Low-rank regularization and solution uniqueness in over-parameterized matrix sensing

Kelly Geyer
Boston University

Anastasios Kyrillidis
Rice University

Amir Kalev
University of Maryland

Abstract

We consider the question whether algorithmic choices in over-parameterized linear matrix factorization introduce implicit low-rank regularization. We focus on the noiseless matrix sensing scenario over low-rank positive semi-definite (PSD) matrices over the reals, with a sensing mechanism that satisfies restricted isometry properties. Surprisingly, it was recently argued that for recovery of PSD matrices, gradient descent over a squared, *full-rank* factorized space introduces implicit low-rank regularization. Thus, a clever choice of the recovery algorithm avoids the need for explicit low-rank regularization. In this contribution, we prove that in fact, under certain conditions, the PSD constraint by itself is sufficient to lead to a unique low-rank matrix recovery, without explicit or implicit regularization. Therefore, under these conditions, the set of PSD matrices that are consistent with the observed data, is a singleton, regardless of the algorithm used. Our numerical study indicates that this result is general and extends to cases beyond the those covered by the proof.

1 INTRODUCTION

We study how *over-parameterization* relates to regularization Zhang et al. (2016). By over-parameterization, we assume that the number of parameters to estimate is larger than the available data, thus leading to an underdetermined system.[1] E.g., deep neural networks are usually designed over-parameterized, with ever growing number of layers, and, eventually, a larger number of parameters (Telgarsky, 2016). What is surprising though is the lack of *overfitting* in such networks: while there could be many different parameter realizations that lead to no training error, the algorithms select models that also generalize well to unseen data, despite over-parameterization (Keskar et al., 2016; Poggio et al., 2017; Soltanolkotabi et al., 2017; Dinh et al., 2017; Cooper, 2018).

The authors of (Li et al., 2017) show that the success of over-parameterization can be theoretically fleshed out in the context of shallow, linear neural networks. They consider the case of low-rank and positive semi-definite (PSD) factorization in matrix sensing (Recht et al., 2010): given measurements $y = \mathcal{A}(X^\star) \in \mathbb{R}^m$ —where $X^\star \in \mathbb{R}^{n\times n}$ has rank $r \ll n$ and is PSD, and $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ satisfies regulatory conditions, such as the restricted isometry property— they prove that a *square and full-rank* factorized gradient descent algorithm over $U \in \mathbb{R}^{n\times n}$, where $X = UU^\top$, converges to $X^\star$. *I.e.*, whereas the algorithm has the expressive power to find any matrix X that is consistent with the noiseless data (and due to over-parametrization there might be infinitely many such X's), in contrast, it automatically converges to the minimum rank solution. This argument was previously conjectured in (Gunasekar et al., 2017).

This could be seen as a first step towards understanding over-parameterization from an algorithmic point of view in general non-linear models, whose objectives are more involved and complex. Such network simplifications have been followed in other recent works in machine learning and theoretical computer science, such as in convolutional neural networks (Du et al., 2017), and landscape characterization of generic objectives (Baldi and Hornik, 1989; Boob and Lan, 2017; Safran and Shamir, 2017).

In this work, we provide a different perspective on the

[1] It helps picturing over-parameterization via a simple linear system of equations: when the number of parameters is more than the number of equations, there is an infinite number of solutions, and which is the one we choose depends on additional regularization bias.

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

interpretation of over-parameterization in matrix sensing. We show that, in the noiseless case, *under certain conditions the PSD constraint by itself is sufficient to lead to a unique matrix recovery from observations, without the use of implicit or explicit low-rankness.* In other words, the set of PSD matrices that is consistent with the (noiseless) measurements is a singleton, irrespective of the algorithm used.

Notation. Vectors are denoted with plain lower case letters; matrices are denoted with capital letters; and mappings, from one Euclidean space to another, are denoted with capital calligraphic letters. Given $x \in \mathbb{R}^n$, its ℓ_1-norm is defined as $\|x\|_1 = \sum_{i=1}^n |x_i|$, where x_i denotes its i-th entry; similarly, we define the ℓ_2-norm as $\|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$. The ℓ_0-pseudonorm, $\|x\|_0$, is defined as the number non-zero entries in x. Given x, $\texttt{diag}(x) \in \mathbb{R}^{n\times n}$ is the diagonal matrix with diagonal entries the vector x. For two matrices X, Y with appropriate dimensions, we define their inner product as $\langle X, Y\rangle = \mathrm{Tr}(X^\top Y)$, where $\mathrm{Tr}(\cdot)$ is the trace operator. Given $X \in \mathbb{R}^{n\times n}$, the nuclear norm is defined as $\|X\|_* = \sum_{i=1}^n \sigma_i(X)$, where $\sigma_i(X)$ is the i-th singular value. The spectral norm is denoted as $\|X\| = \sigma_{\max}(X)$, where $\sigma_{\max}(\cdot)$ is the maximum singular value.

1.1 Related work

Implicit regularization in matrix sensing. This area was initiated by the conjecture in (Gunasekar et al., 2017): The authors suggest that non-convex gradient descent on a full-dimensional factorization $UU^\top$, where $U \in \mathbb{R}^{n\times n}$, converges to the minimum nuclear norm solution. Reference (Li et al., 2017) sheds light on this conjecture: the authors provide theoretical arguments on how implicit low-rank regularization is inserted by algorithms, even beyond learning matrix factorization models, such as one-hidden-layer neural nets with quadratic activation; see also (Du and Lee, 2018).

Implicit regularization beyond matrix sensing. For the general linear regression setting, (Wilson et al., 2017) shows that, under specific assumptions, adaptive gradient methods, like AdaGrad and Adam, converge to a different solution than the simple (stochastic) gradient descent (SGD); see also (Gunasekar et al., 2018). SGD has been shown to converge to the so-called *minimum norm solution*; see also (Soudry et al., 2017) for the case of logistic regression. This behavior is also demonstrated using DNNs in (Wilson et al., 2017), where simple gradient descent generalizes as well as the adaptive methods.

