Low-rank Solutions of Linear Matrix Equations via Procrustes Flow

Stephen Tu, Ross Boczar, Max Simchowitz {STEPHENT,BOCZAR,MSIMCHOW}@BERKELEY.EDU
EECS Department, UC Berkeley, Berkeley, CA.

Mahdi Soltanolkotabi SOLTANOL@USC.EDU
Ming Hsieh Department of Electrical Engineering, USC, Los Angeles, CA.

Benjamin Recht BRECHT@BERKELEY.EDU
EECS Department, UC Berkeley, Berkeley, CA.

Abstract

In this paper we study the problem of recovering a low-rank matrix from linear measurements. Our algorithm, which we call *Procrustes Flow*, starts from an initial estimate obtained by a thresholding scheme followed by gradient descent on a non-convex objective. We show that as long as the measurements obey a standard restricted isometry property, our algorithm converges to the unknown matrix at a geometric rate. In the case of Gaussian measurements, such convergence occurs for a $n_1 \times n_2$ matrix of rank r when the number of measurements exceeds a constant times $(n_1 + n_2)r$.

1. Introduction

Low rank models are ubiquitous in machine learning, and over a decade of research has been dedicated to determining when such models can be efficiently recovered from partial information (Fazel, 2002; Rennie & Srebro, 2005; Candès & Recht, 2009). See (Davenport & Romberg, 2016) for an extended survey on this topic. The simplest such recovery problem concerns how can we can find a low-rank matrix obeying a set of linear equations? What is the computational complexity of such an algorithm? More specifically, we are interested in solving problems of the form

$$\min_{\boldsymbol{M}\in\mathbb{R}^{n_1\times n_2}} \operatorname{rank}(\boldsymbol{M}) \quad \text{s.t.} \quad \mathcal{A}(\boldsymbol{M}) = \boldsymbol{b}\,, \tag{1.1}$$

where $\mathcal{A} : \mathbb{R}^{n_1\times n_2} \longrightarrow \mathbb{R}^m$ is a known affine transformation that maps matrices to vectors. More specifically, the k-th entry of $\mathcal{A}(\boldsymbol{X})$ is $\langle \boldsymbol{A}_k, \boldsymbol{X}\rangle := \operatorname{Tr}(\boldsymbol{A}_k^\top \boldsymbol{X})$, where each $\boldsymbol{A}_k \in \mathbb{R}^{n_1\times n_2}$.

Since the early seventies, a popular heuristic for solving such problems has been to replace $\boldsymbol{M}$ with a low-rank factorization $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{V}^\top$ and solve matrix bilinear equations of the form

$$\operatorname*{find}_{\boldsymbol{U}\in\mathbb{R}^{n_1\times r}, \boldsymbol{V}\in\mathbb{R}^{n_2\times r}} \quad \text{s.t.} \quad \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^\top) = \boldsymbol{b}, \tag{1.2}$$

via a local search heuristic (Ruhe, 1974). Many researchers have demonstrated that such heuristics work well in practice for a variety of problems (Rennie & Srebro, 2005; Funk, 2006; Lee et al., 2010; Recht & Ré, 2013). However, these procedures lack strong guarantees associated with convex programming heuristics for solving (1.1).

In this paper we show that a local search heuristic solves (1.2) under standard restricted isometry assumptions on the linear map $\mathcal{A}$. For standard ensembles of equality constraints, we demonstrate that $\boldsymbol{M}$ can be estimated by such heuristics as long as we have $\Omega((n_1 + n_2)r)$ equations.[1] This is merely a constant factor more than the number of parameters needed to specify a $n_1 \times n_2$ rank r matrix. Specialized to a random Gaussian model and positive semidefinite matrices, our work improves upon recent independent work by Zheng and Lafferty (Zheng & Lafferty, 2015).

2. Algorithms

In this paper we study a local search heuristic for solving matrix bilinear equations of the form (1.2) which consists of two components: (1) a careful initialization obtained by a projected gradient scheme on $n_1 \times n_2$ matrices, and (2) a series of successive refinements of this initial solution via a

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48. Copyright 2016 by the author(s).

[1] Here and throughout we use $f(x) = \Omega(g(x))$ if there is a positive constant C such that $f(x) \geq Cg(x)$ for all x sufficiently large.

gradient descent scheme. This algorithm is a natural extension of the Wirtinger Flow algorithm developed in (Candès et al., 2015) for solving vector quadratic equations. Following (Candès et al., 2015), we shall refer to the combination of these two steps as the Procrustes Flow (PF) algorithm. We shall describe two variants of our algorithm based on whether the sought after solution $\boldsymbol{M}$ is positive semidefinite (PSD) or not. The former is detailed in Algorithm 1, and the latter in Algorithm 2.

The initialization phase of both variants is rather similar and is described in Section 2.1. The successive refinement phase is explained in Section 2.2 for PSD matrices and in Section 2.3 for arbitrary matrices. Throughout this paper when describing the PSD case, we assume the size of the matrix is $\boldsymbol{M}$ is $n \times n$, i.e. $n_1 = n_2 = n$.

2.1. Initialization via low-rank projected gradients

In the initial phase of our algorithm we start from $\widetilde{\boldsymbol{M}}_0 = \boldsymbol{0}_{n_1 \times n_2}$ and apply successive updates of the form

$$\widetilde{\boldsymbol{M}}_{\tau+1} = \mathcal{P}_r\left(\widetilde{\boldsymbol{M}}_\tau - \alpha_{\tau+1}\sum_{k=1}^{m}\left(\langle \boldsymbol{A}_k, \widetilde{\boldsymbol{M}}_\tau\rangle - b_k\right)\boldsymbol{A}_k\right), \tag{2.1}$$

on rank r matrices of size $n_1 \times n_2$. Here, $\mathcal{P}_r$ denotes projection onto either rank-r matrices or rank-r PSD matrices, both of which can be computed efficiently via Lanczos methods. We run (2.1) for T_0 iterations and use the resulting matrix $\boldsymbol{M}_{T_0}$ for initialization purposes. In the PSD case, we set our initialization to an $n \times r$ matrix $\boldsymbol{U}_0$ obeying $\widetilde{\boldsymbol{M}}_{T_0} = \boldsymbol{U}_0\boldsymbol{U}_0^\mathsf{T}$. In the more general case of rectangular matrices we need to use two factors. Let $\widetilde{\boldsymbol{M}}_{T_0} = \boldsymbol{C}_{T_0}\boldsymbol{\Sigma}_{T_0}\boldsymbol{D}_{T_0}^\mathsf{T}$ be the Singular Value Decomposition (SVD) of $\widetilde{\boldsymbol{M}}_{T_0}$. We initialize our algorithm in the rectangular case by setting $\boldsymbol{U}_0 = \boldsymbol{C}_{T_0}\boldsymbol{\Sigma}_{T_0}^{1/2}$ and $\boldsymbol{V}_0 = \boldsymbol{D}_{T_0}\boldsymbol{\Sigma}_{T_0}^{1/2}$.

Updates of the form (2.1) have a long history in compressed sensing/matrix sensing literature (see e.g. (Tropp & Gilbert, 2007; Garg & Khandekar, 2009; Needell & Tropp, 2009; Needell & Vershynin, 2009; Blumensath & Davies, 2009; Meka et al., 2009; Cai et al., 2010)). Furthermore, using the first step of the update (2.1) for the purposes of initialization has also been proposed in previous work (see e.g. (Achlioptas & McSherry, 2007; Keshavan et al., 2010; Jain et al., 2013)).

