Journées SMAI MODE 2018

A Fast Algorithm for Sparse Reduced-Rank Regression

Benjamin DUBOIS
Université Paris-Est, LIGM
École des Ponts - ParisTech
Marne-la-Vallée, France

Jean-François Delmas
Université Paris-Est, CERMICS
École des Ponts - ParisTech
Marne-la-Vallée, France

Guillaume Obozinski
Université Paris-Est, LIGM
École des Ponts - ParisTech
Marne-la-Vallée, France

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.

Abstract. In this work, we consider a descent algorithm for the Sparse Reduced-Rank Regression problem. A recent litterature revisited such nonconvex problems based on an explicit parametrization of the low-rank matrix. However, no general convergence rate result was provided, in particular not in the nondifferentiable case. Reformulating the minimization problem and analyzing its geometry in a neighborhood of the optimal set, we show the Polyak-Łojasiewicz inequality or its extension to the nondifferentiable case are satisfied. Consequently, we establish the linear convergence of the proximal block-coordinate gradient algorithm in this neighborhood.

Key words : Sparse Reduced-Rank Regression, nonconvex optimization, biconvex formulation, proximal gradient descent, Polyak-Łojasiewicz inequality.

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 one particular instance of these : it corresponds to the multiple output linear regression in which all the parameter vectors associated to the different dimensions are constrained to lie in a low-dimensional space. It is formulated as

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

where $Y \in \mathbb{R}^{n,k}$ and $X \in \mathbb{R}^{n,p}$. As described by Velu and Reinsel (2013), (RRR) is one of the few low-rank matrix problems whose solution has an analytical form. Assuming X^TX is invertible, let $W^* = (X^TX)^{-1}X^TY$ denote the full-rank least squares estimator, PSQ^T the

singular value decomposition of $(X^T X)^{-1/2} X^T Y$ and Q_r the first r columns of Q. The solution to (RRR) is then $W_r^* = W^* Q_r Q_r^T$. However this solution is computationally expensive.
A variant of interest called Sparse Reduced-Rank Regression (SRRR) is the problem in which a sparsity-inducing penalty $\|W\|_{1,2} = \sum_i (\sum_j W_{i,j}^2)^{1/2}$ is added as a regularizer[1]. Doing so, one loses the analytical form for the solution. Yet, Bunea et al. (2011, 2012); Chen et al. (2012); Chen and Huang (2012); Ma and Sun (2014); Mukherjee et al. (2015) studied the statistical properties of this estimator and gave numerical algorithms to compute it. In spite of their relative simplicity, these algorithms are costly in the high-dimensional setting.
In the last decade, many optimization problem of the form $\min_{W:\ \mathrm{rank}(W) \leq r} F_w(W)$ with F_w convex have been tackled via the convex relaxation version obtained by replacing the rank constraint with a contraint or a regularization on the trace-norm $\|W\|_*$; these formulations however lead to expensive algorithms and the relaxation induces a bias, typically. A recent literature revisited a number of these problems based on an explicit parameterization of the low-rank matrix, which yields biconvex problems of the form

$$\min_{U \in \mathbb{R}^{p,r},\, V \in \mathbb{R}^{k,r}} F_w(UV^T). \tag{1}$$

In particular, it is possible to reformulate (RRR) in that form.
Among others, iterative first-order algorithms that are classical for the jointly convex setting may be applied to the nonconvex problem (1). A number of recent papers have established stronger theoretical guarantees for these algorithms in the smooth nonconvex case. In particular Park et al. (2016) and Wang et al. (2016) establish convergence rate guarantees, provided an appropriate initialization is used and penalties such as $\frac{1}{4}\|U^T U - V^T V\|_F^2$ are added to the objective as regularizers. Several papers considered the geometry of such problems (Li et al., 2016; Li and Tang, 2016) but do not provide general convergence rate results, in particular not in the non-differentiable case.
In this work, we reformulate the minimization problem (1) in the SRRR setup and apply tools specifically designed for the study of nonconvex problems to analyze the rate of convergence of classical descent methods. Indeed, we propose to apply either a forward-backward algorithm with line-search or possibly block coordinate descent. Despite the lack of strong convexity, there's still hope for an asymptotical linear convergence rate as long as a Polyak-Łojasiewicz-like condition (Polyak, 1963; Karimi et al., 2016) is satisfied in a neighborhood of the optima.
We focus on the objective $F_w(UV^T) = \frac{1}{2}\|Y - XUV^T\|_F^2 + \lambda \|UV^T\|_{1,2}$. First, following Chen and Huang (2016), we impose $V^T V = I_r$. Secondly, expanding the Frobenius norm and using the orthogonal invariance of both $\|\,.\,\|_F^2$ and $\|\,.\,\|_{1,2}$, we obtain :