No spurious local minima. There is a recent line of work, focusing on non-convex problems, that state conditions under which problem formulations actually have no-spurious local minima, when we transform the problem from its convex formulation to a non-convex one. Characteristic examples include that factored gradient descent does not introduce spurious local minima in matrix completion Ge et al. (2016) and matrix sensing Bhojanapalli et al. (2016b); Park et al. (2017), and all local minima are global in some tensor decompositions Ge et al. (2015) and dictionary learning Sun et al. (2016); see Ge et al. (2017); Sun et al. (2015) for a complete overview of these results. Further, there is literature that characterizes the landscape of factorization problems, using the strict saddle property, to indicate that we can escape easily any saddle point Ge et al. (2015); Zhu et al. (2018); in this work, we take a different path showing that by construction the set of solutions that satisfy our observations is a singleton, demystifying the behavior of factored gradient descent in over-parameterized matrix sensing.

2 NONNEGATIVITY AND SPARSITY: THE VECTOR ANALOG OF PSD AND LOW RANKNESS

We briefly describe the work of (Bruckstein et al., 2008), as we borrow ideas from that paper. Consider the problem of finding a non-negative, sparse solution to an over-parameterized linear system of equations: $Ax^\star = b$. Here, the *sensing matrix* is $A \in \mathbb{R}^{m\times n}$, where $m < n$, the unknown $x^\star \in \mathbb{R}^n$ satisfies $x^\star \geq 0$ (entrywise) and is sufficiently sparse $\|x^\star\|_0 \leq k$, and the measurements are $b \in \mathbb{R}^m$.

This scenario suggests the following optimization problem as a solution:

$$\min_{x\in\mathbb{R}^n} \quad f(x) \quad \text{subject to} \quad b = Ax \text{ and } x \geq 0. \tag{1}$$

Here, f is a function metric that measures the quality of the candidate solutions. Examples are $f(x) = \|x\|_2^2$ (*i.e.*, the minimum norm solution that satisfies the constraints), $f(x) = \|x\|_1$ (*i.e.*, the solution that has small ℓ_1-norm, and promotes sparsity), and $f(x) = \|x\|_0$ (*i.e.*, the solution with the smallest number of non-zeros). These tasks have been encountered in statistics, computer vision and signal processing applications (Zass and Shashua, 2007; Shashua and Hazan, 2005; Hazan et al., 2005), and they are popular in the compressed sensing literature (Donoho, 2006; Foucart and Rauhut, 2013), when $x^\star$ is assumed sparse.

Let us disregard for the moment the positivity constraints on x. By definition, an over-parameterized

linear inverse problem has infinite number of solutions. Unless we use the information that $x^\star$ is sparse, its reconstruction using only b and A is an ill-posed problem, and there is no hope in finding the true vector without ambiguity.

Therefore, to reconstruct $x^\star$ in an over-parametrized setting, prior knowledge should be exploited by the optimization solver. Compressed sensing is an example where additional constraints restrict the feasible set to a singleton: under proper assumptions on the sensing matrix A –such as the restricted isometry property (Candes, 2008), or the coherence property (Bruckstein et al., 2008)– and assuming sufficient number of measurements $m < n$, one can show that the feasible set $\{x : Ax = b \text{ and } \|x\|_0 \leq k\}$ contains only one element, for sufficiently small k.

Re-inserting the positivity constraints in our discussion, (Bruckstein et al., 2008) show that, when a sufficiently sparse solution $x^\star$ generates $b = Ax^\star$, and assuming the row-span of A intersects with the positive orthant, then *the non-negative constraint by itself is sufficient to identify the sparse $x^\star$, and reduce the cardinality of the feasible solutions $\{x : Ax = b\}$ to singleton.* In other words, the inclusion of a sparsity inducing f in (1) is not needed, even if we know a priori that $x^\star$ is sparse; non-negativity is sufficient to find a unique solution to the feasibility problem:

$$\text{find} \quad x \quad \text{such that} \quad b = Ax \text{ and } x \geq 0,$$

that matches $x^\star$. This way, we can still use convex optimization solvers –linear programming in this particular case– and avoid hard non-convex problem instances.

3 THE MATRIX SENSING PROBLEM FOR PSD MATRICES

Let us now describe the *matrix sensing* problem, draw the connections with the vector case, and study the over-parametrization $X = UU^\top$, for $U \in \mathbb{R}^{n\times n}$. Following (Li et al., 2017), we consider the PSD-constrained case, where the optimum solution is both low-rank and PSD.

Its rough description is as *a problem of linear system of equations over matrices.* It is derived by the generative model $b = \mathcal{A}(X^\star)$, where $X^\star \in \mathbb{R}^{n\times n}$ is the low-rank, PSD ground truth. Let the true rank of $X^\star$ be $r \ll n$. The mapping $\mathcal{A} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ is such that the i-th entry of $\mathcal{A}(X)$ is given by $(\mathcal{A}(X))_i = \langle A_i, X\rangle$, for $A_i \in \mathbb{R}^{n\times n}$ independently drawn symmetric measurement matrices.

We study the PSD-constrained formulation, where we aim to find $X^\star$ via:

$$\min_{X\in\mathbb{R}^{n\times n}} \quad f(X) \quad \text{subject to} \quad b = \mathcal{A}(X),\ X \succeq 0. \tag{2}$$

$f(X)$ again represents a function metric that promotes low-rankness; standard choices include the nuclear norm $f(X) = \|X\|_*$ (which imposes "sparsity" on the set of singular values and hence low-rankness), and the non-convex $f(X) = \text{rank}(X)$ metric.

Practical methods for this scenario include (i) the PSD-constrained basis pursuit algorithm for matrices (Chen et al., 2001; Goldstein and Setzer, 2010) that solve (2) for $f(X) := \|X\|_*$ using interior-point methods (Liu and Vandenberghe, 2009); and (ii) projected gradient descent algorithms, that solve an equivalent form of (2) for wisely chosen $\lambda > 0$:

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times n}} \quad & g(X) := \tfrac{1}{2}\|b - \mathcal{A}(X)\|_2^2 \\ \text{subject to} \quad & X \succeq 0, f(X) \leq \lambda. \end{aligned} \tag{3}$$

via:

$$X_{i+1} = \Pi_{\mathcal{C}}\left(X_i - \eta\nabla g(X_i)\right),$$

for $\Pi_{\mathcal{C}}(Y) := \arg\min_{X\in\mathcal{C}} \frac{1}{2}\|X - Y\|_F^2$, and $\mathcal{C} := \{X : X \succeq 0, f(X) \leq \lambda\}$ (Kyrillidis and Cevher, 2011, 2014; Khanna and Kyrillidis, 2017). In (3), the objective f appears in the constraint set as $f(X) := \text{rank}(X)$ or $f(X) := \|X\|_*$.