2.2. Successive refinement via gradient descent – positive semidefinite case

We first focus on the PSD case. As mentioned earlier, we are interested in finding a matrix $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ obeying matrix quadratic equations of the form $\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\mathsf{T}) = \boldsymbol{b}$. We wish to refine our initial estimate by minimizing the non-convex function

$$f(\boldsymbol{U}) := \frac{1}{4}\left\|\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\mathsf{T}) - \boldsymbol{b}\right\|_{\ell_2}^2, \tag{2.2}$$

over $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, which minimizes the misfit in our quadratic equations via the square loss. To solve (2.2), starting from our initial estimate $\boldsymbol{U}_0 \in \mathbb{R}^{n\times r}$ we apply the successive updates

$$\begin{aligned}\boldsymbol{U}_{\tau+1} &:= \boldsymbol{U}_\tau - \frac{\mu_{\tau+1}}{\|\boldsymbol{U}_0\|^2}\nabla f(\boldsymbol{U}_\tau)\\ &= \boldsymbol{U}_\tau - \frac{\mu_{\tau+1}}{\|\boldsymbol{U}_0\|^2}\left(\sum_{k=1}^{m}(\langle \boldsymbol{A}_k, \boldsymbol{U}_\tau\boldsymbol{U}_\tau^\mathsf{T}\rangle - b_k)\boldsymbol{A}_k\boldsymbol{U}_\tau\right).\end{aligned} \tag{2.3}$$

Here and throughout, for a matrix $\boldsymbol{X}$, $\sigma_\ell(\boldsymbol{X})$ denotes the ℓ-th largest singular value of $\boldsymbol{X}$, and $\|\boldsymbol{X}\| = \sigma_1(\boldsymbol{X})$ is the operator norm. We note that the update (2.3) is essentially gradient descent with a carefully chosen step size.

Algorithm 1 Procrustes Flow (PF)

```
Require: $\{\boldsymbol{A}_k\}_{k=1}^m, \{b_k\}_{k=1}^m, \{\alpha_\tau\}_{\tau=1}^\infty, \{\mu_\tau\}_{\tau=1}^\infty, T_0 \in \mathbb{N}$.
  // Initialization phase.
  $\widetilde{\boldsymbol{M}}_0 := \boldsymbol{0}_{n\times n}$.
  for $\tau = 0, 1, ..., T_0 - 1$ do
    // Projection onto rank $r$ PSD matrices.
    $\widetilde{\boldsymbol{M}}_{\tau+1} \leftarrow \mathcal{P}_r(\widetilde{\boldsymbol{M}}_\tau - \alpha_{\tau+1}\sum_{k=1}^m(\langle \boldsymbol{A}_k, \widetilde{\boldsymbol{M}}_\tau\rangle - b_k)\boldsymbol{A}_k)$.
  end for
  // SVD of $\widetilde{\boldsymbol{M}}_{T_0}$, with $\boldsymbol{Q} \in \mathbb{R}^{n\times r}, \boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$.
  $\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{Q}^\mathsf{T} := \widetilde{\boldsymbol{M}}_{T_0}$.
  $\boldsymbol{U}_0 := \boldsymbol{Q}\boldsymbol{\Sigma}^{1/2}$.
  // Gradient descent phase.
  repeat
    $\boldsymbol{U}_{\tau+1} \leftarrow \boldsymbol{U}_\tau - \frac{\mu_{\tau+1}}{\|\boldsymbol{U}_0\|^2}\nabla f(\boldsymbol{U}_\tau)$.
  until convergence
```

2.3. Successive refinement via gradient descent – general case

We now consider the general case. Here, we are interested in finding matrices $\boldsymbol{U} \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n_2\times r}$ obeying matrix quadratic equations of the form $\boldsymbol{b} = \mathcal{A}(\boldsymbol{U}\boldsymbol{V}^\mathsf{T})$. In this case, we refine our initial estimate by minimizing the non-convex function

$$g(\boldsymbol{U}, \boldsymbol{V}) := \frac{1}{2}\left\|\mathcal{A}(\boldsymbol{U}\boldsymbol{V}^\mathsf{T}) - \boldsymbol{b}\right\|_{\ell_2}^2 + \frac{1}{16}\left\|\boldsymbol{U}^\mathsf{T}\boldsymbol{U} - \boldsymbol{V}^\mathsf{T}\boldsymbol{V}\right\|_F^2. \tag{2.4}$$

over $\boldsymbol{U} \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n_2\times r}$. Note that this is similar to (2.2) but adds a regularizer to measure mismatch between $\boldsymbol{U}$ and $\boldsymbol{V}$. Given a factorization $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{V}^\mathsf{T}$, for any invertible $r \times r$ matrix $\boldsymbol{P}$, $\boldsymbol{U}\boldsymbol{P}$ and $\boldsymbol{V}\boldsymbol{P}^{-\mathsf{T}}$ is also a valid factorization. The purpose of the second term in (2.4) is to account for this redundancy and put the two factors on "equal footing". To solve (2.4), starting from our initial estimates $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$ we apply the successive updates

$$\boldsymbol{U}_{\tau+1} := \boldsymbol{U}_\tau - \frac{\mu_{\tau+1}}{\|\boldsymbol{U}_0\|^2}\nabla_{\boldsymbol{U}} g(\boldsymbol{U}_\tau, \boldsymbol{V}_\tau) \qquad (2.5)$$

$$\boldsymbol{V}_{\tau+1} := \boldsymbol{V}_\tau - \frac{\mu_{\tau+1}}{\|\boldsymbol{V}_0\|^2}\nabla_{\boldsymbol{V}} g(\boldsymbol{U}_\tau, \boldsymbol{V}_\tau) \qquad (2.6)$$

where $\nabla_{\boldsymbol{U}} g(\boldsymbol{U}_\tau, \boldsymbol{V}_\tau)$ is equal to

$$\sum_{k=1}^{m}(\langle \boldsymbol{A}_k, \boldsymbol{U}_\tau \boldsymbol{V}_\tau^\top\rangle - b_k)\boldsymbol{A}_k \boldsymbol{V}_\tau + \frac{1}{4}\boldsymbol{U}_\tau(\boldsymbol{U}_\tau^\top \boldsymbol{U}_\tau - \boldsymbol{V}_\tau^\top \boldsymbol{V}_\tau)$$

and $\nabla_{\boldsymbol{V}} g(\boldsymbol{U}_\tau, \boldsymbol{V}_\tau)$ is equal to

$$\sum_{k=1}^{m}(\langle \boldsymbol{A}_k, \boldsymbol{U}_\tau \boldsymbol{V}_\tau^\top\rangle - b_k)\boldsymbol{A}_k^\top \boldsymbol{U}_\tau + \frac{1}{4}\boldsymbol{V}_\tau(\boldsymbol{V}_\tau^\top \boldsymbol{V}_\tau - \boldsymbol{U}_\tau^\top \boldsymbol{U}_\tau)\,.$$

Again, (2.5) and (2.6) are essentially gradient descent with a carefully chosen step size.

Algorithm 2 Rectangular Procrustes Flow (RPF)

```
Require: $\{\boldsymbol{A}_k\}_{k=1}^m, \{b_k\}_{k=1}^m, \{\alpha_\tau\}_{\tau=1}^\infty, \{\mu_\tau\}_{\tau=1}^\infty, T_0 \in \mathbb{N}$.
  // Initialization phase.
  $\widetilde{\boldsymbol{M}}_0 := \boldsymbol{0}_{n_1 \times n_2}$.
  for $\tau = 0, 1, ..., T_0 - 1$ do
    // Projection onto rank $r$ matrices.
    $\widetilde{\boldsymbol{M}}_{\tau+1} \leftarrow \mathcal{P}_r(\widetilde{\boldsymbol{M}}_\tau - \alpha_{\tau+1}\sum_{k=1}^m(\langle \boldsymbol{A}_k, \widetilde{\boldsymbol{M}}_\tau\rangle - b_k)\boldsymbol{A}_k)$.
  end for
  // SVD of $\widetilde{\boldsymbol{M}}_{T_0}$, with
  // $\boldsymbol{C} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}, \boldsymbol{D} \in \mathbb{R}^{n_2 \times r}$ .
  $\boldsymbol{C\Sigma D}^\top := \widetilde{\boldsymbol{M}}_{T_0}$.
  $\boldsymbol{U}_0 := \boldsymbol{C\Sigma}^{1/2}$.
  $\boldsymbol{V}_0 := \boldsymbol{D\Sigma}^{1/2}$.
  // Gradient descent phase.
  repeat
    $\boldsymbol{U}_{\tau+1} \leftarrow \boldsymbol{U}_\tau - \frac{\mu_{\tau+1}}{\|\boldsymbol{U}_0\|^2}\nabla_{\boldsymbol{U}} g(\boldsymbol{U}_\tau, \boldsymbol{V}_\tau)$.
    $\boldsymbol{V}_{\tau+1} \leftarrow \boldsymbol{V}_\tau - \frac{\mu_{\tau+1}}{\|\boldsymbol{V}_0\|^2}\nabla_{\boldsymbol{V}} g(\boldsymbol{U}_\tau, \boldsymbol{V}_\tau)$.
  until convergence
```

3. Main Results

For our theoretical results we shall focus on affine maps $\mathcal{A}$ which obey the matrix Restricted Isometry Property (RIP).

