Information and Inference: A Journal of the IMA (2024) **00**, iaae009
https://doi.org/10.1093/imaiai/iaae009

Flat minima generalize for low-rank matrix recovery

LIJUN DING
Wisconsin Institute for Discovery, University of Wisconsin - Madison, Madison, WI 53795, USA

DMITRIY DRUSVYATSKIY*
Department of Mathematics, University of Washington, Seattle, WA 98195, USA
*Corresponding author. Email: ddrusv@uw.edu

MARYAM FAZEL
Electrical and Computer Engineering Department, University of Washington, Seattle, WA 98195, USA

AND

ZAID HARCHAOUI
Department of Statistics, University of Washington, Seattle, WA 98195, USA

[Received on 14 March 2022; revised on 30 December 2023; accepted on 14 March 2024]

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 analyse overparameterized matrix and bilinear sensing, robust principal component analysis, 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 complete the paper with synthetic experiments that illustrate our findings.

Keywords: Flat minimizers; overparameterization; low rank matrix recovery; nuclear norm; matrix sensing; bilinear sensing; matrix completion; covariance matrix estimation; robust PCA; neural network;

2020 Math Subject Classification: 68T09; 68T07; 90C26.

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. 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

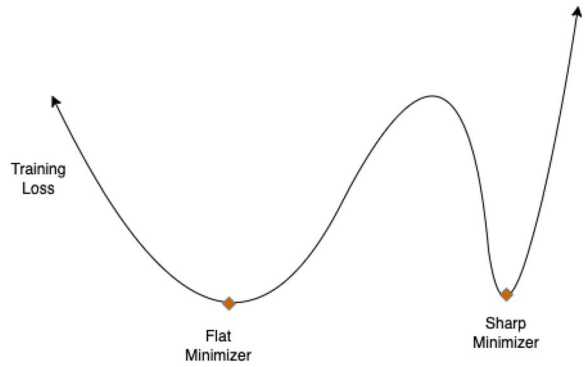


FIG. 1. Flat vs. sharp minima of the training loss.

[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. 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 function grows slowly—generalize well. See Fig. 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?

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

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 $b = \mathscr{A}(M_\natural)$ under a linear measurement map $\mathscr{A}$: $\mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$. A common approach to this task is through the non-convex optimization problem:

$$\min_{L,R} \quad f(L,R) := \left\| \mathscr{A}(LR^\top) - b \right\|_2^2 \qquad \text{with } L \in \mathbb{R}^{d_1 \times k} \text{ and } R \in \mathbb{R}^{d_2 \times k}. \tag{1.1}$$

The set of minimizers of f, which we denote by $\mathscr{M}$, consists of all solutions to the equation $\mathscr{A}(LR^\top) = b$. 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 $(L,R) \in \mathscr{M}$ as follows.

- (L,R) is **norm-minimal** if it minimizes over $\mathscr{M}$ 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 $\mathscr{M}$ 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 $\mathscr{M}$ to the origin in the Frobenius norm. Being balanced amounts to requiring L and R to have the same singular values and right-singular vectors. Flat solutions are defined using the 'scaled trace' of the quadratic form $D^2 f(L,R)$ defined as

$$\mathrm{str}(D^2 f(L,R)) := \tfrac{1}{d_1} \sum_{i \leq d_1, j \in [k]} D^2 f(L,R)\big[e_i e_j^\top\big] + \tfrac{1}{d_2} \sum_{i > d_1, j \in [k]} D^2 f(L,R)\big[e_i e_j^\top\big], \tag{1.2}$$

where e_i and e_j are the unit coordinate vectors in $\mathbb{R}^{d_1+d_2}$ and $\mathbb{R}^k$, respectively. [1] 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 (1.1) *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 $\mathscr{M}$. 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 principal component analysis (PCA), covariance matrix estimation and single hidden layer neural networks with

[1] The input space of f is viewed as $\mathbb{R}^{(d_1+d_2)\times k}$ by stacking the two matrices space $\mathbb{R}^{d_1 \times k}$ and $\mathbb{R}^{d_2 \times k}$.

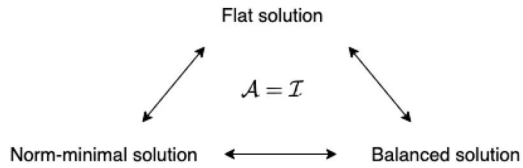


FIG. 2. Equivalence between balanced, minimal norm and flat solutions when $\mathscr{A} = \mathscr{I}$.

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.

Note that we use flat solutions to mean those points that are global minimizers of (1.1) with the smallest scaled trace. One can extend our results to solutions in $\mathscr{M}$ that is only approximately minimal in scaled trace and achieve approximate recovery results. It would be of great theoretical and practical interests in extending the result here to local minimizers of (1.1), as algorithms are not guaranteed to find a global one a priori. However, we note that in neither case, exact recovery is possible. We restrict our attention to solutions in $\mathscr{M}$ with smallest trace for simplicity and crisp statement of flat solutions.

One may interpret our results in three ways: (1) in terms of regularization and generalization property, our result suggests that the measure of flatness, the scaled trace, could be a reasonable regularizer for some learning tasks, due to the generalization power and near norm minimalitly property of flat solutions in overparametrized matrix factorization; (2) algorithmically, our result serves as a theoretical basis for methods favouring flat solutions such as those in [12,20,29,43]; (3) in terms of optimization landscape of (1.1), due to near balancedness, the objective function is well-behaved near flat solutions, meaning the Lipschitz constant of the gradient is small [18]. This implies stable and quicker convergence behaviour of first-order methods near flat solutions.

1.2 *Main results and outline of the paper*

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

An immediate difficulty with analysing flat solutions of the problem (1.1) with a general measurement map $\mathscr{A}$ is that flat solutions are defined as minimizers of a highly non-convex optimization problem corresponding to minimizing the scaled trace over the solution set. In Section 3, we derive a simple convex relaxation of flat minimizers. Setting the notation, let us write $\mathscr{A}$ as $\mathscr{A}(X) = (\langle A_i, X\rangle, \ldots, \langle A_m, X\rangle)$ for some matrices $A_i \in \mathbb{R}^{d_1 \times d_2}$ and define the 'rescaling' matrices

$$D_1 := \left(\frac{1}{md_2}\sum_{j=1}^{m} A_i A_i^\top\right)^{\frac{1}{2}} \quad \text{and} \quad D_2 := \left(\frac{1}{md_1}\sum_{j=1}^{m} A_i^\top A_i\right)^{\frac{1}{2}}. \tag{1.3}$$

We will show in Theorem 3.1 that flat solutions can be identified with minimizers of the problem

$$\min_{X \in \mathbb{R}^{d_1 \times d_2}:\ \mathrm{rank}(X) \le k} \left\| D_1 X D_2 \right\|_* \qquad \text{subject to} \qquad \mathscr{A}(X) = b. \tag{1.4}$$

It is worthwhile to note that without the D_1 and D_2 matrices, the problem (3.6) 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:

$$\min_{X\in\mathbb{R}^{d_1\times d_2}} \left\|D_1 X D_2\right\|_* \qquad \text{subject to} \qquad \mathscr{A}(X) = b. \tag{1.5}$$

Summarizing, verifying that flat solutions exactly recover $M_\natural$ is reduced to showing that $M_\natural$ (which has rank $r_\natural$) is the unique solution of the convex problem (1.5).

In Section 4, we will show that if the map $\mathscr{A}$ satisfies ℓ_2/ℓ_2 or ℓ_1/ℓ_2 restricted isometry properties (RIPs) and the rescaling matrices D_1 and D_2 are sufficiently close to the identity, then $M_\natural$ is the unique solution of (1.5). As a consequence, we deduce that flat solutions exactly recover $M_\natural$ 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 i.i.d. standard Gaussian, 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 $\mathscr{A}$ is generated according to a Gaussian matrix sensing or bilinear sensing model. Then as long as we are in the regime $m \gtrsim r_\natural d_{\max}$ and $d_{\min} \gtrsim \log m$, with high probability, any flat solution (L_f, R_f) satisfies $L_f R_f^\top = M_\natural$, and is nearly norm-minimal and nearly balanced.

Note that our requirement on the sample size $m \gtrsim r_\natural d_{\max}$ matches the known regime for exact recovery with nuclear norm minimization [5,8]. 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 analysing 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 success probability $p \in (0, 1)$. The main difficulty with analysing the matrix completion problem is that the linear map $\mathscr{A}$ does not have good RIPs. Moreover, the existing techniques for analysing the nuclear norm relaxation of the matrix completion problem [9,46] do not directly apply to the problem (1.5) because of the dependence between the rescaling matrices D_1, D_2, and the observation map $\mathscr{A}$. Consequently, we settle for an approximate recovery guarantee.

THEOREM 1.2 (Matrix completion (Informal)) Suppose that $\mathscr{A}$ is generated from the Bernoulli matrix completion model with success probability $p > 0$ and let $\mu > 0$ be an incoherence parameter of $M_\natural$.[2] 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 $\left\|L_f R_f^\top - M_\natural\right\|_* \le \gamma \left\|M_\natural\right\|_*$ 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 $p \gtrsim$

[2] See (5.2) for the definition of the incoherence parameter μ.

$\mu r_\natural \log(\mu r_\natural)\frac{\log(d_{\max})}{d_{\min}}$. 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 PCA in Section 6. Though this problem is not of the form (1.1), 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

$$Y = M_\natural + S_\natural, \tag{1.6}$$

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

$$\min_{L,R} f(L,R) := \operatorname{dist}^2_{\Omega}(Y - LR^\top), \tag{1.7}$$

where $\operatorname{dist}^2_{\Omega}$ is the square Frobenius distance to the sparsity-inducing set $\Omega := \{S \mid \|S\|_{1,1} \le \left\|S_\natural\right\|_{1,1}\}$. 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 define a minimizer of (1.7) to be *flat* if it minimizes the scaled trace $\operatorname{str}(D^2 f_{L,R}(L,R))$ over all $(L,R) \in \mathscr{M}$. We will prove the following theorem, which largely follows from the results of [15].

THEOREM 1.3 (Robust PCA (Informal)) Let μ be the strong incoherence parameter of $M_\natural$.[3] 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 analyses 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

$$b_i = x_i^\top M_\natural x_i, \tag{1.8}$$

where $x_1,\dots,x_m \overset{\text{iid}}{\sim} N(0,I_d)$. Note that in the special case $r_2 = 0$, this problem reduces to covariance matrix estimation [14] 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

[3] See (6.1) for the definition of the strong incoherence parameter μ.

natural optimization formulation of the problem takes the form

$$\min_{U_1\in\mathbb{R}^{d\times k_1},U_2\in\mathbb{R}^{d\times k_2}} f(U_1,U_2)=\frac{1}{m}\left\|\mathscr{A}(U_1U_1^\top-U_2U_2^\top-M_\natural)\right\|_2^2, \tag{1.9}$$

where the sensing matrices are $A_i=x_ix_i^\top$ and $k_i\geq r_i$ for $i=1,2$. It is straightforward to verify the equality $\mathrm{str}(D^2f(U_1,U_2)))=d\,\mathrm{tr}(D^2f(U_1,U_2))$, and therefore we declare a minimizer $(U_{1,f},U_{2,f})$ to be *flat* if it has minimal trace $\mathrm{tr}(D^2f(U_1,U_2))$ among all minimizers of (1.9).

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 (1.9) 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 [14] in terms of r and d. An interesting example of (1.9) arises from a model of shallow neural networks, analysed 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 $M_\natural=U_\natural\mathrm{diag}(v)U_\natural^\top$ from the observations (1.8).

