Beyond Procrustes: Balancing-free Gradient Descent for Asymmetric Low-Rank Matrix Sensing

Cong Ma$^1$, Yuanxin Li$^2$, Yuejie Chi$^2$

$^1$Department of Operations Research and Financial Engineering, Princeton University
$^2$Department of Electrical and Computer Engineering, Carnegie Mellon University

Abstract—Low-rank matrix estimation plays a central role in many applications across science and engineering. Recently, nonconvex formulations based on matrix factorization are provably solved by simple gradient descent algorithms with strong computational and statistical guarantees. However, when the low-rank matrices are asymmetric, existing approaches rely on adding a regularization term to balance the two matrix factors which in practice can be removed safely without hurting the performance when initialized via the spectral method. In this paper, we justify this theoretically for the matrix sensing problem, which aims to recover a low-rank matrix from a small number of linear measurements. As long as the measurement ensemble satisfies a routine procedure to insert a regularizer that balances the two factors $[5]$, $[6]$, $[7]$: $g(X, Y) = \lambda \|X^\top X - Y^\top Y\|_F^2$ (4)

where $\lambda > 0$ is some regularization parameter, and apply gradient descent to the regularized loss function instead:

$$\min_{X,Y} f_{\text{reg}}(X, Y) := f(X, Y) + g(X, Y).$$

For a variety of important problems such as low-rank matrix sensing and matrix completion, it has been established that gradient descent over the regularized loss function, when properly initialized, achieves compelling statistical and computational guarantees.

A. Why balancing is needed in prior work?

To handle such asymmetric factorization, it is common to stack the two factors into one augmented factor $W_2 = \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times r}$ and then seek to estimate $W_2$ directly, by rewriting the loss function with respect to the lifted low-rank matrix:

$$W_2 W_2^\top = \begin{bmatrix} X_2 X_2^\top \\ Y_2 Y_2^\top \end{bmatrix} \in \mathbb{R}^{(n_1+n_2) \times (n_1+n_2)}.$$  

It is obvious that the loss function originally with respect to the asymmetric matrix $X_2 Y_2^\top$ only constrains the off-diagonal blocks of $W_2 W_2^\top$ and not the diagonal ones; correspondingly, the loss function is not (restricted) strongly convex with respect to the augmented factor, unless we appropriately regularize the diagonal blocks, which gives rise to the adoption of the regularization term in (4).

To understand a bit better why this regularization term (4) may help analysis, consider a toy example of factorizing a rank-one matrix $x_2 y_2^\top$, where $f(x, y)$ and $g(x, y)$ respectively are $f(x, y) = \frac{1}{2} \|x y^\top - x_2 y_2^\top\|_F^2$ and $g(x, y) = \frac{1}{2} (\|x\|_2^2 - \|y\|_2^2)^2$. Figure 1 illustrates the landscape of the unregularized loss function $f(x, y)$ and the regularized loss function $f_{\text{reg}}(x, y)$,
respectively, when the arguments are scalar-valued, i.e., \( n_1 = n_2 = 1 \). One can clearly appreciate the value of the regularizer: \( \bar{f}(x, y) \) becomes strongly convex in the local neighborhood around the global optimum \((1, 1)\). In contrast, the Hessian of the unregularized loss function remains rank deficient along the ambiguity set whenever \( xy = 1 \) (colored in red).

B. This paper: balancing-free procedure?

This goal of this paper is to understand the effectiveness of vanilla gradient descent (3) when initialized with balanced factors via the spectral method. Indeed, Figure 2 plots the normalized reconstruction error \( \|X \|_F^2 - \|M \|_F^2 \) for low-rank matrix completion with respect to the iteration count, using either a regularized loss function or an unregularized loss function when initialized by the spectral method. The two iterates converge in almost exactly the same trajectory, suggesting that gradient descent over the unregularized loss function converges almost in the same manner as its regularized counterpart, and perhaps is more natural to use in practice since it eliminates the tuning of regularization parameters.