Definition 3.1 (Restricted Isometry Property (RIP) (Candès & Tao, 2005; Recht et al., 2010)). *The map $\mathcal{A}$ satisfies r-RIP with constant δ_r, if*

$$(1 - \delta_r)\,\|\boldsymbol{X}\|_F^2 \le \|\mathcal{A}(\boldsymbol{X})\|_{\ell_2}^2 \le (1 + \delta_r)\,\|\boldsymbol{X}\|_F^2\,,$$

holds for all matrices $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ of rank at most r.

As mentioned earlier it is not possible to recover the factors $\boldsymbol{U}$ and $\boldsymbol{V}$ in (1.2) exactly. For example, in the PSD case it is only possible to recover $\boldsymbol{U}$ up to a certain rotational factor as if $\boldsymbol{U}$ obeys (3.5), then so does any matrix $\boldsymbol{UR}$ with $\boldsymbol{R} \in \mathbb{R}^{r \times r}$ an orthonormal matrix satisfying $\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}_r$. This naturally leads to defining the distance between two matrices $\boldsymbol{U}, \boldsymbol{X} \in \mathbb{R}^{n \times r}$ as

$$\operatorname{dist}(\boldsymbol{U}, \boldsymbol{X}) := \min_{\boldsymbol{R} \in \mathbb{R}^{r \times r}:\, \boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}_r} \|\boldsymbol{U} - \boldsymbol{XR}\|_F\,. \qquad (3.1)$$

We note that this distance is the solution to the classic *orthogonal Procrustes problem* (hence the name of the algorithm). It is known that the optimal rotation matrix $\boldsymbol{R}$ minimizing $\|\boldsymbol{U} - \boldsymbol{XR}\|_F$ is equal to $\boldsymbol{R} = \boldsymbol{AB}^\top$, where $\boldsymbol{A\Sigma B}^\top$ is the singular value decomposition (SVD) of $\boldsymbol{X}^\top \boldsymbol{U}$. We now have all of the elements in place to state our main results.

3.1. Quadratic measurements

When the low-rank matrix $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ is PSD we are interested in finding a matrix $\boldsymbol{U} \in \mathbb{R}^{n \times r}$ obeying quadratic equations of the form

$$\mathcal{A}(\boldsymbol{UU}^T) = \boldsymbol{b}, \qquad (3.2)$$

where we assume $\boldsymbol{b} = \mathcal{A}(\boldsymbol{M})$ for a planted rank-r solution $\boldsymbol{M} = \boldsymbol{XX}^\top \in \mathbb{R}^{n \times n}$ with $\boldsymbol{X} \in \mathbb{R}^{n \times r}$. We wish to recover $\boldsymbol{X}$. This is of course only possible up to a certain rotational factor as if $\boldsymbol{U}$ obeys (3.5), then so does any matrix $\boldsymbol{UR}$ with $\boldsymbol{R} \in \mathbb{R}^{r \times r}$ an orthonormal matrix satisfying $\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}_r$. Our first theorem shows that Procrustes Flow indeed recovers $\boldsymbol{X}$ up to this ambiguity factor.

Theorem 3.2. *Let $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ be an arbitrary rank-r symmetric positive semidefinite matrix with singular values $\sigma_1(\boldsymbol{M}) \ge \sigma_2(\boldsymbol{M}) \ge ... \ge \sigma_r(\boldsymbol{M}) > 0$ and condition number $\kappa = \sigma_1(\boldsymbol{M})/\sigma_r(\boldsymbol{M})$. Assume $\boldsymbol{M} = \boldsymbol{XX}^\top$ for some $\boldsymbol{X} \in \mathbb{R}^{n \times r}$ and let $\boldsymbol{b} = \mathcal{A}(\boldsymbol{M}) \in \mathbb{R}^m$ be m linear measurements. Furthermore, assume the mapping $\mathcal{A}$ obeys rank-$6r$ RIP with RIP constant $\delta_{6r} \le 1/10$. Also let $\alpha_\tau = 1$ for all $\tau = 1, 2, \ldots$. Then, using $T_0 \ge \log(\sqrt{r}\kappa) + 2$ iterations of the initialization phase of Procrustes Flow as stated in Algorithm 1 yields a solution $\boldsymbol{U}_0$ obeying*

$$\operatorname{dist}(\boldsymbol{U}_0, \boldsymbol{X}) \le \frac{1}{4}\sigma_r(\boldsymbol{X}). \qquad (3.3)$$

Furthermore, take a constant step size $\mu_\tau = \mu$ for all $\tau = 1, 2, \ldots$, with $\mu \le 36/425$. Then, starting from any initial solution obeying (3.3), the τ-th iterate of Algorithm 1 satisfies

$$\operatorname{dist}(\boldsymbol{U}_\tau, \boldsymbol{X}) \le \frac{1}{4}\left(1 - \frac{8}{25}\frac{\mu}{\kappa}\right)^{\frac{\tau}{2}}\sigma_r(\boldsymbol{X}). \qquad (3.4)$$

3.2. Bilinear measurements

In the more general case when the low-rank matrix $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ is rectangular we are interested in finding matrices $\boldsymbol{U} \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{V} \in \mathbb{R}^{n_2 \times r}$ obeying bilinear equations of the form

$$\mathcal{A}(\boldsymbol{U}\boldsymbol{V}^{\mathsf{T}}) = \boldsymbol{b}, \tag{3.5}$$

where we assume $\boldsymbol{b} = \mathcal{A}(\boldsymbol{M})$ for a planted rank-r solution $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ with $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{Y}^{\mathsf{T}}$ where $\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}$. Again we wish to recover the factors $\boldsymbol{X}$ and $\boldsymbol{Y}$. The next theorem shows that we can also provide a guarantee similar to that of Theorem 3.2 for this more general rectangular case.

Theorem 3.3. *Let $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ be an arbitrary rank-r matrix with singular values $\sigma_1(\boldsymbol{M}) \geq \sigma_2(\boldsymbol{M}) \geq ... \geq \sigma_r(\boldsymbol{M}) > 0$ and condition number $\kappa = \sigma_1(\boldsymbol{M})/\sigma_r(\boldsymbol{M})$. Let $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^{\mathsf{T}}$ be the SVD of $\boldsymbol{M}$ and define $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{\Sigma}^{1/2} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{\Sigma}^{1/2} \in \mathbb{R}^{n_2 \times r}$. Also, let $\boldsymbol{b} = \mathcal{A}(\boldsymbol{M}) \in \mathbb{R}^m$ be m linear measurements where the mapping $\mathcal{A}$ obeys rank-$6r$ RIP with RIP constant $\delta_{6r} \leq 1/25$. Also let $\alpha_\tau = 1$ for all $\tau = 1, 2, \ldots$. Then, using $T_0 \geq 3\log(\sqrt{r}\kappa) + 5$ iterations of the initialization phase of Procrustes Flow as stated in Algorithm 2 yields a solution $\boldsymbol{U}_0, \boldsymbol{V}_0$ obeying*

$$\text{dist}\left(\begin{bmatrix}\boldsymbol{U}_0\\ \boldsymbol{V}_0\end{bmatrix}, \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix}\right) \leq \frac{1}{4}\sigma_r(\boldsymbol{X}). \tag{3.6}$$

Furthermore, take a constant step size $\mu_\tau = \mu$ for all $\tau = 1, 2, \ldots$ and assume $\mu \leq 2/187$. Then, starting from any initial solution obeying (3.6), the τ-th iterate of Algorithm 2 satisfies

$$\text{dist}\left(\begin{bmatrix}\boldsymbol{U}_\tau\\ \boldsymbol{V}_\tau\end{bmatrix}, \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix}\right) \leq \frac{1}{4}\left(1 - \frac{4}{25}\frac{\mu}{\kappa}\right)^{\frac{\tau}{2}} \sigma_r(\boldsymbol{X}). \tag{3.7}$$

The above theorem shows that Procrustes Flow algorithm achieves a good initialization under the RIP assumptions on the mapping $\mathcal{A}$. Also, starting from any sufficiently accurate initialization the algorithm exhibits geometric convergence to the unknown matrix $\boldsymbol{M}$. We note that in the above result we have not attempted to optimize the constants. Furthermore, there is a natural tradeoff involved between the upper bound on the RIP constant, the radius in which PF is contractive (3.6), and its rate of convergence (3.7). In particular, as it will become clear in the proofs one can increase the radius in which PF is contractive (increase the constant $1/4$ in (3.6)) and the rate of convergence (increase the constant $4/25$ in (3.7)) by assuming a smaller upper bound on the RIP constant.

