Line-search methods and rank increase on low-rank matrix varieties

André Uschmajew† and Bart Vandereycken‡

†MATHICSE-ANCHP, Section de Mathématiques, École Polytechnique Fédérale de Lausanne,
1015 Lausanne, Switzerland
‡Department of Mathematics, Princeton University, Fine Hall, Princeton, NJ 08544, US
Email: andre.uschmajew@epfl.ch, bartv@math.princeton.edu

Abstract—Based on an explicit characterization of tangent cones one can devise line-search methods to minimize functions on the variety of matrices with rank bounded by some fixed value, thereby extending the Riemannian optimization techniques from the smooth manifold of fixed rank to its closure. This allows for a rank-adaptive optimization strategy where locally optimal solutions of some smaller rank are used as a starting point for an improved approximation with a larger rank.

Contrary to optimization on the smooth manifold of fixed-rank matrices, no special treatment is needed for rank-deficient matrices when optimizing on the variety. Hence, this gives a sound theoretical framework for the analysis of rank-increasing greedy algorithms, which can be more efficient than starting the calculations with large but fixed rank.

1. Introduction

We consider the problem of minimizing a smooth function $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ over

$$\mathcal{M}_{\leq k} = \{X \in \mathbb{R}^{m\times n} \mid \operatorname{rank}(X) \leq k\},$$

the set of matrices of rank at most k. This problem occurs in a number of applications such as low-rank matrix completion; see [2] and the references therein.

The set $\mathcal{M}_{\leq k}$ is the closure of the set of matrices of constant rank,

$$\mathcal{M}_k = \{X \in \mathbb{R}^{m\times n} \mid \operatorname{rank}(X) = k\},$$

which is known to be a real-analytic submanifold of $\mathbb{R}^{m\times n}$ of dimension $(m+n-k)k$; see, e.g, [15, Ex. 1.7]. Hence, the problem

$$\min_{X\in\mathcal{M}_k} f(X) \tag{1}$$

can be in principle treated by well-established techniques from Riemannian optimization [1, 3, 9, 10, 14], the most simplest being projected steepest descent with line search. This class of algorithms constitutes an alternative to block coordinate techniques like alternating direction methods based on the low-rank ansatz $X = UV^T$, which tend to suffer from slow convergence on large problems [16].

A fundamental disadvantage of (1) is, however, that the manifold is not closed in $\mathbb{R}^{m\times n}$. Therefore, even when assuming f to have bounded sublevel sets, we can neither guarantee that (1) has a solution, nor that a function decreasing sequence generated by some optimization algorithm will have cluster points in $\mathcal{M}_k$, making a convergence analysis complicated.

Since $\mathcal{M}_{\leq k}$ is closed, these theoretical problems disappear when we consider the optimization problem

$$\min_{X\in\mathcal{M}_{\leq k}} f(X). \tag{2}$$

For instance, the existence of a global minimum is guaranteed for continuous and coercive f. Furthermore, in the case of rank minimization problems, one is actually interested in solving (2) instead of (1). On the other hand, there seems to be a practical issue with solving (2): the gradient is only defined in points of full rank since at singular points with $\operatorname{rank}(X) < k$, the set $\mathcal{M}_{\leq k}$ is no longer smooth. Hence, smooth line-search algorithms for (2), like projected steepest descent, formally cease to exist and become "non-smooth" and infinitely conditioned in practice.

Fortunately, it has been shown in [12] that the tangent cone of $\mathcal{M}_{\leq k}$ in singular points has a rather simple characterization (see below). Hence the problems of non-closedness of $\mathcal{M}_k$ and of overestimating k—so that the iterative algorithms would become non-smooth—can be successfully addressed by directly optimizing on $\mathcal{M}_{\leq k}$.

