lemma 4.7, in order to ensure exact recovery, it remains to simply choose a large enough l such that the inequality $\frac{\delta_2}{\delta_1} \cdot (\frac{1+\delta}{1-\delta})^2 \leq \sqrt{l}$ holds (recall δ_1, δ_2 are numerical constants). An application of Lemma 4.7 and Theorem 3.3 completes the proof. $\square$

Appendix A generalizes the material in this section to the noisy observation setting, wherein $b = \mathcal{A}(M_\natural) + e$ with $e \sim N(0, \sigma^2 I)$ for some $\sigma^2 > 0$.

5   Matrix completion and approximate recovery

In this section, we focus on the matrix completion problem [9, 46]. This is an instance of (1) where the linear measurement map $\mathcal{A}$ is generated as follows. For each $i \in [d_1]$ and $j \in [d_2]$, let ξ_{ij} be independent Bernoulli random variables with success probability p. The linear map $\mathcal{A} \colon \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ is then defined by the relation

$$[\mathcal{A}(Z)]_{ij} = Z_{ij}\xi_{ij} \qquad \text{for any } (i, j) \in [d_1] \times [d_2]. \tag{33}$$

The difficulty of recovering the matrix $M_\natural$ is typically measured by an incoherence parameter, which we now define. Given a singular value decomposition $M_\natural = U_\natural \Sigma_\natural V_\natural^\top$ with $\Sigma_\natural \in \mathbb{R}^{r_\natural \times r_\natural}$, the *incoherence parameter* is the smallest $\mu > 0$ satisfying

$$\|U_\natural\|_{2,\infty} \leq \sqrt{\frac{\mu r_\natural}{d_1}}, \quad \text{and} \quad \|V_\natural\|_{2,\infty} \leq \sqrt{\frac{\mu r_\natural}{d_2}}. \tag{34}$$

Here $\|A\|_{2,\infty}$ denotes the maximal ℓ_2-norm of the rows of the matrix A. The strategies outlined in the previous section do not directly apply for analyzing flat minima of the matrix completion problem because the linear map $\mathcal{A}(D_1^{-1} \cdot D_2^{-1})$ does not satisfy RIP type conditions. Instead we will settle for a weak recovery result.

Theorem 5.1 (Recovery error of flat solution)**.** *Suppose that $\mathcal{A}$ is a random sampling map described in (33). Then there exist numerical constants $c, C > 0$ such that the following is true. Given any $\gamma \in (0, 1)$, provided we are in the regime*

$$p \geq C \max\left\{\frac{1}{\gamma}\sqrt{\frac{r_\natural \log(d_{\max})}{d_{\min}}}, \frac{\mu r_\natural \log(\mu r_\natural) \log(d_{\max})}{d_{\min}}\right\}, \tag{35}$$

with probability at least $1 - cd_{\min}^{-5}$, any flat solution (L_f, R_f) satisfies

$$\left\|L_f R_f^\top - M_\natural\right\|_* \leq \gamma \|M_\natural\|_*. \tag{36}$$

15

Hence according to Theorem 5.1, in order to conclude that flat solutions achieve a constant relative error, we must be in the regime $p \gtrsim \sqrt{\frac{r_\natural \log d_{\max}}{d_{\min}}}$. This is a stronger requirement than is needed for exact recovery of the ground truth matrix [13], which is $p \gtrsim \mu r_\natural \log(\mu r_\natural) \frac{\log(d_{\max})}{d_{\min}}$. Our numerical experiments, however, suggest that flat solutions exactly recover the ground truth.

As the first step towards proving Theorem 5.1, we estimate the condition numbers of D_1, D_2.

Lemma 5.2 (Condition number). *For any given $\delta \in (0,1)$ and $c \geq 1$, as long as $p \geq \frac{(1+c)\log(d_{\max})}{2\delta^2 d_{\min}}$, with probability at least $1 - 4d_{\min}^{-c}$, the estimate*

$$\sqrt{\frac{p}{d_1 d_2}(1-\delta)} I_{d_i} \preceq D_i \preceq \sqrt{\frac{p}{d_1 d_2}(1+\delta)} I_{d_i} \qquad holds \ for \ i = 1,2. \tag{37}$$

Proof. Let $m = d_1 d_2$ and set the sensing matrices $A_{ij} = \xi_{ij} e_i e_j^\top$ for all pairs $i \in [d_1]$ and $j \in [d_2]$. Therefore the equality $A_{ij} A_{ij}^\top = \xi_{ij} e_i e_i^\top$ holds, and we can write

$$D_1^2 = \frac{1}{d_1 d_2^2} \sum_{i \in [d_1], j \in [d_2]}^{m} \xi_{ij} e_i e_i^\top = \frac{1}{d_1 d_2^2} \operatorname{diag}\left(\sum_{j=1}^{d_2} \xi_{1j}, \ldots, \sum_{j=1}^{d_2} \xi_{d_1 j}\right). \tag{38}$$

Bernstein's inequality [55, Theorem 2.8.1] implies for each index $i \in [d_2]$ the estimate

$$\mathbb{P}\left\{\left|\frac{1}{pd_2}\sum_{j=1}^{d_2} \xi_{ij} - 1\right| \geq \delta\right\} \leq 2\exp\left(-2pd_2\delta^2\right) \qquad \forall \delta > 0.$$

Taking the union bound over $i \in [d_1]$ we deduce that the condition

$$\frac{p}{d_1 d_2}(1-\delta)I \preceq D_1^2 \preceq \frac{p}{d_1 d_2}(1+\delta)I$$

fails with probability at most

$$d_1 \exp\left(-2pd_2\delta^2\right) \leq \exp\left(-2pd_2\delta^2 + \log(d_1)\right) \leq \exp\left(\frac{-pd_2\delta^2}{4}\right) \leq d_2^{-c}.$$

Using the same argument for D_2 and taking a union bound completes the proof. $\qquad\square$

Next, we will show that flat solutions are almost optimal for the standard convex relaxation of the matrix completion problem:

$$\min_{\mathcal{A}(X)=\mathcal{A}(M_\natural)} \|X\|_*. \tag{39}$$

Lemma 5.3 (Flat minima and nuclear norm minimization). *Suppose that $M_\natural$ is a solution of the problem (39) and suppose that the condition numbers of D_1 and D_2 are upper bounded by some constant $\kappa > 0$. Then any flat solution (L_f, R_f) of (1) is nearly optimal for the convex relaxation (39) in the sense that:*

$$\left\|L_f R_f^\top\right\|_* - \|M_\natural\|_* \leq (\kappa^2 - 1)\|M_\natural\|_*. \tag{40}$$

Proof. We successively estimate

$$\left\|L_f R_f^\top\right\|_* \leq \frac{1}{2}\left(\|L_f\|_F^2 + \|R_f\|_F^2\right) \leq \kappa^2 \left(\min_{\mathcal{A}(LR^\top)=\mathcal{A}(M_\natural)} \frac{1}{2}\|L\|_F^2 + \frac{1}{2}\|R\|_F^2\right) = \kappa^2 \|M_\natural\|_*,$$

where the first and last inequalities follow from (10) and the second inequality follows from Theorem 3.3. We therefore deduce $\left\|L_f R_f^\top\right\|_* - \|M_\natural\|_* \leq (\kappa^2 - 1)\|M_\natural\|_*$, as claimed. $\qquad\square$

It remains to translate the suboptimality gap (40) into an estimate on the distance $\left\|L_f R_f^\top - M_\natural\right\|_*$. This is the content of the following lemma.

Lemma 5.4 (Sharp growth in matrix completion). *Suppose the linear map $\mathcal{A}$ is generated according to the matrix completion model. Then there exist constants $c, c_1, C > 0$ such that in the regime $p \geq \frac{C \mu r_\natural \log(\mu r_\natural) \log(d_{\max})}{d_{\min}}$, with probability at least $1 - c_1 d_{\min}^{-5}$, any feasible matrix X of the problem* (39) *satisfies the inequality:*

$$\|X - M_\natural\|_* \leq 8\left(1 + \sqrt{\frac{6 r_\natural}{p}}\right)(\|X\|_* - \|M_\natural\|_*). \tag{41}$$

Proof. Let $M_\natural = U_\natural \Sigma_\natural V_\natural^\top$ be a singular value decomposition of $M_\natural$, where $\Sigma_\natural \in \mathbb{R}^{r_\natural \times r_\natural}$ is a diagonal matrix of singular values. Define the matrix $R := X - M_\natural$ and define the linear map $P_{\mathcal{T}}$ by

$$P_{\mathcal{T}}(Z) := U_\natural U_\natural^\top Z + Z V_\natural V_\natural^\top - U_\natural U_\natural^\top Z V_\natural V_\natural^\top, \tag{42}$$

Set $P_{\mathcal{T}^\perp}(Z) := Z - P_{\mathcal{T}}(Z)$. Observe that we may bound $\|R\|_*$ as follows:

$$\|R\|_* = \|P_{\mathcal{T}^\perp}(R) + P_{\mathcal{T}}(R)\|_* \leq \|P_{\mathcal{T}^\perp}(R)\|_* + \|P_{\mathcal{T}}(R)\|_* \overset{(a)}{\leq} \|P_{\mathcal{T}^\perp}(R)\|_* + \sqrt{3 r_\natural}\|P_{\mathcal{T}}(R)\|_{\mathrm{F}}, \tag{43}$$

where the step (a) is due to the fact that $P_{\mathcal{T}}(R)$ has rank no more than $3 r_\natural$. We now bound $\|P_{\mathcal{T}^\perp}(R)\|_*$ and $\|P_{\mathcal{T}}(R)\|_{\mathrm{F}}$ separately. As verified in [16, Section 6], the premise in [13, Proposition 2] is satisfied with probability at least $1 - c_3 d_1^{-5} - c_3 d_2^{-5}$ for some universal $c_3 > 0$ under the condition $p \geq \frac{C \mu r_\natural \log(\mu r_\natural) \log(d_{\max})}{d_{\min}}$. Hence, the result [13, Proposition 2 and its proof][3] shows that with probability at least $1 - c_3 d_1^{-5} - c_3 d_2^{-5}$, there holds the inequality

$$\|P_{\mathcal{T}^\perp}(R)\|_* \leq 8(\|X\|_* - \|M_\natural\|_*). \tag{44}$$

Moreover, the premise in [13, Lemma 5] is satisfied with probability at least $1 - c_4 d_1^{-5} - c_4 d_2^{-5}$ for some universal constant $c_4 > 0$ as verified in [9, Lemma 4.1] or in [14, Lemma 11]. Hence, [13, Lemma 5 and its proof][4] shows that with probability at least $1 - c_4 d_1^{-5} - c_4 d_2^{-5}$, the inequality

$$\|P_{\mathcal{T}}(R)\|_{\mathrm{F}} \leq \frac{\sqrt{2}}{\sqrt{p}}\|P_{\mathcal{T}^\perp}(R)\|_*. \tag{45}$$

holds. Combining (43),(44), and (45), yields the desired inequality (41). □

Putting all the lemmas together, we can now prove Theorem 5.1.