The most common measurement ensemble which satisfies the isotropy and RIP assumptions is the Gaussian ensemble here each matrix $\boldsymbol{A}_k$ has i.i.d. $\mathcal{N}(0, 1/m)$ entries.[2] For this ensemble to achieve a RIP constant of δ_r, we require at least $m = \Omega(\frac{1}{\delta_r^2} nr)$ measurements. Using Equation (3.7) together with a simple calculation, we can conclude that for $\boldsymbol{M}_\tau = \boldsymbol{U}_\tau \boldsymbol{V}_\tau^{\mathsf{T}}$, we have

$$\begin{aligned}
\|\boldsymbol{M}_\tau - \boldsymbol{M}\|_F &\leq \frac{9}{4}\sqrt{\sigma_1(\boldsymbol{M})} \cdot \text{dist}\left(\begin{bmatrix}\boldsymbol{U}_\tau\\ \boldsymbol{V}_\tau\end{bmatrix}, \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix}\right)\\
&\leq \frac{9}{16}\sqrt{\sigma_1(\boldsymbol{M})\sigma_r(\boldsymbol{M})}\left(1 - \frac{4}{25}\frac{\mu}{\kappa}\right)^{\frac{\tau}{2}}\\
&\leq \frac{9}{16}\|\boldsymbol{M}\|_F\left(1 - \frac{4}{25}\frac{\mu}{\kappa}\right)^{\frac{\tau}{2}}. \qquad (3.8)
\end{aligned}$$

Thus, applying Theorem 3.3 to this measurement ensemble, we conclude that the Procrustes Flow algorithm yields a solution with relative error ($\|\boldsymbol{M}_\tau - \boldsymbol{M}\|_F / \|\boldsymbol{M}\|_F \leq \epsilon$) in $\mathcal{O}(\kappa \log(1/\epsilon))$ iterations using only $\Omega(nr)$ measurements. We would like to note that if more measurements are available it is not necessary to use multiple projected gradient updates in the initialization phase. In particular, for the Gaussian model if $m = \Omega(nr^2\kappa^2)$, then (3.3) will hold after the first iteration ($T_0 = 1$).

How to verify the initialization is complete. Theorems 3.2 and 3.3 require that $T_0 = \Omega(\log(\sqrt{r}\kappa))$, but κ is a property of $\boldsymbol{M}$ and is hence unknown. However, under the same hypotheses regarding the RIP constant in Theorems 3.2 and 3.3, we can use each iterate of initialization to test whether or not we have entered the radius of convergence. The following lemma establishes a sufficient condition we can check using only information from $\widetilde{\boldsymbol{M}}_\tau$. We establish this result only in the symmetric case– the extension to the general case is straightforward.

Lemma 3.4. *Assume the RIP constant of $\mathcal{A}$ satisfies $\delta_{2r} \leq 1/10$. Let $\widetilde{\boldsymbol{M}}_\tau$ denote the τ-th step of the initialization phase in Algorithm 1, and let $\boldsymbol{U}_0 \in \mathbb{R}^{n \times r}$ be the such that $\widetilde{\boldsymbol{M}}_\tau = \boldsymbol{U}_0\boldsymbol{U}_0^{\mathsf{T}}$. Define*

$$e_\tau := \left\|\mathcal{A}(\widetilde{\boldsymbol{M}}_\tau) - \boldsymbol{b}\right\|_{\ell_2} = \left\|\mathcal{A}(\widetilde{\boldsymbol{M}}_\tau - \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}})\right\|_{\ell_2}.$$

Then, if $e_\tau \leq \frac{3}{20}\sigma_r(\widetilde{\boldsymbol{M}}_\tau)$, we have that $\text{dist}(\boldsymbol{U}_0, \boldsymbol{X}) \leq \frac{1}{4}\sigma_r(\boldsymbol{X})$.

One might consider using solely the projected gradient updates (i.e. set $T_0 = \infty$) as in previous approaches (Tropp & Gilbert, 2007; Garg & Khandekar, 2009; Needell & Tropp, 2009; Needell & Vershynin, 2009; Blumensath & Davies, 2009; Meka et al., 2009; Cai et al., 2010). We note that the projected gradient updates in the initialization phase

[2] We note that in the PSD case the so called *spiked Gaussian ensemble* would be the right equivalent. In this case each symmetric matrix $\boldsymbol{A}_k$ has $\mathcal{N}(0, 1/m)$ entries on the diagonal and $\mathcal{N}(0, 1/2m)$ entries elsewhere.

require computing the first r singular vectors of a matrix whereas the gradient updates do not require any singular vector computations. Such singular computations may be prohibitive compared to the gradient updates, especially when n_1 or n_2 is large and for ensembles where matrix-vector multiplication is fast. We would like to emphasize, however, that for small n_1, n_2 and dense matrices using projected gradient updates may be more efficient. Our scheme is a natural interpolation: one could only do projected gradient steps, or one could do one projected gradient step. Here we argue that very few projected gradients provide sufficient initialization such that gradient descent converges geometrically.

4. Related work

There is a vast literature dedicated to low-rank matrix recovery/sensing and semidefinite programming. We shall only focus on the papers most related to our framework.

Recht, Fazel, and Parrilo were the first to study low-rank solutions of linear matrix equations under RIP assumptions (Recht et al., 2010). They showed that if the rank-r RIP constant of $\mathcal{A}$ is less than a fixed numerical constant, then the matrix with minimum trace satisfying the equality constraints coincided with the minimum rank solution. In particular, for the Gaussian ensemble the required number of measurements is $\Omega(nr)$ (Candès & Plan, 2011). Subsequently, a series of papers (Candès & Recht, 2009; Gross, 2011; Recht, 2011; Candès et al., 2014) showed that trace minimization and related convex optimization approaches also work for other measurement ensembles such as those arising in matrix completion and related problems. In this paper we have established a similar result to (Recht et al., 2010). We require the same order of measurements $\Omega(nr)$ but use a more computationally friendly local search algorithm. Also related to this work are projection gradient schemes with hard thresholding (Tropp & Gilbert, 2007; Garg & Khandekar, 2009; Needell & Tropp, 2009; Needell & Vershynin, 2009; Blumensath & Davies, 2009; Meka et al., 2009; Cai et al., 2010). Such algorithms enjoy similar guarantees to that of (Recht et al., 2010) and this work. Indeed, we utilize such results in the initialization phase of our algorithm. However, such algorithms require a rank-r SVD in each iteration which may be expensive for large problem sizes. We would like to emphasize, however, that for small problem sizes and dense matrices (such as Gaussian ensembles) such algorithms may be faster than gradient descent approaches such as ours.

