Fast Algorithms for Sparse Reduced-Rank Regression

Benjamin Dubois, Jean-François Delmas, Guillaume Obozinski

Fast Algorithms for Sparse Reduced-Rank Regression

Benjamin Dubois[*,†] **Jean-François Delmas**[*] **Guillaume Obozinski**[†]
[*]CERMICS, École des Ponts, UPE, Champs-sur-Marne, France
[†]LIGM, UMR 8049, École des Ponts, UPEM, ESIEE Paris, CNRS, UPE, Champs-sur-Marne, France

Abstract

We consider a reformulation of Reduced-Rank Regression (RRR) and Sparse Reduced-Rank Regression (SRRR) as a non-convex non-differentiable function of a single of the two matrices usually introduced to parametrize low-rank matrix learning problems. We study the behavior of proximal gradient algorithms for the minimization of the objective. In particular, based on an analysis of the geometry of the problem, we establish that a *proximal* Polyak-Łojasiewicz inequality is satisfied in a neighborhood of the set of optima under a condition on the regularization parameter. We consequently derive linear convergence rates for the proximal gradient descent with line search and for related algorithms in a neighborhood of the optima. Our experiments show that our formulation leads to much faster learning algorithms for RRR and especially for SRRR.

1 Introduction

In matrix learning problems, an effective way of reducing the number of degrees of freedom is to constrain the rank of the coefficient matrix to be learned. Low-rank constraints lead however to non-convex optimization problems for which the structure of critical points and the behavior of standard optimization algorithms, like gradient descent, stochastic block coordinate gradient descent and their proximal counterparts, are difficult to analyze. These difficulties have lead researchers to either use these algorithms without guarantee or to consider convex relaxations in which the low-rank constraint is replaced by a trace-norm constraint or penalty.

Proceedings of the 22[nd] International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

In the last few years however, a better understanding of the geometry of these problems (Li et al., 2016; Zhu et al., 2017b), new tools from non-convex analysis (Attouch and Bolte, 2009; Frankel et al., 2015; Karimi et al., 2016; Csiba and Richtárik, 2017; Khamaru and Wainwright, 2018) as well as results on the behavior of standard algorithms around saddle points (Lee et al., 2017) were developed under regularity assumptions to analyze their convergence and eventually prove rates of convergence.

Formulations that require to learn a low-rank matrix or its factors appear in many problems in machine learning, from variants of Principal Components Analysis and Canonical Correlation Analysis, to matrix completion problems and multi-task learning formulations. Reduced-Rank Regression (RRR) is a fundamental model of this family. It corresponds to the multiple outputs linear regression in which all the vectors of parameters associated with the different dimensions are constrained to lie in a space of dimension $r \in \mathbb{N}^*$. Precisely, if $X \in \mathbb{R}^{n,p}$ is a design matrix and $Y \in \mathbb{R}^{n,k}$ has columns corresponding to the multiple tasks, then the problem is usually formulated with $\|\cdot\|_F$ the Frobenius norm as

$$\min_{W \in \mathbb{R}^{p,k}:\ \mathrm{rank}(W) \leq r} \frac{1}{2}\|Y - XW\|_F^2. \tag{1}$$

The solution of Problem (1) can be obtained in closed form (Velu and Reinsel, 2013) and requires to project the usual multivariate linear regression parameter estimate on the subspace spanned by the top right singular vectors of the matrix $(X^TX)^{-1/2}X^TY$.

Sparse Reduced-Rank Regression (SRRR) is a variant in which the objective is regularized by the group-Lasso norm $\|W\|_{1,2} = \sum_i (\sum_j W_{ij}^2)^{1/2}$, in order to induce row-wise sparsity in the matrix W, which corresponds to simultaneous variable selection for all tasks. Given $\lambda > 0$, the optimization problem takes the form

$$\min_{W \in \mathbb{R}^{p,k}:\ \mathrm{rank}(W) \leq r} \frac{1}{2}\|Y - XW\|_F^2 + \lambda\|W\|_{1,2}. \tag{2}$$

For this formulation, there is no closed form solution anymore, and the conceptually simple algorithms that

have been proposed to solve Problem (2) are not so computationally efficient.

In the last decade, many optimization problems of the form

$$\min_{W\in\mathbb{R}^{p,k}:\ \mathrm{rank}(W)\leq r} \mathcal{F}(W) \tag{3}$$

with $\mathcal{F}$ a convex function have been tackled via the convex relaxation obtained by replacing the rank constraint with a constraint or a regularization on the trace-norm $\|W\|_*$; unfortunately, these formulations often lead to expensive algorithms and the relaxation induces a bias. A recent literature revisited a number of these problems based on an explicit parameterization of the low-rank matrix, as biconvex problems of the form

$$\min_{U\in\mathbb{R}^{p,r},\, V\in\mathbb{R}^{k,r}} \mathcal{F}(UV^T). \tag{4}$$

In particular, it is natural to formulate Problem (1) and Problem (2) in this form.

In this paper, we additionally impose $V^TV = I_r$ without loss of generality and we reformulate the SRRR problem as a non-convex non-differentiable optimization problem of a single thin matrix U. Based on the geometry of the objective (see Corollary 6), we establish in Corollary 9 a generalized Polyak-Łojasewicz inequality (Polyak, 1963; Karimi et al., 2016) in a neighborhood of the minima which can be leveraged to show in Corollary 10 asymptotic linear convergence of the proximal gradient algorithm and of stochastic block coordinate proximal descent algorithms. Our results are also relevant to solve very large-scale RRR instances for which the direct computation of the closed form solution would not be possible.

The paper is structured as follows. In Section 2, we discuss related work. In Section 3, we reformulate the RRR/SRRR problems. In Section 4, we obtain global convergence results. To analyze the local convergence in Section 5, we review the structure of RRR and establish properties based on the orthogonal invariance of the objective as well as the convexity of its restriction on certain cones in a neighborhood of the optima. Thus, we obtain a Polyak-Łojasiewicz inequality and a generalized Polyak-Łojasiewicz inequality respectively for RRR and SRRR in a neighborhood of the global minima. Finally Section 6 illustrates with numerical experiments the performances of the proposed algorithms.