Recently, we have witnessed a series of works (Zhao et al., 2015; Park et al., 2016a,c; Sun and Luo, 2016; Bhojanapalli et al., 2016b; Park et al., 2016b; Kyrillidis et al., 2018; Ge et al., 2017; Hsieh et al., 2017), that operate directly on the factorization $X = UU^\top$, and do not include any PSD and rank constraints. This is based on the observation that, for any rank-r and PSD X, the factorization $UU^\top$, for $U \in \mathbb{R}^{n\times r}$, guarantees that $X(= UU^\top)$ is at the same time PSD and at most rank-r. This re-parameterizes (2) as:

$$\text{find} \quad U \in \mathbb{R}^{n\times r} \quad \text{subject to} \quad b = \mathcal{A}(UU^\top),$$

and (3) as:

$$\min_{U\in\mathbb{R}^{n\times r}} \quad g(UU^\top) := \tfrac{1}{2}\|b - \mathcal{A}(UU^\top)\|_2^2$$

Observe that in both cases, there are no metrics that explicitly favor low-rankness or any PSD constraints; these are implicitly encoded by the factorization $UU^\top$. Algorithmic solutions for the above criteria include the factorized gradient descent (Bhojanapalli et al., 2016a; Park et al., 2016b) that obeys the following recursion:

$$U_{i+1} = U_i - \eta\nabla g(U_iU_i^\top)\cdot U_i. \tag{4}$$

Current theory (Bhojanapalli et al., 2016a; Park et al., 2016b) assumes that r is known a priori, in order to set the dimensions of the factor $U \in \mathbb{R}^{n\times r}$, accordingly. The only work that deviates from this perspective is

the recent work in (Li et al., 2017), where the authors prove that, even if we use square $U \in \mathbb{R}^{n\times n}$ in (4), we could still converge to the low-rank ground truth $X^\star$, with proper initialization and step size selection. The result relies on restricted isometry assumptions of $\mathcal{A}$. In a manner, this suggests that *operating on the factorized space, the algorithm implicitly favors low-rank solutions, even if there is expressive power to select a full rank-n $\widehat{X} = \widehat{U}\widehat{U}^\top$ as a solution.* The following subsection provides a different perspective on the matter: *the PSD constraint by itself is sufficient, under certain conditions, to reduce the feasibility set to a singleton, irrespective of how one imposes it in an algorithm.*

Restricted isometries. We note that the *Restricted Isometry Property* (RIP) assumption is made in (Zhao et al., 2015; Park et al., 2016a,c; Sun and Luo, 2016; Bhojanapalli et al., 2016b; Park et al., 2016b; Kyrillidis et al., 2018; Ge et al., 2017; Hsieh et al., 2017; Li et al., 2017). There are various versions of RIP, with the most well-known being the RIP-ℓ_2/ℓ_2 (Chen et al., 2015). By construction of the sensing matrices we choose for our theory—i.e., as outer products of Gaussians, we will use a variant of RIP, RIP-ℓ_2/ℓ_1 (Chen et al., 2015). The equivalence or superiority of one RIP definition over the other is not known, to the best of our knowledge. The RIP of linear maps on low rank matrices is key in our disucssion (Candes and Plan, 2011; Liu, 2011):

Definition 1 (RIP in ℓ_2/ℓ_1 Chen et al. (2015)) *A linear map $\mathcal{F} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ satisfies the r-RIP-ℓ_2/ℓ_1 with constant δ_r, if $(1-\delta_r)\|X\|_F \leq \|\mathcal{F}(X)\|_1 \leq (1+\delta_r)\|X\|_F$, is satisfied for all matrices $X \in \mathbb{R}^{n\times n}$ such that rank$(X) \leq r$. Here, $\|\cdot\|_1$ denotes the ℓ_1-norm over matrices.*

Corollary 1 ((Needell and Tropp, 2009)) *Let γ and r be positive integers. Then, $\delta_{\gamma r} \leq \gamma \cdot \delta_{2r}$.*

Sensing mappings comprised of Wishart matrices and their properties. We extend the results in the previous section, and *prove that, under appropriate conditions, the set of solutions $\{X \in \mathbb{R}^{n\times n} : b = \mathcal{A}(X), X \succeq 0\}$ is a singleton.* To generalize, in our theoretical developments, we will consider the case where $\mathcal{A}$ is generated through a Gaussian process.

First, let us define the Wishart distribution.

Definition 2 (Muirhead (2009)) *Let $Z \in \mathbb{R}^{p\times n}$ where each row $z_1, \ldots, z_p \sim \mathcal{N}_n(0, \Sigma)$ (multivariate normal with zero mean). Define $n \times n$ matrix A by $A = \sum_{i=1}^p z_i^T z_i = Z^T Z$. We say that A follows a Wishart distribution with p degrees of freedom and covariance matrix $\Sigma \succeq 0$, which we denote by $A \sim W_n(p, \Sigma)$.*

We consider the sensing map $(\mathcal{A}(X))_i = \langle A_i, X\rangle$, where A_i are non-singular Wishart matrices for $i = 1, \ldots, m$, and m is the total number of measurements.

Wishart matrices are commonly used to estimate covariance in high dimensional statistics (Vershynin, 2012; Chen et al., 2015), and they have properties that we will exploit in our theory. To generate $\mathcal{A}$, we generate Wishart matrices A_i as defined above, for $\Sigma = \sigma^2 I_n \succ 0$, and I_n is the $n \times n$ identity matrix.

The parameter p is user-defined and set to $p > n + 1$. By assumption of Σ, all A_i's are non-singular Wishart matrices (Eaton, 2007). This ensures that, $\forall A_i$, our theory holds by the properties of non-singular Wishart matrices: *i*) the density function of A_i exists, *ii*) A_i^{-1} exists, and *iii*) A_i is positive definite.