In the present note we argue that the tangent cone of $\mathcal{M}_{\leq k}$ can also serve for the "opposite" approach where the rank k is gradually increased and we use low-rank approximations of the gradient of f in full space to warm-start the subsequent problem on $\mathcal{M}_{\leq k}$. Such rank increasing strategies were recently shown to be very effective; see, e.g., [4, 7, 9, 13].

2. Line-search methods and tangent cone

Consider a general line-search method on $\mathcal{M}_{\leq k}$,

$$X_{n+1} = P_{\leq k}(X_n + \alpha_n \Xi_n), \tag{3}$$

where Ξ_n is a search direction in the tangent cone at X_n, α_n is a step-size, and $P_{\leq k}$ is a metric projection onto $\mathcal{M}_{\leq k}$ in some norm. We will use a best rank-k approximation in Frobenius norm. Such a projection is necessary as in general $X_n + \alpha_n\Xi_n$ will not be in $\mathcal{M}_{\leq k}$ anymore.

When applied to the smooth manifold $\mathcal{M}_k$, the definition (3) stays formally the same except that the tangent

cone simplifies to the tangent space to the manifold in that point. However, in this case, the metric projections on $\mathcal{M}_k$ might only be possible for α_n very small. In the worst case, $\alpha_n \to 0$ when the iterates (X_n) want to accumulate to a point with rank less than k, or when they "cross" the subvariety $\mathcal{M}_{\leq k-1}$. The advantage of the line-search method (3) on $\mathcal{M}_{\leq k}$ is that the projection $P_{\leq k}$ is always possible since $\mathcal{M}_{\leq k}$ is closed. Hence, there is no need to impose restrictions on the maximally allowed step-length α_n.

The abstract definition of tangent cone to a closed set $\mathcal{M} \subset \mathbb{R}^n$ at a point $x \in \mathcal{M}$ looks complicated,

$$T_x\mathcal{M} = \{\xi \in \mathbb{R}^n \mid \exists\,(x_n) \subseteq \mathcal{M}, \exists\,(a_n) \subseteq \mathbb{R}_+ \text{ s.t. } x_n \to x,\ a_n \to +\infty,\ a_n(x_n - x) \to \xi\}, \quad (4)$$

see, e.g., [11] for this standard definition. However, [3] and [12] show that for $\mathcal{M}_{\leq k}$ one has a simple explicit characterization.

To derive it, let $X \in \mathcal{M}_{\leq k}$ have rank $s \leq k$. Then

$$X = USV^T,$$

for some $S \in \mathbb{R}^{s\times s}$ and orthonormal $U \in \mathbb{R}^{m\times s}$, $V \in \mathbb{R}^{n\times s}$. The tangent space at X with respect to the smooth manifold $\mathcal{M}_s$ is known (see, e.g., [6]) to consist of elements

$$T_X\mathcal{M}_s \ni \Xi_s = \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} A & B \\ C & 0 \end{bmatrix} \begin{bmatrix} V & V_\perp \end{bmatrix}^T,$$

where $A \in \mathbb{R}^{s\times s}$, $B \in \mathbb{R}^{s\times(n-s)}$, and $C \in \mathbb{R}^{(m-s)\times s}$ are arbitrary. Obviously, $T_X\mathcal{M}_s$ has to be a subset of $T_X\mathcal{M}_{\leq k}$. However, when $s < k$ we can approach X by matrices of rank larger than s in the definition (4), which produces another sort of tangential vectors. As it turns out, it holds

$$T_X\mathcal{M}_{\leq k} = T_X\mathcal{M}_S + \{\Xi_{k-s} \in (T_X\mathcal{M}_S)^\perp \mid \operatorname{rank}(\Xi_{k-s}) \leq k - s\},$$

where the orthogonal complement is taken with respect to the Frobenius inner product. As a consequence, tangent vectors in the tangent cone at X take the form

$$T_X\mathcal{M}_{\leq k} \ni \Xi = \Xi_s + \Xi_{k-s} \quad (5)$$

$$= \begin{bmatrix} U & U_\perp \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V & V_\perp \end{bmatrix}^T, \quad (6)$$

where A, B, and C are still arbitrary, but $\operatorname{rank}(D) \leq k - s$.

