Fast Algorithms for Sparse Reduced-Rank Regression
Benjamin Dubois, Jean-François Delmas, Guillaume Obozinski

To cite this version:
Benjamin Dubois, Jean-François Delmas, Guillaume Obozinski. Fast Algorithms for Sparse Reduced-Rank Regression. Proceedings of Machine Learning Research, In press. hal-02075623

HAL Id: hal-02075623
https://enpc.hal.science/hal-02075623
Submitted on 22 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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.

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 the Frobenius norm as

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

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

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

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

$$\min_{W \in \mathbb{R}^{p \times k}: \text{rank}(W) \leq r} \mathcal{F}(W)$$

(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 $\|W\|_\tau$; 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

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

(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 9), we establish in Corollary 9 a generalized Polyak-Łojasiewicz 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 (2013); 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 (1) (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 $Y^TXU \in \mathbb{R}^{k \times r}$ 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 (1) provided an appropriate initialization is used and penalties such as $\frac{1}{2}\|U^TU - V^TV\|_F^2$ 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); Bousmal 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 Cibari and Richtarik (2017) to prove convergence rates. A parallel thread of research focused on the Kurdyka-Łojasiewicz inequality (KL), 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 KL 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 Tai (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 KL 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

\[
\min_{U \in \mathbb{R}^{p \times r}, V \in \mathbb{R}^{k \times r}} \frac{1}{2} \|Y - XUV^T\|_F^2 + \lambda \|UV^T\|_{1,2},
\]

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 \( V \in \mathbb{R}^{k \times r} \) such that \( V^TV = I_r \), the objective becomes \( \frac{1}{2} \|XU\|_F^2 - \langle Y, XUV^T \rangle + \lambda \|UV^T\|_{1,2} \) where \( \langle \cdot, \cdot \rangle \) is the Frobenius inner product. The value of the orthogonal Procrustes problem

\[
\max_{V \in \mathbb{R}^{k \times r}, VV^T = I_r} \langle Y, XUV^T \rangle
\]

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

\[
f_1(U) = \frac{1}{2} \|XU\|_F^2 \quad \text{and} \quad f_2(U) = \|Y^TXU\|_s
\]

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

\[
\min_{U \in \mathbb{R}^{p \times r}} f(U),
\]

(RRR)

The objectives, as differences of convex functions, are clearly non-convex. However, they are still orthogonal-invariant i.e. for any \( U \in \mathbb{R}^{p \times r} \) and \( R \in \mathbb{R}^{r \times r} \) 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 \( \|W\|_s \), 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 \( PSQT \) be the reduced singular value decomposition of \((X^TX)^{-1}X^TY\). If the latter has rank \( \ell \) then \( P \in \mathbb{R}^{p \times \ell} \) and \( Q \in \mathbb{R}^{k \times \ell} \) have orthonormal columns and \( S \in \mathbb{R}^{\ell \times \ell} \) 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 \( Q_r \in \mathbb{R}^{k \times r} \) be the matrix obtained by keeping the first \( r \) columns of \( Q \), the solution is \( W^* = W^*Q_rQ_r^T \).

3.3 Algorithms and complexity

The algorithms that we consider are essentially subgradient 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 \( U \in \mathbb{R}^{p \times r} \), the idea is to use a subgradient \( z_t \) of \( f_2 \). We assume that \( X^TX \) is invertible but consider a more general case in Appendix D where we detail the computations. Given \( R_1, R_2 \) a singular value decomposition of \( Y^TX \) such that \( \text{Im} R_1 \subset \text{Im} X^TX \), we compute \( z_t = X^TYR_1R_2^T \) with \( R_1 \in \mathbb{R}^{k \times r} \), \( R_2^TR_2 = I_r \), \( D = \text{diag}(d_1 \geq \ldots \geq d_r) \in \mathbb{R}^{r \times r} \) with \( d_r \geq 0 \) and \( R_2 \in O_r \). With a slight abuse of notation, we define \( \nabla f(U) := \nabla f_1(U) - z_t \). 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 \( F^\lambda \) at \( U \),
Fast Algorithms for Sparse Reduced-Rank Regression

\[ \hat{f}_{t,U}(U') := f(U) + \langle \nabla f(U), U' - U \rangle + \frac{1}{2\lambda} \| U' - U \|_F^2 \]
and \[ \tilde{F}_t^\lambda(U') := \hat{f}_{t,U}(U') + \lambda \| U' \|_{1,2}^2 \]. At each iteration of Algorithm 1, \( U \) is updated with Algorithm 2 to \( U_+ \), the unique minimizer of \( \tilde{F}_t^\lambda(U) \) if the line search condition
\[ \tilde{F}_t^\lambda(U_+) \geq F^\lambda(U) \]
is satisfied. Otherwise, \( t \) is decreased by a multiplicative factor \( \beta < 1 \). We explain why Algorithm 2 terminates in Appendix D.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 \( X, Y, t, \) starting point \( U \)
Initialize \( k = 0, U_0 \leftarrow U, t_{-1} \leftarrow t \)
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 \( t \leftarrow \frac{t_{k-1}}{\pi} \) with probability \( \pi \), o/w \( t \leftarrow t_{k-1} \)
\( U_+ \leftarrow \arg\min_{U'} \tilde{F}_t^\lambda(U') \)
while \( \text{(LS)} \) is not satisfied do
\( t \leftarrow \beta t \)
\( U_+ \leftarrow \arg\min_{U'} \tilde{F}_t^\lambda(U') \)
end while
Output: \( t, U_+ \)

