http://www.aimspress.com/journal/mine

Mathematics in Engineering, 5(3): 1–22.
DOI:10.3934/mine.2023053
Received: 22 February 2021
Revised: 19 August 2021
Accepted: 19 August 2021
Published: 27 September 2022

Research article

Reexamining low rank matrix factorization for trace norm regularization[†]

Carlo Ciliberto[1], **Massimiliano Pontil**[1,2,*] **and Dimitrios Stamos**[1]

[1] Department of Computer Science, University College London, London, United Kingdom
[2] Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genoa, Italy

[†] **This contribution is part of the Special Issue:** Mathematical aspects of machine learning
Guest Editors: Ernesto De Vito; Lorenzo Rosasco; Silvia Villa
Link: `www.aimspress.com/mine/article/6026/special-articles`

* **Correspondence:** Email: massimiliano.pontil@iit.it.

Abstract: Trace norm regularization is a widely used approach for learning low rank matrices. A standard optimization strategy is based on formulating the problem as one of low rank matrix factorization which, however, leads to a non-convex problem. In practice this approach works well, and it is often computationally faster than standard convex solvers such as proximal gradient methods. Nevertheless, it is not guaranteed to converge to a global optimum, and the optimization can be trapped at poor stationary points. In this paper we show that it is possible to characterize all critical points of the non-convex problem. This allows us to provide an efficient criterion to determine whether a critical point is also a global minimizer. Our analysis suggests an iterative meta-algorithm that dynamically expands the parameter space and allows the optimization to escape any non-global critical point, thereby converging to a global minimizer. The algorithm can be applied to problems such as matrix completion or multitask learning, and our analysis holds for any random initialization of the factor matrices. Finally, we confirm the good performance of the algorithm on synthetic and real datasets.

Keywords: collaborative filtering; low rank matrix factorization; trace norm regularization

1. Introduction

Learning low rank matrices is a problem of broad interest in machine learning and statistics, with applications ranging from collaborative filtering [1, 2], to multitask learning [3], to computer vision [4], and many more. A principled approach to tackle this problem is via suitable convex relaxations. Perhaps the most successful strategy in this sense is provided by trace (or nuclear) norm regularization [3, 5–7]. However, solving the corresponding optimization problem is computationally expensive for two main reasons. First, many commonly used algorithms require the computation of the proximity operator (e.g., [8]) which entails performing the singular value decomposition at each iteration. Second, and most importantly, the space complexity of convex solvers grows with the matrix size, which makes it prohibitive to employ them for large scale applications.

Due to the above shortcomings, practical algorithms for low rank matrix completion often use an explicit low rank matrix factorization to reduce the number of variables (see e.g., [2, 9] and references therein). In particular, a reduced variational form of the trace norm is used [7]. The resulting problem is however non-convex, and popular methods such as alternate minimization or alternate gradient descent may get struck at poor stationary points. Recent studies [10, 11] have shown that under certain conditions on the data generation process (underling low rank model, RIP, etc.) a particular version of the non-convex problem can be solved efficiently. However such conditions are not verifiable in real applications, and the problem of finding a global solution remains open.

In this paper we characterize the critical points of the non-convex problem and provide an efficient criterion to determine whether a critical point is also a global minimizer. Our analysis is constructive and suggests an iterative meta-algorithm that dynamically expands the parameter space to escape any non-global critical point, thereby converging to a global minimizer. We highlight the potential of the proposed meta-algorithm, by comparing its computational and statistical performance to two state of the art methods [9, 12].

The paper is organized as follows. In Sec. 2 we introduce the trace norm regularization problem and basic notions used throughout the paper. In Sec. 2.2 we present the low rank matrix factorization approach and set the main questions addressed in the paper. In Sec. 3 we present our analysis of the critical points of the low rank matrix factorization problem and the associated meta-algorithm. In Sec. 4 we report results from numerical experiments using our method. The Appendix contains proofs of the results, only stated in the main body of the paper, together with further auxiliary results and empirical observations.