By definition of A_i as positive definite matrices, this results to the following observation:

$$\exists\ \varphi = [\varphi_1,\ \varphi_2,\ \ldots,\ \varphi_m]^\top \quad \text{such that} \tag{5}$$

$$B = \sum_{i=1}^m \varphi_i A_i \in \mathbb{R}^{n\times n} \quad \text{and} \quad B \succ 0. \tag{6}$$

I.e., there exists at least one vector φ such that the weighted sum of A_i's is a positive definite matrix. This can be easily derived from the fact that by construction all A_i's are positive definite; this also relates to the Farkas' Lemma for semidefinite programs (Lovász, 2003).

To proceed, we require the following definitions of Wishart matrices:

- By (Muirhead, 2009), we know that B is a non-singular Wishart matrix satisfying: $B \sim W_n(m \cdot p, \Sigma)$. I.e., the weight sum of Wishart matrices satisfies the Wishart distribution.
- As a non-singular matrix, B^{-1} exists and follows an inverse Wishart distribution. In particular, $B^{-1} \sim W_n^{-1}(mp + n + 1, \Sigma^{-1})$ (Eaton, 2007; Muirhead, 2009).

Further, since $B \succ 0$, there exists a unique $V \in \mathbb{R}^{n\times n}$ such that $B = VV^\top$.

- Regarding the decomposition $B = VV^\top$, we can extract information about V's by Bartlett's Decomposition (Eaton, 2007). In particular, the matrix V is a lower triangular matrix, where the random variables $V_{kj}|k \geq j$ are mutually independent: for $k > j$, V_{kj} follow a normal distribution as $V_{kj} \sim \mathcal{N}(0, \sigma^2)$, and diagonal elements of V follow a chi-squared distribution as $V_{jj}^2 \sim \sigma^2 \cdot \mathcal{X}^2_{m-j+1}$, for all $j = 1, \ldots, n$.

Equivalent reformulation of matrix sensing with fixed trace. Given the above set up, we will

make the following connections, starting with the following change of variables. Given the full rank V such that $B = VV^\top$, and for each A_i, we define a new mapping $\mathcal{M} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$, such that:

$$(\mathcal{M}(X))_i = \langle M_i, X\rangle = \left\langle V^{-1}A_i\left(V^{-1}\right)^\top, X\right\rangle,$$
$$\text{for all } \ i, \quad \text{where } M_i := V^{-1}A_i\left(V^{-1}\right)^\top.$$

Given $X \in \mathbb{R}^{n\times n}$ and $X \succeq 0$, define the auxiliary variable $Y = V^\top XV \in \mathbb{R}^{n\times n}$; observe that, for full rank V, $Y \succeq 0$. Then, for any $X \succeq 0$, we have:

$$\begin{aligned} b_i &= \langle A_i, X\rangle = \langle A_i, XVV^{-1}\rangle \\ &= \langle A_i\left(V^{-1}\right)^\top, XV\rangle = \langle A_i\left(V^{-1}\right)^\top, (V^\top)^{-1}V^\top XV\rangle \\ &= \langle V^{-1}A_i\left(V^{-1}\right)^\top, V^\top XV\rangle = \langle M_i, Y\rangle, \end{aligned}$$

where the last equality is due to the definitions of $(\mathcal{M}(\cdot))_i$ and Y. For the rest of the discussion, we assume that $b = \mathcal{A}(X^\star)$, for rank-r $X^\star$.

The above indicate the one-to-one correspondence between the original feasibility set and the corresponding set, after the change of variables:

$$\begin{aligned} &\{X \in \mathbb{R}^{n\times n} : b = \mathcal{A}(X), X \succeq 0\} \quad \text{and} \\ &\{Y \in \mathbb{R}^{n\times n} : b = \mathcal{M}(Y), Y \succeq 0\}. \end{aligned} \tag{7}$$

Further, the rank of the solutions, $X^\star$ and $Y^\star$, are the same. After the change of variables to $\mathcal{M}$, for X and Y that belong to the above sets, we observe:

$$\begin{aligned} \mathrm{Tr}(Y) &\overset{(i)}{=} \mathrm{Tr}(V^\top XV) = \mathrm{Tr}(XVV^\top) = \mathrm{Tr}(XB) \\ &\overset{(ii)}{=} \mathrm{Tr}\left(X\sum_{i=1}^m \varphi_i A_i\right) = \sum_{i=1}^m \varphi_i \cdot \langle A_i, X\rangle \\ &\overset{(iii)}{=} \sum_{i=1}^m \varphi_i \cdot b_i := c, \quad \text{for constant } \ c. \end{aligned}$$

Here, (i) is due to the definition of $Y = V^\top XV$, (ii) is due to the assumption that the span of $\mathcal{A}$ is strictly positive and equals B, according to (3), and (iii) is due to $b_i = \langle A_i, X\rangle$, for X being in the feasibility set. *This dictates that the trace of matrices in the set* $\{Y \in \mathbb{R}^{n\times n} : b = \mathcal{M}(Y), Y \succeq 0\}$ *is constant and does not depend on* X *directly*; it only depends on the measurement vector b and the vector φ defined above.

Let us focus on the set $\{Y \in \mathbb{R}^{n\times n} : b = \mathcal{M}(Y), Y \succeq 0\}$. By definition, $b = \mathcal{M}(Y^\star)$, where $Y^\star$ is rank-r and relates to $X^\star$ in $Y^\star = V^\top X^\star V$. Assume that $\mathcal{M} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$, is a linear map that satisfies the RIP in Definition 1. Consider the convex optimization criterion with estimate $\widehat{Y}$:

$$\widehat{Y} = \underset{Y\in\mathbb{R}^{n\times n}}{\mathrm{argmin}} \ \|Y\|_* \quad \text{subject to} \quad b = \mathcal{M}(Y). \tag{8}$$

The following result is from (Chen et al., 2015).