Proof of Theorem 5.1. Lemma 5.4 ensures that in the regime $p \geq C \frac{\mu r_\natural \log(\mu r_\natural) \log(d_{\max})}{d_{\min}}$, with probability at least $1 - c_1 d_{\min}^{-5}$, the estimate

$$\|X - M_\natural\|_* \leq 8\left(1 + \sqrt{\frac{6 r_\natural}{p}}\right)(\|X\|_* - \|M_\natural\|_*),$$

holds for all X satisfying $\mathcal{A}(X) = \mathcal{A}(M_\natural)$. In this event, $M_\natural$ is clearly a minimizer of (39). Lemma 5.3 therefore ensures that the matrix $X := L_f R_f^\top$ satisfies

$$\|X\|_* - \|M_\natural\|_* \leq (\kappa^2 - 1)\|M_\natural\|_*,$$

---

[3]Specifically, the first displayed equation above [13, Lemma 5]

[4]In the displayed equation in the statement of the lemma, one can simply replace n^5 by $\frac{1}{\sqrt{p}}$ and set $Z = R$.

where κ is an upper bound on the condition numbers of D_1 and D_2. Lemma 5.2 in turn ensures that for any $\delta \in (0,1)$, in the regime $p \geq \frac{3\log(d_{\max})}{\delta^2 d_{\min}}$, with probability at least $1 - 4d_{\min}^{-5}$, the upper bound $\kappa \leq \sqrt{\frac{1+\delta}{1-\delta}}$ is valid. Algebraic manipulations therefore yield, within these events, the estimate:

$$\left\| L_f R_f^\top - M_\natural \right\|_* \leq C\delta \sqrt{\frac{r_\natural}{p}} \|M_\natural\|_*, \tag{46}$$

for a some numerical constant $C > 0$. To summarize, there exist numerical constants $c_1, c_2, C > 0$ such that the following is true. Given any $\delta \in (0,1)$, provided we are in the regime

$$p \geq c_1 \max\{\mu r_\natural \log(\mu r_\natural), \delta^{-2}\} \cdot \frac{\log(d_{\max})}{d_{\min}}, \tag{47}$$

with probability at least $1 - c_2 d_{\min}^{-5}$, any flat solution (L_f, R_f) satisfies (46). Let us now try to set

$$\delta^{-2} = \max\left\{1, \mu r_\natural \log(\mu r_\natural), \frac{C^2 r_\natural}{\gamma^2 p}\right\}.$$

This choice is consistent with the requirement (47) as long as (35) holds. With this choice of δ, the estimate (46) becomes $\left\| L_f R_f^\top - M_\natural \right\|_* \leq \gamma \|M_\natural\|_*$, as claimed. $\qquad\square$

6 Robust principal component analysis (PCA)

In this section, we focus on problem of principal component analysis (PCA) with outliers, also known as "robust PCA", following the approach in [7, 11]. Though this problem is not of the form (1), we will see that flat solutions (appropriately defined) exactly recover the ground truth under reasonable assumptions. The robust PCA problem asks to find a matrix $M_\natural \in \mathbb{R}^{d_1 \times d_2}$ that has been corrupted by sparse noise $S_\natural$. More precisely, we observe a matrix $Y \in \mathbb{R}^{d_1 \times d_2}$ of the form

$$Y = M_\natural + S_\natural,$$

The matrix $S_\natural$ is assumed to have at most $l_\natural$ many nonzero entries in any column and in any row, and $M_\natural$ has rank $r_\natural$. Moreover, following existing literature we assume that the matrix $M_\natural$ is *strongly incoherent with parameter* μ. That is, given a singular value decomposition $M_\natural = U_\natural \Sigma_\natural V_\natural^\top$ with $\Sigma_\natural \in \mathbb{R}^{r_\natural \times r_\natural}$, we let $\mu > 0$ denote the smallest constant satisfying

$$\|U_\natural\|_{2,\infty} \leq \sqrt{\frac{\mu r_\natural}{d_1}}, \qquad \|V_\natural\|_{2,\infty} \leq \sqrt{\frac{\mu r_\natural}{d_2}}, \qquad \text{and} \qquad \left\| U_\natural V_\natural^\top \right\|_\infty \leq \sqrt{\frac{\mu r_\natural}{d_1 d_2}}, \tag{48}$$

where $\|\cdot\|_\infty$ denotes the entrywise sup-norm.

One common approach for recovering $M_\natural$ is to solve the problem:

$$\min_{L,R} \min_{S \in \Omega} \left\| LR^\top + S - Y \right\|_F^2. \tag{49}$$

where we define the set

$$\Omega := \{S \mid \|S\|_{1,1} \leq \|S_\natural\|_{1,1}\},$$

and $\|\cdot\|_{1,1}$ is entry-wise ℓ_1-norm used to promote sparsity. The factors L and R vary over $\mathbb{R}^{d_1 \times k}$ and $\mathbb{R}^{d_2 \times k}$, respectively. As usual, we focus on the rank overparameterized setting $k \geq r_\natural$. Note that the optimal value of the problem (49) is clearly zero.

Observe that we may express the problem (51) more compactly as

$$\min_{L,R} \; f(L,R) := \operatorname{dist}^2_\Omega(Y - LR^\top), \tag{50}$$

where $\operatorname{dist}^2_\Omega$ denotes the square Frobenius distance to Ω. This is the overparameterized problem that we will focus on. As usual, we let $\mathcal{S}$ denote the set of minimizers of f; note that $\mathcal{S}$ is simply the set of pairs (L,R) satisfying $Y - LR^\top \in \Omega$. Observe that the objective function f is C^1-smooth but not C^2-smooth. Therefore, in order to measure flatness, we proceed via a smoothing technique introduced in [24, Section 4.2], [21, Section 4.3]. Namely, we approximate f near a basepoint $(\tilde{L}, \tilde{S})$ by the local model:

$$f_{\tilde{L},\tilde{R}}(L,R) := \left\| Y - LR^\top - P_\Omega(Y - \tilde{L}\tilde{R}^\top) \right\|^2_{\mathrm{F}}, \tag{51}$$

where P_Ω denotes the nearest point projection onto Ω. It is straightforward to see that the C^2-smooth function $f_{\tilde{L},\tilde{R}}(\cdot, \cdot)$ majorizes f and agrees with $f(\cdot, \cdot)$ up to first order at $(\tilde{L}, \tilde{R})$. We may therefore define a minimizer of (50) to be *flat* if it solves the problem:

$$\min_{(L,R)\in\mathcal{S}} \operatorname{str}(D^2 f_{L,R}(L,R)). \tag{52}$$

The following is the main result of the section.

Theorem 6.1 (Exact recovery in Robust PCA). *There is a numerical constant $c > 0$ such that in the regime $l_\natural \leq \frac{d_{\min}}{\mu r_\natural}$, any flat minimizer (L_f, R_f) of (50) satisfies $L_f R_f^\top = M_\natural$.*

Proof. Let $(L_0, R_0) \in \mathcal{S}$ be a solution of (50). Since $f(L_0, R_0) = 0$, the equality $f_{L_0,R_0}(L_0, R_0) = 0$ holds. In particular, we may write $f_{L_0,R_0}(L,R)$ as

$$f_{L_0,R_0}(L,R) = \left\| LR^\top - W_\natural \right\|^2_{\mathrm{F}},$$

where we define $W_\natural := Y - P_\Omega(Y - L_0 R_0^\top)$. Therefore appealing to Lemma 2.1, we may write the scaled trace as

$$\operatorname{str}(D^2 f_{L_0,R_0}(L_0, R_0)) = 2(\|L_0\|^2_{\mathrm{F}} + \|R_0\|^2_{\mathrm{F}}).$$

Thus any flat solution (L_f, R_f) of (50) solves the problem:

$$\begin{aligned} \min_{L,R} \quad & \|L\|^2_{\mathrm{F}} + \|R\|^2_{\mathrm{F}} \\ \text{subject to} \quad & Y - LR^\top \in \Omega. \end{aligned} \tag{53}$$

Equivalently, the characterization (10) implies that the matrix $X_f = L_f R_f^\top$ solves the problem

$$\begin{aligned} \text{minimize} \quad & \|X\|_* \\ \text{subject to} \quad & Y - X \in \Omega, \\ & \operatorname{rank}(X) \leq k. \end{aligned} \tag{54}$$

On the other hand, the result [14, Theorem 3][5] shows that $M_\natural$ is the unique minimizer of the convex relaxation

$$\begin{aligned} \text{minimize} \quad & \|X\|_* \\ \text{subject to} \quad & Y - X \in \Omega. \end{aligned} \tag{56}$$

---

[5]The result [14, Theorem 3] actually shows that $(M_\natural, S_\natural)$ uniquely solves

$$\begin{aligned} \text{minimize} \quad & \|X\|_* + \lambda\|S\|_{1,1} \\ \text{subject to} \quad & Y = X + S, \end{aligned} \tag{55}$$

for some $\lambda > 0$. Now for any solution X_1 to (56), the pair $(X_1, Y - X_1)$ is feasible for (55) and satisfies $\|X_1\|_* + \lambda\|Y - X_1\|_{1,1} \leq \|M_\natural\|_* + \lambda\|S_\natural\|_{1,1}$, by definition of Ω. Hence by the uniqueness of (55), we know $X_1 = M_\natural$.