2. Background and problem setting

In this work we study trace norm regularized problems of the form

$$\underset{W\in\mathbb{R}^{n\times m}}{\operatorname{minimize}}\, f_\lambda(W), \qquad f_\lambda(W) = \ell(W) + \lambda\,\|W\|_* \tag{2.1}$$

where $\ell : \mathbb{R}^{n\times m} \to \mathbb{R}$ is a twice differentiable convex function with Lipschitz continuous gradient, $\|W\|_*$ denotes the trace norm of a matrix $W \in \mathbb{R}^{n\times m}$, namely the sum of the singular values of W, and λ is a positive parameter. Examples of relevant learning problems that can be formulated as Eq (2.1) are:

Matrix completion. In this setting we wish to recover a matrix $Y \in \mathbb{R}^{n\times m}$ from a small subset of its entries. A typical choice for ℓ is the square error, $\ell(W) = \|M \odot (Y - W)\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm, $\odot$ denotes the Hadamard product (i.e., the entry-wise product) between two matrices and $M \in \mathbb{R}^{n\times m}$ is a binary matrix used to "mask" the entries of Y that are not available.

Multi-task learning. Here, the columns $w_1, \ldots, w_m$ of matrix W are interpreted as the regression vectors of different learning tasks. Given m datasets $(x_j^i, y_j^i)_{i=1}^{n_j}$ with $x_j^i \in \mathbb{R}^n$ and $y_j^i \in \mathbb{R}$, for $j = 1, \ldots, m$, we choose $\ell(W) = \sum_{j=1}^m \frac{1}{n_j} \sum_{i=1}^{n_j} \bar{\ell}(y_j^i, w_j^\top x_j^i)$, where $\bar{\ell} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is a prescribed loss function (e.g., the square or the logistic loss).

Other examples which are captured by problem (2.1) include collaborative filtering with attributes [13] and multiclass classification [5].

2.1. *Proximal gradient methods*

Problem (2.1) can be solved by first-order optimization methods such as the proximal forward-backward (PFB) splitting [8]. Given a starting point $W_0 \in \mathbb{R}^{n \times m}$, PFB produces a sequence $(W_k)_{k \in \mathbb{N}}$ with

$$W_{k+1} = \operatorname{prox}_{\gamma\lambda\|\cdot\|_*}\left(W_k - \gamma \nabla \ell(W_k)\right) \tag{2.2}$$

where $\gamma > 0$ and for any convex function $\phi : \mathbb{R}^{n \times m} \to \mathbb{R}$, the associated proximity operator at $W \in \mathbb{R}^{n \times m}$ is defined as $\operatorname{prox}_\phi(W) = \operatorname{argmin}\{\phi(Z) + \frac{1}{2}\|W - Z\|_F^2 : Z \in \mathbb{R}^{n \times m}\}$. In particular, the proximity operator of the trace norm at $W \in \mathbb{R}^{n \times m}$ corresponds to performing a soft-thresholding on the singular values of W. That is, assuming a *singular value decomposition (SVD)* $W = U\Sigma V^\top$, where $U \in \mathbb{R}^{n \times r}$, $V \in \mathbb{R}^{m \times r}$ have orthonormal columns, $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_r)$, with $\sigma_1 \geq \cdots \geq \sigma_r > 0$ and $r = \operatorname{rank}(W)$, we have

$$\operatorname{prox}_{\gamma\lambda\|\cdot\|_*}(W) = U \operatorname{diag}(h_{\gamma\lambda}(\sigma_1), \dots, h_{\gamma\lambda}(\sigma_r)) V^\top \tag{2.3}$$

where $h_{\gamma\lambda}$ is the *soft-thresholding* operator, defined for $\sigma \geq 0$ as $h_{\gamma\lambda}(\sigma) = \max(0, \sigma - \gamma\lambda)$.