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 $\langle X,Y\rangle=\mathrm{tr}(X^\top Y)$ and the induced Frobenius norm $\|X\|_{\mathrm{F}}=\sqrt{\langle X,X\rangle}$. 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]:

$$\|X\|_*=\min_{X=LR^\top}\|L\|_{\mathrm{F}}\|R\|_{\mathrm{F}}=\min_{X=LR^\top}\tfrac{1}{2}\left(\|L\|_{\mathrm{F}}^2+\left\|R^2\right\|_{\mathrm{F}}\right). \tag{1.10}$$

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

In this section, we focus on the idealized objective (1.1) where the measurement map $\mathscr{A}$ is the identity:

$$\min_{L\in\mathbb{R}^{d_1\times k},R\in\mathbb{R}^{d_2\times k}} f(L,R) = \left\|LR^\top - M_\natural\right\|_{\mathrm{F}}^2. \tag{2.1}$$

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 $\mathscr{M}$ of (2.1) 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 $\mathscr{A}$ 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 $\mathscr{A} = \mathscr{I}$ 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 $(L,R) \in \mathscr{M}$ is the quadratic form, $D^2 f(L,R)[U,V] = 2\left\|LV^\top + UR^\top\right\|_{\mathrm{F}}^2$. Consequently, the scaled trace is simply

$$\mathrm{str}(D^2 f(L,R)) = 2(\|L\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2). \tag{2.2}$$

Proof. The claimed expression for $D^2 f(L,R)$ follows from elementary algebraic manipulations. 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 $D^2 f(L,R)[e_i e_j^\top] = 2\left\|R_j\right\|_{\mathrm{F}}^2$ for $i \leq d_1$ and $D^2 f(L,R))[e_i e_j^\top] = 2\left\|L_j\right\|_{\mathrm{F}}^2$ for $i > d_1$. Therefore, from the definition (1.2), the scaled trace becomes

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

as claimed. □

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

LEMMA 2.2 (Equivalence) Norm-minimal, flat and balanced solutions of (2.1) all coincide.

Proof. The equivalence of flat and norm-minimal solutions follows directly from (2.2). Next, we prove the equivalence between minimal norm and balanced solutions. Suppose $(L,R) \in \mathscr{M}$ is balanced. The equality $L^\top L = R^\top R$ implies that L and R have the same singular values and 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 $\left\|M_\natural\right\|_* = \mathrm{tr}(\Sigma^2)$. Noting the equality $\frac{1}{2}\left(\|L\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2\right) = \mathrm{tr}(\Sigma^2)$ along with (1.10), 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. □

3. Convex relaxation and regularity of flat solutions

In this section, we begin investigating flat minimizers of the problem (1.1) with general linear measurement maps $\mathscr{A}$. It will be convenient to write the linear map $\mathscr{A}(X)$ in coordinates as

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

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

$$D_1 = \left(\frac{1}{md_2}\sum_{i=1}^m A_i A_i^\top\right)^{\frac{1}{2}} \quad \text{and} \quad D_2 = \left(\frac{1}{md_1}\sum_{i=1}^m A_i^\top A_i\right)^{\frac{1}{2}}. \tag{3.1}$$

The section presents two main results: Theorems 3.1 and 3.2. 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. These two results suggest a two-part strategy in achieving our goals: exact recovery and regularity of flat solutions. We detail the strategy in Section 3.3 and discuss how it is used in the following sections.

3.1 *A convex relaxation for flat solutions*

Flat solutions are by definition minimizers of the highly non-convex problem $\min_{(L,R)\in\mathscr{M}} \mathrm{str}(D^2 f(L,R))$. 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 $\mathscr{A} = \mathscr{I}$. Lemma 3.1 provides an analogous statement for general maps $\mathscr{A}$ 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 $(L,R) \in \mathscr{M}$ is the quadratic form, $D^2 f(L,R)[U,V] = 2\langle \mathscr{A}(LV^\top + UR^\top)\|_2^2$. Moreover, the scaled trace can be written as

$$\mathrm{str}(D^2 f(L,R)) = 2m\Big(\|D_1 L\|_{\mathrm{F}}^2 + \|D_2 R\|_{\mathrm{F}}^2\Big). \tag{3.2}$$

Proof. The claimed expression for $D^2 f(L,R)$ follows from elementary algebraic manipulations. Next, we verify (3.2). To this end, the definition of the scaled trace (1.2) yields the expression

$$\mathrm{str}(\mathscr{H}(L,R)) = \frac{2}{d_1}\sum_{i=1}^{d_1}\sum_{j=1}^{k}\left\|\mathscr{A}\Big(e_i e_j^\top R^\top\Big)\right\|_2^2 + \frac{2}{d_2}\sum_{i=1}^{d_2}\sum_{j=1}^{k}\left\|\mathscr{A}\Big(Le_j e_i^\top\Big)\right\|_2^2. \tag{3.3}$$

Note that here e_i is a basis vector in $\mathbb{R}^{d_1}$ in the first summand and a basis vector in $\mathbb{R}^{d_2}$ for the second summand.

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

$$\begin{aligned}\sum_{i=1}^{d_2}\sum_{j=1}^{k}\left\|\mathscr{A}\left(Le_je_i^\top\right)\right\|_2^2 &= \sum_{i=1}^{d_2}\sum_{j=1}^{k}\sum_{l=1}^{m}\left\langle A_l, Le_je_i^\top\right\rangle^2 \\ &= \sum_{i=1}^{d_2}\sum_{j=1}^{k}\sum_{l=1}^{m}\left\langle A_{l,i}, L_j\right\rangle^2 \\ &= \sum_{i=1}^{d_2}\sum_{j=1}^{k}\sum_{l=1}^{m}\left\langle A_{l,i}A_{l,i}^\top, L_jL_j^\top\right\rangle \\ &= \sum_{l=1}^{m}\left\langle\sum_{i=1}^{d_2}A_{l,i}A_{l,i}^\top, \sum_{j=1}^{k}L_jL_j^\top\right\rangle = \sum_{l=1}^{m}\left\langle A_lA_l^\top, LL^\top\right\rangle = md_2\left\|D_1L\right\|_{\mathrm{F}}^2.\end{aligned} \tag{3.4}$$

A similar argument shows $\left\|D_2R\right\|_{\mathrm{F}}^2 = \frac{1}{md_1}\sum_{i=1}^{d_1}\sum_{j=1}^{k}\left\|\mathscr{A}(e_ie_j^\top R^\top)\right\|_2^2$, completing the proof. □

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

$$\min_{L,R}\frac{1}{2}\left(\left\|D_1L\right\|_{\mathrm{F}}^2+\left\|D_2R\right\|_{\mathrm{F}}^2\right)\qquad\text{subject to}\qquad\mathscr{A}(LR^\top)=b. \tag{3.5}$$

In turn, it follows directly from (1.10) that so long as D_1, D_2 are invertible, the problem (3.5) is equivalent to minimizing the nuclear norm over rank constrained matrices:

$$\min_{X\in\mathbb{R}^{d_1\times d_2}:\mathrm{rank}(X)\le k}\|X\|_*\qquad\text{subject to}\qquad\mathscr{A}(D_1^{-1}XD_2^{-1})=b. \tag{3.6}$$

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

$$\min_{X\in\mathbb{R}^{d_1\times d_2}}\|X\|_*\qquad\text{subject to}\qquad\mathscr{A}(D_1^{-1}XD_2^{-1})=b. \tag{3.7}$$

These observations are summarized in the following theorem.

THEOREM 3.1 (Convex relaxation) Suppose the matrices D_1 and D_2 are invertible. Then the problems (3.5) are (3.6) are equivalent in the following sense. Let $l=\min(k,d_{\min})$.

1. the optimal values of (3.5) and (3.6) are equal.
2. if L, R solves (3.5), then $X=D_1LR^\top D_2$ is a minimizer of (3.6).
3. if a solution X of (3.6) has a singular value decomposition $X=U\Sigma V^\top$ for some diagonal matrix $\Sigma\in\mathbb{R}^{l\times l}$ with non-negative entries, then the matrices $L=D_1^{-1}U\sqrt{\Sigma}$ and $R=D_2^{-1}V\sqrt{\Sigma}$ are minimizers of (3.5) 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 (3.5) when $l<k$.

Moreover, if $X = D_1 M_\natural D_2$ is the unique minimizer of the problem (3.7), then any flat solution (L, R) satisfies $LR^\top = M_\natural$.

Section 4 will verify that the convex relaxation (3.7) indeed recovers $M_\natural$ under RIPs on the measurement map $\mathscr{A}$, 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 3.2 (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 (1.1) satisfies the following properties.

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

$$\left\|L_f\right\|_{\mathrm{F}}^2 + \left\|R_f\right\|_{\mathrm{F}}^2 \leq \kappa^2 \cdot \left(\min_{\mathscr{A}(LR^\top)=b} \|L\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2 \right). \tag{3.8}$$

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

$$\left\|L_f^\top L_f - R_f^\top R_f\right\|_* \leq 2\left(\kappa^2 - 1\right) \left\|M_\natural\right\|_*. \tag{3.9}$$

The proof of Theorem 3.2 relies on the following linear algebraic lemma, proved in Appendix B.1.

LEMMA 3.2 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

$$\min_{\tilde{L}, \tilde{R}:\, Q_1 \tilde{L} \tilde{R}^\top Q_2 = X} \frac{1}{2}\left(\left\|Q_1 \tilde{L}\right\|_{\mathrm{F}}^2 + \left\|Q_2 \tilde{R}\right\|_{\mathrm{F}}^2 \right), \tag{3.10}$$

satisfies the inequality:

$$\left\|L^\top L - R^\top R\right\|_* \leq \left(1 - \kappa^{-2}\right)\left(\|L\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2\right). \tag{3.11}$$

We are now ready to prove Theorem 3.2.

Proof of Theorem 3.2. We first prove inequality (3.8). To this end, for any $(L, R) \in \mathscr{M}$, we successively estimate:

$$\frac{\alpha_1^2}{2}\left(\left\|L_f\right\|_{\mathrm{F}}^2 + \left\|R_f\right\|_{\mathrm{F}}^2\right) \leq \frac{1}{2}\left(\left\|D_1 L_f\right\|_{\mathrm{F}}^2 + \left\|D_2 R_f\right\|_{\mathrm{F}}^2\right) \leq \frac{1}{2}\left(\left\|D_1 L\right\|_{\mathrm{F}}^2 + \left\|D_2 R\right\|_{\mathrm{F}}^2\right) \leq \frac{\alpha_2^2}{2}\left(\|L\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2\right),$$

where the second inequality follows from the characterization (3.5) of flat solutions. Taking the infimum over pairs $(L, R) \in \mathscr{M}$ completes the proof of (3.8).

Next, 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

$$\min_{\tilde{L},\tilde{R}:\, D_1\tilde{L}\tilde{R}^\top D_2 = X} \frac{1}{2}\left(\left\|D_1\tilde{L}\right\|_{\mathrm{F}}^2 + \left\|D_2\tilde{R}\right\|_{\mathrm{F}}^2\right). \tag{3.12}$$

Lemma 3.2 together with the already established estimate (3.8) implies

$$\left\|L_f^\top L_f - R_f^\top R_f\right\|_* \le \left(\kappa^2 - 1\right)\left(\|L\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2\right),$$

for all $(L, R) \in \mathscr{M}$. In particular, minimizing the right-hand side over L, R satisfying $M_\natural = LR^\top$ yields an upper bound of $2\left\|M_\natural\right\|_*$. The proof is complete. □

3.3 *Proof strategy for exact recovery and regularity of flat solutions*

The previous two sections, Section 3.1 and Section 3.2, motivate a two-part strategy for achieving our two goals:

1. Exact recovery of flat solutions: via Theorem 3.1, we only need to argue that the convex relaxation (3.7) admits $D_1 M_\natural D_2$ as its unique minimizer.
2. Regularity of flat solutions: via Theorem 3.2, we only need to show that the condition numbers of the scaling matrices D_1 and D_2 are close to one.