2 Related Work

Velu and Reinsel (2013) studied Problem (1) and showed that it is one of the few low-rank matrix problems which has a closed form solution. Baldi and Hornik (1989) studied thoroughly the biconvex version of Problem (1) and identified its critical points to show that its local minima are global. Bunea et al. (2011, 2012); Chen and Huang (2012); Ma and Sun (2014); Mukherjee et al. (2015); She (2017) considered Problem (2) and highlighted the statistical properties of the estimator. The algorithms proposed in these papers all consist essentially in optimizing alternatingly with respect to U and V an objective of the form (4) (and more precisely the objective (5) introduced in Section 3) under the constraint $V^TV = I_r$. The full optimization w.r.t. V requires to compute an SVD of the matrix $Y^TXU \in \mathbb{R}^{k,r}$ which is of reasonable size, but the full optimization w.r.t. U requires to solve a full group-Lasso problem.

Among others, iterative first-order algorithms that are classical for the jointly convex setting may be applied to the non-convex Problem (4). Until recently, precise convergence guarantees were relatively rare but the observation of good empirical rates of convergence motivated a finer analysis. In particular, a number of recent papers established stronger theoretical results for these algorithms in the smooth non-convex case. Notably, Jain et al. (2017) obtained the first *global* linear rate of convergence for the very particular case of the matrix square-root computation. For more general biconvex formulations, Park et al. (2016) and Wang et al. (2016) established convergence rate guarantees for the gradient descent algorithm for Problem (4) provided an appropriate initialization is used and penalties such as $\frac{1}{4}\|U^TU - V^TV\|_F^2$ are added to the objective as regularizers.

As a consequence of the aforementioned performances, there was a regain of interest for the biconvex problems like (4) and their geometry has been studied in numerous papers. Bhojanapalli et al. (2016); Boumal et al. (2016); Ge et al. (2016, 2017); Kawaguchi (2016); Li et al. (2018, 2017); Zhu et al. (2017a) studied critical points and made use of the strict saddle property to show global convergence results for gradient descent and stochastic variants. Some of these works define a partition of the space and characterize the behavior of gradient descent in each region (Li et al., 2016; Zhu et al., 2017b).

Besides, it was shown recently that appropriate first-order algorithms cannot converge to saddle points when the curvature of the objective is strict around them (Lee et al., 2017; Panageas and Piliouras, 2016; Sun et al., 2015). These algorithms actually spend only a limited amount of time near the saddle points if the Hessian is Lipschitz (Du et al., 2017; Jin et al., 2017). However, these papers do not provide general convergence rate results, in particular not in the non-differentiable case.

From the performances of classical first-order algo-

rithms originated attempts to characterize convergence and to possibly prove rates based on the local geometry of non-convex objective functions around minima. In particular, Karimi et al. (2016) reviewed and provided a unified point of view of the recent literature on the Polyak-Łojasiewicz inequality (Polyak, 1963). This type of results was leveraged by Csiba and Richtárik (2017) to prove convergence rates. A parallel thread of research focused on the Kurdyka-Łojasiewicz inequality (KŁ), with the motivation that all semi-algebraic functions satisfy it. Attouch and Bolte (2009); Attouch et al. (2013); Frankel et al. (2015); Ochs et al. (2014) were able to characterize asymptotic convergence rates for the forward-backward algorithm under the KŁ inequality. These types of results were extended for block coordinate descent schemes in Attouch et al. (2010); Bolte et al. (2014); Xu and Yin (2017); Nikolova and Tan (2017), and for accelerated proximal descent algorithms in Chouzenoux et al. (2014); Li and Lin (2015). However, in general, it remains difficult to prove a specific rate for a given problem, because the exact rate depends on the best exponent that can be obtained in the KŁ inequality, and with the exception of some results provided in Li and Pong (2017), determining this exponent remains difficult.

3 Reformulation and algorithm

3.1 A new formulation for RRR and SRRR with a single thin matrix U

We reformulate the biconvex version of SRRR

$$\min_{U\in\mathbb{R}^{p,r},\, V\in\mathbb{R}^{k,r}} \frac{1}{2}\|Y - XUV^T\|_F^2 + \lambda\|UV^T\|_{1,2}, \quad (5)$$

by eliminating V as follows. First, we can impose $V^TV = I_r$ as in Chen and Huang (2012) without loss of generality. Then, expanding the Frobenius norm and using the invariance of the norms to the transformation $U \mapsto UV^T$ with $V \in \mathbb{R}^{k,r}$ such that $V^TV = I_r$, the objective becomes $\frac{1}{2}\|XU\|_F^2 - \langle Y, XUV^T\rangle + \lambda\|U\|_{1,2}$ where $\langle\cdot,\cdot\rangle$ is the Frobenius inner product. The value of the orthogonal Procrustes problem

$$\max_{V\in\mathbb{R}^{k,r}:\, V^TV=I_r} \langle Y, XUV^T\rangle$$

is the trace-norm $\|Y^TXU\|_*$ (*cf.* Fact 25 in Appendix C). So, letting $f(U) := f_1(U) - f_2(U)$ with

$$f_1(U) = \frac{1}{2}\|XU\|_F^2 \quad \text{and} \quad f_2(U) = \|Y^TXU\|_*$$

and $F^\lambda(U) := f(U) + \lambda\|U\|_{1,2}$, RRR and SRRR are respectively reformulated as

$$\min_{U\in\mathbb{R}^{p,r}} f(U), \quad (\text{RRR})$$

$$\min_{U\in\mathbb{R}^{p,r}} F^\lambda(U). \quad (\text{SRRR})$$

The objectives, as differences of convex functions, are clearly non-convex. However, they are still orthogonal-invariant *i.e.* for any $U \in \mathbb{R}^{p,r}$ and $R \in \mathbb{R}^{r,r}$ such that $R^TR = I_r$, we have $f(UR) = f(U)$ and $F^\lambda(UR) = F^\lambda(U)$. Note that the above derivations would still be valid if we replaced the row-wise group-Lasso $\|\cdot\|_{1,2}$ by any regularizer which is invariant when the argument is multiplied on the right by an orthogonal matrix.

Also, note that although f involves a trace-norm, its argument, Y^TXU, is of dimensions $k \times r$ while, in convex relaxations of low-rank formulations like Problem (3), the rank constraint is substituted with a trace-norm regularizer $\|W\|_*$ that is computed for a typically large matrix W of dimensions $p \times k$.