PFB guarantees that for a suitable choice of the descent step γ (e.g., $\gamma < 2/L$ with L the Lipschitz constant of the gradient of ℓ), the sequence $f_\lambda(W_k)$ converges to the global minimum of the problem with a rate of $O(1/k)$ [8] (faster rates can be achieved using accelerated versions of the algorithm [14]). However, at each iteration PFB performs the SVD of an $n \times m$ matrix via Eqs (2.2) and (2.3), which requires $O(\min(n, m)\, nm)$ operations, a procedure that becomes prohibitively expensive for large values of n and m. Other methods such as those based on Frank-Wolfe procedure (e.g., [15]) alleviate this cost but require more iterations. More importantly, these methods need to store in memory the iterates W_k imposing an $O(nm)$ space complexity, a major bottleneck for large scale applications. However, trace norm regularization is typically applied to problems where the solution of problem (2.1) is assumed to be low-rank, namely of the form $AB^\top$ for some $A \in \mathbb{R}^{n \times r}, B \in \mathbb{R}^{m \times r}$ and $r \ll \min(n, m)$. Therefore it would be ideal to have an optimization method capable to capture this aspect and consequently reduce the memory requirement to $O(r(m + n))$ by keeping track of the two factor matrices A and B throughout the optimization rather than their product. This is the idea behind factorization-based methods, which have been observed to lead to remarkable performance in practice and are the subject of our investigation in this work.

2.2. *Matrix factorization approach*

Factorization methods build on the so-called *variational form* of the trace norm. Specifically, the trace norm of a matrix $W \in \mathbb{R}^{n \times m}$ can be characterized as (see e.g., [16] or Lemma 9 in the Appendix)

$$\|W\|_* = \frac{1}{2} \inf \left\{ \|A\|_F^2 + \|B\|_F^2, \ : \ r \in \mathbb{N},\ A \in \mathbb{R}^{n \times r},\ B \in \mathbb{R}^{m \times r},\ W = AB^\top \right\}. \tag{2.4}$$

with the infimum always attained for $r = \operatorname{rank}(W)$. The above formulation leads to the following “factorized” version of the original optimization problem (2.1)

$$\underset{A \in \mathbb{R}^{n \times r}, B \in \mathbb{R}^{m \times r}}{\text{minimize}}\ g_{\lambda,r}(A, B), \qquad g_{\lambda,r}(A, B) = \ell(AB^\top) + \frac{\lambda}{2}\left(\|A\|_F^2 + \|B\|_F^2\right) \tag{2.5}$$

where $r \in \mathbb{N}$ is now a further hyperparameter of the problem. Clearly, f_λ and $g_{\lambda,r}$ are tightly related and a natural question is whether minimizing the latter would allow to recover a solution of the original problem. The following well-known result (of which we provide a proof in the Appendix for completeness) shows that the two problems are indeed equivalent (for sufficiently large r).

Proposition 1 (Equivalence between problems (2.1) and (2.5))**.** *Let $W_* \in \mathbb{R}^{n\times m}$ be a global minimizer of f_λ in Eq (2.1) with $r_* = rank(W_*)$. Then, for every $r \geq r_*$, every global minimizer (A_*, B_*) of $g_{\lambda,r}$ is such that*

$$g_{\lambda,r}(A_*, B_*) = f_\lambda(A_* B_*^\top) = f_\lambda(W_*). \tag{2.6}$$

The above proposition implies that for sufficiently large values of r in Eq (2.5), the optimization of f_λ and $g_{\lambda,r}$ are *equivalent*. Therefore, *we can minimize f_λ by finding a global minimizer for $g_{\lambda,r}$*. This is a well-known approach to trace norm regularization (see e.g., [1, 2]) and can be extremely advantageous in practice. Indeed, we can leverage on a large body of smooth optimization literature to solve such a factorized problem [17]. As an example, if we apply Gradient Descent (GD) from a starting point (A_0, B_0), we obtain the sequence $(A_k, B_k)_{k\in\mathbb{N}}$ with

$$\begin{array}{rcl} A_{k+1} & = & A_k - \gamma(\nabla\ell(A_k B_k^\top) B_k + \lambda A_k) \\ B_{k+1} & = & B_k - \gamma(\nabla\ell(A_k B_k^\top)^\top A_k + \lambda B_k). \end{array} \tag{2.7}$$

This approach is much more appealing than PFB from a computational perspective because: 1) for small values of r, the iterations at Eq (2.7) are extremely fast since they mainly consists of matrix products and therefore require only $O(nmr)$ operations; 2) the space complexity of GD is $O(r(n + m))$, which may be remarkably smaller than the $O(nm)$ of PFB for small values of r; 3) Even if for large values of r, e.g., $r = \min(n, m)$, every iteration has the same time complexity as PFB, we do not need to perform expensive operations such as the SVD of an $n \times m$ matrix at every iteration. This dramatically reduces computational times in practice.