We implement this strategy in Section 4 based on the RIP [4] of $\mathscr{A}$ with two applications: matrix and bilinear sensing. In Sections 5, 6 and 7, we again follow this strategy. However, each of the problems instances needs a different argument than the one in Section 4 based on RIP. The specific reasons and the way we deal with each problem are listed below:

- Matrix completion: due to lacking RIP, we settle for a weak recovery result. We first establish regularity by checking the condition number of D_1 and D_2 directly in Lemma 5.1. Note that this result tells us that the product of the pair in flat solutions is near optimal for the nuclear norm minimization problem (8.2). See Lemma 5.2 for detail. We then utilize this observation and a sharp growth result, formalized in Lemma 5.3, to achieve a weak recovery result.
- Robust PCA: the objective in (1.7) differs from (1.1) and it lacks smoothness. We deal with this issue by redefining the flat solution using a smooth surrogate. We then achieve exact recovery result by analysing the convex relaxation (6.9) (different from (3.7)) using [15, Theorem 3]. The regularity property follows directly as D_1 and D_2 are identity in this case.
- Neural network with quadratic activations: its formulation (1.9) is different from (1.1), hence we need to redefine the flat solution using the trace of the Hessian. It also requires a technical lemma, Lemma 7.3, which mimics the characterization of nuclear norm (1.10) for a symmetric decomposition of symmetric matrices. This lemma then allows us to analyse a different convex relaxation (7.11) and utilize results in Section 4 for exact recovery.

[4] RIP is formally defined in Definition 4.1

We do not further discuss regularity of flat solution for the last case since the definition of minimal norm and balance of solution do not directly apply to this problem. However, we believe they can be properly defined and proved using the analysis in Section 7. We omit the details for simplicity.

4. Flat minima under RIP conditions: matrix and bilinear sensing

In this section, we follow our recipe in Section 3.3 using the restricted isometry property and achieve our goals for flat solutions of the two examples: matrix and bilinear sensing. Specifically, in Section 4.1, we formally define RIP and show in Lemma 4.2 that exact recovery of flat solution is implied by RIP and well conditioning of D_1 and D_2. In Section 4.2, we validate the RIP property and well conditioning of D_1 and D_2 of two important applications: matrix and bilinear sensing. We then apply Lemma 4.2 and Theorem 3.2 to the two examples to achieve our goals: exact recovery and regularity of flat solutions.

4.1 *Exact recovery via RIP and well conditioning of scaling matrices*

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

DEFINITION 4.1 (Restricted isometry property) A linear map $\mathscr{A}\colon \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$ satisfies an ℓ_p/ℓ_2 *RIP with parameters* (r, δ_1, δ_2) if the estimate

$$\delta_1 \|X\|_{\mathrm{F}} \le \frac{\|\mathscr{A}(X)\|_p}{m^{1/p}} \le \delta_2 \|X\|_{\mathrm{F}} \tag{4.1}$$

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

We will mainly focus on the $p = 1$ and $p = 2$ in this paper. Recall that our goal is to show that under RIP conditions, with well conditioning of D_1 and D_2, flat solutions exactly recover the $M_\natural$. We will need the following lemma, whose proof is immediate from definitions.

LEMMA 4.1 Let $\mathscr{A}(\cdot)$ 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 map $\mathscr{A}(Q_1^{-1} \cdot Q_2^{-1})$ 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 $\mathscr{A}(\cdot)$ satisfies RIP, then so does the perturbed map $\mathscr{A}(Q_1 \cdot Q_2)$, provided that the condition numbers of the positive definite matrices Q_1 and Q_2 are sufficiently close to one. We shall apply this lemma in the next section by setting $Q_i = D_i$ for $i = 1, 2$.

LEMMA 4.2 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 $\mathscr{A}(\cdot)$ be a linear map satisfying one of the following conditions.

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

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

$$\min_{X\in\mathbb{R}^{d_1\times d_2}} \|X\|_* \qquad \text{subject to} \qquad \mathscr{A}(Q_1^{-1} X Q_2^{-1}) = b. \tag{4.2}$$

Proof. Define the map $\mathscr{B}(Z) = \alpha_2^2 \mathscr{A}(Q_1^{-1} Z Q_2^{-1})$. Then Lemma 4.1 implies that $\mathscr{B}$ 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 [5, Theorem 2.1] in the second guarantees that $Q_1 M_\natural Q_2$ is the unique solution of (4.2), as claimed. □

4.2 *Applications: matrix and bilinear sensing*

Let us introduce the random ensembles of the two applications: matrix and bilinear sensing.

DEFINITION 4.2 (Matrix and bilinear sensing) We introduce the following definitions.

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

As mentioned earlier, in this paper we will be primarily interested in ℓ_2/ℓ_2 and ℓ_1/ℓ_2 RIPs. In particular, the two random measurement models satisfy these two properties. The following two lemmas are from [8, Theorem 2.3], [47, Theorem 4.2] and [5, Theorem 2.2].

LEMMA 4.3 (ℓ_2/ℓ_2 RIP in matrix sensing) Let $\mathscr{A}$ 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 $m \geq cr(d_1 + d_2)$, with probability at least $1 - \exp(-Cm)$, the map $\mathscr{A}$ satisfies ℓ_2/ℓ_2 RIP with parameters $(2r, 1 - \delta, 1 + \delta)$.

LEMMA 4.4 (ℓ_1/ℓ_2 RIP in bilinear sensing) Let $\mathscr{A}$ be a Gaussian bilinear ensemble. For any positive integer $k \geq 2$, there exist constants $c, C > 0$ depending only on k and numerical constants $\delta_1, \delta_2 > 0$ and such that in the regime $m \geq cr(d_1 + d_2)$, with probability at least $1 - \exp(-Cm)$, the measurement map $\mathscr{A}$ satisfies ℓ_1/ℓ_2 RIP with parameters (kr, δ_1, δ_2).

In order to apply Lemma 4.2 and Theorem 3.2, it remains to estimate the condition number κ of the matrices D_1 and D_2 under RIP (or statistical assumptions). The following lemma shows that D_i are well conditioned under ℓ_2/ℓ_2 RIP.

LEMMA 4.5 (Conditioning of D_i under ℓ_2/ℓ_2 RIP) Suppose that the linear map $\mathscr{A}$ 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 (3.4)), we actually showed the expression:

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

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 $\mathscr{A}$, we deduce $\delta_1^2 \|v\|_2^2 \leq \left\| D_1 v \right\|_2^2 \leq \delta_2^2 \|v\|_2^2$. A similar argument shows that D_2 satisfies the analogous inequality with $v \in \mathbb{R}^{d_2}$. □

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 4.6 (Conditioning of D_i for bilinear sensing) Let $\mathscr{A}$ be a Gaussian bilinear ensemble. Then there exist constants $c_1, c_2, c_3, c_4 > 0$ 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:

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

Proof. First observe $A_i A_i^\top = \|b_i\|_2^2 a_i a_i^\top$ for each index i. Hoeffding inequality 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 see 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

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

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

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

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 $m \geq c_3 \frac{\max\{d_1, d_2\}}{\delta^2}$, 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. □

The following are the two main results of the section as applications of our Lemma 4.2 and Theorem 3.2.

THEOREM 4.3 (Exact recovery in matrix sensing) Suppose that $\mathscr{A}$ 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 $m \geq cr_\natural(d_1 + d_2)$, with probability at least $1 - \exp(-Cm)$, any flat solution (L_f, R_f) of (1.1) satisfies $L_f R_f^\top = M_\natural$ and is automatically nearly norm-minimal and nearly balanced:

$$\left\|L_f\right\|_{\mathrm{F}}^2 + \left\|R_f\right\|_{\mathrm{F}}^2 \leq \left(\frac{1+\delta}{1-\delta}\right)^2 \cdot \left(\min_{\mathscr{A}(LR^\top)=b} \|L\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2\right)$$

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

Proof. Lemma 4.3 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 $m \geq c_1 r(d_1 + d_2)$, with probability at least $1 - \exp(-C_1 m)$, the map $\mathscr{A}$ satisfies ℓ_2/ℓ_2 RIP with parameters $(r, 1-\delta, 1+\delta)$. In this event, Lemma 4.5 ensures that the condition number κ of D_1 and D_2 is bounded by $\frac{1+\delta}{1-\delta}$. Set $r = 5r_\natural$ and choose any $\delta \leq 0.1$ satisfying $(1+\delta)\left(\frac{1+\delta}{1-\delta}\right)^2 \leq \frac{11}{10}$. An application of Lemma 4.2 and Theorem 3.2 completes the proof. □

THEOREM 4.4 (Exact recovery in bilinear sensing) Suppose that $\mathscr{A}$ 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 $m \geq cr_\natural(d_1 + d_2)$ and $\log(m) \leq c_4 d_{\min}$, with probability at least $1 - c_3 \exp(-Cd_{\min})$ any flat solution (L_f, R_f) of (1.1) satisfies $L_f R_f^\top = M_\natural$ and is automatically nearly norm-minimal and nearly balanced:

$$\left\|L_f\right\|_{\mathrm{F}}^2 + \left\|R_f\right\|_{\mathrm{F}}^2 \leq \left(\frac{1+\delta}{1-\delta}\right)^2 \cdot \left(\min_{\mathscr{A}(LR^\top)=b} \|L\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2\right)$$

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

Proof. For any integer $k \in \mathbb{N}$, Lemma 4.4 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 $\mathscr{A}$ satisfies ℓ_1/ℓ_2 RIP with parameters $(lr_\natural, \delta_1, \delta_2)$. Lemma 4.6 in turn ensures there exist constants $c_5, c_6, c_7, c_8 > 0$ such that for any $\delta \in (0, 1)$ as long 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:

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

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 4.2, 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 4.2 and Theorem 3.2 completes the proof. □

Appendix A generalizes the material in this section to the noisy observation setting, wherein $b = \mathscr{A}(M_\natural) + e$ 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 (1.1) where the measurement map $\mathscr{A}$ 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 $\mathscr{A}$: $\mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ is then defined by the relation

$$[\mathscr{A}(Z)]_{ij} = Z_{ij}\xi_{ij} \qquad \text{for any } (i,j) \in [d_1] \times [d_2]. \tag{5.1}$$

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

$$\left\|U_\natural\right\|_{2,\infty} \le \sqrt{\frac{\mu r_\natural}{d_1}}, \quad \text{and} \quad \left\|V_\natural\right\|_{2,\infty} \le \sqrt{\frac{\mu r_\natural}{d_2}}. \tag{5.2}$$

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 analysing flat minima of the matrix completion problem because the linear map $\mathscr{A}(D_1^{-1} \cdot D_2^{-1})$ does not satisfy RIP type conditions. More precisely, even though D_1 and D_2 are well-conditioned diagonal matrices as shown in Lemma 5.1, the linear map $\mathscr{A}$ in matrix completion does not satisfy RIP condition in general by considering matrices with only one non-zero entry.

Instead, we will settle for a weak recovery result.

THEOREM 5.1 (Recovery error of flat solution) Suppose that $\mathscr{A}$ is generated from the matrix completion problem. 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

$$p \ge C \max\left\{\frac{1}{\gamma}\sqrt{\frac{r_\natural \log(d_{\max})}{d_{\min}}}, \frac{\mu r_\natural \log(\mu r_\natural)\log(d_{\max})}{d_{\min}}\right\}, \tag{5.3}$$

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

$$\left\|L_f R_f^\top - M_\natural\right\|_* \le \gamma \left\|M_\natural\right\|_*. \tag{5.4}$$

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. Note this lemmas implies that flat solutions are near norm minimal and balanced due to Theorem 3.2.

LEMMA 5.1 (Condition number) For any $\delta \in (0,1)$ 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

$$\sqrt{\frac{p}{d_1 d_2}(1-\delta)} I_{d_i} \preceq D_i \preceq \sqrt{\frac{p}{d_1 d_2}(1+\delta)} I_{d_i} \qquad \text{holds for } i = 1, 2. \tag{5.5}$$

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

$$D_1^2 = \frac{1}{d_1 d_2^2} \sum_{i \in [d_1], j \in [d_2]}^{m} \xi_{ij} e_i e_i^\top = \frac{1}{d_1 d_2^2} \operatorname{diag}\left(\sum_{j=1}^{d_2} \xi_{1j}, \ldots, \sum_{j=1}^{d_2} \xi_{d_1 j} \right). \tag{5.6}$$

Hoeffding inequality implies for each index $i \in [d_2]$ the estimate

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

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 $d_1 \exp\left(-2 p d_2 \delta^2\right) \leq \exp\left(-2 p d_2 \delta^2 + \log(d_1)\right) \leq \exp\left(\frac{-p d_2 \delta^2}{4}\right) \leq d_2^{-c}$. Using the same argument for D_2 and taking a union bound completes the proof. □

Next, we will show that flat solutions are almost optimal for the standard convex relaxation of the matrix completion problem, $\min_{\mathscr{A}(X) = \mathscr{A}(M_\natural)} \|X\|_*$.