3.2 Characterization of the optima of the classical RRR formulation

Velu and Reinsel (2013) characterized the closed form solution of Problem (1) when X^TX is invertible as follows. Let $W^* := (X^TX)^{-1}X^TY$ denote the full-rank least squares estimator. Let PSQ^T be the reduced singular value decomposition of $(X^TX)^{-\frac{1}{2}}X^TY$. If the latter has rank ℓ then $P \in \mathbb{R}^{p,\ell}$ and $Q \in \mathbb{R}^{k,\ell}$ have orthonormal columns and $S \in \mathbb{R}^{\ell,\ell}$ is the diagonal matrix with singular values $s_1 \geq \ldots \geq s_\ell > 0$. The solution of Problem (1) is unique if $s_r > s_{r+1}$: let $Q_r \in \mathbb{R}^{k,r}$ be the matrix obtained by keeping the first r columns of Q, the solution is $W_r^* := W^*Q_rQ_r^T$.

3.3 Algorithms and complexity

The algorithms that we consider are essentially proximal gradient algorithms with line search, except for the fact that f_2 is not differentiable when Y^TXU is not full-rank, which entails that f is not differentiable everywhere. To address this issue, and given that f is a difference of a smooth convex function and a continuous convex function, we consider the subgradient-type algorithms proposed in Khamaru and Wainwright (2018).

Given $U \in \mathbb{R}^{p,r}$, the idea is to use a subgradient z_U of f_2. We assume that X^TX is invertible but consider a more general case in Appendix D.1.2 where we detail the computations. Given $R_1DR_2^T$ a singular value decomposition of Y^TXU such that $\mathrm{Im}\, R_1 \subset \mathrm{Im}\, Y^TX$, we compute $z_U = X^TYR_1R_2^T$ with $R_1 \in \mathbb{R}^{k,r}$, $R_1^TR_1 = I_r$, $D = \mathrm{diag}(d_1 \geq \ldots \geq d_r) \in \mathbb{R}^{r,r}$ with $d_r \geq 0$ and $R_2 \in \mathcal{O}_r$. With a slight abuse of notation, we define $\nabla f(U) := \nabla f_1(U) - z_U$. Note that this is the gradient of the natural DC programming upper bound. We introduce for any $t > 0$ the t-approximation functions of f and F^λ at U,

$\tilde{f}_{t,U}(U') := f(U) + \langle \nabla f(U), U' - U \rangle + \frac{1}{2t}\|U' - U\|_F^2$ and $\tilde{F}^{\lambda}_{t,U}(U') := \tilde{f}_{t,U}(U') + \lambda\|U'\|_{1,2}$. At each iteration of Algorithm 1, U is updated with Algorithm 2 to U_+ the unique minimizer of $\tilde{F}^{\lambda}_{t,U}$ if the line search condition

$$\tilde{F}^{\lambda}_{t,U}(U_+) \geq F^{\lambda}(U_+) \tag{LS}$$

is satisfied. Otherwise, t is decreased by a multiplicative factor $\beta < 1$. We explain why Algorithm 2 terminates in Appendix E.2. The obtained algorithm is almost a gradient descent algorithm when $\lambda = 0$ and a proximal gradient descent algorithm when $\lambda > 0$ (see Appendix D.2). In practice, our algorithms stay away from points where f is non-differentiable and reduce to plain gradient descent and plain proximal gradient descent respectively. This motivated us to also consider for the experiments the accelerated proximal gradient algorithm of Li and Lin (2015), designed for the non-convex setting. We adapt in Section 4 parts of the global convergence results of Khamaru and Wainwright (2018) to our algorithms.

Algorithm 1 Proximal Gradient Descent with LSP

Input: data X, Y, $\bar{t}$, starting point $\bar{U}$
Initialize $k = 0$, $U_0 \leftarrow \bar{U}$, $t_{-1} \leftarrow \bar{t}$
while not converged **do**
  Compute t, U_+ with t_{k-1}, U_k and Algorithm 2
  $t_k \leftarrow t$
  $U_{k+1} \leftarrow U_+$
  $k = k + 1$
end while

Algorithm 2 Line Search Procedure (LSP)

Input: t_{k-1}, U_k, parameters $\beta \in (0, 1)$, $\pi \in (0, 1]$
Set $t \leftarrow \frac{t_{k-1}}{\beta}$ with probability π, o/w $t \leftarrow t_{k-1}$
$U_+ \leftarrow \operatorname{argmin}_{U'} \tilde{F}^{\lambda}_{t,U_k}(U')$
while (LS) is not satisfied **do**
  $t \leftarrow \beta t$
  $U_+ \leftarrow \operatorname{argmin}_{U'} \tilde{F}^{\lambda}_{t,U_k}(U')$
end while
Output: t, U_+

To discuss the complexity of the algorithm, we assume that X^TX and Y^TX are computed in advance. Although the computation of z_U requires an SVD of Y^TXU, the latter costs only $O(kr^2)$. Computing $\nabla f(U)$ has then a complexity of $O(p^2r + pkr)$. The biconvex formulation of Park et al. (2016) leads to iterations with the same theoretical complexity for RRR but it is incompatible with SRRR. Additionally, experiments show that our algorithm is faster (*cf.* Section 6 and Appendix M).

4 Global convergence results

Although recent papers such as Lee et al. (2017) have shown that the gradient descent algorithm escapes saddle points by leveraging the strict saddle property, global convergence for Algorithm 1 is not obvious because f is not smooth. Besides, to the best of our knowledge, none of the papers that exclude convergence towards saddle points deals with regularizers or line search.

4.1 Convergence to a critical point for RRR

For RRR, results of Khamaru and Wainwright (2018) apply to our formulation and show that our algorithm converges towards a critical point. Precisely, f_1 is continuously differentiable with Lipschitz gradients, f_2 is continuous and convex and the difference f is bounded below by $-\frac{1}{2}\|Y\|_F^2$. Besides, as a difference of semi-algebraic functions, f satisfies the Kurdyka-Łojasiewicz property whose definition is given in Appendix B.4. Therefore, for gradient descent, our setting satisfies the conditions of Theorems 1 and 3 of Khamaru and Wainwright (2018) and we can prove that our algorithm converges from any initial point to a critical point in the sense of Definition 21 in Appendix B.5. This is more formally stated in Appendix F.1.