C. Notations and organization of this paper

We use boldface lowercase (resp. uppercase) letters to represent vectors (resp. matrices). We denote by \( |x|_2 \) the \( l_2 \) norm of a vector \( x \), and \( X^\top, X^{-1}, \|X\| \) and \( \|X\|_F \) the transpose, the inverse, the spectral norm and the Frobenius norm of a matrix \( X \), respectively. Furthermore, we denote \( X^{-\top} = (X^{-1})^{-1} = (X^\top)^{-1} \) for an invertible matrix \( X \). The \( k \)th largest singular value of a matrix \( X \) is denoted by \( \sigma_k(X) \).

The inner product between two matrices \( X \) and \( Y \) is defined as \( \langle X, Y \rangle = Tr(Y^\top X) \), where \( Tr(\cdot) \) is the trace. Denote \( \mathcal{O}(r \times r) \) as the set of \( r \times r \) orthonormal matrices. In addition, we use \( c \) and \( C \) with different subscripts to represent positive numerical constants, whose values may change from line to line.

II. MAIN RESULTS

Let the object of interest \( M_2 \in \mathbb{R}^{n_1 \times n_2} \) be a rank-\( r \) matrix with the Singular Value Decomposition (SVD) given as

\[
M_2 = U_2 \Sigma_2 V_2^\top,
\]

where \( U_2 \in \mathbb{R}^{n_1 \times r} \), \( V_2 \in \mathbb{R}^{n_2 \times r} \) and \( \Sigma_2 \in \mathbb{R}^{r \times r} \). Without loss of generality, we denote the ground truth factors as

\[
X_2 = U_2 \Sigma_2^{1/2} \quad \text{and} \quad Y_2 = V_2 \Sigma_2^{1/2}. \tag{6}
\]

Let \( \sigma_{\max} := \sigma_1(M_2) \) and \( \sigma_{\min} := \sigma_r(M_2) \) be the largest and smallest nonzero singular value of \( M_2 \). The condition number of \( M_2 \) is defined as \( \kappa := \sigma_{\max}/\sigma_{\min} \).

Since the factors are identifiable up to invertible transforms since \( (X_2 P)(Y_2 P^{-\top})^\top = Z_2 Y_2^\top \) for any invertible matrix \( P \in \mathbb{R}^{r \times r} \), we measure the distance between two pairs of factors \( Z = (X, Y) \) and \( Z_2 = (X_2, Y_2) \) as:

\[
\text{dist}(Z, Z_2) = \min_{P \in \mathbb{R}^{r \times r}, \text{invertible}} \sqrt{\|XP - X_2\|_F^2 + \|YP^{-\top} - Y_2\|_F^2}. \tag{7}
\]

A. Low-rank matrix sensing

Suppose we are given a set of \( m \) measurements as follows

\[
y_i = \langle A_i, M_2 \rangle = \langle A_i, X_2 Y_2^\top \rangle, \quad i = 1, \ldots, m, \tag{8}
\]

where \( A_i \in \mathbb{R}^{n_1 \times n_2} \) is the \( i \)th sensing matrix, \( i = 1, \ldots, m \).

For convenience, we define \( A : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m \) as an affine
Algorithm 1 Gradient Descent with Spectral Initialization (unregularized Procrustes Flow)

**Input:** Measurements $y = \{y_i\}_{i=1}^m$, and sensing matrices $\{A_i\}_{i=1}^m$.

**Parameters:** Step size $\eta_t$, rank $r$, and number of iterations $T$.

**Initialization:** Initialize $X_0 = U\Sigma^{1/2}$ and $Y_0 = V\Sigma^{1/2}$, where $U\Sigma V^T$ is the rank-$r$ SVD of the surrogate matrix $K = \frac{1}{m} \sum_{i=1}^m y_i A_i$.