LEMMA 5.2 (Flat minima and nuclear norm minimization) Suppose that $M_\natural$ solves the nuclear-norm minimization problem $\min_{\mathscr{A}(X) = \mathscr{A}(M_\natural)} \|X\|_*$ 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 (1.1) satisfies:

$$\left\| L_f R_f^\top \right\|_* - \left\| M_\natural \right\|_* \leq (\kappa^2 - 1) \left\| M_\natural \right\|_*. \tag{5.7}$$

Proof. We successively estimate

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

where the first and last inequalities follow from (1.10) and the second inequality follows from Theorem 3.2. We therefore deduce $\left\| L_f R_f^\top \right\|_* - \left\| M_\natural \right\|_* \leq (\kappa^2 - 1) \left\| M_\natural \right\|_*$, as claimed. □

It remains to translate the suboptimality gap (5.7) into an estimate on $\left\|L_f R_f^\top - M_\natural\right\|_*$. This is the content of the following lemma, whose proof appears in Appendix B.2.

LEMMA 5.3 (Sharp growth in matrix completion) Suppose that the linear map $\mathscr{A}$ 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 matrix X with $\mathscr{A}(X) = \mathscr{A}(M_\natural)$ satisfies:

$$\left\|X - M_\natural\right\|_* \leq 8\left(1 + \sqrt{\tfrac{6r_\natural}{p}}\right)\left(\|X\|_* - \left\|M_\natural\right\|_*\right). \tag{5.8}$$

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

Proof of Theorem 5.1. Lemma 5.3 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

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

holds for all X satisfying $\mathscr{A}(X) = \mathscr{A}(M_\natural)$. In this event, Lemma 5.2 ensures that the matrix $X := L_f R_f^\top$ satisfies

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

where κ is an upper bound on the condition numbers of D_1 and D_2. Lemma 5.1 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

$$\left\|L_f R_f^\top - M_\natural\right\|_* \leq C\delta\sqrt{\frac{r_\natural}{p}}\left\|M_\natural\right\|_*, \tag{5.9}$$

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

$$p \geq c_1 \max\{\mu r_\natural \log(\mu r_\natural), \delta^{-2}\} \cdot \frac{\log(d_{\max})}{d_{\min}}, \tag{5.10}$$

with probability at least $1 - c_2 d_{\min}^{-5}$, any flat solution (L_f, R_f) satisfies (5.9). 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 (5.10) as long as (5.3) holds. With this choice of δ, the estimate (5.9) becomes $\left\|L_f R_f^\top - M_\natural\right\|_* \leq \gamma\left\|M_\natural\right\|_*$, as claimed. □

6. Robust PCA

In this section, we focus on the problem of PCA with outliers, also known as 'robust PCA', following the approach in [7,11]. Though this problem is not of the form (1.1), 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 non-zero 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

$$\left\|U_\natural\right\|_{2,\infty} \leq \sqrt{\frac{\mu r_\natural}{d_1}}, \qquad \left\|V_\natural\right\|_{2,\infty} \leq \sqrt{\frac{\mu r_\natural}{d_2}} \qquad \text{and} \qquad \left\|U_\natural V_\natural^\top\right\|_\infty \leq \sqrt{\frac{\mu r_\natural}{d_1 d_2}}, \tag{6.1}$$

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

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

$$\min_{L,R} \min_{S \in \Omega} \left\|LR^\top + S - Y\right\|_{\mathrm{F}}^2, \tag{6.2}$$

where we define the set $\Omega := \{S \mid \|S\|_{1,1} \leq \left\|S_\natural\right\|_{1,1}\}$, 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 overparameterized setting $k \geq r_\natural$. Note that the optimal value of (6.2) is clearly zero.

Observe that we may express the problem (6.2) more compactly as

$$\min_{L,R} f(L,R) := \mathrm{dist}_\Omega^2(Y - LR^\top), \tag{6.3}$$

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

$$f_{\tilde{L},\tilde{R}}(L,R) := \left\|Y - LR^\top - P_\Omega(Y - \tilde{L}\tilde{R}^\top)\right\|_{\mathrm{F}}^2, \tag{6.4}$$

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 (6.3) to be *flat* if it solves the problem:

$$\min_{(L,R)\in\mathscr{M}} \operatorname{str}(D^2 f_{L,R}(L,R)). \tag{6.5}$$

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 (6.3) satisfies $L_f R_f^\top = M_\natural$. Moreover, any flat solution (L_f, R_f) is norm-minimal and balanced.

Proof. Let $(L_0, R_0) \in \mathscr{M}$ be a solution of (6.3). Since we have $f(L_0, R_0) = 0$, the equality $f_{L_0,R_0}(L_0, R_0) = 0$ holds. In particular, we may write $f_{L_0,R_0}(L,R) = \left\| LR^\top - W_\natural \right\|_{\mathrm{F}}^2$, where we define $W_\natural := Y - P_\Omega(Y - L_0 R_0^\top)$. Therefore appealing to Lemma 2.1, we may write

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

Thus any flat solution (L_f, R_f) of (6.3) solves the problem:

$$\min_{L,R} \quad \|L\|_{\mathrm{F}}^2 + \|R\|_{\mathrm{F}}^2 \qquad \text{subject to} \qquad Y - LR^\top \in \Omega. \tag{6.6}$$

Equivalently, the characterization (1.10) implies that the matrix $X_f = L_f R_f^\top$ solves the problem

$$\min \quad \|X\|_* \qquad \text{subject to} \quad Y - X \in \Omega, \operatorname{rank}(X) \leq k. \tag{6.7}$$

On the other hand, [15, Theorem 3][5] shows that $M_\natural$ is the unique minimizer of the convex relaxation

$$\min \quad \|X\|_* \qquad \text{subject to} \quad Y - X \in \Omega. \tag{6.9}$$

Hence, we know $M_\natural$ also uniquely solves (6.7) and we conclude $M_\natural = X_f = L_f R_f^\top$, as claimed. That (L_f, R_f) is norm minimal and balanced follow from Lemma 3.2 by setting $Q_i = I$ for $i = 1, 2$ and $X = M_\natural$. □

[5] The result [15, Theorem 3] actually shows that $(M_\natural, S_\natural)$ uniquely solves

$$\begin{aligned} &\text{minimize} \quad \|X\|_* + \lambda \|S\|_{1,1} \\ &\text{subject to} \quad Y = X + S, \end{aligned} \tag{6.8}$$

for some $\lambda > 0$. Now for any solution X_1 to (6.9), the pair $(X_1, Y - X_1)$ is feasible for (6.8) 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 (6.8), we know $X_1 = M_\natural$.

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 analysing the energy landscape around saddle points. Though this problem is not in the form (1.1), 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 set-up, 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

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

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

$$\hat{y}(U, x) = u^\top q(U^\top x). \tag{7.1}$$

Thus the overparameterized problem we aim to solve is

$$\min_{U\in\mathbb{R}^{d\times k}} f(U) := \frac{1}{m}\sum_{i=1}^{m}(\hat{y}(U, x_i) - y_i)^2. \tag{7.2}$$

As usual, we define the solution set $\mathscr{M} = \{U \in \mathbb{R}^{d\times k} : f(U) = 0\}$. We will see shortly that $\mathscr{M}$ is non-empty and therefore coincides with the set of minimizers of f. Naturally, we declare a matrix $U_f \in \mathscr{M}$ to be *flat* if it solves the problem $\min_{U\in\mathscr{M}} tr(D^2 f(U))$. In this section, we aim to show

with high probability over the training set $\{(x_i, y_i)\}_{i=1,\dots,n}$ 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 (7.2) to low-rank matrix factorization. To see this, we can write $\hat{y}(U, x) - y(U_\natural, x)$ as

$$\begin{aligned}
\hat{y}(U, x) - y(U_\natural, x) &= u^\top q(U^\top x) - v^\top q(U_\natural^\top x) \\
&= \Big\langle \underbrace{U\mathrm{diag}(u)U^\top}_{U_1U_1^\top - U_2U_2^\top}, xx^\top \Big\rangle - \Big\langle \underbrace{U_\natural \mathrm{diag}(v) U_\natural^\top}_{=:M_\natural}, xx^\top \Big\rangle \\
&= \left\langle U_1U_1^\top - U_2U_2^\top - M_\natural, xx^\top \right\rangle.
\end{aligned} \tag{7.3}$$

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 (7.3), we may rewrite the objective of (7.2) as

$$f(U_1, U_2) = \frac{1}{m}\left\|\mathscr{A}(U_1U_1^\top - U_2U_2^\top - M_\natural)\right\|_2^2, \tag{7.4}$$

where the linear map $\mathscr{A}$ is defined as $\mathscr{A}\colon \mathbb{R}^{d\times d} \to \mathbb{R}^m$ with

$$[\mathscr{A}(Z)]_i = \left\langle Z, x_i x_i^\top \right\rangle \text{ for any } Z \in \mathbb{R}^{d\times d}. \tag{7.5}$$

In particular, from the second equation in (7.3) and our assumption on v, there always exists a matrix $U = [U_1, U_2]$ satisfying $U_1U_1^\top - U_2U_2^\top = M_\natural$. Therefore, the set of minimizers of f is non-empty and it coincides with $\mathscr{M}$. Note that in the special case $r_2 = k_2 = 0$, the problem (7.4) becomes covariance matrix estimation [14] and further reduces to phase retrieval when $k_1 = r_1 = 1$ [10].

Summarizing, 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 $U_1U_1^\top - U_2U_2^\top$ 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 $m \geq Cr_\natural d$ 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 (7.2) satisfies

$$U_{f,1}U_{f,1}^\top - U_{f,2}U_{f,2}^\top = M_\natural, \tag{7.6}$$

and achieves zero generalization error, $\mathbb{E}_{x\sim N(0,I)}(\hat{y}(U_f, x) - y(U_\natural, x)) = 0$.

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 (whose proof can be found in Section B.3) that expresses the trace of the Hessian in the same spirit as Lemma 3.1. With this in mind, we define the matrix

$$D_0 := \left(\frac{1}{md}\sum_{i=1}^m A_iA_i^\top\right)^{\frac{1}{2}}. \tag{7.7}$$

LEMMA 7.1 (Trace) For any matrix $[U_1, U_2] \in \mathscr{M}$, the trace of $D^2f(U_1, U_2)$ can be written as

$$\mathrm{tr}(D^2f(U_1, U_2)) = 4d\left\|D_0U_1\right\|_{\mathrm{F}}^2 + 4d\left\|D_0U_2\right\|_{\mathrm{F}}^2. \tag{7.8}$$

In particular, Lemma 7.1 implies that flat solutions are exactly the minimizers of the problem

$$\min_{U_1,U_2} \frac{1}{2}\left(\left\|D_0U_1\right\|_{\mathrm{F}}^2 + \left\|D_0U_2\right\|_{\mathrm{F}}^2\right) \qquad \text{such that} \qquad \mathscr{A}(U_1U_1^\top - U_2U_2^\top) = \mathscr{A}(M_\natural). \tag{7.9}$$

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 that are in the spirit of (1.10). Their proofs can be found in Section B.3.

LEMMA 7.2 (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 $X = U_1U_1^\top - U_2U_2^\top$ 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.

LEMMA 7.3 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:

$$\|X\|_* = \min_{X=U_1U_1^\top - U_2U_2^\top, U_i\in\mathbb{R}^{d\times k_i}} \|U_1\|_{\mathrm{F}}^2 + \|U_2\|_{\mathrm{F}}^2 .$$

Lemmas 7.2-7.3 imply that the problem (7.9), characterizing flat solutions, is equivalent to:

$$\begin{aligned} &\text{minimize} && \|X\|_* \\ &\text{subject to} && \mathscr{A}(D^{-1}XD^{-1}) = \mathscr{A}(M_\natural), X \text{ is symmetric,} \\ & && X \text{ has at most } k_1 \text{ positive eigenvalues and } k_2 \text{ negative eigenvalues.} \end{aligned} \tag{7.10}$$

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

$$\text{minimize} \quad \|X\|_* \qquad \text{subject to} \quad \mathscr{A}(D^{-1}XD^{-1}) = \mathscr{A}(M_\natural), X \in \mathbb{R}^{d\times d} \text{ is symmetric.} \tag{7.11}$$

The following theorem summarizes these observations.