Theorem 1 (Chen et al. (2015), Informal) *Assume $\mathcal{M}(\cdot)$ satisfies the RIP-ℓ_2/ℓ_1 for some $\delta_{\gamma r} < 1$, and for some integer $r \geq 1$. Then,* (8)*, in the absence of noise, allows perfect recovery of the unique $Y^\star$ that satisfies the measurements, with exponentially high probability, provided we have enough samples $O(\gamma nr)$.*

Let us interpret and use this theorem. Assume that $\mathrm{rank}(Y^\star) = r$, and $\delta_{\gamma r} < 1$. Under these assumptions, the minimizer $\widehat{Y}$ of (8) is identical to the unique, rank-r matrix $Y^\star$ that satisfies the the set of observations $b = \mathcal{M}(Y^\star)$. Taking into account the PSD nature of Y, we also have $\|Y^\star\|_* = \mathrm{Tr}(Y^\star)$; further, since $Y^\star$ is unique, we see that $\mathrm{Tr}(Y^\star) = c$ for some c. Note that we do not include the constraint $\mathrm{Tr}(Y) = c$ in the optimization, because any feasible solution should satisfy this condition, as we note above. Also, we do not include the PSD constraint; the problem (8) is sufficient to guarantee uniqueness.

By the above theorem, any other PSD solution, $Y^\sharp$, that satisfies the measurements b must have a nuclear norm larger than $\|\widehat{Y}\|_* = \|Y^\star\|_*$. Being PSD, this also means $\mathrm{Tr}(Y^\sharp) > c$, which implies that any other PSD solution is not in the feasible set $\{Y \in \mathbb{R}^{n\times n} : b = \mathcal{M}(Y), Y \succeq 0\}$. Hence, this set contains only one element, by contradiction.

Due to the one-to-one correspondence between the sets in (7) then, we infer that the first set is also a singleton. This further implies that *the inclusion of any metric f that favors low-rankness in* (2)-(3) *or restricting U to be a tall matrix with wisely chosen r in* (4) *makes no difference, as there is only one matrix that fits measurements b.*

RIP-ℓ_2/ℓ_1 for the new sensing mapping $\mathcal{M}(\cdot)$. Key assumption in our discussion so far is that a form of RIP holds for the transformed sensing map $\mathcal{M}(\cdot)$ in order to guarantee the uniqueness of the solution, through the convex optimization objective—and not the original sensing map $\mathcal{A}$. Thus, in general, it is required to find such transformation between $\mathcal{A}$ and $\mathcal{M}$.

We know that $M_i := V^{-1}A_i\left(V^{-1}\right)^\top$, where $B = \sum_{i=1}^m \varphi_i A_i = VV^\top$. By construction through the Wishart distribution, we know that M_i, by Theorem 3.2.4 from (Muirhead, 2009), satisfies:

$$M_i := V^{-1}A_i(V^{-1})^T \sim W_n(p, \Sigma_M),$$

where $\Sigma_M = V^{-1}\Sigma(V^{-1})^T = \sigma^2(VV^T)^{-1} = \sigma^2 B^{-1}$.

Using the following definition of sub-exponential random variables, we show Lemma 1.

Definition 3 ***(Sum of Sub-Exponential Random Variables, (Wainwright, 2015))*** *Suppose that*

$X_1, \ldots, X_n$ are independent (τ_i^2, b_i)-sub-Exponential random variables. Then, the sum $\sum_{i=1}^n X_i$ is $(\sum_{i=1}^n \tau_i^2, b^\star)$-sub-Exponential, where $b^\star = \max_i\{b_i\}$.

Lemma 1 *Non-singular Wishart matrices are sub-exponential matrices.*

Proof. Let A be a non-singular Wishart random matrix. According to our constructions so far, A may be characterized by $A \sim W_n(p, \Sigma)$, $p > n$, and $\Sigma = \sigma^2 I_n \succ 0$. Recall that $A := Z^T Z$, where $Z \in \mathbb{R}^{p\times n}$, where each row of Z is generated from a multivariate normal distribution with zero mean. It is well known that each row Z_i, $1 \leq i \leq p$, is a sub-Gaussian random vector (Vershynin, 2017), and that all elements of these vectors are sub-Gaussian random variables. From Rinaldo (2019), we know that both the square of a sub-Gaussian random variable, and the product of independent sub-Gaussian random variables, are sub-Exponential; see also Lemma 7 in Ahmed and Romberg (2014).

We then can use the following result from Foucart and Subramanian (2019):

Theorem 2 ((Foucart and Subramanian, 2019)) *Given sub-exponential sensing matrices M_i for the matrix sensing setting over rank-r matrices, the RIP-ℓ_2/ℓ_1 requirement in Definition 1 is satisfied with probability:*

$$\mathbb{P}\left(|\|\mathcal{M}(X)\|_1 - \|X\|_F| > \delta\|X\|_F\right) \leq 2e^{-\kappa\delta^2 m};$$

for κ constant related to the subexponential distribution, and $m \geq c(\delta)nr$, for constant $c(\delta)$.

This completes the proof: In plain words, using Gaussian-generated sensing matrices, we can generate Wishart sensing maps $\mathcal{A}$, such that the solution cardinality that satisfies the observations is a singleton.

Remark 1 *The above show that the specific RIP-ℓ_2/ℓ_1 assumption on $\mathcal{M}$ is a sufficient,* but not a necessary*, condition to guarantee that the feasibility set $\{X \in \mathbb{R}^{n\times n} : b = \mathcal{A}(X), X \succeq 0\}$ is a singleton, $X^\star$. It remains an open question to find necessary conditions and possibly different sufficient conditions –such as the incoherence condition in (Bruckstein et al., 2008) for matrices, or the RIP-ℓ_2/ℓ_2– that also lead to a singleton set. In (Baldwin et al., 2015) the authors find particular instances of sensing maps that, while not satisfy RIP condition, lead to a singleton set. It is interesting to study the sensing construction in the current paper for RIP-ℓ_2/ℓ_2 conditions using Theorem 3:*

Theorem 3 ((Chen et al., 2015), Theorem 5) *Let $\mathcal{P}(\cdot) : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ be a sensing map, with individual matrices $P_i \in \mathbb{R}^{n\times n}$. Suppose that for all $1 \leq i \leq m$, $\|P_i\|_2 \leq K$, $\|\mathbb{E}\left[\mathcal{P}_i^*\mathcal{P}_i\right] - \mathcal{I}\|_2 \leq \frac{c_5}{n}$, hold for some quantity $K \leq n^2$. For any small constant $\delta > 0$, if $m > c_0 r K^2 \log^7 n$, then with probability at least $1 - 1/n^2$, one has $\mathcal{P}$ satisfies RIP-ℓ_2/ℓ_2 w.r.t. all matrices of rank at most r and obeys $\delta_r \leq \delta$.*

4 EXPERIMENTS

The aim of the following experiments is twofold: in Section 4.1, to show that the theory applies in practice; in Section 4.2, to show that a sensing map $\mathcal{A}$ beyond the Wishart distribution can be sufficient to lead to a good approximation of $X^\star$, without the use of explicit regularization for low-rankness.