**Gradient loop:** For $t = 0 : 1 : T - 1$, do

$$X_{t+1} = X_t - \frac{\eta_t}{\|Y_0\|^2} \left[ \sum_{i=1}^m (A_i X_t Y_i^T - y_i) A_i Y_t \right];$$

(10a)

$$Y_{t+1} = Y_t - \frac{\eta_t}{\|X_0\|^2} \left[ \sum_{i=1}^m (A_i X_t Y_i^T - y_i) A_i X_t \right].$$

(10b)

**Output:** $X_T$ and $Y_T$.

Algorithm 1 describes the gradient descent algorithm initialized by the spectral method for minimizing (9). Compared to the Procrustes Flow (PF) algorithm [5], which minimizes the regularized loss function in (5), the new algorithm does not require the Procrustes Flow (PF) algorithm [5], which is on the order of \(\log(1/\epsilon)\) iterations, which is order-wise equivalent to the regularized PF algorithm in [5]. Comparing to [5], which requires \(\delta_{4r} \leq c\). Theorem 1 only requires a weaker assumption \(\delta_{2r} \leq c\). However, the basin of attraction allowed by Theorem 1 is smaller than that in [5], which is \(\min_{R \in O_r \times r} \|Z_0 R - Z_0\|_F \leq c_0 \sigma_{\min}(X_z)\).

We still need to find a good initialization that satisfies (11). In general, one could initialize with the balanced factors of the output after running multiple iterations of projected gradient descent (over the low-rank matrix), i.e.

$$M_{t+1} = \mathcal{P}_r \left( M_t - \frac{1}{m} \sum_{i=1}^m (A_i M_t - y_i) A_i \right),$$

where \(\mathcal{P}_r\) is the projection to the best rank-$r$ approximation. The spectral initialization specified in Algorithm 1 can be regarded as the output at the first iteration, initialized at zero $M_0 = 0$. Based on [10], [5], the iterates satisfy

$$\min_{R \in O_r \times r} \|Z_0 R - Z_0\|_F \leq c_2 (2\delta_{4r})^2 \frac{\|M_0\|_F}{\sigma_{\min}(X_z)}$$

for some constant $c_2$. Thus, to achieve the required initialization condition, if we use the spectral method specified in Algorithm 1, which corresponds to setting $\tau = 1$ in (12), we need

$$\delta_{4r} \leq c_2 \frac{1}{\kappa \sqrt{2}} \frac{\|M_0\|_F}{\sigma_{\min}(X_z)}.$$
To the best of our knowledge, the balancing regularization term (4) was first introduced in [5] to deal with non-square matrix factorization, and has become a standard approach to deal with asymmetric low-rank matrix estimation [6], [7], [8], [28], [29], [30]. A major benefit of adding the regularization term is to reduce the ambiguity set from invertible transforms to orthonormal transforms, so that the distance defined in (7) is minimized over \( P \in O^{r \times r} \). For the special rank-one matrix recovery problem, there are some evidence in the prior literature that a balancing regularization is not needed, for example, Ma et al. [23] established that vanilla gradient descent works for blind deconvolution at a near-optimal sample complexity with spectral initialization. In [31], the trajectory of gradient descent is studied for asymmetric matrix factorization with an infinitesimal and diminishing step size; in contrast, we consider the case when the step size is constant for low-rank matrix estimation with incomplete observations.

Finally, we remark that a similar regularization term (4) is also adopted when analyzing the optimization landscape of low-rank matrix estimation, e.g. [32], [33], [34], [35]. Without such a regularization term, the landscape of matrix factorization no longer possesses the intriguing property “all saddle points are strict saddle” and therefore one cannot invoke theory such as [36] to argue the global convergence of gradient descent using an unregularized loss function. Our work partially bridges this gap and suggests the benign behavior of gradient descent even in the absence of local strong convexity.