More recently, there has been a few results using non-convex optimization schemes for matrix recovery problems. In particular, theoretical guarantees for matrix completion have been established using manifold optimization (Keshavan et al., 2010) and alternating minimization (Keshavan, 2012) (albeit with the caveat of requiring a fresh set of samples in each iteration). See also (Hardt, 2014; Sun & Luo, 2015). Later on, Jain et.al. (Jain et al., 2013) analyzed the performance of alternating minimization under similar modeling assumptions to (Recht et al., 2010) and this paper. However, the requirements on the RIP constant in (Jain et al., 2013) are more stringent compared to (Recht et al., 2010) and ours. In particular, the authors require $\delta_{4r} \leq c/r$ whereas we only require $\delta_{6r} \leq c$. Specialized to the Gaussian model, the results of (Jain et al., 2013) require $\Omega(nr^3\kappa^2)$ measurements.[3]

Our algorithm and analysis are inspired by the recent paper (Candès et al., 2015) by Candes, Li and Soltanolkotabi. See also (Soltanolkotabi, 2014; Cai et al., 2015) for some stability results. In (Candès et al., 2015) the authors introduced a local regularity condition to analyze the convergence of a gradient descent-like scheme for phase retrieval. We use a similar regularity condition but generalize it to ranks higher than one. Recently, independent of our work, Zheng and Lafferty (Zheng & Lafferty, 2015) provided an analysis of gradient descent using (2.2) via the same regularity condition. Zheng and Lafferty focus on the Gaussian ensemble, and establish a sample complexity of $m = \Omega(nr^3\kappa^2 \log n)$. In comparison we only require $\Omega(nr)$ measurements removing both the dependence on κ in the sample complexity and improving the asymptotic rate. We would like to emphasize that the improvement in our result is not just due to the more sophisticated initialization scheme. In particular, Zheng and Lafferty show geometric convergence starting from any initial solution obeying $\text{dist}(\boldsymbol{U}_0, \boldsymbol{X}) \leq c \cdot \sigma_r(\boldsymbol{X})$ as long as the number of measurements obeys $m = \Omega(nr\kappa^2 \log n)$. In contrast, we establish geometric convergence starting from the same neighborhood of $\boldsymbol{U}_0$ with only $\Omega(nr)$ measurements. Our results also differs in terms of the convergence rate. We establish a convergence rate of the form $1 - \frac{\mu}{\kappa}$ whereas (Zheng & Lafferty, 2015) establishes a slower convergence rate of the form $1 - \frac{\mu}{nr^2\kappa^2}$. Moreover, the theory of restricted isometries in our work considerably simplifies the analysis.

Finally, we would also like to mention (Sa et al., 2015) for guarantees using stochastic gradient algorithms. The results of (Sa et al., 2015) are applicable to a variety of models; focusing on the Gaussian ensemble, the authors require $\Omega\left((nr \log n)/\epsilon\right)$ samples to reach a relative error of ϵ. In contrast, our sample complexity is independent of the desired relative error ϵ. However, their algorithm only requires a random initialization.

[3] The authors also propose a stage-wise algorithm with improved sample complexity of $\Omega(nr^3\tilde{\kappa}^2)$ where $\tilde{\kappa}$ is a local condition number defined as the ratio of the maximum ratio of two successive eigenvalues. We note, however, that in general $\tilde{\kappa}$ can be as large as κ.

Since the first version of this paper appeared on arXiv, a few recent papers have also studied low-rank recovery from RIP measurements via Procrustes Flow type schemes (Bhojanapalli et al., 2015; Zhao et al., 2015; Chen & Wainwright, 2015). We would like to point out that the results presented in these papers are suboptimal compared to ours. For example, by utilizing some of the results of the previous version of this paper, (Bhojanapalli et al., 2015) provides a similar convergence rate to ours. However, this convergence occurs in a smaller radius around the planted solution so that the required number of measurements is significantly higher. Furthermore, the results of (Bhojanapalli et al., 2015) only apply when the matrix is PSD and do not work for general rectangular matricies. Similarly, result in (Chen & Wainwright, 2015) holds only for PSD matrices, and the convergence rate has a high-degree polynomial dependence on condition number. The algorithm from (Zhao et al., 2015) does generalize to rectangular matrices, but the sample complexity is of the order of $\mathcal{O}(nr^3 \log n)$ rather than the complexity $\mathcal{O}(nr)$ we establish here. Moreover, our analysis of both the PSD and rectangular cases is far more concise.

5. Proof ideas

We first sketch the proof outline for the symmetric PSD case (Theorem 3.2) However, whenever possible we will state lemmas in the more general setting. Full proofs can be found in the extended version of this paper (Tu et al., 2016).

Recall in this setting that we assume a fixed symmetric PSD $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ of rank r, which admits a factorization $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{X}^\mathsf{T}$ for $\boldsymbol{X} \in \mathbb{R}^{n\times r}$. Before we dive into the details of the proofs, we would like to mention that we will prove our results using the update

$$\boldsymbol{U}_{\tau+1} = \boldsymbol{U}_\tau - \frac{\mu}{\|\boldsymbol{X}\|^2}\nabla f(\boldsymbol{U}_\tau), \tag{5.1}$$

in lieu of the PF update

$$\boldsymbol{U}_{\tau+1} = \boldsymbol{U}_\tau - \frac{\mu_{\mathrm{PF}}}{\|\boldsymbol{U}_0\|^2}\nabla f(\boldsymbol{U}_\tau). \tag{5.2}$$

We will prove that our initial solution obeys $\mathrm{dist}(\boldsymbol{U}_0, \boldsymbol{X}) \leq \sigma_r(\boldsymbol{X})/4$. Hence, applying the triangle inequality we conclude that $\|\boldsymbol{U}_0\|^2 \leq \frac{25}{16}\|\boldsymbol{X}\|^2$, and similarly, $\|\boldsymbol{U}_0\|^2 \geq \frac{9}{16}\|\boldsymbol{X}\|^2$. Thus, any result proven for the update (5.1) will automatically carry over to the PF update with a simple rescaling of the upper bound on the step size via the former inequality. Furthermore, we can upper bound the convergence rate of gradient descent using the PF update in terms of properties of $\boldsymbol{X}$ instead of $\boldsymbol{U}_0$ via the latter.

5.1. Preliminaries

We start with a well known characterization of RIP.

Lemma 5.1. *(Candès, 2008) Let $\mathcal{A}$ satisfy $2r$-RIP with constant δ_{2r}. Then, for all matrices $\boldsymbol{X}, \boldsymbol{Y}$ of rank at most r, we have*

$$|\langle \mathcal{A}(\boldsymbol{X}), \mathcal{A}(\boldsymbol{Y})\rangle - \langle \boldsymbol{X}, \boldsymbol{Y}\rangle| \leq \delta_{2r}\|\boldsymbol{X}\|_F\|\boldsymbol{Y}\|_F \ .$$

Next, we state a recent result which characterizes the convergence rate of projected gradient descent onto general non-convex sets specialized to our problem. See (Meka et al., 2009) for related results using singular value hard thresholding. Throughout, $\mathcal{P}_r(\boldsymbol{M})$ denotes projection onto rank-r matrices. For a symmetric PSD matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ denotes projection onto the rank-r PSD matrices and for a rectangular matrix $\boldsymbol{M} \in \mathbb{R}^{n_1\times n_2}$ it denotes projection onto rank-r matrices.

Lemma 5.2. *(Oymak et al., 2015) Let $\boldsymbol{M} \in \mathbb{R}^{n_1\times n_2}$ be an arbitrary matrix of rank r. Also let $\boldsymbol{b} = \mathcal{A}(\boldsymbol{M}) \in \mathbb{R}^m$ be m linear measurements. Consider the iterative updates*

$$\boldsymbol{Z}_{\tau+1} \leftarrow \mathcal{P}_r\left(\boldsymbol{Z}_\tau - \sum_{k=1}^{m}(\langle \boldsymbol{A}_k, \boldsymbol{Z}_\tau\rangle - b_k)\boldsymbol{A}_k\right) .$$

Then

$$\|\boldsymbol{Z}_\tau - \boldsymbol{M}\|_F \leq \rho(\mathcal{A})^\tau \|\boldsymbol{Z}_0 - \boldsymbol{M}\|_F ,$$

holds. Here, $\rho(\mathcal{A})$ is defined as

$$\rho(\mathcal{A}) := 2 \sup_{\substack{\|\boldsymbol{X}\|_F=1,\mathrm{rank}(\boldsymbol{X})\leq 2r, \\ \|\boldsymbol{Y}\|_F=1,\mathrm{rank}(\boldsymbol{Y})\leq 2r}} |\langle \mathcal{A}(\boldsymbol{X}), \mathcal{A}(\boldsymbol{Y})\rangle - \langle \boldsymbol{X}, \boldsymbol{Y}\rangle|.$$

We shall make repeated use of the following lemma which upper bounds $\left\|\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\|_F$ by some factor of $\mathrm{dist}(\boldsymbol{U}, \boldsymbol{X})$, which is immediate from two applications of the triangle inequality.