Hence, we know $M_\natural$ also uniquely solves (54) and we conclude $M_\natural = X_f = L_f R_f^\top$, as claimed. $\quad\square$

7 Neural networks with quadratic activations and covariance matrix estimation

In this section, we investigate flat minimizers of a one hidden layer neural network, considered in the work [34, 51] for the purpose of analyzing the energy landscape around saddle points. Though this problem is not in the form (1), we will see that flat minimizers (naturally defined) exactly recover the ground truth under reasonable statistical assumptions. As a special case, we will obtain guarantees for flat minimizers of the overparameterized covariance matrix estimation problem.

The problem setup, following [34, 51], is as follows. We suppose that given an input vector $x \in \mathbb{R}^d$ a response vector $y(x)$ is given by the function

$$y(U_\natural, x) = v^\top q(U_\natural^\top x).$$

We assume that the output weight vector $v \in \mathbb{R}^r$ has r_1 positive entries and r_2 negative entries, the hidden layer weight matrix $U_\natural$ has dimensions $d \times r_\natural$, and we use a quadratic activation $q(s) = s^2$ applied coordinatewise. We get to observe a set of m pairs $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, where the features x_i are drawn as $x_i \overset{\text{iid}}{\sim} N(0, I_d)$ and the output values y_i are given by

$$y_i = y(x_i) \qquad \forall i = 1, \ldots, m.$$

We aim to fit the data with an overparameterized neural network with a single hidden layer with weights $U \in \mathbb{R}^{d \times k}$ and an output layer with weights $u = (\mathbf{1}_{k_1}, -\mathbf{1}_{k_2})$, where $k_1 \geq r_1$, and $k_2 \geq r_2$. The prediction $\hat{y}$ of the neural network on input x is thus given by

$$\hat{y}(U, x) = u^\top q(U^\top x). \tag{57}$$

Thus the overparameterized problem we aim to solve is

$$\min_{U \in \mathbb{R}^{d \times k}} \ f(U) := \frac{1}{m} \sum_{i=1}^m (\hat{y}(U, x_i) - y_i)^2. \tag{58}$$

As usual, we define the solution set $\mathcal{S} = \{U \in \mathbb{R}^{d \times k} : f(U) = 0\}$. We will see shortly that $\mathcal{S}$ is nonempty and therefore coincides with the set of minimizers of f. Naturally, we declare a matrix $U_f \in \mathcal{S}$ to be *flat* if it minimizes the trace of the Hessian of f over the set of the minimizers of f, i.e., it solves the problem

$$\min_{U \in \mathcal{S}} \ \mathrm{tr}(D^2 f(U)).$$

In this section, we aim to show:

> with high probability over the training set $\{(x_i, y_i)\}_{i=1,\ldots,n}$ flat solutions U_f
> achieve zero generalization error, that is $\mathbb{E}_{x \sim N(0, I)}(\hat{y}(U_f, x) - y(U_\natural, x)) = 0$.

Indeed, we will prove a stronger result by relating the problem (58) to low-rank matrix factorization. To see this, we can write $\hat{y}(U, x) - y(U_\natural, x)$ as:

$$\begin{aligned}
\hat{y}(U, x) - y(U_\natural, x) &= u^\top q(U^\top x) - v^\top q(U_\natural^\top x) \\
&= \langle \underbrace{U \operatorname{diag}(u) U^\top}_{U_1 U_1^\top - U_2 U_2^\top}, xx^\top \rangle - \langle \underbrace{U_\natural \operatorname{diag}(v) U_\natural^\top}_{=: M_\natural}, xx^\top \rangle \\
&= \langle U_1 U_1^\top - U_2 U_2^\top - M_\natural, xx^\top \rangle.
\end{aligned} \tag{59}$$

Here, we write $U = [U_1, U_2]$ with $U_1 \in \mathbb{R}^{d \times k_1}$ and $U_2 \in \mathbb{R}^{d \times k_2}$. Note that the matrix $M_\natural$ is symmetric. Using (59), we may rewrite the objective of (58) as

$$f(U_1, U_2) = \frac{1}{m} \left\| \mathcal{A}(U_1 U_1^\top - U_2 U_2^\top - M_\natural) \right\|_2^2, \tag{60}$$

where the linear map $\mathcal{A}$ is defined as $\mathcal{A} : \mathbb{R}^{d \times d} \to \mathbb{R}^m$ with $[\mathcal{A}(Z)]_i = \langle Z, x_i x_i^\top \rangle$ for any $Z \in \mathbb{R}^{d \times d}$. In particular, from the second equation in (59) and our assumption on v, there always exists a matrix $U = [U_1, U_2]$ satisfying $U_1 U_1^\top - U_2 U_2^\top = M_\natural$. Therefore, the set of minimizers of f is nonempty and it coincides with $\mathcal{S}$. Note that in the special case $r_2 = k_2 = 0$, the problem (60) becomes covariance matrix estimation [15] and further reduces to phase retrieval when $k_1 = r_1 = 1$ [10].

Summarizing, the task of finding a matrix U with a small generalization error, $|\mathbb{E}_{x \sim N(0,I)}[\hat{y}(U, x) - y(U_\natural, x)]|$, amounts to implicitly recovering the symmetric matrix $M_\natural$, but with the parameterization $U_1 U_1^\top - U_2 U_2^\top$ instead of the usual $LR^\top$ parameterization. The following is the main theorem of the section.

Theorem 7.1 (Exact recovery). *There exist numerical constant $c, C > 0$, such that in the regime $m \geq C r_\natural d$ and $d \geq C \log m$, with probability at least $1 - C \exp(-cd)$, any flattest solution $U_f = (U_{f,1}, U_{f,2})$ of (58) satisfies*

$$U_{f,1} U_{f,1}^\top - U_{f,2} U_{f,2}^\top = M_\natural, \tag{61}$$

and achieves zero generalization error:

$$\mathbb{E}_{x \sim N(0,I)}(\hat{y}(U_f, x) - y(U_\natural, x)) = 0. \tag{62}$$

The rest of the section is devoted to the proof of Theorem 7.1. The general strategy is very similar to the one pursued in Section 4. We begin with the following lemma that expresses the trace of the Hessian in the same spirit as Lemma 3.1. With this in mind, we define the matrix

$$D := \left(\frac{1}{md} \sum_{i=1}^m A_i A_i^\top \right)^{\frac{1}{2}}. \tag{63}$$

Lemma 7.2 (Trace). *The second order derivative of the function f at any matrix $[U_1, U_2] \in \mathcal{S}$ is the quadratic form:*

$$D^2 f(U_1, U_2)(V_1, V_2) = \frac{2}{m} \| \mathcal{A}(U_1 V_1^\top + V_1 U_1^\top - U_2 V_2^\top - V_2 U_2^\top) \|_2^2$$

Moreover, the trace can be written as

$$\mathrm{tr}(D^2 f(U_1, U_2)) = 4d \| DU_1 \|_F^2 + 4d \| DU_2 \|_F^2. \tag{64}$$

Proof. The expression for $D^2 f(U_1, U_2)[V_1, V_2]$ follows immediately from algebraic manipulations. The trace of the Hessian therefore can be written as

$$\mathrm{tr}(D^2 f(U_1, U_2))$$
$$= \frac{2}{m} \left(\sum_{i=1}^d \sum_{j=1}^k \left\| \mathcal{A}(U_1 e_j e_i^\top + e_i e_j^\top U_1^\top) \right\|_2^2 \right) + \frac{2}{m} \left(\sum_{i=1}^d \sum_{j=1}^k \left\| \mathcal{A}(e_i e_j^\top U_2^\top + U_2 e_i^\top e_j) \right\|_2^2 \right). \tag{65}$$

Using the symmetry of the matrices A_i, the first term can be written as

$$\sum_{i=1}^d \sum_{j=1}^k \left\| \mathcal{A}(U_1 e_j e_i^\top + e_i e_j^\top U_1^\top) \right\|_2^2 = 2 \sum_{i=1}^d \sum_{j=1}^k \sum_{l=1}^m \langle A_l, U_1 e_j e_i^\top \rangle^2. \tag{66}$$

Following exactly the same computation as (18) completes the proof. $\qquad \square$

21

In particular, Lemma 7.2 implies that flat solutions are exactly the minimizers of the problem

$$\min_{U_1, U_2} \frac{1}{2}\left(\|DU_1\|_F^2 + \|DU_2\|_F^2\right) \qquad \text{such that} \qquad \mathcal{A}(U_1 U_1^\top - U_2 U_2^\top) = \mathcal{A}(M_\natural). \qquad (67)$$

We would like to next rewrite this problem in terms of minimizing a nuclear norm of a $d \times d$ matrix. With this in mind, we will require the following two lemmas in the spirit of the characterization of the nuclear norm (10).