$$\min_{U,V,\, V^T V = I_r} \frac{1}{2}\|XU\|_F^2 - \langle Y, XUV^T \rangle + \lambda \|U\|_{1,2} \tag{2}$$

Recognizing a Procrustres problem (Higham and Papadimitriou, 1995) $\max_{V,\, V^T V = I_r} \langle Y, XUV^T \rangle = \|Y^T XU\|_*$ where $\|\,.\,\|_*$ is the trace-norm, we can reformulate the problem as :

$$\min_U f(U) + \lambda \|U\|_{1,2} \text{ where } f(U) = \frac{1}{2}\|XU\|_F^2 - \|Y^T XU\|_*. \tag{SRRR}$$

The objective is invariant to the transformation $U \leftarrow UR$ where R belongs to the Stiefel manifold $\mathcal{O}_r = \left\{R \in \mathbb{R}^{r,r}, R^T R = I_r\right\}$. Besides, note that the trace-norm of $Y^T XU$ which is of

[1] The results we show could be extended for any penalty ψ that satisfies $\psi(MR) = \psi(M)$ for all M, R such that $R^T R = I$.

dimension (k, r) is much cheaper to compute than for the matrix W of dimension (p, k) for the convex relaxation. To keep the discussion simple, we assume from now on that $X^T X$ is invertible. In order to apply a proximal gradient algorithm, we get rid of the non-differentiablity due to the trace-norm by applying Nesterov smoothing $\|Y^T XU\|_* \leftarrow \inf_{U'} \|Y^T XU'\|_* + \frac{\epsilon}{2}\|U' - U\|_{F^2}$ for a small $\epsilon > 0$. Assuming the spectrum of $Y^T XU$ is bounded below by a positive constant in a neighborhood of the optima, this transformation results only in the addition of a constant term around the set of optima. This assumption seems reasonable as it is true for $\lambda = 0$ whenever $r \leq \text{rank}(X^T Y)$ and we can show the set of optima varies continuously as a function of λ (Bonnans and Shapiro, 1998, Thm. 6.4).

Slightly modifying the results Baldi and Hornik (1989) obtained for the biconvex formulation of RRR we identify the set of optima Ω_0^* for (SRRR) when $\lambda = 0$. Let's define $\tilde{I} = (1_{i=j})_{i,j} \in \mathbb{R}^{\ell,r}$. The set Ω_0^* is the image by a linear transformation from $\mathbb{R}^{r,r}$ to $\mathbb{R}^{p,r}$ of $\mathcal{O}_r$:

$$\Omega_0^* = \left\{ (X^T X)^{-\frac{1}{2}} PS\tilde{I}R, R \in \mathcal{O}_r \right\} \tag{3}$$

where PSQ^T is the singular value decomposition of $(X^T X)^{-1/2} X^T Y$. While f is not convex, computing its Hessian, we obtain :

Theorem 1. *There exists $L > \mu > 0$ and a sublevel set $\mathcal{V}$ of the function f that can be partitioned into disjoint convex elements $\mathcal{V}_R$ such that $\mathcal{V} = \cup_{R \in \mathcal{O}_r} \mathcal{V}_R$, f is L-smooth on $\mathcal{V}$ and the restriction of f on each $\mathcal{V}_R$ is μ-strongly convex.*