The convergence analysis for the line-search method (3) has been conducted in [12] for sufficiently gradient-related tangential search directions Ξ_n and Armijo backtracking for selection of the step-size α_n, and shall not be repeated here. Let us only mention that any metric projection

$$G_{\leq k}(X) \in \operatorname*{argmin}_{\Xi \in T_X\mathcal{M}_{\leq k}} \|-\nabla f(X) - \Xi\|_F$$

of the negative gradient on the tangent cone is such a search direction. When $\operatorname{rank}(X) = k$, $G_{\leq k}(X)$ is just the orthogonal projection on $T_X\mathcal{M}_k$ which is given by

$$G_k(X) = UU^T\nabla f(X)VV^T - UU^\perp\nabla f(X) - \nabla f(X)VV^\perp,$$

and can be efficiently implemented, also for large m, n, if $\nabla f(X)$ itself possesses some low-rank or sparse structure. If $\operatorname{rank}(X) = s < k$, one first calculates a metric projection Ξ_s of $-\nabla f(X)$ on $T_X\mathcal{M}_s$, and afterwards a best rank $(k - s)$ approximation Ξ_{k-s} of $-\nabla f(X) - \Xi_s$ to obtain

$$G_{\leq k}(X) = \Xi_s + \Xi_{k-s}.$$

When computing matrix vector products with $-\nabla f(X) - \Xi_s$, one can exploit that Ξ_s is a low-rank matrix. This makes the computation of Ξ_{k-s} possible using methods form large-scale and sparse SVD calculations or randomized low-rank techniques, see, e.g., [8, 5].

3. Rank increasing algorithm

If an iterate X_n of the line-search method (3) happens to have rank $s < k$, any search direction Ξ_n of the form (5) with nonzero D will increase the rank for the next iterate by $\operatorname{rank}(D) \leq k - s$. This observation can be used to devise a systematic rank increasing strategy. Given a locally optimal solution $X_n \in \mathcal{M}_{\leq k}$, we can use X_n as a (then rank-deficient) starting point for a line-search method on $\mathcal{M}_{k+\ell}$ without modifying the form the algorithm. It is the first update on $\mathcal{M}_{\leq k+\ell}$ which then will or will not increase the rank as desired.

A particular case enters when X_n has rank $s \leq k$, and is already a critical point of f on the smooth manifold $\mathcal{M}_s$, that is, $\nabla f(X) \in (T_{X_n}\mathcal{M}_s)^\perp$. We embed X_n in $\mathcal{M}_{\leq k+\ell}$ and choose the projection $G_{\leq k+\ell}(X_n)$ of the negative gradient on $T_{X_n}\mathcal{M}_{\leq k+\ell}$ as a search direction Ξ_n. By the considerations above, as the projection on $T_{X_n}\mathcal{M}_s$ is zero, Ξ_n is just given by the best rank $(k + \ell - s)$ approximation of $-\nabla f(X_n)$. As a consequence, if $\Xi_n = G_{\leq k+\ell}(X_n)$ is zero, then $\nabla f(X_n) = 0$ and we can terminate. Otherwise, the next iterate $X_{n+1} = X_n + \alpha_n\Xi_n$ will have a rank of exactly $s + \operatorname{rank}(\Xi_n)$.

The result is a rank-adaptive algorithm that starts with a rank-one matrix, listed as Alg. 1.