Lemma 7.3 (Decompositions and pos/neg eigenvalues). *The following two statements are equivalent for any symmetric matrix $X \in \mathbb{R}^{d \times d}$.*

1. *X admits a decomposition $X = U_1 U_1^\top - U_2 U_2^\top$ for some matrices $U_i \in \mathbb{R}^{d \times k_i}$,*

2. *X has at most k_1 non-negative eigenvalues and k_2 non-positive eigenvalues.*

Proof. The implication $2 \Rightarrow 1$ follows immediately from an eigenvalue decomposition of X. Conversely, suppose that 1 holds. Observe that 1 clearly is equivalent to being able to write $X = A - B$ with $A, B \succeq 0$, $\text{rank}(A) \leq k_1$, and $\text{rank}(B) \leq k_2$. Let r_1 be the number of strictly positive eigenvalues of X and let r_2 be the number of strictly negative eigenvalues of X. We now prove $r_1 \leq k_1$ by contradiction. A similar arguments yields $r_2 \leq k_2$. Suppose indeed $r_1 > k_1$ and consider the matrix $A := X + B$. Let $\mathcal{U}$ be the span of eigenspaces corresponding to the top r_1 eigenvalues of X. Note that $\mathcal{U}$ has dimension r_1. Cauchy's interlacing theorem implies that the r_1-th largest eigenvalue of $X + B$ satisfies that $\lambda_{r_1}(X + B) \geq \min_{v \in \mathcal{U} \setminus \{0\}} \frac{v^\top (X+B)v}{v^\top v}$. Since $B \succeq 0$, for any $v \in \mathcal{U} \setminus \{0\}$ we estimate $\frac{v^\top (X+B)v}{v^\top v} = \frac{v^\top X v + v^\top B v}{v^\top v} \geq \frac{v^\top X v}{v^\top v} \geq \lambda_{r_1}(X) > 0$. We conclude that that rank of A is at least r_1, which is a contradiction since A has rank at most k_1. $\qquad \square$

Lemma 7.4. *If a symmetric matrix $X \in \mathbb{R}^{d \times d}$ admits a decomposition $X = \bar{U}_1 \bar{U}_1^\top - \bar{U}_2 \bar{U}_2^\top$ for some matrices $\bar{U}_i \in \mathbb{R}^{d \times k_i}$, then equality holds:*

$$\|X\|_* = \min_{X = U_1 U_1^\top - U_2 U_2^\top, U_i \in \mathbb{R}^{d \times k_i}} \|U_1\|_F^2 + \|U_2\|_F^2.$$

Proof. First, using triangle inequality, for any (U_1, U_2) such that $X = U_1 U_1^\top - U_2 U_2^\top$, we have $\|X\|_* = \|U_1 U_1^\top - U_2 U_2^\top\|_* \leq \|U_1 U_1^\top\|_* + \|U_2 U_2^\top\|_* = \text{tr}(U_1 U_1^\top) + \text{tr}(U_2 U_2^\top) = \|U_1\|_F^2 + \|U_2\|_F^2$. Conversely, suppose that we may write $X = U_1 U_1^\top - U_2 U_2^\top$ for some U_1, U_2. Then Lemma 7.3 implies that X has at most k_1 non-negative eigenvalues and k_2 non-positive eigenvalues. Thus, we may write any eigenvalue decomposition of X as $X = P_1 \Lambda_1 P_1^\top - P_2 \Lambda_2 P_2^\top$, where the diagonal matrices $\Lambda_i \in \mathbb{R}^{r_i \times r_i}$ have positive entries. By taking $U_i = [P_i \sqrt{\Lambda_i}, 0_{k_i - r_i}]$, we see that equality $\|X\|_* = \|U_1\|_F^2 + \|U_2\|_F^2$ holds and the proof is complete. $\qquad \square$

Lemma 7.3 and 7.4 directly imply that the problem (67), which characterizes flat solutions, is equivalent to the rank constrained problem:

$$\begin{aligned} \text{minimize} \quad & \|X\|_*. \\ \text{subject to} \quad & \mathcal{A}(D^{-1} X D^{-1}) = \mathcal{A}(M_\natural), \\ & X \text{ is symmetric and has at most } k_1 \text{ positive eigenvalues and } k_2 \text{ negative eigenvalues.} \end{aligned}$$
$$(68)$$

Therefore a natural convex relaxation simply drops the requirements on the eigenvalues:

$$\begin{aligned} \text{minimize} \quad & \|X\|_* \\ \text{subject to} \quad & \mathcal{A}(D^{-1} X D^{-1}) = \mathcal{A}(M_\natural), \\ & X \in \mathbb{R}^{d \times d} \text{ is symmetric.} \end{aligned} \qquad (69)$$

The following theorem summarizes these observations.

Theorem 7.5 (Convex relaxation). *Suppose that the matrix D is invertible. Then the problems* (67) *and* (68) *are equivalent in the following sense.*

1. *The optimal values of* (67) *and* (68) *are equal.*

2. *If $[U_1, U_2]$ solves* (67), *then $X = D(U_1 U_1^\top - U_2 U_2^\top)D$ is optimal for* (68).

3. *If a minimizer X of* (68) *has an eigenvalue decomposition $X = P_1 \Lambda_1 P_1^\top - P_2 \Lambda_2 P_2^\top$ for some diagonal matrices $\Lambda_i \in \mathbb{R}^{r_i \times r_i}$ with positive entries, then the matrices $U_i = [D^{-1} P_i \sqrt{\Lambda_i}, 0_{d,(k_i - r_i)}]$, $i = 1, 2$ are minimizers of* (67).

Moreover, if $X = DM_\natural D$ is the unique minimizer of the problem (69), *then any flat solution $[U_1, U_2]$ satisfies $U_1 U_1^\top - U_2 U_2^\top = M_\natural$.*

Next, we aim to understand when the problem (69) exactly recovers $D^{-1} M_\natural D$. The difficulty is that even in the case $k_2 = 0$, corresponding to covariance matrix estimation, the linear map $\mathcal{A}$ satisfies ℓ_1/ℓ_2 RIP only if $m \gtrsim r_\natural^2 d$ [15], which is suboptimal by a factor of $r_\natural$. We will sidestep this issue by relating the program (68) to one with a different linear map that does satisfy the ℓ_1/ℓ_2-RIP condition over all rank $r_\natural$ matrices in the optimal regime $m \gtrsim r_\natural d$. The reduction we use is inspired by [5, Equation (0.36)].

Lemma 7.6. *Define the linear map $\mathcal{A}_1 : \mathbb{R}^{d \times d} \to \mathbb{R}^{\lfloor m \rfloor/2}$ by*

$$[\mathcal{A}_1(Z)]_i = \left(\frac{x_{2i-1} + x_{2i}}{\sqrt{2}} \right)^\top Z \left(\frac{x_{2i-1} - x_{2i}}{\sqrt{2}} \right), \qquad for\ i = 1, \ldots, \lfloor m \rfloor/2, \tag{70}$$

and consider the convex optimization program

$$\min_{\mathcal{A}_1(D^{-1} X D^{-1}) = \mathcal{A}_1(M_\natural)} \|X\|_*. \tag{71}$$

If $DM_\natural D$ is the unique solution of (71), *then it is also the unique solution of* (69).

Proof. We first note the equalities

$$\begin{aligned} [\mathcal{A}_1(Z)]_i &= \left(\frac{x_{2i-1} + x_{2i}}{\sqrt{2}} \right)^\top Z \left(\frac{x_{2i-1} - x_{2i}}{\sqrt{2}} \right) \\ &= \frac{1}{2} \langle x_{2i-1} x_{2i-1}^\top, Z \rangle - \frac{1}{2} \langle x_{2i} x_{2i}^\top, Z \rangle \\ &= \frac{1}{2} ([\mathcal{A}_{2i-1}(Z)] - [\mathcal{A}_{2i}(Z)]). \end{aligned} \tag{72}$$

It follows immediately that any X that is feasible for (69) is also feasible for (71). Consequently, if the symmetric matrix $DM_\natural D$ is a unique minimizer of (71), then it must also be a unique minimizer of (69). This completes the proof. $\qquad\square$

We are now ready to prove Theorem 7.1.

Proof of Theorem 7.1. Notice that the two vectors, $\frac{x_{2i-1} + x_{2i}}{\sqrt{2}}$ and $\frac{x_{2i-1} - x_{2i}}{\sqrt{2}}$, are jointly normal and uncorrelated, and therefore are independent. Consequently, we see that the map $\mathcal{A}_1$ defined in (70) follows the bilinear sensing model (Definition 4.1). Therefore Lemma 4.5 implies that for any positive integer $k \geq 2$, there exist constants $c, C > 0$ depending only on k and numerical constants

23

$\delta_1, \delta_2 > 0$ and such that in the regime $m \geq cr_\natural d$, with probability at least $1 - \exp(-Cm)$, the measurement map $\mathcal{A}_1$ satisfies ℓ_1/ℓ_2 RIP with parameters $(kr_\natural, \delta_1, \delta_2)$. In this event, Lemma 4.6 with $Q_1 = Q_2 = D^{-1}$ implies that $\mathcal{A}_1(D^{-1} \cdot D^{-1})$ satisfies ℓ_1/ℓ_2 RIP with parameters $(kr_\natural, \alpha_2^{-2}\delta_1, \alpha_1^{-2}\delta_2)$, where $\alpha_1 > 0$ is a lower bound on the minimal eigenvalue of D and $\alpha_2 > 0$ is an upper-bound on the maximal eigenvalue of D. In order to estimate κ, we may write

$$D^2 = \frac{1}{md} \sum_{i=1}^{m} A_i A_i^\top = \frac{1}{md} \sum_{i=1}^{m} \|x_i\|^2 x_i x_i^\top.$$

Exactly the same proof as that of Lemma 4.9 ensures that there exist constants $c_1, c_2, c_3, c_4 > 0$ such that as long as we are in the regime, $m \geq c_3 d$ and $\log(m) \leq c_4 d$, the estimate holds:

$$\mathbb{P}\left\{\frac{1}{2}I_d \preceq D \preceq \frac{3}{2}I_d\right\} \geq 1 - c_2 e^{-c_1 d}.$$

Consequently, in this regime we may upper bound the condition number κ of D by 3. In light of Lemma 4.7, in order to ensure that $DM_\natural D$ is the unique minimizer of (71), it remains to simply choose k such that the inequality $\frac{9\delta_2}{\delta_1} \leq \sqrt{k}$ holds. Using Lemmas 7.5-7.6 completes the proof. $\square$

8 Numerical experiments

Recall that we have proved that for a variety of overparameterized problems, under standard statistical assumptions, in the noiseless setting, (1) flat solutions recover the ground truth and (2) flat solutions are nearly norm-minimal and nearly-balanced (but not exactly). In this section, we numerically validate both of the claims, in order. Note that finding flat solutions in these examples, amounts to solving a convex optimization problem as long as the number of measurements is sufficiently large.