One can show the set of optima is continuously modified (Bonnans and Shapiro, 1998, Thm. 6.4) as a function of λ. Consequently, there exists $\bar{\lambda}$ such that for $\lambda < \bar{\lambda}$, the optima stay in $\mathcal{V}$ and the previous results holds for (SRRR) as well. Let's fix such a λ. A direct corollary of Theorem 1 is as follows:

Corollary 2. *Let f^* and F^* denote respectively the global minima of f and F. For all $U \in \mathcal{V}$, f satisfies the Polyak-Łojasiewicz inequality:*

$$\frac{1}{2}\|\nabla f(U)\|_F^2 \geq \mu(f(U) - f^*). \tag{4}$$

Moreover, there is μ_λ and a neighborhood $\mathcal{V}_\lambda$ of the optima for problem (SRRR) such that for all $U \in \mathcal{V}_\lambda$, f satisfies the following so-called proximal Polyak-Łojasiewicz inequality:

$$-L \min_{U'} [\langle \nabla f(U), U' - U \rangle + \frac{L}{2}\|U' - U\|^2 + \lambda\|U'\|_{1,2} - \lambda\|U\|_{1,2}] \geq \mu_\lambda(F(U) - F^*) \tag{5}$$

We propose a direct proof of Corollary 2 based on Theorem 1; it should be noted that the geometric structure leveraged in Theorem 1 can also be used to obtain the first part (4) of Corollary 2 as a consequence of Theorem 3.2 in Li and Pong (2017).

The proximal generalization of the Polyak-Łojasiewicz inequality proven in this corollary was considered in Karimi et al. (2016) and Csiba and Richtárik (2017) to establish linear convergence of proximal gradient and proximal block-coordinate gradient algorithms.

In particular, using the framework of Csiba and Richtárik (2017), that we slightly extend to allow for line-searches, we obtain a bound at each iteration on the optimal gap between the values of F. In particular, given a line search parameter β such that $0 < \beta < 1$, and given a set of rows to update $\mathcal{S}_k \subset [\![1, r]\!]$, the chosen algorithm finds at iteration k a step-size $t_k > \frac{\beta}{L}$ and U_{k+1} such that

$$U_{k+1} = \underset{U', \forall i \notin \mathcal{S}_k, U'_{i,:} = U_{i,:}}{\operatorname{argmin}} f(U) + \langle \nabla f(U), U' - U \rangle + \frac{1}{2t} \|U' - U\|^2 + \lambda \|U'\|_{1,2}. \tag{6}$$

In $\mathcal{V}$, the convergence is linear and we have precisely the following theorem:

Theorem 3. *At iteration k, if $U_k \in \mathcal{V}$ and a set of rows $\mathcal{S}_k$ is selected, then $U_{k+1} \in \mathcal{V}$ and*

$$F(U^{k+1}) - F^* \leq [1 - \rho_k](F(U^k) - F^*), \tag{7}$$

where $\rho_k = \frac{|\mathcal{S}_k|}{p} \min(\frac{1}{2}, \beta \frac{\mu}{L})$.

Led in the same conditions as in Bunea et al. (2012), numerical experiments confirmed that in the case $\lambda = 0$, the proposed algorithm can be faster than their alternated exact minimization procedure and that iterative methods presented in Park et al. (2016). As expected, we observed similar convergence rates in the SRRR setting.

References

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

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

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, K., Chan, K.-S., and Stenseth, N. C. (2012). Reduced rank stochastic regression with a sparse singular value decomposition. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):203–221.

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.

Chen, L. and Huang, J. Z. (2016). Sparse reduced-rank regression with covariance estimation. *Statistics and Computing*, 26(1-2):461–470.

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

Higham, N. and Papadimitriou, P. (1995). Matrix procrustes problems. *Technical Report, University of Manchester.*

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.

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, Q. and Tang, G. (2016). The nonconvex geometry of low-rank matrix optimizations with general objective functions. *arXiv preprint arXiv:1611.03060*.

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.

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.

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*.