**IV. PROOF SKETCH OF THEOREM 1**

In this section, we provide a proof sketch of Theorem 1. We first discuss some basic properties of aligning two low-rank factors via invertible transforms, then prove a similar result for a warm-up case of low-rank matrix factorization, of which our problem of interest can be regarded as a perturbed version.

**A. Alignment via invertible transforms**

For \( Z = [X^T, Y^T]^T \) and \( Z_t = [X_t^T, Y_t^T]^T \), we define the optimal alignment matrix \( Q \) as

\[
Q := \text{argmin}_{P \in R^{r \times r}} \sqrt{\|XP - X_t\|_F^2 + \|YP^T - Y_t\|_F^2}.
\]

Furthermore, we call \( Z \) and \( Z_t \) are aligned if the corresponding optimal alignment matrix \( Q = I \). Throughout the paper, we assume the optimal alignment matrix between the \( t \)th iterate \( Z_t = [X_t^T, Y_t^T]^T \) and \( Z_0 \) is denoted as \( Q_t \). Below we provide some basic understandings of this alignment operation.

**Lemma 1:** Given two matrices \( Z \) and \( Z_t \), and their optimal alignment matrix \( Q \), we have

\[
\tilde{X}^T(\tilde{X} - X_t) = (\tilde{Y} - Y_t)^T \tilde{Y},
\]

where \( \tilde{X} = XQ \) and \( \tilde{Y} = YQ^{-T} \) are the matrices after the alignment.

\[1\] It is guaranteed with high probability that the minimum is attained for \( Z_t \).

**Lemma 2:** Let \( Q \) be the optimal alignment matrix between \( Z \) and \( Z_t \). Suppose there exists a matrix \( P \) with \( 1/2 \leq \sigma_{\min}(P) \leq \sigma_{\max}(P) \leq 3/2 \) such that

\[
\max \left\{ \|XP - X_t\|_F, \|YP^T - Y_t\|_F \right\} \leq \delta \leq \frac{1}{4}\sigma_{\min}(X_t).
\]

Then one has

\[
\|P - Q\| \leq \|P - Q\|_F \leq \frac{10\delta}{\sigma_{\min}(X_t)}.
\]

Both lemmas provide basic understandings on the solution of solving the alignment problem with invertible transformations, which can be regarded as a generalization of the classical orthogonal Procrustes problem which only considers orthonormal transforms. Clearly, this generalized problem is much more challenging and our work provides some first understandings into it, to the best of our knowledge. These lemmas provide the basis for the subsequent analyses.

**B. A warm-up: low-rank matrix factorization**

We consider the following minimization problem

\[
f_{MF}(X, Y) = \frac{1}{2} \|XY^T - Z_t\|_F^2, \tag{14}
\]

where \( X \in \mathbb{R}^{n_1 \times r} \) and \( Y \in \mathbb{R}^{n_2 \times r} \). The gradient descent updates with an initialization \((X_0, Y_0)\) can be written as

\[
X_{t+1} = X_t - \eta \sigma_{\max}\nabla_X f_{MF}(X_t, Y_t)
= X_t - \frac{\eta}{\sigma_{\max}} (X_tY_t^T - M_t)Y_t;
\]

\[
Y_{t+1} = Y_t - \eta \sigma_{\max}\nabla_Y f_{MF}(X_t, Y_t)
= Y_t - \frac{\eta}{\sigma_{\max}} (X_tY_t^T - M_t)^T X_t.
\]

We have the following theorem regarding the performance of (15), which parallels with Theorem 1.