The strategy of minimizing $g_{\lambda,r}$ instead of f_λ was originally proposed in [7] and its empirical advantages have been extensively documented in previous literature, e.g., [2]. However, this approach opens important theoretical and practical questions that have not been addressed by previous work:

- **How to choose r?** By Prop. 1 we know that for suitably large values of r the minimization of $g_{\lambda,r}$ and f_λ are equivalent. However, a lower bound for such an r cannot be recovered analytically from the functional itself and, so, it is not clear how to choose r in practice.
- **Global convergence**. The function $g_{\lambda,r}$ is not jointly convex in the two variables A and B. This opens the question of whether GD (or other optimization methods) converge to a global minimizer for sufficiently large values of r.

Investigating such issues is the main focus of this work.

3. Analysis

In this section we study the questions outlined above and provide a meta-algorithm to minimize the function $g_{\lambda,r}$ while incrementally searching for a rank r for which Prop. 1 is verified. Our analysis builds upon the following keypoints:

- (Prop. 2) We characterize all critical points of $g_{\lambda,r}$, namely those points to which iterative optimization methods applied to Eq (2.5) (e.g., GD) could in principle converge to.
- (Thm. 3) We derive an efficient criterion to determine whether a critical point of $g_{\lambda,r}$ is a global minimizer (typically an NP hard problem for non-convex functions).
- (Prop. 4) We show that for any critical point (A, B) of $g_{\lambda,r}$ which is not a global minimizer, it is always possible to *constructively* find a descent direction for $g_{\lambda,r+1}$ from the point $([A\ 0], [B\ 0]) \in \mathbb{R}^{n\times(r+1)} \times \mathbb{R}^{m\times(r+1)}$.
- (Thm. 5) By combining the above results we show that for $r \geq \min(n, m)$, every critical point of $g_{\lambda,r}$ is either a global minimizer or a so-called *strict saddle point*, namely a point where the Hessian of the target function has at least a negative direction. We can then appeal to [18] to show that descent methods such as GD avoid strict saddle points and hence convergence to a global minimizer.

The above discussion suggests a natural "meta-algorithm" (which is presented more formally in Sec. 3.3) to address the minimization of f_λ via $g_{\lambda,r}$ while increasing r incrementally:

1) Initialize $r = 1$. Choose $A'_0 \in \mathbb{R}^n$ and $B'_0 \in \mathbb{R}^m$.

2) Starting from (A'_{r-1}, B'_{r-1}), converge to a critical point (A_r, B_r) for $g_{\lambda,r}$.

3) If (A_r, B_r) satisfies our criterion for global optimality (see Thm. 3) stop, otherwise:

4) Perform a step in a descent direction for $g_{\lambda,r+1}$ from $([A_r\ 0], [B_r\ 0])$ to a point (A'_r, B'_r), $A'_r \in \mathbb{R}^{n\times r+1}, B'_r \in \mathbb{R}^{m\times r+1}$; Increase r to $r + 1$ and go back to Step 2.

From our analysis in the following, the procedure above is guaranteed to stop *at most* after $r = \min(n, m)$ iterations. However, Prop. 1, together with our criterion for global optimality, suggests that this meta-algorithm could stop much earlier if f_λ admits a low-rank minimizer (which is indeed the case in our experiments). This has two main advantages: 1) by exploring candidate ranks incrementally, we can expect significantly faster computations and convergence if our optimality criterion activates for $r \ll \min(n, m)$ and 2) we automatically recover the rank of a minimizer for f_λ without the need to perform expensive operations such as SVD.

Remark 1. *The meta-algorithm considered in this paper is related to the optimization strategy recently proposed in [19], where the authors study convex problems for which a non-convex "factorized" formulation exists, including the setting considered in this work as a special case. However, by adopting such a general perspective, the resulting optimization strategy is less effective when applied to the specific minimization of $g_{\lambda,r}$. In particular:* 1) *the optimality criterion derived in [19] is only a sufficient but not necessary condition;* 2) *the upper bound on r is much larger than the one provided in this work, i.e., $r = nm$ rather than $r = \min(n, m)$;* 3) *convergence guarantees to a global optimum cannot be provided.*

By focusing exclusively on the question of minimizing f_λ via its factorized form $g_{\lambda,r}$ and, by leveraging on the specific structure of the problem, it is instead possible to provide a further analysis of the behavior of the proposed meta-algorithm.