4.2 Convergence to a critical point for SRRR

In addition to the properties of f_1 and f_2 discussed above in Section 4.1, the norm $\|\cdot\|_{1,2}$ is clearly proper, lower semi-continuous and convex so our setting for proximal gradient descent satisfies the conditions of the first part of Theorem 2 in Khamaru and Wainwright (2018). The latter can be adapted to prove that all limit points of the sequence are critical points in the sense of Definition 21 in Appendix B.5. However, to prove actual convergence of the sequence, their Theorem 4 formally requires that f_2 is a function with locally Lipschitz gradient, which is not true when Y^TXU is not full-rank.

Actually, an inspection of the proof of Theorem 4 in Khamaru and Wainwright (2018) shows that the local smoothness condition is only required in a neighborhood of the limit points of the sequence. We prove in Appendix F.2 that if all groups of at least r rows of X^TY are assumed full-rank, which holds almost surely if X and Y contain for example continuous additive noise, and unless local minima are so sparse that the number of selected variables is strictly smaller than r, then any local minimum $U \in \mathbb{R}^{p,r}$ is such that Y^TXU is full-rank. As a consequence, if we assume that the limit points of the sequence produced by the algorithm are a subset of the local minima, then these limit points

are contained within a compact set where the function is smooth and the proof of Theorem 4 of Khamaru and Wainwright (2018) can be adapted in a straightforward manner to obtain global convergence.

5 Local convergence analysis

In this section, we prove linear convergence rates in a neighborhood of the global minima for RRR and under a condition on the regularization parameter λ for SRRR. Precisely, we first study the geometry around the optima of (RRR) via a change of variables. Then, a continuity argument shows that the structure remains approximately the same for (SRRR) with a small $\lambda > 0$. Finally, we introduce and leverage Polyak-Łojasiewicz inequalities to prove local linear convergence.

5.1 A key reparameterization for RRR

The relation between RRR and PCA and the form of the analytical solution given by Velu and Reinsel (2013) will allow us to show that our study of the objective of RRR can be reduced to the study of the particular case in which X and Y are full-rank diagonal matrices, via a linear change of variables based on the singular value decomposition PSQ^T introduced in Section 3.2 of the matrix $(X^TX)^{-\frac{1}{2}}X^TY$. From now on, we assume that the rank parameter r is smaller than the rank of X^TY *i.e.* $r \leq \ell := \operatorname{rank}(X^TY)$. It makes sense to assume that the imposed rank is less than the rank of the optimum for the unconstrained problem, otherwise the rank constraint is essentially useless. We also assume[1] that $s_1 > \ldots > s_\ell$ and that X^TX is invertible.

With the notations of Section 3.2, let $P^\perp \in \mathbb{R}^{p,p-\ell}$ be a matrix such that ${P^\perp}^T P^\perp = I_{p-\ell}$ and $P^TP^\perp = 0$, and consider the linear transformation $U = \tau(A, C)$ where

$$\tau : \begin{cases} \mathbb{R}^{\ell,r} \times \mathbb{R}^{p-\ell,r} \to \mathbb{R}^{p,r} \\ (A, C) \mapsto (X^TX)^{-\frac{1}{2}}(PSA + P^\perp C) \end{cases}. \tag{6}$$

Defining $f_a(A) = \frac{1}{2}\|SA\|_F^2 - \|S^2A\|_*$, we show in Appendix G.1 that

$$(f \circ \tau)(A, C) = f_a(A) + \frac{1}{2}\|C\|_F^2. \tag{7}$$

Since τ is invertible, the minimization in (RRR) w.r.t. U is equivalent to the minimization of $f \circ \tau$ w.r.t. (A, C). We can therefore study the original optimization problem by studying f_a.

[1] These assumptions are also reasonable and will hold in particular if (X, Y) are drawn from a continuous density. We discuss the case where X^TX is not invertible in Appendix G and in Appendix H.2, we show why these assumptions are needed.

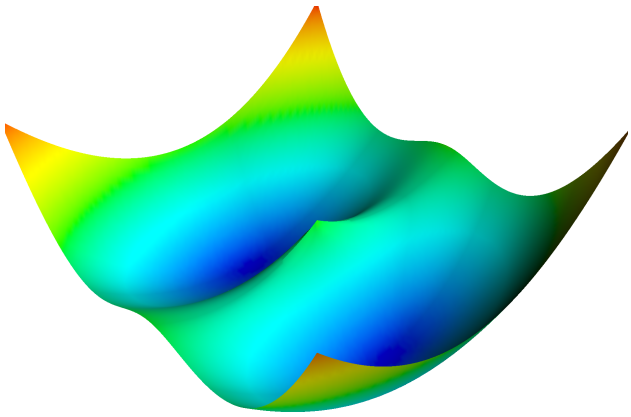

Figure 1: Graph of f_a for $A \in \mathbb{R}^{2,1}$. In this particular case, $\Omega_a^* = \{(1; 0), (-1; 0)\}$ and $\mathcal{O}_1 = \{-1, 1\}$.

Similarly to Baldi and Hornik (1989), we characterize the minima of f_a using the connexion between PCA and RRR, with a proof given in Appendix G.2.

Lemma 1. *The set of minima of f_a is*

$$\Omega_a^* := \left\{\tilde{I}R \mid R \in \mathcal{O}_r\right\} \quad \textit{with} \quad \tilde{I} := \begin{bmatrix} I_r \\ 0_{\ell-r,r} \end{bmatrix} \in \mathbb{R}^{\ell,r}.$$

In words, Ω_a^* is the image by the linear transformation $R \mapsto \tilde{I}R$ of the Stiefel manifold $\mathcal{O}_r := \left\{R \in \mathbb{R}^{r,r}, R^TR = I_r\right\}$. In particular, Ω_a^* has two connected components. We also classify the critical points of f_a in Appendix G.3 :

Lemma 2. *Rank-deficient matrices cannot be critical points of f_a. Critical points of f_a among full-rank matrices are differentiable points and either global minima or saddle points. Therefore, all local minima of f_a are global.*