4.1 Using Wishart matrices in simulated matrix sensing problems

The following example shows some preliminary results, shown in Figure 1. Here, we compare *i*) least squares in X with no constraints (CVX - second order method), *ii*) least squares in X with PSD constraints (CVX - second order method), and *iii*) least squares in X with PSD constraints (using projected gradient descent (PGD)). The plot assumes small $n = 15$ for proof of concept, where $X \in \mathbb{R}^{n\times n}$, due to the computational restrictions the second order method poses; same behavior is observed for any value we tested. The degrees of freedom are $n(n+1)/2$. While, criterion *i*) finds a good solution after observing $n(n+1)/2$ samples, as expected, criteria *ii*) $-$ *iii*) find a relatively good solution well before that, which implies that the PSD constraint alone is sufficient to find the solution, only from a limited set of random measurements.

We generate $X^\star$ as a rank-1 PSD matrix, where $X^\star = \lambda vv^\top$ for random scalar λ and vector v. The measurement matrices A_i are generated from a Gaussian distribution: $b \overset{i.i.d.}{\sim} \mathcal{N}_n(0, I_n)$ and $A_i = \frac{1}{2\sqrt{n}} bb^\top$. Therefore the A_i are symmetric and PSD, according to the construction described in the main text. The measurements $y \in \mathbb{R}^m$ are generated by $y_i = \mathrm{tr}(A_i^T X^\star)$ for $i = 1, \ldots, m$.

4.2 Beyond Wishart matrices: quantum state tomography

In this subsection, we consider the setting of quantum state tomography (QST): We generate measurements according to $b_i = \langle A_i, X^\star\rangle$, $i = 1, \ldots, m$, where $A_i = \left(I \pm \otimes_{j=1}^q s_j\right)/2$ for $s_{i,j} \in \{\sigma_x, \sigma_y, \sigma_z\}$ a random Pauli operator. In all settings, for simplicity, we assume $X^\star \in \mathbb{C}^{2^q\times 2^q}$ is rank-1, PSD and normalized $\mathrm{Tr}(X^\star) = 1$, to

	$\min \|X\|_*$		$\min \|X\|_F^2$		$\min \frac{1}{2}\|b-\mathcal{A}(X)\|_2^2$	
(n^2, m)	$\text{dist}(\widehat{X}, X^\star)$	$\text{dist}(\widehat{X}_1, X^\star)$	$\text{dist}(\widehat{X}, X^\star)$	$\text{dist}(\widehat{X}_1, X^\star)$	$\text{dist}(\widehat{X}, X^\star)$	$\text{dist}(\widehat{X}_1, X^\star)$
$(256, 128)$	$3.58 \cdot 10^{-5}$	$3.45 \cdot 10^{-5}$	$4.66 \cdot 10^{-3}$	$4.49 \cdot 10^{-3}$	$1.54 \cdot 10^{-4}$	$1.43 \cdot 10^{-4}$
$(1024, 288)$	$1.65 \cdot 10^{-5}$	$1.63 \cdot 10^{-5}$	$1.68 \cdot 10^{-3}$	$1.64 \cdot 10^{-3}$	$8.08 \cdot 10^{-5}$	$7.61 \cdot 10^{-5}$
$(4096, 640)$	$1.84 \cdot 10^{-5}$	$1.82 \cdot 10^{-5}$	$1.51 \cdot 10^{-3}$	$1.48 \cdot 10^{-3}$	$1.04 \cdot 10^{-4}$	$9.81 \cdot 10^{-5}$
$(16384, 1536)$	$1.28 \cdot 10^{-5}$	$1.27 \cdot 10^{-5}$	$1.00 \cdot 10^{-3}$	$9.98 \cdot 10^{-3}$	$5.45 \cdot 10^{-5}$	$5.20 \cdot 10^{-5}$

Table 1: Experimental results for (11). $\text{dist}(\widehat{X}, X^\star)$ defines the entrywise distance $\|\widehat{X} - X^\star\|_F$.

	$\min \frac{1}{2}\|b-\mathcal{A}(X)\|_2^2$		$U \in \mathbb{C}^{n\times r}$		$U \in \mathbb{C}^{n\times n}$	
(n^2, m)	$\text{dist}(\widehat{X}, X^\star)$	$\text{dist}(\widehat{X}_1, X^\star)$	$\text{dist}(\widehat{X}, X^\star)$	$\text{dist}(\widehat{X}_1, X^\star)$	$\text{dist}(\widehat{X}, X^\star)$	$\text{dist}(\widehat{X}_1, X^\star)$
$(256, 128)$	$1.54 \cdot 10^{-4}$	$1.43 \cdot 10^{-4}$	$9.52 \cdot 10^{-5}$	-	$3.12 \cdot 10^{-2}$	$2.82 \cdot 10^{-2}$
$(1024, 288)$	$8.08 \cdot 10^{-5}$	$7.61 \cdot 10^{-5}$	$4.47 \cdot 10^{-5}$	-	$1.87 \cdot 10^{-2}$	$1.76 \cdot 10^{-2}$
$(4096, 640)$	$1.04 \cdot 10^{-4}$	$9.81 \cdot 10^{-5}$	$4.07 \cdot 10^{-5}$	-	$2.51 \cdot 10^{-2}$	$2.37 \cdot 10^{-2}$
$(16384, 1536)$	$5.45 \cdot 10^{-5}$	$5.20 \cdot 10^{-5}$	$2.47 \cdot 10^{-5}$	-	$1.41 \cdot 10^{-2}$	$1.35 \cdot 10^{-2}$

Table 2: Results for $UU^\top$ parameterization. $\text{dist}(\widehat{X}, X^\star)$ defines the entrywise distance $\|\widehat{X} - X^\star\|_F$.