Experiment setup. We consider four problems described earlier in the paper: (a) matrix sensing, (b) bilinear sensing, (c) matrix completion, and (d) neural networks with quadratic activation. For each setting, we consider different combination of the dimension $d = d_1 = d_2$ and the number of measurements m (p for matrix completion). For each combination (d, m) ((d, p) for matrix completion), we randomly generate a rank 2 ground truth unit Frobenius norm matrix $M_\natural$ (rank 3 for the setting of neural network with quadratic activation), then repeatedly generate the linear measurement map $\mathcal{A}$ and solve ten times the convex relaxation associated with being a flat solution and the nuclear norm minimization problem.

Exact Recovery. To measure the success of exact recovery, for a solution $\hat{X}$ from the convex relaxation of the scaled trace problem (or from the nuclear norm minimization), we measure the Frobenius norm error $\left\|D_1^{-1}\hat{X}D_2^{-1} - M_\natural\right\|_F$ (or $\left\|\hat{X} - M_\natural\right\|_F$ for the nuclear norm minimization). Our criterion for exact recovery is whether this error is smaller than 10^{-6} or not. Figure 3 shows the empirical probability of success recovery (averaging over ten times) for each combination of dimension and number of measurements. The figure is in gray scale and the whiter color indicates higher success probability. We observe that the frequency of exact recovery by flat solutions almost matches the frequency of exact recovery by nuclear norm minimization. Notice moreover that flat solutions exactly recover the ground truth matrix, though we are only able to show weak recovery for matrix completion.
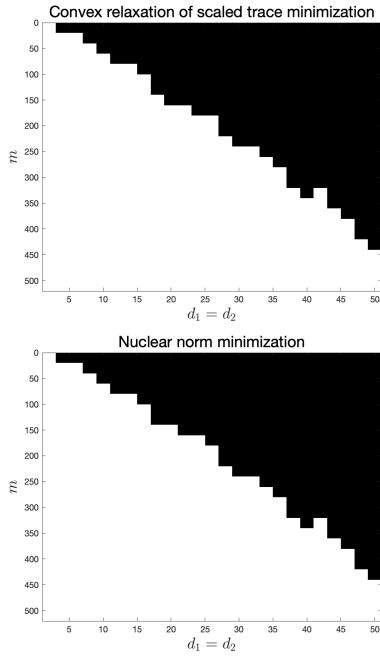
(a) Matrix Sensing
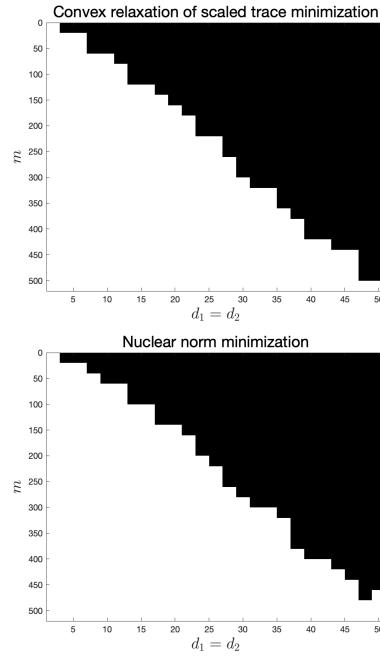
(b) Bilinear Sensing
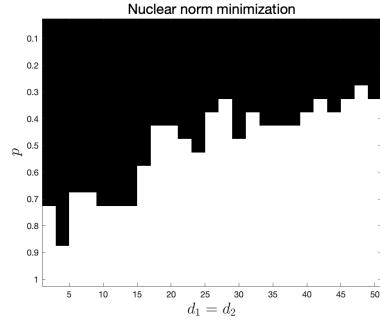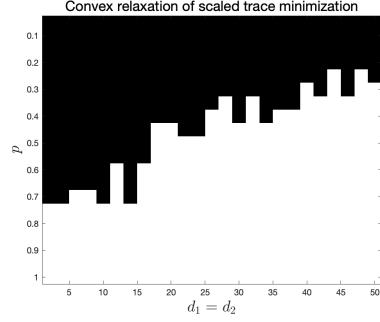
(c) Matrix Completion
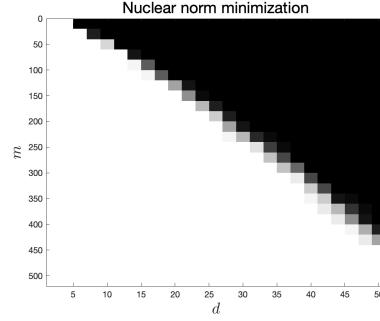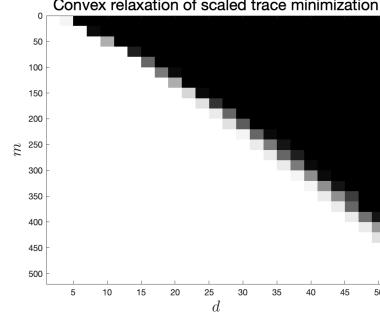
(d) NN with quadratic activation

Figure 3: The empirical probability of successful recovery of $M_\natural$ for different combination of dimension d and number of measurements m (p for matrix completion). We use gray scale and the whiter the color, the higher probability of success.

Regularity. Next we test the regularity of flat solutions for the (a) matrix sensing, (b) bilinear sensing, (c) matrix completion problems. We only consider the pairs (d, m) such that the matrices D_1, D_2 are nonsingular. Let $\hat{X}$ be the solution of the convex relaxation for being a flat solution and let $\hat{X}_{\mathtt{nuc}}$ be the solution to the nuclear norm minimization problem. We compute the factors $L_f = D_1^{-1}U\sqrt{\Sigma}$ and $R_f = D_2^{-1}V\sqrt{\Sigma}$ using the full SVD of $\hat{X} = U\Sigma V^\top$. We then use the quantity $\dfrac{\|L_f\|_{\mathrm{F}}^2 + \|R_f\|_{\mathrm{F}}^2}{2\|\hat{X}_{\mathtt{nuc}}\|_*}$ to measure the norm-minimality of flat solutions, and the quantity $\left\|L_f^\top L_f - R_f^\top R_f\right\|_* / \|M_{\natural}\|_*$ to measure balancedness. Whenever one of the matrices D_1, D_2 is singular, we set both measures to be 10^{20}. The result (in $\log 10$ scale) is shown in Figure 4. We observe that whenever flat solutions exactly recover the ground truth, both measures are small but not exactly zero. In particular, the norm-minimal and flat solutions are distinct.


(a) Matrix Sensing

(b) Bilinear Sensing
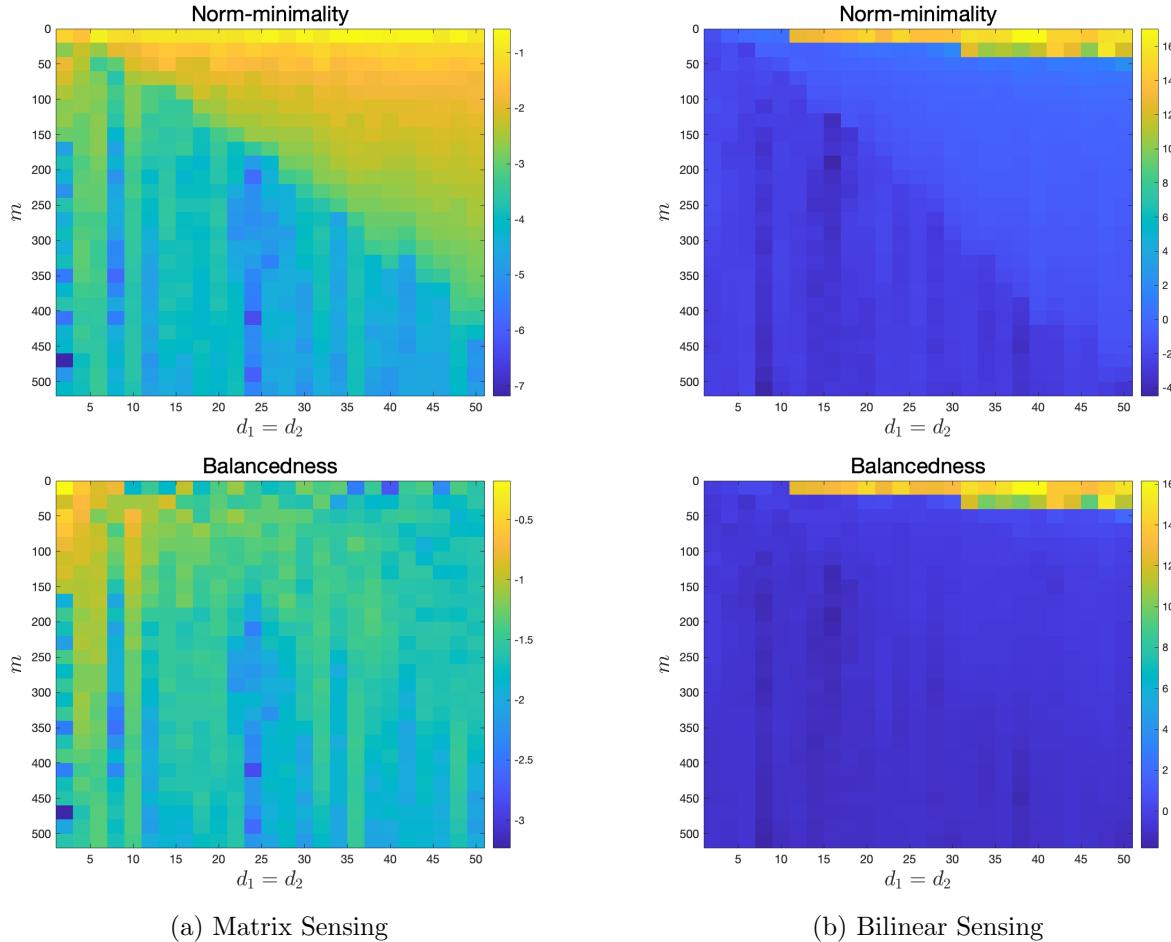
Figure 4: The empirical average (over 10 trials) of the regularity measure : (1) minimal norm, $\dfrac{\|L_f\|_{\mathrm{F}}^2 + \|R_f\|_{\mathrm{F}}^2}{2\|\hat{X}_{\mathtt{cvx}}\|_*} - 1$, and (2) balance, $\dfrac{\|L_f^\top L_f - R_f^\top R_f\|_*}{\|M_{\natural}\|_*}$. Both measures are presented in $\log 10$ scales.