Increasing the rank by using rank-one approximations of the gradient has been proposed in [10]. Similar ideas are used in [13] and for tensors in [4]. It has been noted that such a rank increasing step can be regarded as a perturbed steepest descent step in full space. In fact, the best rank-one approximation Ξ_n of $-\nabla f(X_n)$ satisfies the angle condition

$$\langle -\nabla f(X_n), \Xi_n\rangle_F \geq \frac{1}{\sqrt{\min(m,n)}}\|\nabla f(X_n)\|_F\|\Xi_n\|_F,$$

which follows from considering the singular value decomposition. Hence, when f is a positive definite quadratic function and an exact line-search is used in the rank-increase step, then this step reduces the error in the energy norm to the full-space solution by some fixed factor.

Algorithm 1: Line-search with rank-increase on $\mathcal{M}_{\leq k}$

Input: Starting guess $X_0 \in \mathcal{M}_{\leq 1}$.

1 $n \leftarrow 0, k \leftarrow 1$
2 **while** *not converged* **do**
3 *Perform inner optimization*
4 **while** $\|G_{\leq k}(X_n)\|_F \geq$ `tol` **do**
5 Choose search-direction $\Xi_n \in T_X\mathcal{M}_{\leq k}$
6 Perform line-search for step size α_n
7 $X_{n+1} \leftarrow P_{\leq k}(X_n + \alpha_n \Xi_n)$
8 $n \leftarrow n + 1$
9 **end**
10 *Increase rank and warm-start*
11 $k \leftarrow k + \ell$
12 $\Xi_n \leftarrow G_{\leq k+\ell}(X_n)$
13 Choose step-size α_n
14 $X_{n+1} \leftarrow P_{\leq k}(X_n + \alpha_n \Xi_n)$
15 $n \leftarrow n + 1$
16 **end**

4. Application

We illustrate the effectiveness of Alg. 1 for the low-rank matrix completion problem

$$\min \operatorname{rank}(X) \qquad \text{s.t.} \quad P_\Omega(X) = P_\Omega(A), \tag{7}$$

where $A \in \mathbb{R}^{m\times n}$ is some unknown low-rank matrix that is only known on a subset Ω of all its entries. We focus on the case where A has exponentially decaying singular values since it is known that existing fixed-rank approaches perform very badly in this setting (see also below).

Under certain conditions (see [2] for more details), problem (7) has a unique solution X_*. By writing

$$\min f(X) = \tfrac{1}{2}\|P_\Omega(X) - P_\Omega(A)\|_F^2 \qquad \text{s.t.} \quad \operatorname{rank}(X) \leq k_*$$

where $k_* = \operatorname{rank}(X_*)$, we can solve (7) by Alg. 1. We stress that Alg. 1 does not require us to know k_*.

For all experiments, the inner optimization in Alg. 1 is solved using LRGeomCG [14] with a strong Wolfe line-search. The same line-search procedure is used in line 13 of Alg. 1. As stopping condition (line 4) we take, for example, a reduction of 10^{-4} of the relative gradient $\|G_{\leq k}(X)\|_F/\|X\|_F$. The reported relative errors for iterate X_n are computed as

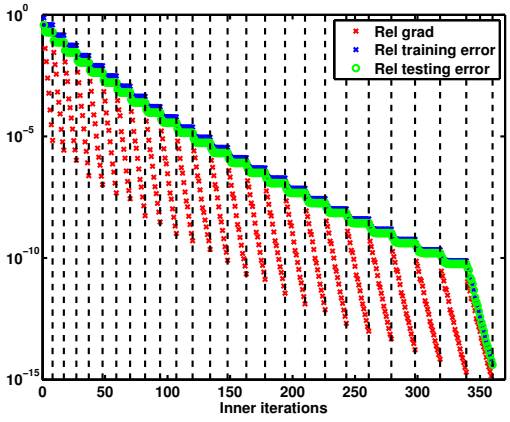
$$\text{rel. error} = \|P_\Lambda(X_n) - P_\Lambda(A)\|_F/\|P_\Lambda(A)\|_F,$$

where training error uses $\Lambda = \Omega$, and testing error uses another randomly chosen set of indices $\Lambda = \Gamma$ with $|\Gamma| = |\Omega|$.