To discuss the complexity of the algorithm, we assume that \( X^T X \) and \( Y^T X \) are computed in advance. Although the computation of \( X^T X \) requires an SVD of \( Y^T X \), the latter costs only \( O(kr^2) \). Computing \( \nabla f(U) \) has then a complexity of \( O(p^2 r + pkr) \). 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 \( -\frac{1}{2} \| Y \|_2^2 \). 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 \( X^T X \) 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 \( X^T Y \) 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 \( U \in \mathbb{R}^{p \times r} \) is such that \( Y^T X U \) 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 $\lambda$ for SRRR. Precisely, we first study the geometry around the optimum 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 [Vieu 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 $(X^TX)^{-\frac{1}{2}}XY$. From now on, we assume that the rank parameter $r$ is smaller than the rank of $X^TY$ i.e. $r \leq \ell := \text{rank}(X^TY)$. 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 that $s_1 > \ldots > s_\ell$ and that $X^TX$ is invertible.

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

$$
\tau \colon \mathbb{R}^{\ell \times r} \times \mathbb{R}^{p-r, r} \to \mathbb{R}^{\ell \times r}, \\
(A, C) \mapsto (X^TX)^{-\frac{1}{2}}(PSA + P^\perp C) \cdot
$$

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

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

Since $\tau$ 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 $f_a$.

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

$$
\Omega_a := \{ \tilde{I}R \mid R \in \mathcal{O}_r \} \quad \text{with} \quad \tilde{I} := \left[ I_r \right] \in \mathbb{R}^{\ell \times r}.
$$

In words, $\Omega_a$ is the image by the linear transformation $R \mapsto \tilde{I}R$ of the Stiefel manifold $\mathcal{O}_r := \{ R \in \mathbb{R}^{r \times r}, R^T R = I_r \}$. In particular, $\Omega_a$ 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 $A \in \mathbb{R}^{p \times r}$, let

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

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

$$
C_a(R) := \{ A \in \mathbb{R}^{\ell \times r} \mid \tilde{I}R \in \Pi_{\Omega_a^*}(A) \}.
$$

$C_a(R)$ 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 $C_a(R)$ are actually convex cones that are images of each other by orthogonal matrices; this result is essentially a consequence of the polar
Fast Algorithms for Sparse Reduced-Rank Regression

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

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

$$C_a(I_r) = \left\{ \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \mid A_1 \in S^+_r, A_2 \in \mathbb{R}^{r-r,r} \right\}, \quad (8)$$

$$C_a(R) = \{ AR \mid A \in C_a(I_r) \} \quad \text{and} \quad \bigcup_{R \in \mathcal{O}_r} C_a(R) = \mathbb{R}^{r,r}.$$ 

Note that the cones $C_a(R)$ do not form a partition of $\mathbb{R}^{r,r}$ 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 5 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 $\Omega_a^*$ 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 $f_a|_{C_a(R)}$ of $f_a$ on the cone $C_a(R)$. The next result states in particular that $f_a|_{C_a(R)}$ is smooth and strongly convex in a neighborhood of $I_R$.

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

Via $\tau$ these properties of $f_a$ carry over to $f$. Let $\nu_X$ and $L_X$ be respectively the smallest and largest eigenvalues of $X^T X$ and $C(T) := \tau(C_a(R),\mathbb{R}^{r-r,r})$ with $\tau$ defined in Equation 9.

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

To extend partially the previous result to SRRR, we apply Theorem 6.4 of Bonnans and Shapiro (1998) \footnote{The definitions of $\mu$-strong convexity, $L$-smoothness and sublevel sets are recalled in Appendix B} given that (a) the objective $F^\lambda$ of SRRR is locally strongly convex on the cone $C(I_r)$ around the minimum, (b) for every fixed $\lambda$ in some interval $[0, \lambda_f)$, $f$ is locally Lipschitz with a constant that does not depend on $\lambda$ and, (c) $F^\lambda - F^0 = \lambda \left\| \cdot \right\|_{1,2}$ is locally Lipschitz with a constant $\sqrt{\lambda}$ which is $O(\lambda)$, then by Bonnans and Shapiro (1998, Theorem 6.4), there exists $\lambda_f > 0$ such that for all $0 \leq \lambda < \lambda_f$, the minimum of $F^\lambda$ in $C(I_r)$ is a continuous function of $\lambda$. This is detailed in Appendix H.3.