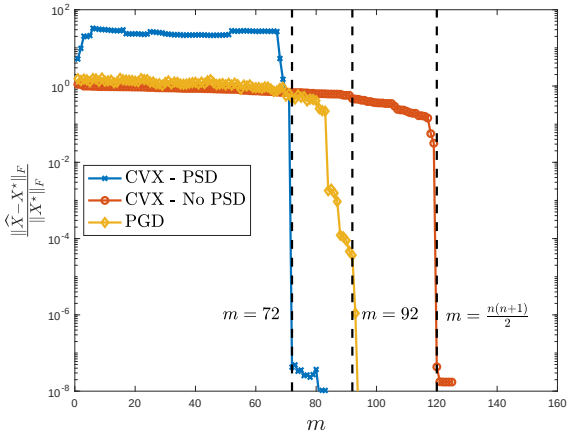

Figure 1: Results for simulation using Wishart sensing matrices A_i.

satisfy the QST setting. Given b and $\mathcal{A}$, we consider:

$$\begin{aligned} \min_{X\in\mathbb{C}^{n\times n}} \quad & \|X\|_* \\ \text{subject to} \quad & b = \mathcal{A}(X), \\ & X \succeq 0. \end{aligned} \tag{9}$$

$$\begin{aligned} \min_{X\in\mathbb{C}^{n\times n}} \quad & \|X\|_F \\ \text{subject to} \quad & b = \mathcal{A}(X), \\ & X \succeq 0. \end{aligned} \tag{10}$$

$$\begin{aligned} \min_{X\in\mathbb{C}^{n\times n}} \quad & \tfrac{1}{2}\|b-\mathcal{A}(X)\|_2^2 \\ \text{subject to} \quad & X \succeq 0. \end{aligned} \tag{11}$$

I.e., (9) is the *nuclear-norm minimization* problem, with explicit regularization towards low-rank solutions (Recht et al., 2010); (10) is the *minimum-norm solution* problem, where the objective regularizes towards X with the minimum Frobenius norm; (11) is the *PSD constrained, least-squares* problem, where the task is to fit the data subject to PSD constraints. In the two latter settings, there is no explicit regularization towards low-rankness.

We use the CVX Matlab implementation, in its low-precision setting, to solve all problems in (11) (Grant and Boyd, 2008, 2014). The results are presented in Table 1: $\text{dist}(\widehat{X}, X^\star)$ denotes the entrywise distance $\|\widehat{X} - X^\star\|_F$. Since the estimates $\widehat{X}$ in all criteria in (11) are only approximately low-rank[2], we also report the entrywise distance between $X^\star$ and the best rank-1 approximation of $\widehat{X}$, denoted as $\widehat{X}_1$. We consider four different settings for (n^2, m) parameters; our experiments are restricted to small values of q in $n^2 = (2^q)^2$, due to the high computational complexity of the CVX solvers (by default we use the `SDPT3` solver (Toh et al., 1999)).

Table 1 support our claim: *All three criteria, and for all cases, lead to the same solution, while they all use different "regularization" in optimization.* Small differences are due to numerical precision, and not equivalent initial conditions. We observe consistently that, using the nuclear-norm bias, we obtain a better approximation of $X^\star$. Thus, *using explicit regularization helps.*

4.3 Behavior of first-order, non-convex solvers on $UU^\top$ parameterization

In view of the previous results, here we study the behavior of first-order, non-convex solvers, that utilize the re-parameterization of X as $UU^\top$. We borrow the iteration in (Bhojanapalli et al., 2016a; Park et al., 2016b), where: $U_{i+1} = U_i - \eta \nabla g(U_i U_i^\top) \cdot U_i$, for $g(UU^\top) := \frac{1}{2}\|b - \mathcal{A}(UU^\top)\|_2^2$. We consider two cases: *i)* $U \in \mathbb{C}^{n\times r}$ where r is the rank of $X^\star$, and is assumed known *a priori*; this is the case in (Bhojanapalli et al., 2016a; Park et al., 2016b) and has explicit regulariza-

[2]Due to numerical precision limits, non of the solutions are $X^\star$ nor rank-1 in the strict sense.

tion, as the algorithm operates only on the space of rank-r matrices. *ii*) $U \in \mathbb{C}^{n\times n}$ where we can operate over the whole space $\mathbb{C}^{n\times n}$; this is the case studied in (Li et al., 2017).

In both cases, the initialization U_0 and step size η follow the prescriptions in (Park et al., 2016b), and they are computed using the same procedures for both cases. Table 2 reports our findings. To ease comparison, we repeat the results of the least-squares objective in (11). We observe that all algorithms converge close to $X^\star$: obviously, using the *a priori* information that $X^\star$ is rank-1 biases towards a low-rank estimate, where faster convergence rates are observed. In the contrary, using $U \in \mathbb{C}^{n\times n}$ shows slower convergence towards the vicinity of $X^\star$; nevertheless, the reported results suggests that still one can achieve a small distance to $X^\star$ ($\|\widehat{X} - X^\star\|_F \lesssim 10^{-2}$). Finally, while $\widehat{X}$ could be full-rank, most of the energy is contained in a small number of principal components, indicating that all algorithms favor (approximately) low-rank solutions.

5 CONCLUSION

We prove that in overparameterized PSD, low-rank matrix sensing, the solution set is a singleton, under RIP assumptions and appropriate transformations on the sensing map $\mathcal{A}$. The PSD constraint itself provides guarantees for unique matrix recovery. The question whether the above can be generalized to less restrictive linear sensing mappings $\mathcal{A}$ remains open. RIP is a sufficient but not a necessary condition; we believe that generalizing our work to more broad settings and assumptions is an interesting research direction.

Acknowledgements

Kelly Geyer was partially supported by NSF award 1547433 RTG: Cross-training in Statistics and Computer Science.

References

A. Ahmed and J. Romberg. Compressive multiplexing of correlated signals. *IEEE Transactions on Information Theory*, 61(1):479–498, 2014.

P. Baldi and K. Hornik. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

C. Baldwin, I. Deutsch, and A. Kalev. Informational completeness in bounded-rank quantum-state tomography. *arXiv preprint arXiv:1510.02736*, 2015.

S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi. Dropping convexity for faster semi-definite optimization. In *Conference on Learning Theory*, pages 530–582, 2016a.

S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global optimality of local search for low rank matrix recovery. In *Advances in Neural Information Processing Systems*, pages 3873–3881, 2016b.