THEOREM 7.2 (Convex relaxation) Suppose that the matrix D is invertible. Then the problems (7.9) and (7.10) are equivalent in the following sense.

1. The optimal values of (7.9) and (7.10) are equal.
2. If $[U_1, U_2]$ solves (7.9), then $X = D(U_1U_1^\top - U_2U_2^\top)D$ is optimal for (7.10).
3. If a minimizer X of (7.10) has an eigenvalue decomposition $X = P_1\Lambda_1P_1^\top - P_2\Lambda_2P_2^\top$ for some diagonal matrices $\Lambda_i \in \mathbb{R}^{r_i\times r_i}$ with positive entries, then the matrices $U_i = [D^{-1}P_i\sqrt{\Sigma_i}, 0_{d,(k_i-r_i)}]$, $i = 1, 2$ are minimizers of (7.9).

Moreover, if $X = DM_\natural D$ is the unique minimizer of the problem (7.11), then any flat solution $[U_1, U_2]$ satisfies $U_1U_1^\top - U_2U_2^\top = M_\natural$.

Next, we aim to understand when the problem (7.11) exactly recovers $D^{-1}M_\natural D$. The difficulty is that even in the case $k_2 = 0$, the linear map $\mathscr{A}$ satisfies ℓ_1/ℓ_2 RIP only if $m \gtrsim r_\natural^2 d$ [14], which is suboptimal by a factor of $r_\natural$. We will sidestep this issue by relating the program (7.10) 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 [6, Equation (0.36)].

LEMMA 7.4 Define the linear map $\mathscr{A}_1 : \mathbb{R}^{d\times d} \to \mathbb{R}^{\lfloor m\rfloor/2}$ by

$$[\mathscr{A}_1(Z)]_i = \left(\frac{x_{2i-1}+x_{2i}}{\sqrt{2}}\right)^\top Z \left(\frac{x_{2i-1}-x_{2i}}{\sqrt{2}}\right), \qquad \text{for } i = 1,\dots,\lfloor m\rfloor/2, \tag{7.12}$$

and consider the convex optimization program

$$\min_{\mathscr{A}_1(D^{-1}XD^{-1})=\mathscr{A}_1(M_\natural)} \|X\|_* . \tag{7.13}$$

If $DM_\natural D$ is the unique solution of (7.13), then it is also the unique solution of (7.11).

Proof. We first note the equalities

$$[\mathscr{A}_1(Z)]_i = \frac{1}{2}\left\langle x_{2i-1}x_{2i-1}^\top, Z\right\rangle - \frac{1}{2}\left\langle x_{2i}x_{2i}^\top, Z\right\rangle = \frac{1}{2}\left([\mathscr{A}_{2i-1}(Z)] - [\mathscr{A}_{2i}(Z)]\right). \tag{7.14}$$

Consequently, any X that is feasible for (7.11) is feasible for (7.13). Thus, if $DM_\natural D$ is a unique minimizer of (7.13), then it must also be a unique minimizer of (7.11). This completes the proof. □

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, the map $\mathscr{A}_1$ defined in (7.12) follows the bilinear sensing model (Definition 4.2). Therefore Lemma 4.4 implies that for any positive integer $k \geq 2$, there exist constants $c, C > 0$ depending only on k and numerical constants $\delta_1, \delta_2 > 0$ and such that in the regime $m \geq cr_\natural d$, with probability at least $1 - \exp(-Cm)$, the measurement map $\mathscr{A}_1$ satisfies ℓ_1/ℓ_2 RIP with parameters $(kr_\natural, \delta_1, \delta_2)$. In this event, Lemma 4.1 with $Q_1 = Q_2 = D^{-1}$ implies that $\mathscr{A}_1(D^{-1}\cdot D^{-1})$ 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 $D^2 = \frac{1}{md}\sum_{i=1}^m A_i A_i^\top = \frac{1}{md}\sum_{i=1}^m \|x_i\|^2 x_i x_i^\top$. The same proof as that of Lemma 4.6 ensures that there exist constants $c_1, c_2, c_3, c_4 > 0$ such in the regime, $m \geq c_3 d$ and $\log(m) \leq c_4 d$, the estimate $\mathbb{P}\left\{\frac{1}{2}I_d \preceq D \preceq \frac{3}{2}I_d\right\} \geq 1 - c_2 e^{-c_1 d}$ holds. Consequently, in this regime we may upper bound the condition number κ of D by 3. In light of Lemma 4.2, in order to ensure that $DM_\natural D$ is the unique minimizer of (7.13), it remains to simply choose k such that the inequality $\frac{9\delta_2}{\delta_1} \leq \sqrt{k}$ holds. Using Lemmas 7.2-7.4 completes the proof. □

8. Numerical experiments

Recall that we have proved that for a variety of overparameterized problems, (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 set-up. We consider three problems described earlier in the paper: (a) matrix sensing, (b) matrix completion and (c) neural networks with quadratic activation. For each setting, we consider different combinations of the dimension $d = d_1 = d_2$ and the number of measurements m (p for matrix

completion). For each combination, 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 we take 10 samples of $\mathscr{A}$ from the corresponding random ensemble, and for each $\mathscr{A}$ solve two problems: the convex relaxation associated with being a flat solution, and the nuclear norm minimization problem. More precisely, the convex relaxation associated with being a flat solution for (a) matrix sensing, (b) matrix completion and (c) neural networks with quadratic activation is of the form

$$\text{minimize}_{X\in\mathscr{E}}\|X\|_* \qquad \text{subject to} \qquad \mathscr{A}(D_1^{-1}XD_2^{-1}) = b. \tag{8.1}$$

For (a) matrix sensing, (b) matrix completion and (c) neural networks, the nuclear norm minimization is of the form

$$\text{minimize}_{X\in\mathscr{E}} \quad \|X\|_* \qquad \text{subject to} \quad \mathscr{A}(X) = \mathscr{A}(M_\natural). \tag{8.2}$$

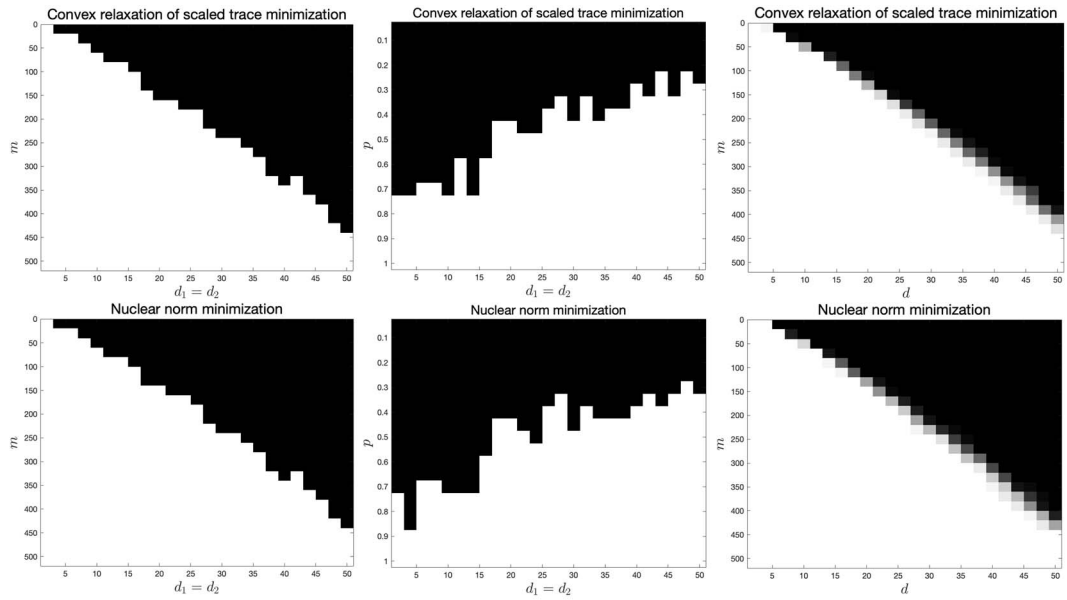
Here $\mathscr{A}$ is defined in Definition 4.2, Equation (5.1) and Equation (7.5), respectively. The Euclidean space $\mathscr{E}$ is $\mathbb{R}^{d_1\times d_2}$ for (a) matrix sensing and (b) matrix completion, and is the set of symmetric matrices of size $d\times d$ for (c) neural networks with quadratic activation. Finally we check their solutions against the ground truth.

Exact recovery. To measure the success rate for 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 $\left\|D_1^{-1}\hat{X}D_2^{-1} - M_\natural\right\|_{\text{F}}$ (or $\left\|\hat{X} - M_\natural\right\|_{\text{F}}$ 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 successful recovery (averaging over ten times) for each combination of dimension and number of measurements. The figure is in grey scale and the whiter colour 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.

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 non-singular. Let $\hat{X}$ be the solution of the convex relaxation for being a flat solution and let $\hat{X}_{\text{nuc}}$ 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 $\left(\|L_f\|_{\text{F}}^2 + \|R_f\|_{\text{F}}^2\right)/2\|\hat{X}_{\text{nuc}}\|_*$ to measure how norm-minimality of the flat solutions, and the quantity $\|L_f^\top L_f - R_f^\top R_f\|_*/\|M_\natural\|_*$ to measure balancedness. The result (in $\log 10$ scale) is shown in Fig. 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.

9. Conclusion and discussion on depth

In this paper, we analysed 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.



(a) Matrix Sensing (b) Matrix Completion (c) NNQA

FIG. 3. The empirical probability of successful recovery of $M_\natural$ for different combinations of dimension d and number of measurements m. The whiter colour indicates higher probability of success. NNQA stands for neural network with quadratic activation.

Matrix factorization problems are suggestive of the behaviour 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 $r_\natural$ non-zero 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 RIP: there exist (δ_1, δ_2) such that

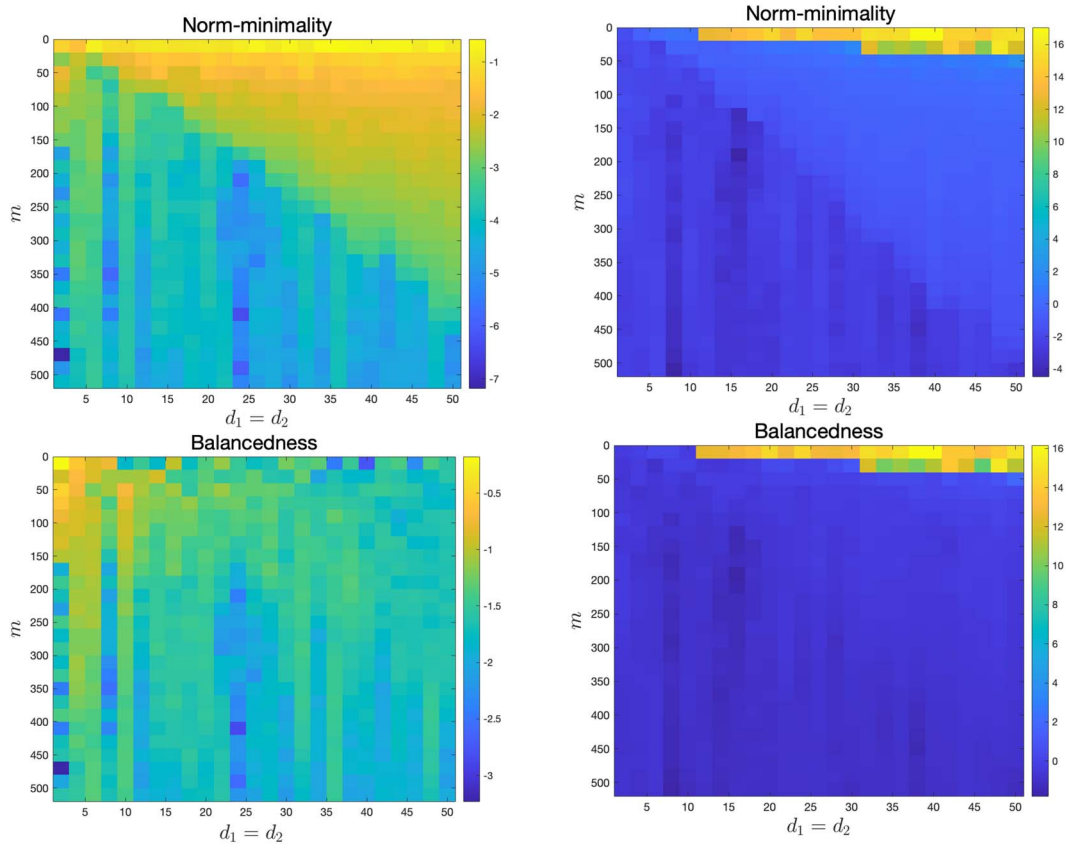
$$\delta_1 \|x\|_2 \le \tfrac{1}{\sqrt{m}} \|Ax\|_2 \le \delta_2 \|x\|_2, \tag{9.1}$$