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

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 (PL) and Kurdyka-Łojasiewicz inequalities (KL) 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, PL generalizes the fact that, for a differentiable and $\mu$-strongly convex function $f$ with optimal value $f^*$,

$$f(x) - f^* \leq \frac{1}{2\mu} \left\| \nabla f(x) \right\|^2. \quad \text{(PL)}$$

Karimi et al. (2016) and Csiba and Richtárik (2017) proposed a generalization to a proximal PL inequality of relevance of 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 \( F^\lambda = f + \lambda h \) 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 \( \tilde{F}_{t,x}^\lambda \) of \( f \) and \( F^\lambda \) at \( x \) as in Section 3.3. The \( t \)-decrease function is defined as

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

(9)

Given \( x \), assume that the minimum in Equation (9) is attained at a point \( x^+ \) for \( t > 0 \) such that the (LS) condition \( \tilde{F}_{t,x}^\lambda(x^+) \geq F^\lambda(x^+) \) is satisfied. Then the decrease in the objective value \( F^\lambda(x^+) - F^\lambda(x^+) \) is lower bounded by \( \gamma_t(x) \), hence the name \( t \)-decrease function (see Fact 35 in Appendix E.1). We make use of a natural generalization of the proximal PL 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 \( F^\lambda \), 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 \( x^+ = \arg\min_x \tilde{F}_{t,x}^\lambda(x') - F^\lambda(x) \), 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 \( F^\lambda \) satisfies the \( t \)-strong proximal PL inequality in a set \( V \subset \mathbb{R}^d \) if there exists \( \alpha(t) > 0 \) such that for any \( x \in V \) where \( F^\lambda(x) > F^\lambda \), we have

\[
\alpha_t(x) \geq \alpha(t). 
\]  

(t-strong proximal PL)

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

### 5.4 Proving local linear convergence

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

---

**Corollary 8.** Let \( 0 \leq \mu < \nu X (1 - \frac{\mu^2}{\nu^2}) \) and \( \nu^0 \) as in Corollary 5. For all \( U \in \mathbb{V}^0 \), \( f(U) - f^* \leq \frac{1}{2\nu} \| \nabla f(U) \|_2^2 \).

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

**Corollary 9.** Let \( 0 \leq \mu < \nu X (1 - \frac{\mu^2}{\nu^2}) \) and \( 0 \leq \lambda < \lambda \). For any \( t > 0 \), \( F^\lambda \) satisfies the \( t \)-strong proximal PL inequality with \( \alpha(t) := \min(\frac{1}{2t}, \mu) \).

In other words, for any \( t > 0 \) and \( U \in \mathbb{V}^\lambda \), we have

\[
\gamma_t(U) \geq \alpha(t) [F^\lambda(U) - F^\lambda]\]

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

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

**Corollary 10.** Let \( 0 \leq \lambda < \lambda \) and \( k \geq 0 \). Assume that \( t_{k-1} > \frac{\beta}{L_X} \) and \( U_{k+1} \) is generated as in Algorithm 4 from \( U_k \in \mathbb{V}^\lambda \). Then \( U_{k+1} \in \mathbb{V}^\lambda \), \( t_k > \frac{\beta}{L_X} \) and denoting \( \rho := \min(\frac{1}{2}, \beta \frac{L_X}{\lambda}) \), we have

\[
F^\lambda(U_{k+1}) - F^\lambda \leq (1 - \rho) [F^\lambda(U_k) - F^\lambda].
\]

As explained in Fact 35 in Appendix E.2, there is only a finite number of steps at the beginning of Algorithm 4 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-Lojasiewicz 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 \( F(UV^\top) + g(U, V) \), with \( F(W) = \frac{\nu}{2} \| Y - XV \|_F^2 \) and \( g(U, V) = \frac{\nu}{2} \| U^\top U - V^\top V \|_F^2 \); this objective has the same optimal value as \( F(UV^\top) \), but the regularizer \( g \) was shown to improve the convergence rate of the algorithm (see Appendix M.1). Note that the formulation
Fast Algorithms for Sparse Reduced-Rank Regression

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 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 $\Sigma$ where $\Sigma_{i,j} = \rho^{|i-j|}$ and $\rho \in (0,1)$. We set $n = 10^4$, $p = 300$ and $k = 200$. We let $W_0 = U_0 V_0^\top$ for $U_0 \in \mathbb{R}^{p \times r}$ and $V_0 \in \mathbb{R}^{k \times r}$ 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 C for additional values of $\rho$ 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.

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 Hachem et al. (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 $\lambda$ 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.
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