5.2 Local strong convexity on cones

Although f_a is not convex even in the neighborhood of its minima, we will show that it is locally convex around them in the subspace orthogonal to the set of minima. For any $A \in \mathbb{R}^{p,r}$, let

$$\Pi_{\Omega_a^*}(A) := \operatorname*{argmin}_{B \in \Omega_a^*} \|B - A\|_F^2$$

be the closest minima to A, and for any $R \in \mathcal{O}_r$, let $\mathcal{C}_a(R)$ be defined as follows

$$\mathcal{C}_a(R) := \{A \in \mathbb{R}^{\ell,r} \mid \tilde{I}R \in \Pi_{\Omega_a^*}(A)\}.$$

$\mathcal{C}_a(R)$ is the set of points that are associated with the same minimum parameterized by $\tilde{I}R$. As shown in the following lemma, the sets $\mathcal{C}_a(R)$ are actually convex cones that are images of each other by orthogonal matrices; this result is essentially a consequence of the polar

decomposition and of the orthogonal invariance of f_a. Let $\mathcal{S}_r^+ \subset \mathbb{R}^{r,r}$ denote the set of positive-semidefinite matrices.

Lemma 3. *For each $R \in \mathcal{O}_r$, $\mathcal{C}_a(R)$ is a cone in $\mathbb{R}^{\ell,r}$ and*

$$\mathcal{C}_a(I_r) = \left\{ \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \mid A_1 \in \mathcal{S}_r^+,\, A_2 \in \mathbb{R}^{\ell-r,r} \right\}, \qquad (8)$$

$$\mathcal{C}_a(R) = \{AR \mid A \in \mathcal{C}_a(I_r)\} \;\; \textit{and} \bigcup_{R\in\mathcal{O}_r} \mathcal{C}_a(R) = \mathbb{R}^{\ell,r}.$$

Note that the cones $\mathcal{C}_a(R)$ do not form a partition of $\mathbb{R}^{\ell,r}$ because if A_1 is not invertible, its polar decomposition is not unique so $[A_1^T\ A_2^T]^T$ is in several cones. However the relative interiors of all the cones partition the set of matrices $[A_1^T\ A_2^T]^T$ such that A_1 is invertible (*cf.* Fact 51 in Appendix H.1). The decomposition on these cones is motivated by the fact that for $r \geq 2$, the function f_a in a neighborhood of each of the two connected components of Ω_a^* can be informally thought of as having the shape of the base of a glass bottle with a punt. This is illustrated on Figure 2.

Thus, given $R \in \mathbb{R}^{r,r}$, we focus on the restriction $f_a|_{\mathcal{C}_a(R)}$ of f_a on the cone $\mathcal{C}_a(R)$. The next result states in particular that $f_a|_{\mathcal{C}_a(R)}$ is smooth and strongly convex[2] in a neighborhood of $\tilde{I}R$.

Theorem 4. *For any $0 < \mu_a < s_\ell^2(1 - \frac{s_r^2}{s_{r+1}^2})$, there exist a non-empty sublevel set $\mathcal{V}_a \subset \mathbb{R}^{\ell,r}$ of f_a such that f_a is s_1^2-smooth in $\mathcal{V}_a$ and for any $R \in \mathcal{O}_r$, the restriction $f_a|_{\mathcal{C}_a(R)}$ is μ_a-strongly convex in $\mathcal{V}_a \cap \mathcal{C}_a(R)$.*

Via τ these properties of f_a carry over to f. Let ν_X and L_X be respectively the smallest and largest eigenvalues of X^TX and $\mathcal{C}(R) := \tau(\mathcal{C}_a(R), \mathbb{R}^{p-\ell,r})$ with τ defined in Equation (6).

Corollary 5. *For any $0 < \mu < \nu_X(1 - \frac{s_{r+1}^2}{s_r^2})$, there exist a non-empty sublevel set $\mathcal{V}^0$ of the function f that can be partitioned into disjoint convex elements $\{\mathcal{C}(R) \cap \mathcal{V}^0\}_{R\in\mathcal{O}_r}$ such that f is L_X-smooth on $\mathcal{V}^0$ and is μ-strongly convex on every $\mathcal{V}^0 \cap \mathcal{C}(R)$.*

To extend partially the previous result to (SRRR), we apply Theorem 6.4 of Bonnans and Shapiro (1998) : given that (a) the objective F^λ of (SRRR) is locally strongly convex on the cone $\mathcal{C}(I_r)$ around the minimum, (b) for every fixed λ in some interval $[0, \tilde{\lambda})$, f is locally Lipschitz with a constant that does not depend on λ and, (c) $F^\lambda - F^0 = \lambda\|\cdot\|_{1,2}$ is locally Lipschitz with a constant $\sqrt{p}\lambda$ which is $O(\lambda)$, then by Bonnans and Shapiro (1998, Theorem 6.4), there exists $\check{\lambda} > 0$ such that for all $0 \leq \lambda < \check{\lambda}$, the minimum of F^λ in $\mathcal{C}(I_r)$ is a continuous function of λ. This is detailed in Appendix H.4.

[2]The definitions of μ-strong convexity, L-smoothness and sublevel sets are recalled in Appendix B.

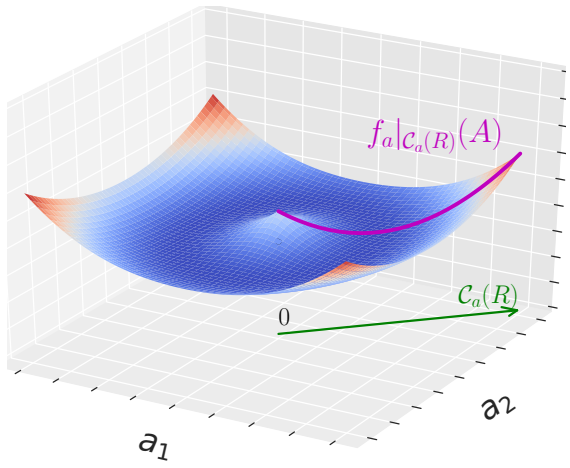

Figure 2: Schematic 2D graph of f_a around one of the connected components of Ω_a^* when $r \geq 2$. Here, the component of Ω_a^* is a circle and the cones are half-lines with extreme points at the origin.