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

$$\min_{x \in \mathbb{R}^d} \quad g(x) := \frac{1}{m} \|Ax - b\|_2^2 . \tag{9.2}$$

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 (9.2) becomes

$$\min_{v_1,\ldots,v_k} \quad f(v) := \frac{1}{m} \left\| A(v_1 \odot \cdots \odot v_k) - b \right\|_2^2 \qquad \text{with } v_i \in \mathbb{R}^d,\ i = 1, \ldots, k. \tag{9.3}$$



(a) Matrix Sensing (b) Bilinear Sensing

FIG. 4. The empirical average (over 10 trials) of the regularity measure: (1) minimal norm, $\frac{\|L_f\|_{\mathrm{F}}^2+\|R_f\|_{\mathrm{F}}^2}{2\|\hat{X}_{\mathrm{CVX}}\|_*} - 1$, and (2) balance, $\frac{\|L_f^\top L_f - R_f^\top R_f\|_*}{\|M_\natural\|_*}$. Both measures are presented in $\log 10$ scales. Whenever one of the D_i is singular, we set both measures to be 10^{20}.

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

$$\min_{v_i \in \mathbb{R}^d, i=1,\dots,k} \mathrm{tr}(D^2 f(v_1,\dots,v_k)) \quad \text{subject to} \quad A(v_1 \odot \cdots \odot v_k) = b. \tag{9.4}$$

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

$$\mathrm{tr}\left(D^2 f\left([v_i]_{i=1}^k\right)\right) = \sum_{i=1}^k \left\| \sqrt{D}\left(v_1 \odot \dots v_{i-1} \odot v_{i+1} \cdots \odot v_k\right) \right\|_2^2, \tag{9.5}$$

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

$$D := \frac{1}{m}\mathrm{diag}\big(a_1^\top a_1, \ldots, a_d^\top a_d\big). \tag{9.6}$$

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 $\frac{1}{m}\left\|Ae_i\right\|_2^2 = \frac{1}{m}a_i^\top a_i$. Since e_i is a sparse vector with only one non-zero, using the $(1-\delta, 1+\delta)$ RIP, we have $\frac{1}{m}a_i^\top a_i = \frac{1}{m}\left\|Ae_i\right\|_2^2 \in [(1-\delta)^2\left\|e_i\right\|_2, (1+\delta)^2\left\|e_i\right\|_2^2] = [(1-\delta)^2, (1+\delta)^2]$ and our proof is complete. □

The next lemma shows that the following optimization problem is equivalent to the optimization problem defining flat solutions (9.4).

$$\min_{x\in\mathbb{R}^d} \sum_{i=1}^d |D_{ii}||x_i|^{2-\frac{2}{k}} \quad \text{s.t.} \quad Ax = b. \tag{9.7}$$

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 (9.4) is equivalent to Problem (9.7) in the following sense:

- If x solves (9.7), 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 (9.4).
- If $v_1, \ldots, v_k$ solves (9.4), then $x = v_1 \odot \cdots \odot v_k$ solves (9.7).

Proof. According to (9.5), the trace of the Hessian is

$$\begin{aligned}
\mathrm{tr}(D^2 f([v_i]_{i=1}^k)) &= \sum_{i=1}^k \left\|\sqrt{D}(v_1 \odot \ldots v_{i-1} \odot v_{i+1} \cdots \odot v_k)\right\|_2^2 \\
&= \sum_{i=1}^k \sum_{j=1}^d D_{jj} \prod_{h\neq i} v_{h,j}^2 \\
&= \sum_{j=1}^d D_{jj} \sum_{i=1}^k \prod_{h\neq i} v_{h,j}^2 \\
&\overset{(a)}{\ge} \sum_{j=1}^d D_{jj} k \left(\prod_{i=1}^k v_{i,j}^{2(k-1)}\right)^{\frac{1}{k}} \\
&= \sum_{j=1}^d D_{jj} \left(\prod_{i=1}^k |v_{i,j}|\right)^{2-\frac{2}{k}}.
\end{aligned} \tag{9.8}$$

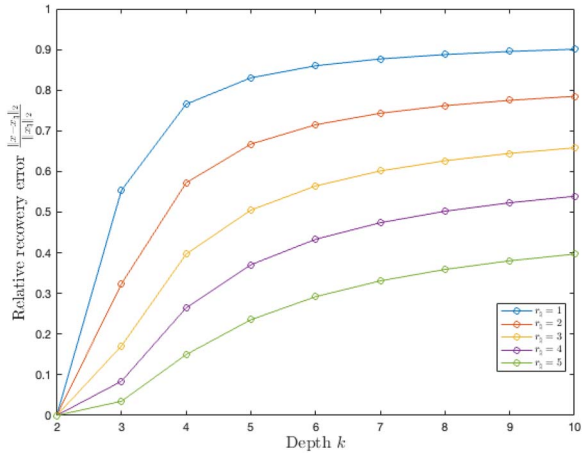


FIG. 5. The effect of depth for different choice of sparsity $r_\natural$.

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 $x = v_1 \odot \cdots \odot v_k$. □

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 satisfies $(1-\delta, 1+\delta)$ RIP with $0 < \delta < c$, then for $k = 2$, any solution (v_1, v_2) to (9.4) 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 (9.7) is close to $\|x\|_2^2$, which should give a dense solution in general. Indeed, in Fig. 5, we plot the solution performance of (9.7) 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.

Data Availability Statement

The data underlying this article are available in the article and in its online supplementary material.

REFERENCES

1. AHMED, A., RECHT, B. & ROMBERG, J. (2013) Blind deconvolution using convex programming. *IEEE Trans. Inf. Theory*, **60**(3), 1711–1732.

[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 (9.7) for each A. The performance metric $\frac{\|x-x_\natural\|_2}{\|x_\natural\|_2}$ is averaged over these 25 trials.

2. Bartlett, P. L. (1998) The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Trans. Inf. Theory*, **44**(2), 525–536.
3. Bartlett, P. L. & Mendelson, S. (2002) Rademacher and gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, **3**(Nov), 463–482.
4. Belkin, M., Hsu, D., Ma, S. & Mandal, S. (2019) Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proc. Natl. Acad. Sci.*, **116**(32), 15849–15854.
5. Cai, T. T. & Zhang, A. (2015a) ROP: matrix recovery via rank-one projections. *Ann. Stat.*, **43**(1), 102–138.
6. Cai, T. T. & Zhang, A. (2015b) Supplement to ROP: matrix recovery via rank-one projections. *Ann. Stat.*, **43**, 1–25.
7. Candès, E. J., Li, X., Ma, Y. & Wright, J. (2011) Robust principal component analysis?. *J. ACM*, **58**(3), 1–37.
8. Candes, E. J. & Plan, Y. (2011) Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theory*, **57**(4), 2342–2359.
9. Candès, E. J. & Recht, B. (2009) Exact matrix completion via convex optimization. *Found. Comput. Math.*, **9**(6), 717–772.
10. Candes, E. J., Strohmer, T. & Voroninski, V. (2013) Phaselift: exact and stable signal recovery from magnitude measurements via convex programming. *Comm. Pure Appl. Math.*, **66**(8), 1241–1274.
11. Chandrasekaran, V., Sanghavi, S., Parrilo, P. A. & Willsky, A. S. (2011) Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.*, **21**(2), 572–596.
12. Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L. & Zecchina, R. (2019) Entropy-sgd: biasing gradient descent into wide valleys. *J. Stat. Mech: Theory Exp.*, **2019**(12), 124018.
13. Chen, Y. (2015) Incoherence-optimal matrix completion. *IEEE Trans. Inf. Theory*, **61**(5), 2909–2923.
14. Chen, Y., Chi, Y. & Goldsmith, A. J. (2015) Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Trans. Inf. Theory*, **61**(7), 4034–4059.
15. Chen, Y., Jalali, A., Sanghavi, S. & Caramanis, C. (2013) Low-rank matrix recovery from errors and erasures. *IEEE Trans. Inf. Theory*, **59**(7), 4324–4337.
16. Ding, L. & Chen, Y. (2020) Leave-one-out approach for matrix completion: primal and dual analysis. *IEEE Trans. Inf. Theory*, **66**(11), 7274–7301.
17. Dinh, L., Pascanu, R., Bengio, S. & Bengio, Y. (2017) Sharp minima can generalize for deep nets. *International Conference on Machine Learning*. PMLR, pp. 1019–1028.
18. Du, S. S., Hu, W. & Lee, J. D. (2018) Algorithmic regularization in learning deep homogeneous models: layers are automatically balanced. *Advances in Neural Information Processing Systems* (Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. & Garnett, R. eds.), **31**. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/fe131d7f5a6b38b23cc967316c13dae2-Paper.pdf.
19. Dziugaite, G. K. & Roy, D. M. (2017) Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. arXiv preprint arXiv:1703.11008.
20. Foret, P., Kleiner, A., Mobahi, H. & Neyshabur, B. (2020) Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412.
21. Ge, R., Jin, C. & Zheng, Y. (2017) No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *Proceedings of the 34th International Conference on Machine Learning* (Precup, D., & Teh, Y. W. eds.) PMLR, pp. 1233–1242.
22. Gunasekar, S., Lee, J. D., Soudry, D. & Srebro, N. (2018) Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems* (Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. & Garnett, R. eds.), **31**. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/0e98aeeb54acf612b9eb4e48a269814c-Paper.pdf.
23. Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B. & Srebro, N. (2017) Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems* (Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. & Garnett, R. eds.) Curran Associates, Inc. **30**. https://proceedings.neurips.cc/paper_files/paper/2017/file/58191d2a914c6dae66371c9dcdc91b41-Paper.pdf.