3.1. Critical points and a criterion for global optimality

Since $g_{\lambda,r}$ is a non-convex smooth function, in principle we can expect optimization algorithms based on first or second order methods to converge only to critical points, i.e., points (A_*, B_*) for which

$\nabla g_{\lambda,r}(A_*, B_*) = 0$. The following result provides a characterization of such critical points and plays a key role in our analysis.

Proposition 2 (Characterization of critical points of $g_{\lambda,r}$)**.** *Let $(A_*, B_*) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r}$ be a critical point of $g_{\lambda,r}$. Let $s \leq \min(n, m)$ and let $U \in \mathbb{R}^{n\times s}$ and $V \in \mathbb{R}^{m\times s}$ be two matrices with orthonormal columns corresponding to the left and right singular vectors of $\nabla\ell(A_*B_*^\top) \in \mathbb{R}^{n\times m}$ with singular value equal to λ. Then, there exists $C \in \mathbb{R}^{s\times r}$, such that $A_* = UC$ and $B_* = -VC$.*

This result is instrumental in deriving a necessary and sufficient condition to determine whether a stationary point for $g_{\lambda,r}$ is actually a global minimizer. Indeed, since f_λ is convex, we can leverage on the first order condition for global optimality, stating that the matrix $W_* = A_*B_*^\top$ is a global minimizer for f_λ (and by Prop. 1 also (A_*, B_*) for $g_{\lambda,r}$) if and only if the zero matrix belongs to its subdifferential (see e.g., [8]). Studying this inclusion leads to the following theorem.

Theorem 3 (A criterion for global optimality)**.** *Let (A_*, B_*) be a critical point of $g_{\lambda,r}$. Then $A_*B_*^\top$ is a minimizer for f_λ if and only if $\|\nabla\ell(A_*B_*^\top)\| \leq \lambda$.*

This result provides a natural strategy to determine whether a descent method minimizing $g_{\lambda,r}$ has converged to a global minimizer, that is we evaluate the operator norm of the gradient, denoted $\|\nabla\ell(A_*B_*^\top)\|$, and then check whether it is larger than λ. For large matrices this operation can be performed efficiently by using approximation methods, e.g., power iteration [20]. Note that in general it is an NP-hard problem to determine whether a critical point of a non-convex function is actually a global minimizer [21]; it is only because of the relation with the convex function f_λ that in this case it is possible to perform such check in polynomial time.

3.2. Escape directions and global convergence

In this section, we observe that for any critical point of $g_{\lambda,r}$ which is not a global minimizer, it is always possible to either find a direction to escape from it or alternatively to increase r by one and to find a decreasing direction for $g_{\lambda,r+1}$. This strategy is suggested by the following result.

Proposition 4 (Escape direction from critical points)**.** *With the same notation of Prop. 2, assume $rank(A_*) = rank(B_*) < r$ and $\|\nabla\ell(A_*B_*^\top)\| = \mu > \lambda$. Then, (A_*, B_*) is a so-called strict saddle point for $g_{\lambda,r}$, namely the Hessian of $g_{\lambda,r}$ at (A_*, B_*) has at least one negative eigenvalue. In particular, there exists $q \in \mathbb{R}^r$ such that $A_*q = 0$, $B_*q = 0$ and if $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ are the left and right singular vectors of $\nabla\ell(A_*B_*^\top)$ with singular value equal to μ, then $g_{\lambda,r}$ decreases locally at (A_*, B_*) along the direction $(uq^\top, -vq^\top)$.*