Corollary 6. *There exists $\bar{\lambda} > 0$ such that for any $0 \leq \lambda < \bar{\lambda}$ and $0 \leq \mu < \nu_X(1 - \frac{s_{r+1}^2}{s_r^2})$, there exists a non-empty sublevel set $\mathcal{V}^\lambda$ of F^λ that can be partitioned into disjoint convex elements $\{\mathcal{C}(R) \cap \mathcal{V}^\lambda\}_{R\in\mathcal{O}_r}$ so that f is L_X-smooth on $\mathcal{V}^\lambda$ and F^λ is μ-strongly convex on every $\mathcal{C}(R) \cap \mathcal{V}^\lambda$.*

These characterizations of the geometry in a neighborhood of the optima immediately lead to Polyak-Łojasiewicz inequalities that entail the linear convergence of first-order algorithms.

5.3 Polyak-Łojasiewicz inequalities and proofs for linear convergence rates

Polyak-Łojasiewicz (PŁ) and Kurdyka-Łojasiewicz inequalities (KŁ) were introduced to generalize to non-convex functions (or just not strongly convex) proofs of rates of convergence for first-order methods (Attouch and Bolte, 2009; Karimi et al., 2016, and references therein). In particular, PŁ generalizes the fact that, for a differentiable and μ-strongly convex function f with optimal value f^*,

$$f(x) - f^* \leq \frac{1}{2\mu}\|\nabla f(x)\|^2. \qquad (\text{PŁ})$$

Karimi et al. (2016) and Csiba and Richtárik (2017) proposed a generalization to a *proximal* PŁ inequality of relevance for forward-backward algorithms applied to non-differentiable functions. In this section , we summa-

rize an immediate extension allowing a line search procedure, of results established for first-order algorithms to prove locally a linear rate of convergence. Consider $d \in \mathbb{N}^*$ and a function[3] $F^\lambda = f + \lambda h$ defined on $\mathbb{R}^d$ and with optimal value $F^{\lambda,*}$, where f is an L-smooth function and h is a proper lower semi-continuous convex function. We define the t-approximation $\tilde{f}_{t,x}$ and $\tilde{F}^\lambda_{t,x}$ of f and F^λ at x as in Section 3.3. The t-decrease function is defined as

$$\gamma_t(x) := -\frac{1}{t} \min_{x' \in \mathbb{R}^d} \left[\tilde{F}^\lambda_{t,x}(x') - F^\lambda(x) \right]. \tag{9}$$

Given x, assume that the minimum in Equation (9) is attained at a point x^+ for $t > 0$ such that the (LS) condition $\tilde{F}^\lambda_{t,x}(x^+) \geq F^\lambda(x^+)$ is satisfied. Then the decrease in the objective value $F^\lambda(x) - F^\lambda(x^+)$ is lower bounded by $t\gamma_t(x)$, hence the name t-decrease function (see Fact 33 in Appendix E.1). We make use of a natural generalization of the *proximal* PŁ inequality proposed by Karimi et al. (2016) and Csiba and Richtárik (2017). For x such that $F^\lambda(x) > F^{\lambda,*}$, with $F^{\lambda,*}$ the minimum of F^λ, we define the t-proximal forcing function :

$$\alpha_t(x) := \frac{\gamma_t(x)}{F^\lambda(x) - F^{\lambda,*}}.$$

We can now state the following theorem that bounds the optimal gap for our algorithm iteratively.

Theorem 7. *(From Lemma 13 in Csiba and Richtárik, 2017) Let $x \in \mathbb{R}^d$ and x^+ be defined by $x^+ = \operatorname{argmin}_{x'}[\tilde{F}^\lambda_{t,x}(x') - F^\lambda(x)]$, where t is chosen so that the line search condition (LS) is satisfied. Then we have*

$$F^\lambda(x^+) - F^{\lambda,*} \leq [1 - t\,\alpha_t(x)] \left[F^\lambda(x) - F^{\lambda,*} \right].$$

Given $t > 0$, we say that F^λ satisfies the (t-strong *proximal* PŁ) inequality in a set $\mathcal{V} \subset \mathbb{R}^d$ if there exists $\alpha(t) > 0$ such that for any $x \in \mathcal{V}$ where $F^\lambda(x) > F^{\lambda,*}$, we have

$$\alpha_t(x) \geq \alpha(t). \qquad \text{(t-strong } \textit{proximal} \text{ PŁ)}$$

If $\lambda h = 0$, then $\gamma_t(x) = \frac{1}{2}\|\nabla f(x)\|^2$ and it is easy to see that (t-strong *proximal* PŁ) boils down to (PŁ) with $\mu = \alpha(t)$.

5.4 Proving local linear convergence

We now return to the functions f and F^λ defined for (RRR) and (SRRR) with minimal values f^* and $F^{\lambda,*}$, and we establish the (PŁ) and (t-strong *proximal* PŁ) inequalities in a neighborhood of their respective global minima.

[3] In this section we use a general variable x but we keep using f and F^λ.

Corollary 8. *Let $0 \leq \mu < \nu_X(1 - \frac{s^2_{r+1}}{s^2_r})$ and $\mathcal{V}^0$ as in Corollary 5. For all $U \in \mathcal{V}^0$, $f(U) - f^* \leq \frac{1}{2\mu}\|\nabla f(U)\|^2_F$.*

In light of Corollary 6, we can also prove the (t-strong *proximal* PŁ) inequality for F^λ with small values of λ. To this end, we consider $\bar{\lambda} > 0$ as in Corollary 6.