24. Ha, W., Liu, H. & Barber, R. F. (2020) An equivalence between critical points for rank constraints versus low-rank factorizations. *SIAM J. Optim.*, **30**(4), 2927–2955.
25. Heckel, R. & Soltanolkotabi, M. (2020) Compressive sensing with un-trained neural networks: Gradient descent finds a smooth approximation. *Proceedings of the 37th International Conference on Machine Learning* (Hal, D. III, & Singh, A. eds.), **119**, pp. 4149–4158. PMLR.
26. Hochreiter, S. & Schmidhuber, J. (1997) Flat minima. *Neural Comput.*, **9**(1), 1–42.
27. Hoffer, E., Hubara, I. & Soudry, D. (2017) Train longer, generalize better: closing the generalization gap in large batch training of neural networks. *Advances in Neural Information Processing Systems* (Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. & Garnett, R. eds.) Curran Associates, Inc. **30**. https://proceedings.neurips.cc/paper_files/paper/2017/file/a5e0ff62be0b08456fc7f1e88812af3d-Paper.pdf.
28. Huang, Y., Cheng, Y., Bapna, A., Firat, O., Chen, D., Chen, M., Lee, H., Ngiam, J., Le, Q. V., Wu, Y., et al. (2019) Gpipe: efficient training of giant neural networks using pipeline parallelism. *Adv. Neural Inf. Process. Syst.*, **32**, 103–112.
29. Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D. & Wilson, A. G. (2018) Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407.
30. Jacot, A., Gabriel, F. & Hongler, C. (2018) Neural tangent kernel: convergence and generalization in neural networks. *Advances in Neural Information Processing Systems* (Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. & Garnett, R. eds.), **31**. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.
31. Jastrzebski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y. & Storkey, A. (2017) Three factors influencing minima in sgd. arXiv preprint arXiv:1711.04623.
32. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S. & Houlsby, N. (2020) Big transfer (bit): general visual representation learning. *Computer Vision–ECCV 2020* (Vedaldi, A., Bischof, H., Brox, T. & Frahm, J-M. eds.) pp. 491–507. Cham: Springer International Publishing.
33. Li, H., Xu, Z., Taylor, G., Studer, C. & Goldstein, T. (2018) Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems* (Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. & Garnett, R. eds.), **31**. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/a41b3bb3e6b050b6c9067c67f663b915-Paper.pdf.
34. Li, Y., Ma, T. & Zhang, H. (2018) Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *Proceedings of the 31st Conference On Learning Theory* (Bubeck, s., Perchet, V. & Rigollet, P. eds.), **75**, pp. 2–47. PMLR.
35. Ling, S. & Strohmer, T. (2015) Self-calibration and biconvex compressive sensing. *Inverse Probl.*, **31**(11), 115002.
36. Liu, T., Li, Y., Wei, S., Zhou, E. & Zhao, T. (2021) Noisy Gradient Descent Converges to Flat Minima for Nonconvex Matrix Factorization. *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics* (Banerjee, A, & Fukumizu, K. eds.), **130**, pp. 1891–1899. PMLR.
37. Ma, C., Li, Y. & Chi, Y. (2021) Beyond Procrustes: balancing-free gradient descent for asymmetric low-rank matrix sensing. *IEEE Trans. Signal Process.*, **69**, 867–877.
38. Masters, D. & Luschi, C. (2018) Revisiting small batch training for deep neural networks. arXiv preprint arXiv:1804.07612.
39. Mulayoff, R. & Michaeli, T. (2020) Unique Properties of Flat Minima in Deep Networks. *Proceedings of the 37th International Conference on Machine Learning* (Hal, D. III, & Singh, A. eds.), **119**, pp. 7108–7118. PMLR.
40. Neyshabur, B., Bhojanapalli, S., McAllester, D. & Srebro, N. (2017) Exploring generalization in deep learning. *Advances in Neural Information Processing Systems* (Guyon, I., Von Luxburg, U., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. & Garnett, R. eds.), **30**. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/10ce03a1ed01077e3e289f3e53c72813-Paper.pdf.
41. Neyshabur, B., Tomioka, R. & Srebro, N. (2014) In search of the real inductive bias: on the role of implicit regularization in deep learning. arXiv preprint arXiv:1412.6614.

42. NitishShirish, K., Dheevatsa, M., Jorge, N., Mikhail, S. & Ping Tak Peter, T. (2016) On large-batch training for deep learning: generalization gap and sharp minima. arXiv preprint arXiv:1609.04836.
43. Norton, M. D. & Royset, J. O. (2023) Diametrical risk minimization: theory and computations. *Mach. Learn.*, **112**, 2933–2951.
44. Ongie, G. & Willett, R. (2022) The role of linear layers in nonlinear interpolating networks. arXiv preprint arXiv:2202.00856.
45. Ongie, G., Willett, R., Soudry, D. & Srebro, N. (2019) A function space view of bounded norm infinite width ReLU nets: the multivariate case. arXiv preprint arXiv:1910.01635.
46. Recht, B. (2011) A simpler approach to matrix completion.. *J. Mach. Learn. Res.*, **12**, 3413–3430.
47. Recht, B., Fazel, M. & Parrilo, P. A. (2010) Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, **52**(3), 471–501.
48. Savarese, P., Evron, I., Soudry, D. & Srebro, N. (2019) How do infinite width bounded norm networks look in function space?. *Proceedings of Machine Learning Research*, **99**, 32nd, Annual Conference on Learning Theory. **99**, pp. 1–24.
49. Shamir, O. (2018) Are resnets provably better than linear predictors? *Advances in Neural Information Processing Systems*, (Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N. & Garnett, R. eds.), **31**. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2018/file/26e359e83860db1d11b6acca57d8ea88-Paper.pdf
50. Smith, S. L. & Le, Q. V. (2017) A bayesian perspective on generalization and stochastic gradient descent. arXiv preprint arXiv:1710.06451.
51. Soltanolkotabi, M., Javanmard, A. & Lee, J. D. (2018) Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Trans. Inf. Theory*, **65**(2), 742–769.
52. Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S. & Srebro, N. (2018) The implicit bias of gradient descent on separable data. *J. Mach. Learn. Res.*, **19**(1), 2822–2878.
53. Srebro, N. & Shraibman, A. (2005) Rank, trace-norm and max-norm. In: *Learning Theory. COLT 2005. Lecture Notes in Computer Science* (Auer, P., & Meir, R. eds.) vol. **3559**. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11503415_37.
54. Tan, M. & Le, Q. (2019) Efficientnet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning* (Chaudhuri, K., & Salakhutdinov, R. eds.), **97**, pp. 6105–6114. PMLR.
55. Vershynin, R. (2018) *High-dimensional probability: An introduction with applications in data science*, vol. **47**. Cambridge university press.
56. Wainwright, M. J. (2019) *High-dimensional statistics: A non-asymptotic viewpoint*, vol. **48**. Cambridge university press.
57. Ye, T. & Du, S. S. (2021) Global convergence of gradient descent for asymmetric low-rank matrix factorization. *Advances in Neural Information Processing Systems* (Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P. S. & Vaughan, W. J. eds.), **34**, 1429–1439. Curran Associates, Inc.
58. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. (2021) Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, **64**(3), 107–115.

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:

$$b = \mathscr{A}(M_\natural) + e, \quad e \sim N(0, \sigma^2 I_m), \tag{A.1}$$

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:

$$\left\{(L,R) \mid \left\| \mathscr{A}(LR^\top) - b \right\|_2 \le \|e\|_2 \right\}. \tag{A.2}$$

The reason for this choice is that in the noisy observation setting, the global solution of (1.1) (with $k = \min\{d_1, d_2\}$) has the potential of overfitting no matter what regularization has been enforced. Indeed, consider the simplest case $\mathscr{A} = \mathscr{I}$, i.e. the map $\mathscr{A}$ 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.

$$\min_{L \in \mathbb{R}^{d_1 \times k},\, R \in \mathbb{R}^{d_2 \times k}} \mathrm{str}(D^2 f(L,R)) \quad \text{subject to} \quad \left\| \mathscr{A}(LR^\top) - b \right\|_2 \le \|e\|_2 . \tag{A.3}$$

The goal of the section is to prove the following.

Theorem A.1 (Noisy matrix sensing) Suppose that $\mathscr{A}$ 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 $m \ge C r_\natural d_{\max}$, with probability at least $1 - C\exp(-c(d_1+d_2))$, any solution (L_f, R_f) of (A.3) satisfies

$$\left\| L_f R_f^\top - M_\natural \right\|_{\mathrm{F}} \le C\sigma \sqrt{\frac{r_\natural (d_1+d_2)}{m}}. \tag{A.4}$$

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 (3.5) and (3.6) in Section 3.1, we see that (A.3) is equivalent to (in the sense of Theorem 3.1) minimizing the nuclear norm over rank constrained matrices so long as D_1, D_2 matrices are invertible[7] :

$$\min_{X \in \mathbb{R}^{d_1 \times d_2} : \mathrm{rank}(X) \le k} \|X\|_* \qquad \text{subject to} \qquad \|\mathscr{A}(D_1^{-1} X D_2^{-1}) - b\| \le \|e\|. \tag{A.5}$$

Let $\hat{Y}$ be any minimizer of (A.5). Also denote the scaled linear map $\tilde{\mathscr{A}}(\cdot) = \mathscr{A}(D_1^{-1} \cdot D_2^{-1})$, 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 (A.4) 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 4.5.

[7] Note that the condition $\mathscr{A}(LR^\top) = b$ is not needed for (3.2) to hold, which is critical for the step (3.5) to hold in the noisy case.

A.1.1 *Bound* $\|\Delta\|_F$ Our proof is based on the argument in [47]. Starting with the feasibility of $\hat{Y}$, we have

$$\begin{aligned} &\frac{1}{2m}\left\{\tilde{\mathscr{A}}(\hat{Y})-b\right\|_2^2 \le \frac{1}{2m}\|e\|_2^2 \overset{(a)}{\iff} \frac{1}{2m}\left\|\tilde{\mathscr{A}}(\Delta)\right\|_2^2+\frac{1}{m}\left\langle\tilde{\mathscr{A}}(\Delta),e\right\rangle \le 0 \\ \implies &\frac{1}{2m}\left\|\tilde{\mathscr{A}}(\Delta)\right\|_2^2 \le -\frac{1}{m}\left\langle\Delta,\tilde{\mathscr{A}}^* e\right\rangle \overset{(b)}{\le} \left\|\tilde{\mathscr{A}}^* e\right\|_{\mathrm{op}}\|\Delta\|_* . \end{aligned} \tag{A.6}$$

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 $\frac{1}{2m}\left\|\tilde{\mathscr{A}}(\Delta)\right\|_2^2$ and upper bound $\left\|\tilde{\mathscr{A}}^* e\right\|_{\mathrm{op}}$ and $\|\Delta\|_*$.

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

LEMMA A.1 [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) $\mathrm{rank}(B_1) \le \mathrm{rank}(A)$, (3) $AB_2^\top = 0$ and $A^\top B_2 = 0$, (4) $\|A+B_2\|_* = \|A\|_* + \|B_2\|_*$ and (5) $\left\langle B_1, B_2\right\rangle = 0$.

Using Lemma A.1, we can decompose $\Delta = R_0 + R_c$ such that $Y_0 R_c^\top = 0, Y_0^\top R_c = 0$, $R_0 \le 2r_\natural$, $\left\langle R_0, R_c\right\rangle = 0$, and $\|Y_0 + R_c\|_* = \|Y_0\|_* + \|R_c\|_*$. Hence, we have

$$\left\|\hat{Y}\right\|_* = \|Y_0+\Delta\|_* \ge \|Y_0+R_c\|_* - \|R_0\|_* = \|Y_0\|_* + \|R_c\|_* - \|R_0\|_* . \tag{A.7}$$

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

$$\|R_c\|_* \le \|R_0\|_* . \tag{A.8}$$

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

$$\|\Delta\|_* \le \|R_c\|_* + \|R_0\|_* \le 2\|R_0\|_* \overset{(b)}{\le} 2\sqrt{2r_\natural}\|R_0\|_F \overset{(c)}{\le} 2\sqrt{2r_\natural}\|\Delta\|_F . \tag{A.9}$$

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

Lower bound on $\frac{1}{m}\left\|\tilde{\mathscr{A}}(\Delta)\right\|_2^2$**.** 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 $R_c = U\mathrm{diag}(\sigma)V'$ be the singular value decomposition of R_c. For each $i \ge 1$ define the index set $I_i = \{3r_\natural(i-1)+1, \ldots, 3r_\natural i\}$, and let $R_i := U_{I_i}\mathrm{diag}(\sigma_{I_i})V'_{I_i}$. Using the fact that $\left\langle R_c, R_0\right\rangle = 0$ and the construction of R_i, $i \ge 1$, we also have

$$\left\langle R_i, R_j\right\rangle = 0, \quad \forall\, i \ne j, i, j \ge 0. \tag{A.10}$$

By construction, we have

$$\sigma_k \leq \frac{1}{3r_\natural}\sum_{j\in I_i}\sigma_j \qquad \forall j \in I_{i+1}, \tag{A.11}$$