9   Conclusion and discussion on depth

In this paper, we analyzed a variety of low rank matrix recovery problems in rank-overparameterized settings. We considered overparameterized matrix and bilinear sensing, robust PCA, covariance

matrix estimation, and single hidden layer neural networks with quadratic activation functions. In all cases, we showed that flat minima, measured by the scaled trace of the Hessian, *exactly recover* the ground truth under standard statistical assumptions. For matrix completion, we established weak recovery, although empirical evidence suggests exact recovery holds here as well.

Matrix factorization problems are suggestive of the behavior one may expect for two layer neural networks. Therefore, an appealing question is to consider the effect that depth may have on generalization properties of flat solutions. In this section, we argue that depth may not bode well for generalization of flat solutions. As a simple model, we consider the setting of sparse recovery under a "deep" overparameterization. Namely, consider a ground truth vector $x_\natural \in \mathbb{R}^d$ with at most $r_\natural$ nonzero coordinates. The goal is to recover $x_\natural$ from the observed measurements $b = Ax_\natural$ under a linear map $A \colon \mathbb{R}^d \to \mathbb{R}^m$. We assume that A satisfies the restricted isometry property (RIP): there exist (δ_1, δ_2) such that

$$\delta_1 \|x\|_2 \leq \tfrac{1}{\sqrt{m}} \|Ax\|_2 \leq \delta_2 \|x\|_2, \tag{73}$$

for all $x \in \mathbb{R}^d$ that have at most $2r$ nonzero coordinates. The simple least square formulation for finding consistent dense signals is

$$\min_{x \in \mathbb{R}^d} \quad g(x) := \frac{1}{m} \|Ax - b\|_2^2. \tag{74}$$

We introduce overparameterization by parameterizing the variable x as the Hadamard product $\odot$ of k factors $x = v_1 \odot v_2 \odot \cdots \odot v_k$ with $v_i \in \mathbb{R}^d$. Thus, the problem (74) becomes

$$\min_{v_1, \ldots, v_k} \quad f(v) := \frac{1}{m} \|A(v_1 \odot \cdots \odot v_k) - b\|_2^2 \qquad \text{with } v_i \in \mathbb{R}^d, \ i = 1, \ldots, k. \tag{75}$$

The flat solutions are naturally defined as those $(v_i)_{i=1}^k$ solving the following problem:

$$\min_{v_i \in \mathbb{R}^d, i=1,\ldots,k} \operatorname{tr}(D^2 f(v_1, \ldots, v_k)) \quad \text{subject to} \quad A(v_1 \odot \cdots \odot v_k) = b. \tag{76}$$

To compute the Hessian $\operatorname{tr}(D^2 f(v_1, \ldots, v_k))$, let a_i be the i-th column of A. Following a similar calculation as in Lemma 3.1 yields the expression

$$\operatorname{tr}(D^2 f([v_i]_{i=1}^k)) = \sum_{i=1}^k \left\| \sqrt{D}(v_1 \odot \ldots v_{i-1} \odot v_{i+1} \cdots \odot v_k) \right\|_2^2, \tag{77}$$

for any $(v_i)_{i=1}^k \in \mathbb{R}^{d \times k}$ where

$$D := \frac{1}{m} \operatorname{diag}(a_1^\top a_1, \ldots, a_d^\top a_d). \tag{78}$$

The following lemma shows that D is close to the identity matrix.

Lemma 9.1. *Suppose that the linear map A satisfies $(1 - \delta, 1 + \delta)$ RIP for some $\delta \in (0, 1)$. Then the matrix D satisfies*

$$(1 - \delta)^2 I \preceq D \preceq (1 + \delta)^2 I.$$

Proof. Indeed, since D is diagonal, we only need to show $\frac{1}{m} a_i^\top a_i \in [(1 - \delta)^2, (1 + \delta)^2]$ for each index i. Note that $\frac{1}{m} \|Ae_i\|_2^2 = \frac{1}{m} a_i^\top a_i$. Since e_i is a sparse vector with only one nonzero, using the $(1 - \delta, 1 + \delta)$ RIP, we have $\frac{1}{m} a_i^\top a_i = \frac{1}{m} \|Ae_i\|_2^2 \in [(1 - \delta)^2 \|e_i\|_2, (1 + \delta)^2 \|e_i\|_2^2] = [(1 - \delta)^2, (1 + \delta)^2]$ and our proof is complete. $\qquad\square$

The next lemma shows that the following optimization problem is equivalent to the optimization problem defining flat solutions (76).

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{d} |D_{ii}||x_i|^{2-\frac{2}{k}} \quad \text{s.t.} \quad Ax = b. \tag{79}$$

Denote by $v_{h,j}$ the j-th component of the vector variable v_h for $1 \le h \le k$.

Lemma 9.2. *Problem* (76) *is equivalent to Problem* (79) *in the following sense:*

- *If x solves* (79)*, then any v_i satisfying $x = v_1 \odot \cdots \odot v_k$ and $|v_{1,j}| = \cdots = |v_{k,j}|$ for any $1 \le j \le d$ solves* (76)*.*

- *If $v_1, \ldots, v_k$ solves* (76)*, then $x = v_1 \odot \cdots \odot v_k$ solves* (79)*.*

Proof. According to (77), the trace of the Hessian is

$$
\begin{aligned}
\operatorname{tr}(D^2 f([v_i]_{i=1}^{k})) &= \sum_{i=1}^{k} \left\| \sqrt{D}(v_1 \odot \ldots v_{i-1} \odot v_{i+1} \cdots \odot v_k) \right\|_2^2 \\
&= \sum_{i=1}^{k} \sum_{j=1}^{d} D_{jj} \prod_{h \ne i} v_{h,j}^2 \\
&= \sum_{j=1}^{d} D_{jj} \sum_{i=1}^{k} \prod_{h \ne i} v_{h,j}^2 \\
&\overset{(a)}{\ge} \sum_{j=1}^{d} D_{jj} k \left(\prod_{i=1}^{k} v_{i,j}^{2(k-1)} \right)^{\frac{1}{k}} \\
&= \sum_{j=1}^{d} D_{jj} \left(\prod_{i=1}^{k} |v_{i,j}| \right)^{2-\frac{2}{k}}.
\end{aligned}
\tag{80}
$$

In the step (a), we use the well-known AM-GM inequality. The equality holds if and only if $|v_{1,j}| = \cdots = |v_{k,j}|$ for any $1 \le j \le d$. The rest follows by letting $x = v_1 \odot \cdots \odot v_k$. $\square$

For the case $k = 2$, the objective is $\sum_{i=1}^{d} |D_{ii}||x_i| = \|Dx\|_1$ which is the rescaled ℓ_1 norm for a near-identity matrix D. Hence, an argument similar to those in Section 4 reveals the minimizer is uniquely $x_\natural$. We state this result formally below.

Lemma 9.3. *There is a universal constant $c > 0$ such that if the linear map A satisfying $(1-\delta, 1+\delta)$ RIP with $0 < \delta < c$. Then for $k = 2$, any solution (v_1, v_2) to* (76) *satisfies $x_\natural = v_1 \odot v_2$.*

On the other hand, higher values of k *do not* encourage sparsity. In the extreme case $k \to \infty$, the objective function in (79) is close to $\|x\|_2^2$ which should give a dense solution in general. Indeed, in Figure 5, we plot the solution performance of (79) for different values $k = \{2, 3, \ldots, 10\}$ and $r_\natural = \{1, 2, 3, 4, 5\}$ measured by the relative error $\frac{\|x - x_\natural\|_2}{\|x_\natural\|_2}$. [6] Indeed, exact recovery is observed for $k = 2$, while the relative error degrades significantly as k increases.

---

[6] We set $d = 1000$ and $m = 3\lceil r_\natural \log d \rceil$ and generate the signal $x_\natural$ with first k components being 1 and zero otherwise. For each configuration of $(k, r_\natural)$, we randomly generate 25 realizations of the Gaussian sensing matrix A and solve (79) for each A. The performance metric $\frac{\|x - x_\natural\|_2}{\|x_\natural\|_2}$ is averaged over these 25 trials.
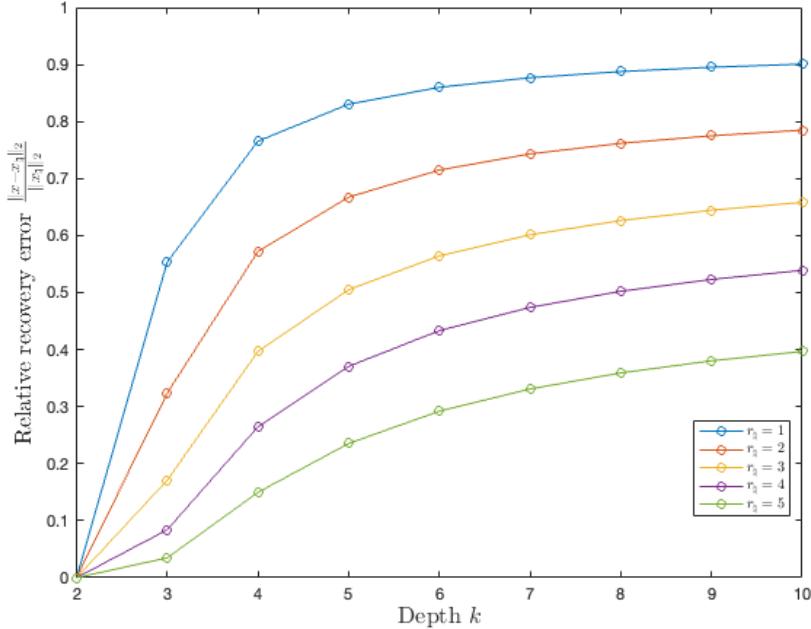
Figure 5: The effect of depth for different choice of sparsity $r_\natural$

References

[1] Ali Ahmed, Benjamin Recht, and Justin Romberg. Blind deconvolution using convex programming. *IEEE Transactions on Information Theory*, 60(3):1711–1732, 2013.