**Theorem 2:** Let \( Z_0 \) be an initialization which satisfies

\[
\min_{R \in O^{r \times r}} \|Z_0R - Z_t\|_F \leq c_0 \frac{1}{\kappa^{3/2}} \sigma_{\min}(X_t),
\]

for some small enough constant \( c_0 \). There exists some \( c_1 \) such that as long as \( \eta \leq c_1 \), the iterates of GD satisfy

\[
\text{dist}(Z_t, Z_0) \leq \left(1 - \frac{\eta}{2\kappa}\right) \text{dist}(Z_0, Z_t).
\]

**C. Analysis for matrix sensing**

We now extend the technique used in the proof of Theorem 2 to the matrix sensing case by leveraging the RIP. Suppose that the initialization \( Z_0 \) satisfies (11). By a similar argument as in [5], it is sufficient to consider the following update rule:

\[
X_{t+1} = X_t - \frac{\eta}{\sigma_{\max}} [A^*A(X_tY_t^T - M_t)] Y_t;
\]

\[
Y_{t+1} = Y_t - \frac{\eta}{\sigma_{\max}} [A^*A(X_tY_t^T - M_t)^T X_t].
\]

Compared with (15), the update rule for matrix sensing differs by the operation of \( A^*A \) when forming the gradient. Therefore, we expect the GD has similar behaviors as earlier as long as
A behaves as a near isometry on low-rank matrices. This can be supplied by the following consequence of the RIP.

**Lemma 3:** Suppose $A$ satisfies $2r$-RIP with constant $\delta_{2r}$. Then, for all matrices $M_1$ and $M_2$ of rank at most $r$, we have
\[
|\langle A(M_1), A(M_2) \rangle - \langle M_1, M_2 \rangle| \leq \delta_{2r} \|M_1\|_F \|M_2\|_F.
\]

**V. Conclusions**

This paper establishes the local linear convergence of gradient descent for rectangular low-rank matrix sensing without explicit regularization of factor balancedness under the standard RIP assumption, as long as a balanced initialization is provided in the basin of attraction, which can be found by the spectral method. Different from previous work, we analyzed a new error metric that takes into account the ambiguity due to invertible transforms, and showed that it contracts linearly even without local restricted strong convexity. We believe our technique can be used for other low-rank matrix estimation problems. To conclude, we outline a few exciting future research directions.

- **Low-rank matrix completion.** We believe it is possible to extend our analysis to study rectangular matrix completion without regularization, by combining the leave-one-out technique in [23], [30] to carefully bound the incoherence of the iterates for both factors even without explicit balancing.

- **Improving dependence on $\kappa$ and $r$.** The current paper does not try to optimize the dependence with respect to $\kappa$ and $r$ in terms of sample complexity and the size of the basin of attraction, which are slightly worse than their regularized counterparts. A finer analysis will likely lead to better dependencies, which we leave to the future work.

**References**

[1] E. Candès and T. Tao, “The power of convex relaxation: Near-optimal matrix completion,” *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053–2080, May 2010.

[2] Y. Chen and Y. Chi, “Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization,” *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 14 – 31, 2018.

[3] M. A. Davenport and J. Romberg, “An overview of low-rank matrix recovery from incomplete observations,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 608–622, 2016.

[4] S. Burer and R. Monteiro, “A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization,” *Mathematical Programming*, vol. 95, no. 2, pp. 329–357, 2003.

[5] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, “Low-rank solutions of linear matrix equations via proxies flow,” in *International Conference Machine Learning*, 2016, pp. 964–973.

[6] Q. Zheng and J. Lafferty, “Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent,” *arXiv preprint arXiv:1605.07051*, 2016.

[7] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, “Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably,” *SIAM Journal on Imaging Sciences*, vol. 11, no. 4, pp. 2165–2204, 2018.

[8] X. Yi, D. Park, Y. Chen, and C. Caramanis, “Fast algorithms for robust PCA via gradient descent,” in *Advances in neural information processing systems*, 2016, pp. 4152–4160.

[9] B. Recht, M. Fazel, and P. A. Parrilo, “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[10] S. Oymak, B. Recht, and M. Soltanolkotabi, “Sharp time–data tradeoffs for linear inverse problems,” *IEEE Transactions on Information Theory*, vol. 64, no. 6, pp. 4129–4158, 2018.