D. Boob and G. Lan. Theoretical properties of the global optimizer of two layer neural network. *arXiv preprint arXiv:1710.11241*, 2017.

A. Bruckstein, M. Elad, and M. Zibulevsky. On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations. *IEEE Transactions on Information Theory*, 54(11):4813–4820, 2008.

E. Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

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

S. S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1): 129–159, 2001.

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

Y. Cooper. The loss landscape of overparameterized neural networks. *arXiv preprint arXiv:1804.10200*, 2018.

L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets. *arXiv preprint arXiv:1703.04933*, 2017.

D. Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

S. Du and J. Lee. On the power of over-parametrization in neural networks with quadratic activation. *arXiv preprint arXiv:1803.01206*, 2018.

S. Du, J. Lee, and Y. Tian. When is a convolutional filter easy to learn? *arXiv preprint arXiv:1709.06129*, 2017.

Morris L. Eaton. *Chapter 8: The Wishart Distribution*, volume Volume 53 of *Lecture Notes–Monograph Series*, pages 302–333. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007. doi: 10.1214/lnms/1196285114. URL `https://doi.org/10.1214/lnms/1196285114`.

S. Foucart and H. Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

S. Foucart and S. Subramanian. Iterative hard thresholding for low-rank recovery from rank-one projections. *Linear Algebra and its Applications*, 2019.

R. Ge, F. Huang, C. Jin, and Y. Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. In *Conference on Learning Theory*, pages 797–842, 2015.

R. Ge, J. Lee, and T. Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

R. Ge, C. Jin, and Y. Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

T. Goldstein and S. Setzer. High-order methods for basis pursuit. *UCLA CAM Report*, pages 10–41, 2010.

M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008.

M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1, March 2014.

S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6152–6160, 2017.

S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

T. Hazan, S. Polak, and A. Shashua. Sparse image coding using a 3D non-negative tensor factorization. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 50–57. IEEE, 2005.

Y.-P. Hsieh, Y.-C. Kao, R. Karimi Mahabadi, Y. Alp, A. Kyrillidis, and V. Cevher. A non-euclidean gradient descent framework for non-convex matrix factorization. Technical report, Institute of Electrical and Electronics Engineers, 2017.

N. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

R. Khanna and A. Kyrillidis. IHT dies hard: Provable accelerated iterative hard thresholding. *arXiv preprint arXiv:1712.09379*, 2017.

A. Kyrillidis and V. Cevher. Recipes on hard thresholding methods. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), 2011 4th IEEE International Workshop on*, pages 353–356. IEEE, 2011.

A. Kyrillidis and V. Cevher. Matrix recipes for hard thresholding methods. *Journal of mathematical imaging and vision*, 48(2):235–265, 2014.

A. Kyrillidis, A. Kalev, D. Park, S. Bhojanapalli, C. Caramanis, and S. Sanghavi. Provable quantum state tomography via non-convex methods. *npj Quantum Information*, 4(36), 2018.

Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parameterized matrix recovery. *arXiv preprint arXiv:1712.09203*, 2017.

Y.-K. Liu. Universal low-rank matrix recovery from Pauli measurements. In *Advances in Neural Information Processing Systems*, pages 1638–1646, 2011.

Z. Liu and L. Vandenberghe. Interior-point method for nuclear norm approximation with application to system identification. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1235–1256, 2009.

László Lovász. Semidefinite programs and combinatorial optimization. In *Recent advances in algorithms and combinatorics*, pages 137–194. Springer, 2003.

Robb J Muirhead. *Aspects of multivariate statistical theory*, volume 197. John Wiley & Sons, 2009.

D. Needell and J. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and computational harmonic analysis*, 26(3): 301–321, 2009.

D. Park, A. Kyrillidis, S. Bhojanapalli, C. Caramanis, and S. Sanghavi. Provable Burer-Monteiro factorization for a class of norm-constrained matrix problems. *arXiv preprint arXiv:1606.01316*, 2016a.

D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Finding low-rank solutions to matrix problems, efficiently and provably. *arXiv preprint arXiv:1606.03168*, 2016b.

D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. *arXiv preprint arXiv:1609.03240*, 2016c.

D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi. Non-square matrix sensing without spurious local minima via the burer-monteiro approach. In *Artificial Intelligence and Statistics*, pages 65–74, 2017.

T. Poggio, K. Kawaguchi, Q. Liao, B. Miranda, L. Rosasco, X. Boix, J. Hidary, and H. Mhaskar. Theory of deep learning III: explaining the non-overfitting puzzle. *arXiv preprint arXiv:1801.00173*, 2017.

B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3): 471–501, 2010.

Alessandro Rinaldo. Lecture notes in 36-709: Advanced statistical theory, February 2019.

I. Safran and O. Shamir. Spurious local minima are common in two-layer ReLU neural networks. *arXiv preprint arXiv:1712.08968*, 2017.

A. Shashua and T. Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*, pages 792–799. ACM, 2005.

M. Soltanolkotabi, A. Javanmard, and J. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *arXiv preprint arXiv:1707.04926*, 2017.

D. Soudry, E. Hoffer, and N. Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

J. Sun, Q. Qu, and J. Wright. When are nonconvex problems not scary? *arXiv preprint arXiv:1510.06096*, 2015.

J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2016.

R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016.

M. Telgarsky. Benefits of depth in neural networks. *arXiv preprint arXiv:1602.04485*, 2016.

K.-C. Toh, M. Todd, and R. Tütüncü. SDPT3?a MATLAB software package for semidefinite programming, version 1.3. *Optimization methods and software*, 11 (1-4):545–581, 1999.

Roman Vershynin. How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686, 2012.

Roman Vershynin. Four lectures on probabilistic methods for data science, 2017.

Martin Wainwright. Lecture notes in stat 201b: Mathematical statistics, January 2015.

A. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht. The marginal value of adaptive gradient methods in machine learning. In *Advances in Neural Information Processing Systems*, pages 4151–4161, 2017.

R. Zass and A. Shashua. Nonnegative sparse PCA. In *Advances in neural information processing systems*, pages 1561–1568, 2007.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In *Advances in Neural Information Processing Systems*, pages 559–567, 2015.

Z. Zhu, Q. Li, G. Tang, and M. Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.