Corollary 9. *Let $0 \leq \mu < \nu_X(1 - \frac{s^2_{r+1}}{s^2_r})$ and $0 \leq \lambda < \bar{\lambda}$. For any $t > 0$, F^λ satisfies the (t-strong* proximal *PŁ) inequality with $\alpha(t) := \min(\frac{1}{2t}, \mu)$. In other words, for any $t > 0$ and $U \in \mathcal{V}^\lambda$, we have*

$$\gamma_t(U) \geq \alpha(t) \left[F^\lambda(U) - F^{\lambda,*} \right]$$

$$\textit{with } \gamma_t(U) := -\frac{1}{t} \min_{U' \in \mathbb{R}^{p,r}} \left[\tilde{F}^\lambda_{t,U}(U') - F^\lambda(U) \right].$$

So, leveraging Theorem 7 and Corollary 8/9 for (RRR)/(SRRR), we obtain the linear rate of convergence which is proved in Appendix J.3. Indeed, if L_X denotes the largest eigenvalue of X^TX and β the step-size decrease factor in Algorithm 2, then we have the following result :

Corollary 10. *Let $0 \leq \lambda < \bar{\lambda}$ and $k \geq 0$. Assume that $t_{k-1} > \frac{\beta}{L_X}$ and U_{k+1} is generated as in Algorithm 1 from $U_k \in \mathcal{V}^\lambda$. Then $U_{k+1} \in \mathcal{V}^\lambda$, $t_k > \frac{\beta}{L_X}$ and denoting $\rho = \min(\frac{1}{2}, \beta\frac{\mu}{L_X})$, we have*

$$F^\lambda(U_{k+1}) - F^{\lambda,*} \leq (1 - \rho) \left[F^\lambda(U_k) - F^{\lambda,*} \right].$$

As explained in Fact 35 in Appendix E.2, there is only a finite number of steps at the beginning of Algorithm 1 for which the assumption $t_k > \frac{\beta}{L_X}$ may fail. The convergence is therefore linear. We propose a direct proof of Corollary 10 based on Corollary 9 and Theorem 7. It should be noted that the geometric structure leveraged for Corollary 9 can also be used to obtain Corollary 10 as a consequence of the Kurdyka-Łojasiewicz inequality (*cf.* Appendix L).

6 Experiments on RRR and SRRR

We perform experiments on simulated data both for RRR and SRRR to assess the performance of the algorithms in terms of speed.

For RRR, we compare gradient descent algorithms in U space and in (U, V) space. In the former case, we just minimize (RRR), whereas in the latter, following Park et al. (2016), we minimize $\mathcal{F}(UV^\top) + g(U, V)$, with $\mathcal{F}(W) = \frac{1}{2}\|Y - XW\|^2_F$ and $g(U, V) = \frac{1}{4}\|U^\top U - V^\top V\|^2_F$; this objective has the same optimal value as $\mathcal{F}(UV^\top)$, but the regularizer g was shown to improve the convergence rate of the algorithm (see Appendix M.1). Note that the formulation

of Park et al. (2016) does not apply to SRRR because the regularizer g is not compatible with the use of the group-Lasso norm.

For SRRR, we implement proximal gradient descent algorithms and compare in speed with the RRR case and with the alternating optimization algorithm proposed[4] in Bunea et al. (2012). In each case we consider variants of these first-order methods with and without line search.

For the alternated procedure, each inner minimization of the matrix U is stopped when a duality gap becomes smaller than the desired precision 10^{-4}. Since it takes more than seconds to optimize, it justifies the relevance of RRR/SRRR.

As in Bunea et al. (2012), we sample the rows of X from a zero-mean Gaussian with a Toeplitz covariance matrix Σ where $\Sigma_{i,j} = \rho^{|i-j|}$ and $\rho \in (0, 1)$. We set $n = 10^3$, $p = 300$ and $k = 200$. We let $W_0 = U_0 V_0^\top$ for $U_0 \in \mathbb{R}^{p,r}$ and $V_0 \in \mathbb{R}^{k,r}$ uniformly drawn from the set of orthonormal matrices with $r_0 = 30$ columns. For SRRR, each row of W_0 is then set to zero with probability p_0. Then we compute $Y = XW_0 + E$ for E a matrix of i.i.d. centered scalar Gaussians with standard deviation $\sigma = 0.1$. Finally, we solve all formulations for a matrix W of rank $r = 20$. For all algorithms, we initialize U (and V if relevant) at random.

We report results for $\rho = 0.6$ in Figure 3 and in Appendix M for additional values of ρ and p_0. For RRR, these results show that the algorithms based on our proposed formulation are significantly faster, both in terms of the number of function/gradient evaluations and in terms of time; moreover they benefit more from the line search. We do not report curves with both line search and acceleration because this combination does not yield any speed increase.

For (SRRR) and (RRR) all algorithms exhibit at least linear convergence. Compared with (RRR), the convergence for (SRRR) typically seems as fast or faster. Additionally, the line search plays a significant role in accelerating the convergence of the algorithm.

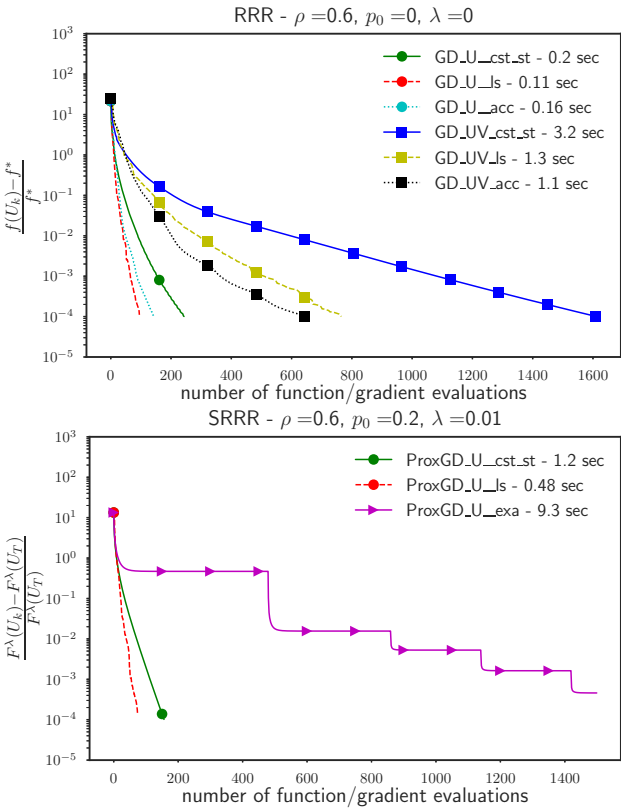

Figure 3: *(Top)* RRR : Convergence of $f(U_k) - f^*$ for gradient descent on our formulation in U with constant step size (`GD_U_cst_st`), with line search (`GD_U_ls`), with the acceleration (`GD_U_acc`) proposed by Li and Lin (2015) and gradient descent for the formulation of (Park et al., 2016) with constant step size, line search and acceleration (`GD_UV_cst_st`, `GD_UV_ls`, `GD_UV_acc`). *(Bottom)* SRRR with $\lambda = 0.01$: Convergence for T large of $F^\lambda(U_k) - F^\lambda(U_T)$ for proximal gradient descent on our formulation with and without line search (`ProxGD_U__ls`, `ProxGD_U__cst_st`), compared with the alternating optimization algorithm (`ProxGD_U__exa`) proposed in Bunea et al. (2012). The running time to reach a precision of 10^{-4} is given at the top right.