[2] Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

[3] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[4] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[5] T Tony Cai and Anru Zhang. Supplement to "rop: Matrix recovery via rank-one projections". DOI: 10.1214/14-AOS1267SUPP.

[6] T Tony Cai and Anru Zhang. Rop: Matrix recovery via rank-one projections. *The Annals of Statistics*, 43(1):102–138, 2015.

[7] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011.

[8] Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011.

[9] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[10] Emmanuel J Candes, Thomas Strohmer, and Vladislav Voroninski. Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[11] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[12] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018, 2019.

[13] Yudong Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, 2015.

[14] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Constantine Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7):4324–4337, 2013.

[15] Yuxin Chen, Yuejie Chi, and Andrea J Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[16] Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 66(11):7274–7301, 2020.

[17] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pages 1019–1028. PMLR, 2017.

[18] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *arXiv preprint arXiv:1806.00900*, 2018.

[19] Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

[20] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

[21] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *International Conference on Machine Learning*, pages 1233–1242. PMLR, 2017.

[22] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. *arXiv preprint arXiv:1806.00468*, 2018.

[23] Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–10. IEEE, 2018.

[24] Wooseok Ha, Haoyang Liu, and Rina Foygel Barber. An equivalence between critical points for rank constraints versus low-rank factorizations. *SIAM Journal on Optimization*, 30(4):2927–2955, 2020.

[25] Reinhard Heckel and Mahdi Soltanolkotabi. Compressive sensing with un-trained neural networks: Gradient descent finds a smooth approximation. In *International Conference on Machine Learning*, pages 4149–4158. PMLR, 2020.

[26] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

[27] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *arXiv preprint arXiv:1705.08741*, 2017.

[28] Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. *Advances in neural information processing systems*, 32:103–112, 2019.

[29] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

[30] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*, 2018.

[31] Stanislaw Jastrzebski, Zachary Kenton, Devansh Arpit, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. Three factors influencing minima in sgd. *arXiv preprint arXiv:1711.04623*, 2017.

[32] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

[33] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*, 2017.

[34] Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in overparameterized matrix sensing and neural networks with quadratic activations. In *Conference On Learning Theory*, pages 2–47. PMLR, 2018.

[35] Shuyang Ling and Thomas Strohmer. Self-calibration and biconvex compressive sensing. *Inverse Problems*, 31(11):115002, 2015.

[36] Tianyi Liu, Yan Li, Song Wei, Enlu Zhou, and Tuo Zhao. Noisy gradient descent converges to flat minima for nonconvex matrix factorization. In *International Conference on Artificial Intelligence and Statistics*, pages 1891–1899. PMLR, 2021.

[37] Cong Ma, Yuanxin Li, and Yuejie Chi. Beyond procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing. *IEEE Transactions on Signal Processing*, 69:867–877, 2021.

[38] Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

[39] Rotem Mulayoff and Tomer Michaeli. Unique properties of flat minima in deep networks. In *International Conference on Machine Learning*, pages 7108–7118. PMLR, 2020.

[40] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring generalization in deep learning. *arXiv preprint arXiv:1706.08947*, 2017.

[41] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[42] Keskar Nitish Shirish, Mudigere Dheevatsa, Nocedal Jorge, Smelyanskiy Mikhail, and Tang Ping Tak Peter. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[43] Matthew D Norton and Johannes O Royset. Diametrical risk minimization: Theory and computations. *Machine Learning*, pages 1–19, 2021.

[44] Greg Ongie and Rebecca Willett. The role of linear layers in nonlinear interpolating networks. *arXiv preprint arXiv:2202.00856*, 2022.

[45] Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

[46] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(12), 2011.

[47] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[48] Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pages 2667–2690. PMLR, 2019.

[49] Ohad Shamir. Are resnets provably better than linear predictors? *arXiv preprint arXiv:1804.06739*, 2018.

[50] Samuel L Smith and Quoc V Le. A bayesian perspective on generalization and stochastic gradient descent. *arXiv preprint arXiv:1710.06451*, 2017.

[51] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

[52] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[53] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005.

[54] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[55] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

[56] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

[57] Tian Ye and Simon S Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems*, 34, 2021.

[58] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

A   Extension to noisy observation

This section considers an extension of the flat solution concept to the setting where the observations are corrupted by noise:

$$b = \mathcal{A}(M_\natural) + e, \quad e \sim N(0, \sigma^2 I_m), \tag{81}$$

where $\sigma > 0$ is the noise level and $N(0, I_m)$ is the standard m-dimensional Gaussian. Our discussion in the rest of the paper focused on the simpler case $\sigma = 0$. We define the flat solution in this setting as follows. We continue to use the scaled trace $\mathrm{str}(D^2 f(L, R))$ as the flatness measure of the objective function. However, instead of considering all solutions (L, R) that interpolate the data, we consider those pairs (L, R) that are in the sublevel set:

$$\{(L, R) \mid \left\| \mathcal{A}(LR^\top) - b \right\|_2 \leq \|e\|_2 \}. \tag{82}$$

The reason for this choice is that in the noisy observation setting, the global solution of (1) (with $k = \min\{d_1, d_2\}$) has the potential of overfitting no matter what regularization has been enforced. Indeed, consider the simplest case $\mathcal{A} = \mathcal{I}$, i.e., the map $\mathcal{A}$ is the identity map. In this setting, any global minimizer $LR^\top$ is simply the observation $b = M_\natural + e$ itself. With the above preparation, we define the flat solutions to be the minimizers of the following problem.

$$\min_{L \in \mathbb{R}^{d_1 \times k}, \, R \in \mathbb{R}^{d_2 \times k}} \mathrm{str}(D^2 f(L, R)) \quad \text{subject to} \quad \left\| \mathcal{A}(LR^\top) - b \right\|_2 \leq \|e\|_2. \tag{83}$$

The goal of the section is to prove the following.

Theorem A.1 (Noisy matrix sensing). *Suppose that $\mathcal{A}$ is a Gaussian ensemble and the noise follows $e \sim N(0, \sigma^2 I_m)$. Then there exists universal constants c, C such that for any $m \geq C r_\natural d_{\max}$, with probability at least $1 - C \exp(-c(d_1 + d_2))$, any solution (L_f, R_f) of (83) satisfies*

$$\left\| L_f R_f^\top - M_\natural \right\|_F \leq C\sigma \sqrt{\frac{r_\natural(d_1 + d_2)}{m}}. \tag{84}$$

Note that the bound $\sigma \sqrt{\frac{r_\natural(d_1 + d_2)}{m}}$ is minimax optimal according to [8].

A.1 Proof of Theorem A.1

Following (19) and (20) in Section 3.1, we see that (83) is equivalent to (in the sense of Theorem 3.2) minimizing the nuclear norm over rank constrained matrices so long as D_1, D_2 matrices are invertible[7]:

$$\min_{X \in \mathbb{R}^{d_1 \times d_2}:\ \text{rank}(X) \le k} \|X\|_* \qquad \text{subject to} \qquad \|\mathcal{A}(D_1^{-1} X D_2^{-1}) - b\| \le \|e\|. \tag{85}$$

Let $\hat{Y}$ be any minimizer of (85). Also denote the scaled linear map $\tilde{\mathcal{A}}(\cdot) = \mathcal{A}(D_1^{-1} \cdot D_2^{-1})$, the scaled ground truth $Y_0 = D_1 M_\natural D_2$, and the difference $\Delta = \hat{Y} - Y_0$. Our task is to show $\|\Delta\|_{\mathrm{F}} \lesssim \sigma \sqrt{\frac{r_\natural (d_1 + d_2)}{m}}$. The bound (84) then immediately follows using $D_1 L R^\top D_2 = X$ and the fact that D_1 and D_2 are near identity from Lemma 4.8.

A.1.1 Bound $\|\Delta\|_{\mathrm{F}}$

Our proof is based on the argument in [47]. Starting with the feasibility of $\hat{Y}$, we have

$$\frac{1}{2m}\left\|\tilde{\mathcal{A}}(\hat{Y}) - b\right\|_2^2 \le \frac{1}{2m}\|e\|_2^2 \overset{(a)}{\Longleftrightarrow} \frac{1}{2m}\left\|\tilde{\mathcal{A}}(\Delta)\right\|_2^2 + \frac{1}{m}\langle \tilde{\mathcal{A}}(\Delta), e\rangle \le 0$$
$$\implies \frac{1}{2m}\left\|\tilde{\mathcal{A}}(\Delta)\right\|_2^2 \le -\frac{1}{m}\langle \Delta, \tilde{\mathcal{A}}^* e\rangle \overset{(b)}{\le} \left\|\tilde{\mathcal{A}}^* e\right\|_{\mathrm{op}} \|\Delta\|_*. \tag{86}$$

Here the step (a) is due to expanding the square for the left hand side, and the step (b) is due to Hölder's inequality. We next try to lower bound $\frac{1}{2m}\left\|\tilde{\mathcal{A}}(\Delta)\right\|_2^2$ and upper bound $\left\|\tilde{\mathcal{A}}^* e\right\|_{\mathrm{op}}$ and $\|\Delta\|_*$.

Upper bound $\|\Delta\|_*$ First, let us introduce a lemma that decomposes Δ.

Lemma A.2. [47, Lemma 2.3 and 3.4] For any $A, B \in \mathbb{R}^{d \times d}$, there exists B_1, B_2 such that (1) $B = B_1 + B_2$, (2) $\text{rank}(B_1) \le \text{rank}(A)$, (3) $A B_2^\top = 0$ and $A^\top B_2 = 0$, (4) $\|A + B_2\|_* = \|A\|_* + \|B_2\|_*$, and (5) $\langle B_1, B_2\rangle = 0$.