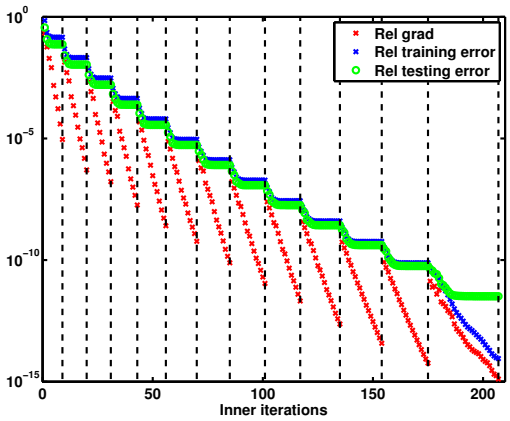
To compute Ξ_{k-s} for $G_{\leq k+\ell}(X_n)$, we use the randomized power algorithm from [5] which in our case requires $5(\ell+2)$ matrix vector products with $-\nabla f(X) - \Xi_s$.

Random matrix Let $U, V \in \mathbb{R}^{n\times k_*}$ be random Gaussian matrices with $n = 1000$ and $k_* = 26$. Define the unknown matrix as

$$A = U\Sigma V^T, \ \Sigma = \operatorname{diag}(10^i), \qquad i = 0, -\tfrac{2}{5}, -\tfrac{4}{5}, \ldots, -\tfrac{50}{5}$$



Total time: 4.63 sec.



Total time: 2.76 sec.

Figure 1: Convergence of Alg. 1 with $\ell = 1$ and $\ell = 2$ for a random matrix with exponentially decaying singular values.

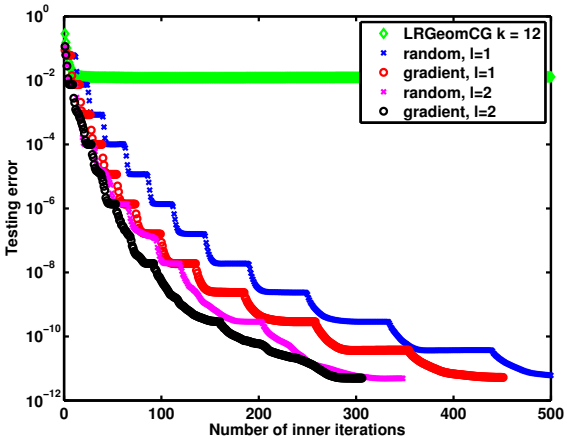
and take Ω as $2.5 \cdot 2nk_*$ uniform random samples. This coincides with an oversampling of about 2.5 compared to the degrees of freedom in the rank 26 matrix A.

In Fig. 1, we see the training and testing error of Alg. 1 for $\ell = 1$ and $\ell = 2$. In both cases, the matrix A is recovered up to a relative error of 10^{-12} and there is a good agreement between testing and training error. For each inner iteration, the relative residual is reduced by a factor of 10^{-5} in less than 30 iterations, showing that the warm-start using $G_{\leq k+\ell}(X_n)$ is very effective.

Bivariate function As next example we take A as the discretized bivariate function

$$f(x, y) = \frac{1}{1 + (x - y)^2}, \qquad (x, y) \in [0, 1]^2,$$

where x and y are uniformly discretized with $m = n = 1000$ points between 0 and 1. Even though A has full rank, one can show that the singular values of A decay exponentially as f is real-analytic. In fact, the numerical rank of A equals $k_* = 20$. We construct Ω as above, but now for $k_* = 20$, to again achieve an oversampling factor of about 2.5.



ℓ	strategy	its.	time (sec.)
1	random	532 (6.83)	15 (0.55)
1	gradient	453 (1.84)	13 (0.24)
2	random	388 (126)	11 (3.6)
2	gradient	296 (6.51)	8.1 (0.34)

Figure 2: Convergence of Alg. 1 where Ξ_n is enlarged using a low-rank approximation of the gradient or using random vectors. The number between brackets is the standard deviation.