Lemma 5.3. *For any $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ obeying $\mathrm{dist}(\boldsymbol{U}, \boldsymbol{X}) \leq \frac{1}{4}\|\boldsymbol{X}\|$, we have*

$$\left\|\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\|_F \leq \frac{9}{4}\|\boldsymbol{X}\|\,\mathrm{dist}(\boldsymbol{U}, \boldsymbol{X}).$$

Finally, we also need the following lemma which upper bounds $\mathrm{dist}(\boldsymbol{U}, \boldsymbol{X})$ by some factor of $\left\|\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\|_F$.

Lemma 5.4. *For any $\boldsymbol{U}, \boldsymbol{X} \in \mathbb{R}^{n\times r}$, we have*

$$\mathrm{dist}^2(\boldsymbol{U}, \boldsymbol{X}) \leq \frac{1}{2(\sqrt{2}-1)\sigma_r^2(\boldsymbol{X})}\left\|\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\|_F^2 .$$

We would like to point out that the dependence on $\sigma_r^2(\boldsymbol{X})$ in the lemma above is unavoidable.

5.2. Proof of convergence of gradient descent updates (Equation (3.4))

We first outline the general proof strategy. See Sections 2.3 and 7.9 of (Candès et al., 2015) for related arguments. We first will show that gradient descent on an approximate estimate of the function f converges. The approximate function we use is $F(\boldsymbol{U}) := \frac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\|_F^2$. When the map $\mathcal{A}$ is random and isotropic in expectation, $F(\boldsymbol{U})$ can be interpreted as the expected value of $f(\boldsymbol{U})$, but we stress that our result is a purely deterministic result. We demonstrate that $F(\boldsymbol{U})$ exhibits geometric convergence in a small neighborhood around $\boldsymbol{X}$. The standard approach in optimization to show this is to prove that the function exhibits strong convexity. However, due to the rotational degrees of freedom for any optimal point, it is not possible for $F(\boldsymbol{U})$ to be strongly convex in any neighborhood around $\boldsymbol{X}$ except in the special case when $r = 1$. Thus, we rely on the approach used by (Candès et al., 2015), which establishes a sufficient condition that only relies on first-order information along certain trajectories. After showing the sufficient condition holds on $F(\boldsymbol{U})$, we use standard RIP results to show that this condition also holds for the function $f(\boldsymbol{U})$.

To begin our analysis, we start with the following formulas for the gradient of $f(\boldsymbol{U})$ and $F(\boldsymbol{U})$

$$\begin{aligned}\nabla f(\boldsymbol{U}) &= \sum_{k=1}^{m} \langle \boldsymbol{A}_k, \boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\rangle \boldsymbol{A}_k \boldsymbol{U} \\ &= \mathcal{A}^*\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}) \cdot \boldsymbol{U}\,, \\ \nabla F(\boldsymbol{U}) &= (\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T})\boldsymbol{U}\,.\end{aligned}$$

Above, $\mathcal{A}^* : \mathbb{R}^m \to \mathbb{R}^{n\times n}$ is the adjoint operator of $\mathcal{A}$, i.e. $\mathcal{A}^*(z) = \sum_{i=1}^m \boldsymbol{A}_k z_k$. Throughout the proof $\boldsymbol{R}$ is the solution to the orthogonal Procrustes problem. That is,

$$\boldsymbol{R} = \mathop{\arg\min}_{\widetilde{\boldsymbol{R}} \in \mathbb{R}^{n\times n}:\, \widetilde{\boldsymbol{R}}^\mathsf{T}\widetilde{\boldsymbol{R}} = \boldsymbol{I}_r} \left\|\boldsymbol{U} - \boldsymbol{X}\widetilde{\boldsymbol{R}}\right\|_F,$$

with the dependence on $\boldsymbol{U}$ omitted for sake of exposition. The following definition defines a notion of strong convexity along certain trajectories of the function.

Definition 5.5. *(Regularity condition, (Candès et al., 2015)) Let $\boldsymbol{X} \in \mathbb{R}^{n\times r}$ be a global optimum of a function f. Define the set $B(\delta)$ as*

$$B(\delta) := \{\boldsymbol{U} \in \mathbb{R}^{n\times r} : \operatorname{dist}(\boldsymbol{U}, \boldsymbol{X}) \le \delta\}\,.$$

The function f satisfies a regularity condition, *denoted by* $\mathsf{RC}(\alpha, \beta, \delta)$*, if for all matrices $\boldsymbol{U} \in B(\delta)$ the following inequality holds:*

$$\langle \nabla f(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{X}\boldsymbol{R}\rangle \ge \frac{1}{\alpha}\|\boldsymbol{U} - \boldsymbol{X}\boldsymbol{R}\|_F^2 + \frac{1}{\beta}\|\nabla f(\boldsymbol{U})\|_F^2\,.$$

If a function satisfies $\mathsf{RC}(\alpha, \beta, \delta)$, then as long as gradient descent starts from a point $\boldsymbol{U}_0 \in B(\delta)$, it will have a geometric rate of convergence to the optimum $\boldsymbol{X}$. This is formalized by the following lemma.

Lemma 5.6. *(Candès et al., 2015) If f satisfies* $\mathsf{RC}(\alpha, \beta, \delta)$ *and $\boldsymbol{U}_0 \in B(\delta)$, then the gradient descent update*

$$\boldsymbol{U}_{\tau+1} \leftarrow \boldsymbol{U}_\tau - \mu\nabla f(\boldsymbol{U}_\tau),$$

with step size $0 < \mu \le 2/\beta$ obeys $\boldsymbol{U}_\tau \in B(\delta)$ and

$$\operatorname{dist}^2(\boldsymbol{U}_\tau, \boldsymbol{X}) \le \left(1 - \frac{2\mu}{\alpha}\right)^\tau \operatorname{dist}^2(\boldsymbol{U}_0, \boldsymbol{X})\,,$$

for all $\tau \ge 0$.

The proof is complete by showing that the regularity condition holds. To this end, we first show in Lemma 5.7 below that the function $F(\boldsymbol{U})$ satisfies a slightly stronger variant of the regularity condition from Definition 5.5. We then show in Lemma 5.8 that the gradient of f is always close to the gradient of F, and in Lemma 5.9 that the gradient of f is Lipschitz around the optimal value $\boldsymbol{X}$.