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

$$\sum_{j\geq 2}\|R_j\|_F \leq \frac{1}{\sqrt{3r_\natural}}\sum_{j\geq 1}\|R_j\|_* = \frac{1}{\sqrt{3r_\natural}}\|R_c\|_* \overset{(a)}{\leq} \frac{1}{\sqrt{3r_\natural}}\|R_0\|_* \overset{(b)}{\leq} \frac{\sqrt{2r_\natural}}{\sqrt{3r_\natural}}\left\|R_0\right\|_{\mathrm{F}}, \tag{A.12}$$

where the step (a) is due to (A.7), and the step (b) is due to the fact that $\mathrm{rank}(R_0) \leq 2r_\natural$. From this inequality, we also have

$$\|\Delta\|_{\mathrm{F}} \leq \left\|R_0+R_1\right\|_{\mathrm{F}} + \sum_{j\geq 2}\left\|R_j\right\|_{\mathrm{F}} \leq \left\|R_0+R_1\right\|_{\mathrm{F}} + 2\left\|R_0\right\|_{\mathrm{F}} \leq 3\left\|R_0+R_1\right\|_{\mathrm{F}} \tag{A.13}$$

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

$$\begin{aligned}
\frac{1}{\sqrt{m}}\|\tilde{\mathscr{A}}(\Delta)\|_2 &\overset{(a)}{\geq} \frac{1}{\sqrt{m}}\|\tilde{\mathscr{A}}(R_0+R_1)\|_2 - \sum_{j\geq 2}\frac{1}{\sqrt{m}}\|\tilde{\mathscr{A}}(R_j)\|_2 \\
&\geq (1-\delta)\left\|R_0+R_1\right\|_{\mathrm{F}} - (1+\delta)\sum_{j\geq 2}\left\|R_j\right\|_{\mathrm{F}} \\
&\overset{(b)}{\geq} (1-\delta)\left\|R_0+R_1\right\|_{\mathrm{F}} - \sqrt{\tfrac{2}{3}}(1+\delta)\left\|R_0\right\|_{\mathrm{F}} \\
&\overset{(c)}{\geq} c_1\left\|R_0+R_1\right\|_{\mathrm{F}} \\
&\overset{(d)}{\geq} c_2\|\Delta\|_{\mathrm{F}}.
\end{aligned} \tag{A.14}$$

Here, in the step (a), we use the reversed triangle inequality. In the step (b), we use (A.12). In the step (c), we use $\left\langle R_0, R_1\right\rangle = 0$ and the choice of δ. The last step (d) is due to (A.13). This finishes the lower bound on $\frac{1}{\sqrt{m}}\|\tilde{\mathscr{A}}(\Delta)\|_2^2$.

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

$$\frac{1}{m}\left\|\tilde{A}(e)\right\|_{\mathrm{op}} \leq C'\sigma\sqrt{\frac{d_1+d_2}{m}}. \tag{A.15}$$

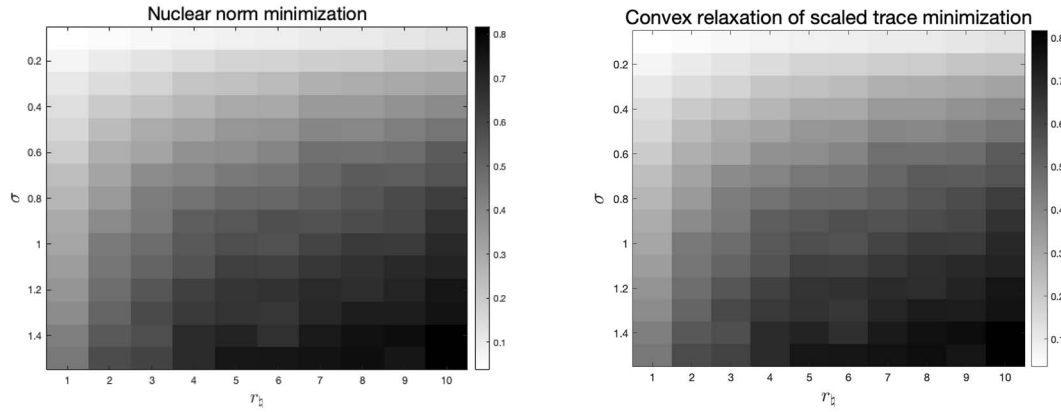


FIG. A6. Matrix sensing with nuclear norm minimization (left) and the convex relaxation (or $k = d$) of Problem (A.5) (right) for different configurations of $(r_\natural, \sigma)$.

Combining (A.6), (A.9), (A.14) and (A.15), we conclude $\|\Delta\|_{\mathrm{F}} \lesssim \sigma\sqrt{\frac{r_\natural(d_1+d_2}{m}}$, 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 (A.5) for the case $k = \min\{d_1, d_2\}$ and the solution $\hat{X}_{\mathrm{nuc}}$ of the nuclear norm minimization (Problem (A.5) with D_1, D_2 being the identity).

Experiment set-up 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 $\mathscr{A}$ 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 (A.5) and the nuclear norm minimization problem. We then average the error $\left\|D_1^{-1}\hat{X}D_2^{-1} - M_\natural\right\|_{\mathrm{F}}$ and $\left\|\hat{X} - M_\natural\right\|_{\mathrm{F}}$ over the 25 trials for each configuration of $r_\natural$ and σ.

Recovery Performance We plot the error in Fig. A6. The white colour indicates small error and the dark colour 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 (A.5). This result validates our theoretical results in Theorem A.1.

B. Missing proofs

B.1 *Proof of Lemma 3.2*

Proof of Lemma 3.2 Lemma 2.2 implies that the pair $(Q_1 L, Q_2 R)$ is balanced, that is $L^\top Q_1^2 L = R^\top Q_2^2 R$. Hence, we may decompose $L^\top L - R^\top R$ in the following way:

$$L^\top L - R^\top R = \left(L^\top L - \frac{L^\top Q_1^2 L}{\alpha_2^2}\right) + \left(\frac{R^\top Q_2^2 R}{\alpha_2^2} - R^\top R\right). \tag{B.1}$$

We bound the first term on the right as follows:

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

Here, (a) and (b) follow, respectively, from the basic inequalities: $\|FG\|_* \leq \|F\|_{\mathrm{F}} \|G\|_{\mathrm{F}}$ and $\|FG\|_F \leq \|F\|_{\mathrm{op}} \|G\|_{\mathrm{F}}$, which hold for all matrices F and G with compatible dimensions. A similar argument yields the inequality $\left\|R^\top R - \frac{R^\top Q_2^2 R}{\alpha_2^2}\right\|_* \leq \left(1-\kappa^{-2}\right)\|R\|_{\mathrm{F}}^2$. The estimate (3.11) follows immediately. □

B.2 *Proof of Lemma 5.3*

Proof of Lemma 5.3. 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 map $P_{\mathscr{T}}$ by

$$P_{\mathscr{T}}(Z) := U_\natural U_\natural^\top Z + Z V_\natural V_\natural^\top - U_\natural U_\natural^\top Z V_\natural V_\natural^\top, \tag{B.3}$$

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

$$\|R\|_* = \left\|P_{\mathscr{T}^\perp}(R) + P_{\mathscr{T}}(R)\right\|_* \leq \left\|P_{\mathscr{T}^\perp}(R)\right\|_* + \left\|P_{\mathscr{T}}(R)\right\|_* \overset{(a)}{\leq} \left\|P_{\mathscr{T}^\perp}(R)\right\|_* + \sqrt{3r_\natural}\left\|P_{\mathscr{T}}(R)\right\|_{\mathrm{F}}, \tag{B.4}$$

where the step (a) is due to the fact that $P_{\mathscr{T}}(R)$ has rank no more than $3r_\natural$. We now bound $\left\|P_{\mathscr{T}^\perp}(R)\right\|_*$ and $\left\|P_{\mathscr{T}}(R)\right\|_{\mathrm{F}}$ 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][8] shows that with probability at least $1 - c_3 d_1^{-5} - c_3 d_2^{-5}$, the inequality

$$\left\|P_{\mathscr{T}^\perp}(R)\right\|_* \leq 8(\|X\|_* - \left\|M_\natural\right\|_*) \tag{B.5}$$

holds. 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 [15, Lemma 11]. Hence, [13,

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

Lemma 5 and its proof][9] shows that with probability at least $1 - c_4 d_1^{-5} - c_4 d_2^{-5}$, the inequality

$$\left\| P_{\mathscr{T}}(R) \right\|_{\mathrm{F}} \leq \frac{\sqrt{2}}{\sqrt{p}} \left\| P_{\mathscr{T}^{\perp}}(R) \right\|_{*} \tag{B.6}$$

holds. Combining (B.4), (B.5) and (B.6) yields the desired inequality (5.8). □

B.3 *Missing proofs in Section 7*

Proof of Lemma 7.1. Elementary algebraic manipulations show the expression for the trace:

$$\begin{aligned} &\mathrm{tr}(D^2 f(U_1, U_2)) \\ =&\frac{2}{m}\left(\sum_{i=1}^{d}\sum_{j=1}^{k}\left\|\mathscr{A}(U_1 e_j e_i^{\top} + e_i e_j^{\top} U_1^{\top})\right\|_2^2\right) + \frac{2}{m}\left(\sum_{i=1}^{d}\sum_{j=1}^{k}\left\|\mathscr{A}(e_i e_j^{\top} U_2^{\top} + U_2 e_i^{\top} e_j)\right\|_2^2\right). \end{aligned} \tag{B.7}$$

Here e_i is a standard basis vector in $\mathbb{R}^d$ and e_j is a standard basis vector in $\mathbb{R}^k$. Using the symmetry of the matrices A_i, the first term can be written as

$$\sum_{i=1}^{d}\sum_{j=1}^{k}\left\|\mathscr{A}(U_1 e_j e_i^{\top} + e_i e_j^{\top} U_1^{\top})\right\|_2^2 = 2\sum_{i=1}^{d}\sum_{j=1}^{k}\sum_{l=1}^{m}\left\langle A_l, U_1 e_j e_i^{\top}\right\rangle^2. \tag{B.8}$$

Following exactly the same computation as (3.4) completes the proof. □

Proof of Lemma 7.2. The implication $2 \Rightarrow 1$ follows from an eigenvalue decomposition of X. Conversely, suppose 1 holds. Observe that 1 clearly is equivalent to being able to write $X = A - B$ with $A, B \succeq 0$, $\mathrm{rank}(A) \leq k_1$ and $\mathrm{rank}(B) \leq k_2$. 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 $\mathscr{U}$ be the span of eigenspaces corresponding to the top r_1 eigenvalues of X. Note that $\mathscr{U}$ has dimension r_1. Cauchy's interlacing theorem implies that the r_1-th largest eigenvalue of $X+B$ satisfies that $\lambda_{r_1}(X+B) \geq \min_{v \in \mathscr{U}\setminus\{0\}} \frac{v^{\top}(X+B)v}{v^{\top}v}$. Since $B \succeq 0$, for any $v \in \mathscr{U} \setminus \{0\}$ we estimate $\frac{v^{\top}(X+B)v}{v^{\top}v} = \frac{v^{\top}Xv + v^{\top}Bv}{v^{\top}v} \geq \frac{v^{\top}Xv}{v^{\top}v} \geq \lambda_{r_1}(X) > 0$. We conclude that that rank of A is at least r_1, which is a contradiction since A has rank at most k_1. □

Proof of Lemma 7.3. 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 $\|X\|_* = \left\|U_1 U_1^{\top} - U_2 U_2^{\top}\right\|_* \leq \left\|U_1 U_1^{\top}\right\|_* + \left\|U_2 U_2^{\top}\right\|_* = \mathrm{tr}(U_1 U_1^{\top}) + \mathrm{tr}(U_2 U_2^{\top}) = \left\|U_1\right\|_{\mathrm{F}}^2 + \left\|U_2\right\|_{\mathrm{F}}^2$. 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 7.2 implies that X has at most k_1 non-negative eigenvalues and k_2 non-positive eigenvalues. Thus, we may

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

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 $\|X\|_* = \left\|U_1\right\|_{\mathrm{F}}^2 + \left\|U_2\right\|_{\mathrm{F}}^2$ holds. □