Conclusion

We considered a reformulation of RRR and SRRR problems as non-convex and non-differentiable optimization problems w.r.t. to a matrix U with r columns. We proposed to apply subgradient-type algorithms proposed by Khamaru and Wainwright (2018), which correspond essentially to gradient descent for RRR and proximal gradient descent for SRRR.

The algorithms are provably convergent to critical points under reasonable assumptions. We show that for a certain range of regularization coefficients λ the objective satisfies a Polyak-Łojasiewicz inequality in a neighborhood of the global minima, which entails local linear convergence if the algorithm converges to them.

For RRR, gradient descent converges to a critical point and if a global minimum of the original objective has been found, it can easily be certified.

Future work could try to determine if convergence to saddle points of SRRR can be excluded and if global linear convergence results can be obtained. Another interesting direction of research is to extend these types of results to other matrix optimization problems with low-rank constraints.

[4] Ma and Sun (2014) consider a similar algorithm.

Acknowledgements

The authors would like to thank Virgine Dordonnat and Vincent Lefieux for useful discussions on this work. This research is funded by RTE France.

References

Attouch, H. and Bolte, J. (2009). On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16.

Attouch, H., Bolte, J., Redont, P., and Soubeyran, A. (2010). Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457.

Attouch, H., Bolte, J., and Svaiter, B. F. (2013). Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods. *Mathematical Programming*, 137(1-2):91–129.

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

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

Bolte, J., Sabach, S., and Teboulle, M. (2014). Proximal alternating linearized minimization or nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494.

Bonnans, J. F. and Shapiro, A. (1998). Optimization problems with perturbations: A guided tour. *SIAM review*, 40(2):228–264.

Boumal, N., Voroninski, V., and Bandeira, A. (2016). The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pages 2757–2765.

Bunea, F., She, Y., and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, pages 1282–1309.

Bunea, F., She, Y., Wegkamp, M. H., et al. (2012). Joint variable and rank selection for parsimonious estimation of high-dimensional matrices. *The Annals of Statistics*, 40(5):2359–2388.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.

Chouzenoux, E., Pesquet, J.-C., and Repetti, A. (2014). Variable metric forward-backward algorithm for minimizing the sum of a differentiable function and a convex function. *Journal of Optimization Theory and Applications*, 162(1):107–132.

Csiba, D. and Richtárik, P. (2017). Global convergence of arbitrary-block gradient methods for generalized Polyak-Łojasiewicz functions. *arXiv preprint arXiv:1709.03014*.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. (2017). Gradient descent can take exponential time to escape saddle points. In *Advances in Neural Information Processing Systems*, pages 1067–1077.

Frankel, P., Garrigos, G., and Peypouquet, J. (2015). Splitting methods with variable metric for Kurdyka-Łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900.

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

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

Jain, P., Jin, C., Kakade, S., and Netrapalli, P. (2017). Global convergence of non-convex gradient descent for computing matrix squareroot. In *Artificial Intelligence and Statistics*, pages 479–488.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. (2017). How to escape saddle points efficiently. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1724–1732.

Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer.

Kawaguchi, K. (2016). Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, pages 586–594.

Khamaru, K. and Wainwright, M. (2018). Convergence guarantees for a class of non-convex and non-smooth optimization problems. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2601–2610.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. (2017). First-order methods almost always avoid saddle points. *arXiv preprint arXiv:1710.07406*.

Li, G. and Pong, T. K. (2017). Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of Computational Mathematics*, pages 1–34.

Li, H. and Lin, Z. (2015). Accelerated proximal gradient methods for nonconvex programming. In *Advances in neural information processing systems*, pages 379–387.

Li, Q., Zhu, Z., and Tang, G. (2017). Geometry of factored nuclear norm regularization. *arXiv preprint arXiv:1704.01265*.

Li, Q., Zhu, Z., and Tang, G. (2018). The non-convex geometry of low-rank matrix optimization. *Information and Inference: A Journal of the IMA*.

Li, X., Wang, Z., Lu, J., Arora, R., Haupt, J., Liu, H., and Zhao, T. (2016). Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296*.

Ma, Z. and Sun, T. (2014). Adaptive sparse reduced-rank regression. *arXiv*, 1403.

Mukherjee, A., Chen, K., Wang, N., and Zhu, J. (2015). On the degrees of freedom of reduced-rank estimators in multivariate regression. *Biometrika*, 102(2):457–477.

Nikolova, M. and Tan, P. (2017). Alternating proximal gradient descent for nonconvex regularised problems with multiconvex coupling terms. *HAL-01492846, 2017*.

Ochs, P., Chen, Y., Brox, T., and Pock, T. (2014). iPiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419.

Panageas, I. and Piliouras, G. (2016). Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *arXiv preprint arXiv:1605.00405*.

Park, D., Kyrillidis, A., Caramanis, C., and Sanghavi, S. (2016). Finding low-rank solutions via non-convex matrix factorization, efficiently and provably. *arXiv preprint arXiv:1606.03168*.

Polyak, B. T. (1963). Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653.

She, Y. (2017). Selective factor extraction in high dimensions. *Biometrika*, 104(1):97–110.

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

Velu, R. and Reinsel, G. C. (2013). *Multivariate reduced-rank regression: theory and applications*, volume 136. Springer Science & Business Media.

Wang, L., Zhang, X., and Gu, Q. (2016). A unified computational and statistical framework for nonconvex low-rank matrix estimation. *arXiv preprint arXiv:1610.05275*.

Xu, Y. and Yin, W. (2017). A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72(2):700–734.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. (2017a). The global optimization geometry of low rank matrix optimization. *arXiv preprint arXiv:1703.01256*.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. (2017b). The global optimization geometry of nonsymmetric matrix factorization and sensing. *arXiv preprint arXiv:1703.01256*.