Lemma 5.7. *Let $F(\boldsymbol{U}) = \frac{1}{4}\left\|\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\|_F^2$. For all $\boldsymbol{U}$ obeying*

$$\|\boldsymbol{U} - \boldsymbol{X}\boldsymbol{R}\| \le \frac{1}{4}\sigma_r(\boldsymbol{X}),$$

we have

$$\begin{aligned}&\langle \nabla F(\boldsymbol{U}), \boldsymbol{U} - \boldsymbol{X}\boldsymbol{R}\rangle \ge \\ &\qquad \frac{1}{20}\left(\left\|\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\|_F^2 + \left\|(\boldsymbol{U} - \boldsymbol{X}\boldsymbol{R})\boldsymbol{U}^\mathsf{T}\right\|_F^2\right) \\ &\qquad + \frac{\sigma_r^2(\boldsymbol{X})}{4}\|\boldsymbol{U} - \boldsymbol{X}\boldsymbol{R}\|_F^2 + \frac{1}{5}\left\|\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\|_F^2\,.\end{aligned} \tag{5.3}$$

Lemma 5.8. *Let $\mathcal{A}$ be a linear map obeying rank-4r RIP with constant δ_{4r}. For any $\boldsymbol{H} \in \mathbb{R}^{n\times r}$ and any $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ obeying $\operatorname{dist}(\boldsymbol{U}, \boldsymbol{X}) \le \frac{1}{4}\|\boldsymbol{X}\|$, we have*

$$|\langle \nabla F(\boldsymbol{U}) - \nabla f(\boldsymbol{U}), \boldsymbol{H}\rangle| \le \delta_{4r}\left\|\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\|_F \left\|\boldsymbol{H}\boldsymbol{U}^\mathsf{T}\right\|_F\,.$$

This immediately implies that for any $\boldsymbol{U} \in \mathbb{R}^{n\times r}$ obeying $\operatorname{dist}(\boldsymbol{U}, \boldsymbol{X}) \le \frac{1}{4}\|\boldsymbol{X}\|$, we have

$$\|\nabla f(\boldsymbol{U}) - \nabla F(\boldsymbol{U})\|_F \le \delta_{4r}\left\|\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\|_F \|\boldsymbol{U}\|\,.$$

Lemma 5.9. *Let $\mathcal{A}$ be a linear map obeying rank-$6r$ RIP with constant δ_{6r}. Suppose that $\delta_{6r} \le 1/10$. Then for all $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, we have that*

$$\left\|\boldsymbol{U}\boldsymbol{U}^\mathsf{T} - \boldsymbol{X}\boldsymbol{X}^\mathsf{T}\right\|_F^2 \ge \frac{10}{17}\frac{1}{\|\boldsymbol{U}\|^2}\|\nabla f(\boldsymbol{U})\|_F^2\,.$$

We now explain how the regularity condition follows from these three lemmas. To begin, note that

$$\begin{aligned}
&\langle \nabla F(\boldsymbol{U}), \boldsymbol{U}-\boldsymbol{X}\boldsymbol{R}\rangle - \langle \nabla f(\boldsymbol{U}), \boldsymbol{U}-\boldsymbol{X}\boldsymbol{R}\rangle \qquad (5.4)\\
&\quad = \langle \nabla F(\boldsymbol{U}) - \nabla f(\boldsymbol{U}), \boldsymbol{U}-\boldsymbol{X}\boldsymbol{R}\rangle\\
&\quad \overset{(a)}{\le} \frac{1}{10}\left\|\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}-\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\right\|_F \left\|(\boldsymbol{U}-\boldsymbol{X}\boldsymbol{R})\boldsymbol{U}^{\mathsf{T}}\right\|_F\\
&\quad \overset{(b)}{\le} \frac{1}{20}\left(\left\|\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}-\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\right\|_F^2 + \left\|(\boldsymbol{U}-\boldsymbol{X}\boldsymbol{R})\boldsymbol{U}^{\mathsf{T}}\right\|_F^2\right)
\end{aligned} \tag{5.5}$$

where (a) holds from Cauchy-Schwarz followed by Lemma 5.8, using the fact that $\delta_{6r} \le \frac{1}{10}$ as assumed in the statement of Theorem 3.2 and (b) follows from $2ab \le a^2+b^2$.

Combining (5.5) with Lemma 5.7 for any $\boldsymbol{U}$ obeying $\|\boldsymbol{U}-\boldsymbol{X}\boldsymbol{R}\| \le \frac{1}{4}\sigma_r(\boldsymbol{X})$, we have

$$\begin{aligned}
&\langle \nabla f(\boldsymbol{U}), \boldsymbol{U}-\boldsymbol{X}\boldsymbol{R}\rangle\\
&\quad \ge \frac{\sigma_r^2(\boldsymbol{X})}{4}\|\boldsymbol{U}-\boldsymbol{X}\boldsymbol{R}\|_F^2 + \frac{1}{5}\left\|\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}-\boldsymbol{X}\boldsymbol{X}^{\mathsf{T}}\right\|_F^2\\
&\quad \overset{(a)}{\ge} \frac{\sigma_r^2(\boldsymbol{X})}{4}\|\boldsymbol{U}-\boldsymbol{X}\boldsymbol{R}\|_F^2 + \frac{2}{17}\frac{1}{\|\boldsymbol{U}\|^2}\|\nabla f(\boldsymbol{U})\|_F^2\\
&\quad \overset{(b)}{\ge} \frac{\sigma_r^2(\boldsymbol{X})}{4}\|\boldsymbol{U}-\boldsymbol{X}\boldsymbol{R}\|_F^2 + \frac{32}{425}\frac{1}{\|\boldsymbol{X}\|^2}\|\nabla f(\boldsymbol{U})\|_F^2\ ,
\end{aligned} \tag{5.6}$$

where (a) follows from Lemma 5.9 and (b) follows from the fact that $\|\boldsymbol{U}\| \le \frac{5}{4}\|\boldsymbol{X}\|$ when $\operatorname{dist}(\boldsymbol{U},\boldsymbol{X}) \le \frac{1}{4}\|\boldsymbol{X}\|$. Equation (5.6) shows that $f(\boldsymbol{U})$ obeys $\mathsf{RC}(4/\sigma_r^2(\boldsymbol{X}), \frac{425}{32}\|\boldsymbol{X}\|^2, \frac{1}{4}\sigma_r(\boldsymbol{X}))$. The convergence result in Equation (3.4) now follows from Lemma 5.6.

5.3. Rectangular case

We now turn our attention to the general case where the matrices are rectangular. Recall that in this case, we want to recover a fixed but unknown rank-r matrix $\boldsymbol{M} \in \mathbb{R}^{n_1\times n_2}$ from linear measurements. Assume that $\boldsymbol{M}$ has a singular value decomposition of the form $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{B}^{\mathsf{T}}$. Define $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{\Sigma}^{1/2} \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{\Sigma}^{1/2} \in \mathbb{R}^{n_2\times r}$. With this piece of notation the iterates $\boldsymbol{U}_\tau \in \mathbb{R}^{n_1\times r}, \boldsymbol{V}_\tau \in \mathbb{R}^{n_2\times r}$ in Algorithm 2 can be thought of as estimates of $\boldsymbol{X}$ and $\boldsymbol{Y}$. The proof of the correctness of the initialization phase of Procrustes Flow (Theorem 3.3, Equation (3.6)) in the rectangular case is similar to the PSD case (Theorem 3.2, Equation (3.3)). In this section we shall describe the main ideas of the proof.

To simplify exposition we aggregate the pairs of matrices $(\boldsymbol{U},\boldsymbol{V})$, $(\boldsymbol{X},\boldsymbol{Y})$, and $(\boldsymbol{X},-\boldsymbol{Y})$ into larger "lifted" matrices as follows

$$\boldsymbol{W} := \begin{bmatrix}\boldsymbol{U}\\ \boldsymbol{V}\end{bmatrix}, \quad \boldsymbol{Z} := \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix}, \quad \text{and} \quad \widetilde{\boldsymbol{Z}} := \begin{bmatrix}\boldsymbol{X}\\ -\boldsymbol{Y}\end{bmatrix}.$$

To prove Theorem 3.3, Equation (3.7), we will demonstrate that the function $g(\boldsymbol{W}) := g(\boldsymbol{U},\boldsymbol{V})$ over the variable $\boldsymbol{W}$ has similar form to $f(\boldsymbol{U})$ over the variable $\boldsymbol{U}$. As in the proof for the PSD case, the crux of Theorem 3.3 lies in establishing that the regularity condition

$$\begin{aligned}
&\langle \nabla g(\boldsymbol{W}), \boldsymbol{W}-\boldsymbol{Z}\boldsymbol{R}\rangle\\
&\quad \ge \frac{\sigma_r(\boldsymbol{M})}{8}\|\boldsymbol{W}-\boldsymbol{Z}\boldsymbol{R}\|_F^2 + \frac{16}{1683\|\boldsymbol{M}\|}\|\nabla g(\boldsymbol{W})\|_F^2\ ,
\end{aligned} \tag{5.7}$$

holds for all $\boldsymbol{W} \in \mathbb{R}^{(n_1+n_2)\times r}$ obeying $\operatorname{dist}(\boldsymbol{W},\boldsymbol{Z}) \le \frac{1}{2\sqrt{2}}\sigma_r^{1/2}(\boldsymbol{M})$. Assuming that this condition holds, we have that $g(\boldsymbol{W})$ obeys $\mathsf{RC}(8/\sigma_r(\boldsymbol{M}), \frac{1683}{16}\|\boldsymbol{M}\|, \frac{1}{2\sqrt{2}}\sigma_r^{1/2}(\boldsymbol{M}))$, and hence Theorem 3.3, Equation (3.7) immediately follows by appealing to Lemma 5.6.