A direct consequence of the result above is that an optimization strategy can remain trapped only at global minimizers of $g_{\lambda,r}$ or at critical points (A_*, B_*) for which A_* and B_* have full rank (since we can always escape from rank deficient ones). If the latter happens, Prop. 4 suggests the strategy adopted in this work, namely to increase the problem dimension to $r + 1$ and consider the "inflated" point $([A_*\ 0], [B_*\ 0])$. Indeed, at such point, $g_{\lambda,r+1}$ attains the same value as $g_{\lambda,r}(A_*B_*^\top)$ and it is straightforward to verify that $([A_*\ 0], [B_*\ 0])$ is still a critical point for $g_{\lambda,r+1}$. Since matrices $[A_*\ 0]$ and $[B_*\ 0]$ have now rank $< r + 1$, we can apply Prop. 4 to find a direction along which $g_{\lambda,r+1}$ decreases. This procedure will stop for $r > \min(n, m)$ since $\text{rank}(A_*) \leq \min(n, m) < r$ and we can therefore apply Prop. 4 to always escape from critical points until we reach a global minimizer (this fact can be actually improved to hold also for $r = \min(n, m)$ as we see in the following).

Note however that in general, if the number of non-global critical points is infinite, it is not guaranteed that such strategy will converge to the global minimizer. However, since by Prop. 4 every such critical point is a *strict* saddle point, we can leverage on previous results from the non-convex optimization literature (see [18] and references therein) in order to prove the following result.

Theorem 5 (Convergence to global minimizers of $g_{\lambda,r}$**).** *Let* $r \geq \min(n, m)$*. Then the set of starting points* (A_0, B_0) *for which GD does not converge to a global minimizer of* $g_{\lambda,r}$ *has measure zero.*

In particular, Thm. 5 suggests to initialize the optimization method used to minimize $g_{\lambda,r}$ by applying a small perturbation to the initial point (A_0, B_0) via additive noise according to a distribution that is absolutely continuous with respect to the Lebesgue measure of $\mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r}$ (e.g., a Gaussian). This perturbation guarantees such initial point to not be in the set of points for which GD converges to strict saddle point and therefore that the meta-algorithm considered in this work converges to a global minimizer. We make this statement more precise in the next section.

3.3. A meta-algorithm to minimize f_λ

We can now formally present the meta-algorithm outlined at the beginning of Sec. 3 to find a solution of the trace norm regularization problem (2.1) by minimizing $g_{\lambda,r}$ in Eq (2.5) for increasing values of r.

Algorithm 1 Meta-algorithm

```
Input: $\lambda > 0$, $\epsilon_{\text{conv}} > 0$ convergence tolerance, $\epsilon_{\text{crit}} > 0$ global criterion tolerance.
Initialize: Set $r = 1$. Sample $A'_0 \in \mathbb{R}^n$ and $B'_0 \in \mathbb{R}^m$ randomly.
For $r = 1$ to $\min(n, m)$
  $(A_r, B_r) = \text{OptimizationAlgorithm}(A'_{r-1}, B'_{r-1}, g_{\lambda,r}, \epsilon_{\text{conv}})$
  If $\|\nabla\ell(A_r B_r^\top)\| \leq \lambda + \epsilon_{\text{crit}}$
    Break
  $(A'_r, B'_r) = ([A_r\ u], [B_r\ v])$ with $u \in \mathbb{R}^n, v \in \mathbb{R}^m$ sampled randomly.
  $r = r + 1$
End
Return $(A_r, B_r)$
```

Algorithm 1 proceeds by iteratively applying the descent method Optimization Algorithm (e.g., GD) to minimize $g_{\lambda,r}$ while increasing the estimated rank r one step at the time. Whenever the optimization algorithm converges to a critical point (A_r, B_r) of $g_{\lambda,r}$ (within a certain tolerance ϵ_{conv}), Algorithm 1 verifies whether the global optimality criterion has been activated (again within a certain tolerance ϵ_{crit}). If this is the case, (A_r, B_r) is a global minimizer and we stop the algorithm. Otherwise we "inflate" the two factor matrices by one column each and repeat the procedure. The new column is initialized randomly, since by Prop. 4 we know that we will not converge again to $([A_r\ 0], [B_r\ 0])$ because it is not full rank. A more refined approach would be to choose u and $-v$ to be the singular vectors of $\nabla\ell(A_r B_r^\top)$ associated to the highest singular value. For the sake of brevity we provide an example of such strategy in the Appendix. However note that we still need to apply a random perturbation to the step in the descent direction in order to invoke Thm. 5 and be guaranteed to not converge to strict saddle points. As a direct corollary to Thm. 5 we have

Corollary 6. *Algorithm 1 with GD as* OptimizationAlgorithm *converges to a point* (A_*, B_*) *such that* $A_*B_*^\top$ *is a global minimizer for* f_λ *with probability* 1.