Using Lemma A.2, we can decompose $\Delta = R_0 + R_c$ such that $Y_0 R_c^\top = 0, Y_0^\top R_c = 0$, $R_0 \le 2r_\natural$, $\langle R_0, R_c\rangle = 0$, and $\|Y_0 + R_c\|_* = \|Y_0\|_* + \|R_c\|_*$. Hence, we have

$$\left\|\hat{Y}\right\|_* = \|Y_0 + \Delta\|_* \ge \|Y_0 + R_c\|_* - \|R_0\|_* = \|Y_0\|_* + \|R_c\|_* - \|R_0\|_*. \tag{87}$$

Using the optimality of $\left\|\hat{Y}\right\|_* \le \|Y_0\|_*$, we have that

$$\|R_c\|_* \le \|R_0\|_*. \tag{88}$$

Using the fact that R_0 has rank no more than $2r_\natural$ and $\langle R_0, R_c\rangle = 0$, we have

$$\|\Delta\|_* \le \|R_c\|_* + \|R_0\|_* \le 2\|R_0\|_* \overset{(b)}{\le} 2\sqrt{2r_\natural}\|R_0\|_{\mathrm{F}} \overset{(c)}{\le} 2\sqrt{2r_\natural}\|\Delta\|_{\mathrm{F}}. \tag{89}$$

Here in the first step, we use the triangle inequality for $\|\Delta\|_*$. This finishes the upper bound of $\|\Delta\|_*$.

---

[7]Note that the condition $\mathcal{A}(LR^\top) = b$ is not needed for (16) to hold, which is critical for the step (19) to hold in the noisy case.

Lower bound on $\frac{1}{m}\left\|\tilde{\mathcal{A}}(\Delta)\right\|_2^2$. Next we partition R_c into a sum of matrices $R_1, R_2, \ldots$ each of rank at most $3r_\natural$ as in [47, Theorem 3.3]. Let $R_c = U \operatorname{diag}(\sigma)V'$ be the singular value decomposition of R_c. For each $i \geq 1$ define the index set $I_i = \{3r_\natural(i-1)+1, \ldots, 3r_\natural i\}$, and let $R_i := U_{I_i} \operatorname{diag}(\sigma_{I_i})V'_{I_i}$. Using the fact that $\langle R_c, R_0 \rangle = 0$ and the construction of R_i, $i \geq 1$, we also have

$$\langle R_i, R_j \rangle = 0, \quad \forall\, i \neq j, i, j \geq 0. \tag{90}$$

By construction, we have

$$\sigma_k \leq \frac{1}{3r_\natural} \sum_{j \in I_i} \sigma_j \qquad \forall\, j \in I_{i+1}, \tag{91}$$

which implies $\|R_{i+1}\|_F^2 \leq \frac{1}{3r_\natural}\|R_i\|_*^2$. We can then compute the following bound

$$\sum_{j \geq 2} \|R_j\|_F \leq \frac{1}{\sqrt{3r_\natural}} \sum_{j \geq 1} \|R_j\|_* = \frac{1}{\sqrt{3r_\natural}}\|R_c\|_* \overset{(a)}{\leq} \frac{1}{\sqrt{3r_\natural}}\|R_0\|_* \overset{(b)}{\leq} \frac{\sqrt{2r_\natural}}{\sqrt{3r_\natural}}\|R_0\|_F \tag{92}$$

where the step (a) is due to (87), and the step (b) is due to the fact that $\operatorname{rank}(R_0) \leq 2r_\natural$. From this inequality, we also have

$$\|\Delta\|_F \leq \|R_0 + R_1\|_F + \sum_{j \geq 2}\|R_j\|_F \leq \|R_0 + R_1\|_F + 2\|R_0\|_F \leq 3\|R_0 + R_1\|_F \tag{93}$$

The last equality is due to that $\langle R_0, R_1 \rangle = 0$. Hence, we have that

$$
\begin{aligned}
\frac{1}{\sqrt{m}}\|\tilde{\mathcal{A}}(\Delta)\|_2 &\overset{(a)}{\geq} \frac{1}{\sqrt{m}}\|\tilde{\mathcal{A}}(R_0 + R_1)\|_2 - \sum_{j \geq 2} \frac{1}{\sqrt{m}}\|\tilde{\mathcal{A}}(R_j)\|_2 \\
&\geq (1 - \delta)\|R_0 + R_1\|_F - (1 + \delta) \sum_{j \geq 2}\|R_j\|_F \\
&\overset{(b)}{\geq} (1 - \delta)\|R_0 + R_1\|_F - \sqrt{\tfrac{2}{3}}(1 + \delta)\|R_0\|_F \\
&\overset{(c)}{\geq} c_1\|R_0 + R_1\|_F \\
&\overset{(d)}{\geq} c_2\|\Delta\|_F.
\end{aligned}
\tag{94}
$$

Here, in the step (a), we use the reversed triangle inequality. In the step (b), we use (92). In the step (c), we use $\langle R_0, R_1 \rangle = 0$ and the choice of δ. The last step (d) is due to (93). This finishes the lower bound on $\frac{1}{\sqrt{m}}\|\tilde{\mathcal{A}}(\Delta)\|_2^2$.

Estimating $\left\|\tilde{\mathcal{A}}^* e\right\|_{\mathrm{op}}$. Note that $\tilde{\mathcal{A}}^*(e) = D_1^{-1}(\mathcal{A}^* e)D_2^{-2}$. Hence $\left\|\tilde{\mathcal{A}}^*(e)\right\|_{\mathrm{op}} = \left\|D_1^{-1}\mathcal{A}^*(e)D_2^{-2}\right\|_{\mathrm{op}} \leq \left\|D_1^{-1}\right\|_{\mathrm{op}}\|\mathcal{A}^* e\|_{\mathrm{op}}\left\|D_2^{-1}\right\|_{\mathrm{op}}$. From [56, Proof of Corollary 10.10], we know that with probability at least $1 - c\exp(-c(d_1 + d_2))$, the inequality $\frac{1}{m}\|\mathcal{A}^* e\|_{\mathrm{op}} \leq C\sigma\sqrt{\frac{d_1+d_2}{m}}$ holds. Since $0.9I \preceq D_i \preceq 1.1I$, we have

$$\frac{1}{m}\left\|\tilde{\mathcal{A}}(e)\right\|_{\mathrm{op}} \leq C'\sigma\sqrt{\frac{d_1 + d_2}{m}}. \tag{95}$$

Combining (86), (89), (94), and (95), we conclude $\|\Delta\|_F \lesssim \sigma\sqrt{\frac{r_\natural(d_1+d_2)}{m}}$, as claimed.

A.2   A numerical demonstration

Finally, we validate Theorem A.1 via a numerical experiments. We compare the performance of the minimizer $\hat{X}$ of Problem (85) for the case $k = \min\{d_1, d_2\}$ and the solution $\hat{X}_{\text{nuc}}$ of the nuclear norm minimization (Problem (85) with D_1, D_2 being the identity).

Experiment setup   We set $d = d_1 = d_2 = 25$ and $m = 1000$. We generate the underlying unit Frobenius norm ground truth matrix $M_\natural$ randomly with rank $r_\natural = \{1, 2, 3, \ldots, 10\}$. We vary the noise level $\sigma = \{0.1, 0.2, \ldots, 1.3, 1.4, 1.5\}$. For each rank $r_\natural$, we generate the sensing Gaussian ensemble $\mathcal{A}$ with $m = 1000$ and use the same one for different noise levels. Then for each noise level, we generate 25 realization of the noise e following $N(0, \sigma^2 I)$, and solve the corresponding Problem (85) and the nuclear norm minimization problem. We then average the error $\left\| D_1^{-1} \hat{X} D_2^{-1} - M_\natural \right\|_{\text{F}}$ and $\left\| \hat{X} - M_\natural \right\|_{\text{F}}$ over the 25 trials for each configuration of $r_\natural$ and σ.

Recovery Performance   We plot the error in Figure 6. The white color indicates small error and the dark color indicates large error. It can be seen that it is hard to differentiate the performance of the solution to the nuclear norm minimization problem and the solution to Problem (85). This result validates our theoretical results in Theorem A.1.
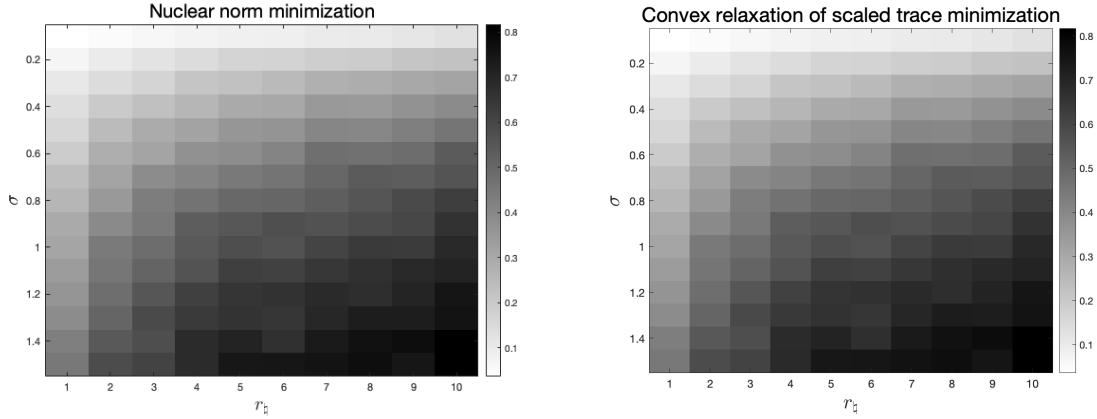

Figure 6: Matrix Sensing with nuclear norm minimization (left) and the convex relaxation (or $k = d$) of Problem (85) (right) for different configurations of $(r_\natural, \sigma)$.