In Fig. 2, we see the testing error for $\ell = 1$ and $\ell = 2$ when we warm-start using the gradient (as explained in §3) in comparison to using random vectors orthogonal to X_n. In both cases, we run Alg. 1 until a maximum rank of 12 since for higher ranks the line-search for the sub-problems usually failed which makes the numerical results less interpretable. The table shows the mean and standard deviation of the number of inner iterations required to decrease the Frobenius norm of the projected gradient in every fixed-rank problem (including rank 12) to 10^{-4} for 10 random initializations. The corresponding error history for one realization is plotted in the curves above. In addition, we have also depicted the error of LRGeomCG when optimizing directly on a rank-12 manifold.

It is clear from the figure that the rank-adaptive strategy greatly improves the standard fixed-rank approach. Also, using a low-rank approximation of the gradient performs the best in terms of the number of iterations and total computational time. In addition, from the table, we also see that random vectors result in less predictable convergence compared to the gradient strategy, the latter certainly being better understandable from a theoretical perspective.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, "Optimization Algorithms on Matrix Manifolds," Princeton University Press, Princeton, NJ, 2008.

[2] E. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, pp. 717–772, 2009.

[3] T. P. Cason, P.-A. Absil, and P. Van Dooren, "Iterative methods for low rank approximation of graph similarity matrices," *Linear Algebra Appl.*, vol. 438, pp. 1863–1882, 2013.

[4] S. V. Dolgov and D. V. Savostyanov, "Alternating minimal energy methods for linear systems in higher dimensions. Part I: SPD systems," *ArXiv e-print*, arXiv:1301.6068, 2013.

[5] N. Halko, P. G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Rev.*, vol. 53, pp. 217–288, 2011.

[6] U. Helmke and M. A. Shayman, "Critical points of matrix least squares distance functions," Linear Algebra Appl., vol. 215, pp. 1–19, 1995.

[7] D. Kressner, M. Steinlechner, and A. Uschmajew, "Low-rank tensor methods with subspace correction for symmetric eigenvalue problems," *EPFL MATHICSE Preprint 40.2013*, 2013.

[8] R. M. Larsen, "PROPACK—Software for large and sparse SVD calculations," `http://soi.stanford.edu/~rmunk/PROPACK`, 2004.

[9] B. Mishra, G. Meyer, F. Bach, and R. Sepulchre, "Low-rank optimization with trace norm penalty," *SIAM J. Optim.*, vol. 23, pp. 2124–2149, 2013.

[10] B. Mishra, G. Meyer, S. Bonnabel, and R. Sepulchre, "Fixed-rank matrix factorizations and Riemannian low-rank optimization," *Comput. Statist.*, 2013 (in press).

[11] R. T. Rockafellar and R. J.-B. Wets, "Variational Analysis," Springer-Verlag, Berlin, 1998.

[12] R. Schneider and A. Uschmajew, "Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality," *ArXiv e-print*, arXiv:1402.5284, 2014.

[13] M. Tan, I. W. Tsang, L. Wang, B. Vandereycken, and S. J. Pan, "Riemannian pursuit for big matrix recovery," *Proceedings of ICML14*, 2014.

[14] B. Vandereycken, "Low-rank matrix completion by Riemannian optimization," *SIAM J. Optim.*, vol. 23, pp. 1214–1236, 2013.

[15] R. O. Wells, "Differential Analysis on Complex Manifolds," Springer, New York, 2008.

[16] Z. Wen, W. Yin, and Y. Zhang, "Solving a low-rank factorization model for matrix completion by a non-linear successive over-relaxation algorithm," *Math. Progr. Comput.*, vol. 4, pp. 333–361, 2012