3.4. Convergence rates

In this section, we touch upon the question of convergence rates for optimization schemes applied to problem (2.5). For simplicity, we focus on GD, but our result generalizes to other first order methods. We provide upper bounds to the number of iterations required to guarantee that GD iterates are ϵ close to a critical point of the target function. By standard convex analysis results (see [22]) it is known that GD applied to a differentiable convex function is guaranteed to have sublinear convergence of $O(1/\epsilon)$ comparable to that of PFB. However, since $g_{\lambda,r}$ is non-convex, here we need to rely on more recent results that investigated the application of first order methods to functions satisfying the so-called *Kurdyka-Lojasiewicz (KL) inequality* [23, 24].

Definition 7 (Kurdyka-Lojasiewicz inequality)**.** *A differentiable function* $g : \mathbb{R}^d \to \mathbb{R}$ *is said to satisfy the Kurdyka-Lojasiewicz inequality at a critical point* $x_* \in \mathbb{R}^d$ *if there exists a neighborhood* $U \subseteq \mathbb{R}^d$ *and constants* $\gamma, \epsilon > 0$ *and* $\alpha \in [0, 1)$ *such that, for all* $x \in U \cap \{x : g(x_*) < g(x) < g(x_*) + \epsilon\}$,

$$\gamma |g(x) - g(x_*)|^\alpha < \|\nabla g(x)\|_F. \tag{3.1}$$

The KL inequality is a measure of how large is the gradient of the function in the neighborhood of a critical point. This allows us to derive convergence rates for GD methods that depend on the constant α. In particular, as a corollary to [18] (see also [23]) we obtain the following result.

Corollary 8 (Convergence rate of gradient descent)**.** *Let* $(A_k, B_k)_{k\in\mathbb{N}}$ *a sequence produced by GD method applied to* $g_{\lambda,r}$*. If* $g_{\lambda,r}$ *satisfies the KL inequality for some* $\alpha \in [0, 1)$*, then there exists a critical point* (A_*, B_*) *of* $g_{\lambda,r}$ *and constants* $C > 0$*,* $b \in (0, 1)$ *such that*

$$\|(A_k, B_k) - (A_*, B_*)\|_F^2 \leq \begin{cases} Cb^k & \text{if } \alpha \in (0, 1/2], \\ Ck^{-\frac{1-\alpha}{2\alpha-1}} & \text{if } \alpha \in (1/2, 1). \end{cases} \tag{3.2}$$

Furthermore, if $\alpha = 0$ *convergence is achieved in a finite number of steps.*

This result shows that depending on the constant $\alpha \in [0, 1)$ appearing in the KL inequality, we can expect different convergence rates for GD applied to problem (2.5). Although, it is a challenging task to identify such constant or even provide an upper bound in specific instances, the class of functions satisfying the KL inequality is extremely large and includes both analytic and *semi-algebraic* functions, see e.g., [25]. In the Appendix, we argue that if a function $\ell : \mathbb{R}^{n\times m} \to \mathbb{R}$ is a semi-algebraic, then also $\ell_r : \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r} \to \mathbb{R}$ such that $\ell_r(A, B) = \ell(AB^\top)$ is semi-algebraic. Therefore, in order to apply Cor. 8 it is sufficient that the error ℓ and the regularizer are semi-algebraic, a property which is verified by many commonly used error functions (or the associated loss functions, e.g., square, logistic) and regularizers (in particular the squared Frobenius norm).

4. Experiments

We report an empirical analysis of the meta-algorithm studied in this work. All experiments were conducted on an Intel Xeon E5-2697 V3 2.60Ghz CPU with 32GB RAM. We consider a matrix completion setting, with the loss $\ell(W) = \|M \odot (Y - W)\|_F^2$, where Y is the matrix that we aim to recover and M a binary matrix masking entries not available at training time. Below we briefly described the datasets used.