To prove (5.7), we make use of the similarity of the expressions with the PSD case. We start, as before, by defining a reference function $F(\boldsymbol{W}) := \frac{1}{4}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}-\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}\right\|_F^2$ with gradient $\nabla F(\boldsymbol{W}) = (\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}-\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}})\boldsymbol{W}$. We now state two lemmas relating g and F, which together immediately imply (5.7). The first lemma relates the regularity condition of g to that of F by utilizing RIP. The second lemma provides a Lipschitz type property for the gradient of g.

Lemma 5.10. *Assume the linear mapping $\mathcal{A}$ obeys $4r$-RIP with constant δ_{4r}. Then g obeys the following regularity condition for any $\boldsymbol{W} \in \mathbb{R}^{(n_1+n_2)\times r}$ and $\boldsymbol{R} \in \mathbb{R}^{r\times r}$,*

$$\begin{aligned}
&\langle \nabla g(\boldsymbol{W}), \boldsymbol{W}-\boldsymbol{Z}\boldsymbol{R}\rangle\\
&\quad \ge -\frac{\delta_{4r}}{2}\left\|\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}-\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}\right\|_F \left\|(\boldsymbol{W}-\boldsymbol{Z}\boldsymbol{R})\boldsymbol{W}^{\mathsf{T}}\right\|_F\\
&\qquad + \frac{1}{4}\langle \nabla F(\boldsymbol{W}), \boldsymbol{W}-\boldsymbol{Z}\boldsymbol{R}\rangle + \frac{1}{8\|\boldsymbol{M}\|}\left\|\widetilde{\boldsymbol{Z}}\widetilde{\boldsymbol{Z}}^{\mathsf{T}}\boldsymbol{W}\right\|_F^2 .
\end{aligned} \tag{5.8}$$

Lemma 5.11. *Let $\mathcal{A}$ be a linear map obeying rank-$6r$ RIP with constant $\delta_{6r} \le 1/10$. Then for all $\boldsymbol{W} \in \mathbb{R}^{(n_1+n_2)\times r}$ satisfying $\operatorname{dist}(\boldsymbol{W},\boldsymbol{Z}) \le \frac{1}{4}\|\boldsymbol{Z}\|$, we have that*

$$\begin{aligned}
&\frac{21}{400}\|\boldsymbol{W}\boldsymbol{W}^{\mathsf{T}}-\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}\|_F^2 + \frac{1}{8\|\boldsymbol{M}\|}\left\|\widetilde{\boldsymbol{Z}}\widetilde{\boldsymbol{Z}}^{\mathsf{T}}\boldsymbol{W}\right\|_F^2\\
&\qquad \ge \frac{16}{1683}\frac{1}{\|\boldsymbol{M}\|}\|\nabla g(\boldsymbol{W})\|_F^2 .
\end{aligned} \tag{5.9}$$

With these lemmas in place we have all the elements to prove (5.7). By applying Lemma 5.7 to $\langle \nabla F(\boldsymbol{W}), \boldsymbol{W}-\boldsymbol{Z}\boldsymbol{R}\rangle$ and combining (5.8) and (5.9), Equation (5.7) follows after some simple manipulations. This concludes the proof of Theorem 3.3.

Acknowledgements

BR is generously supported by ONR awards N00014-11-1-0723 and N00014-13-1-0129, NSF awards CCF-1148243 and CCF-1217058, AFOSR award FA9550-13-1-0138, and a Sloan Research Fellowship. RB is generously supported by ONR award N00014-11-1-0723 and the NDSEG Fellowship. This research is supported in part by NSF CISE Expeditions Award CCF-1139158, LBNL Award 7076018, and DARPA XData Award FA8750-12-2-0331, and gifts from Amazon Web Services, Google, SAP, The Thomas and Stacey Siebel Foundation, Adatao, Adobe, Apple, Inc., Blue Goji, Bosch, C3Energy, Cisco, Cray, Cloudera, EMC2, Ericsson, Facebook, Guavus, HP, Huawei, Informatica, Intel, Microsoft, NetApp, Pivotal, Samsung, Schlumberger, Splunk, Virdata and VMware.

References

Achlioptas, D. and McSherry, F. Fast computation of low-rank matrix approximations. *Journal of the ACM*, 54(2), 2007.

Bhojanapalli, S., Kyrillidis, A., and Sanghavi, S. Dropping convexity for faster semi-definite optimization. *arXiv*, arXiv:1509.03917, 2015.

Blumensath, T. and Davies, M. E. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

Cai, J. F., Candès, E. J., and Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Cai, T. T., Li, X., and Ma, Z. Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *arXiv*, arXiv:1506.03382, 2015.

Candès, E. J. The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Academie des Sciences*, 2008.

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

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Candès, E. J. and Tao, T. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51 (12):4203–4215, 2005.

Candès, E. J., Li, X., and Soltanolkotabi, M. Phase retrieval from coded diffraction patterns. *Applied and Computational Harmonic Analysis*, 39(2):277–299, 2014.

Candès, E. J., Li, X., and Soltanolkotabi, M. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Chen, Y. and Wainwright, M. J. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv*, arXiv:1509.03025, 2015.

Davenport, M. A. and Romberg, J. An overview of low-rank matrix recovery from incomplete observations. *arXiv*, arXiv:1601.06422, 2016.

Fazel, M. *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University, 2002.

Funk, S. Netflix update: Try this at home, December 2006. URL http://sifter.org/~simon/journal/20061211.html.

Garg, R. and Khandekar, R. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, 2009.

Gross, D. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

Hardt, Moritz. Understanding alternating minimization for matrix completion. In *FOCS*, 2014.

Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. In *STOC*, 2013.

Keshavan, R. H. *Efficient algorithms for collaborative filtering*. PhD thesis, Stanford University, 2012.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

Lee, J., Recht, B., Srebro, N., Tropp, J. A., and Salakhutdinov, R. Practical large-scale optimization for max-norm regularization. In *NIPS*, 2010.

Meka, R., Jain, P., and Dhillon, I. S. Guaranteed rank minimization via singular value projection. *arXiv*, arXiv:0909.5457, 2009.

Needell, D. and Tropp, J. A. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

Needell, D. and Vershynin, R. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics*, 9(3):317–334, 2009.

Oymak, S., Recht, B., and Soltanolkotabi, M. Sharp time-data tradeoffs for linear inverse problems. *arXiv*, arXiv:1507.04793, 2015.

Recht, B. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

Recht, B. and Ré, C. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, pp. 201–226, 2013.

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

Rennie, J. and Srebro, N. Fast maximum margin matrix factorization for collaborative prediction. In *ICML*, 2005.

Ruhe, A. Numerical computation of principal components when several observations are missing. Technical report, University of Umea, Institute of Mathematics and Statistics Report, 1974.

Sa, C. De, Olukotun, K., and Ré, C. Global convergence of stochastic gradient descent for some nonconvex matrix problems. In *ICML*, 2015.

Soltanolkotabi, M. *Algorithms and Theory for Clustering and Nonconvex Quadratic Programming*. PhD thesis, Stanford University, 2014.

Sun, R. and Luo, Z. Guaranteed matrix completion via non-convex factorization. In *FOCS*, 2015.

Tropp, J. A. and Gilbert, A. C. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12): 4655–4666, 2007.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. Low-rank solutions of linear matrix equations via procrustes flow. *arXiv*, arXiv:1507.03566, 2016.

Zhao, T., Wang, Z., and Liu, H. A nonconvex optimization framework for low rank matrix estimation. In *NIPS*, 2015.

Zheng, Q. and Lafferty, J. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *NIPS*, 2015.