Synthetic. We generated a 100×100 matrix $Y = AB^\top + E$ as a rank 10 matrix product of two 100×10 matrices A, B plus additive noise E; the entries for A, B and E were independently sampled from a standard normal distribution.

Movielens. This dataset [26] consists of different datasets for movie recommendation of increasing size. They all comprise a number of ratings (from 1 to 5) given by n users on a database of m movies, which are recorded as a $Y \in \mathbb{R}^{n \times m}$ matrix with missing entries. In this work we considered three such datasets of increasing size, namely Movielens $100k$ ($ml100k$) with 100 thousand ratings from $n = 943$ users on $m = 1682$ movies, Movielens $1m$ ($ml1m$) with $\sim$ 1 million ratings, $n = 6040$ users and $m = 3900$ movies and Movielens $10m$ ($ml10m$), with $\sim$ 10 millions ratings, $n = 71567$ users and 10681 movies.

4.1. The criterion for global optimality and the estimated rank

The result in Thm. 3 provides a necessary and sufficient criterion to determine whether Algorithm 1 has achieved a global minimum for f_λ. A natural question is to ask whether, in practice, such criterion will be satisfied for a much smaller rank r than the one at which we are guaranteed convergence, namely $r = \min(n, m)$. To address this question we compared the solution achieved by our approach with the one obtained by iterative soft thresholding (ISTA) (or proximal forward backward, see e.g., [8]) on both synthetic and real datasets. Figure 1 reports the value of r for which our meta-algorithm satisfied the criterion for global optimality and compares it with the rank of the ISTA solution for different values of λ. For the Synthetic dataset (Figure 1 Left) we considered only 20% of the generated matrix Y entries for the optimization of f_λ. For the real dataset (Figure 1 Right) we considered $ml100k$ and sampled 50% of each user's available ratings. For both synthetic and real experiments our meta-algorithm recovers the same rank than as that found by ISTA. However, our algorithm reaches such rank incrementally, exploiting the low rank factorization. As we will see in the next section this results in a much faster convergence speed in practice.

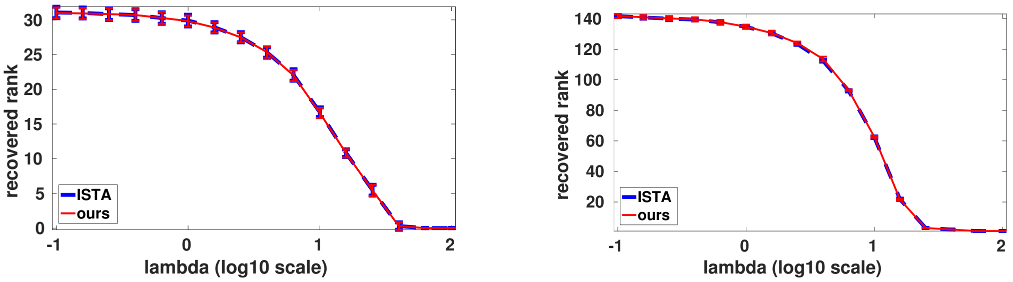

Figure 1. Rank of the solution for iterative soft thresholding algorithm (ISTA) and our method in Algorithm 1 varying lambda, on synthetic data (Left) and ml100k (Right).

4.2. Large scale matrix completion

We compared the performance of our meta-algorithm with two state of the art methods, Active Newton (ALT) [12] and Alternating Least Squares Soft Impute (ASL-SI) [9] on the three Movielens datasets. We used $50\%, 25\%$ and 25% of each user's ratings for training, validation and testing and repeated our experiments across 5 separate trials to account for statistical variability in the sampling. Test error was measured in terms of the Normalized Mean Absolute Error (NMAE), namely the mean (entry-wise) absolute error on the test set, normalized by the maximum discrepancy $\max(Y_{ij}) - \min(Y_{ij})$ between entires in Y. As a reference of the behavior of the different methods, Figure 2, reports on a single trial the decrease of f_λ on training data and NMAE on the test set with respect to time for the best λ chosen by validation. All methods where run until convergence and attained the same value of the objective function and same test error in all our trials. However, as it can be noticed, our meta-algorithm and ALS-SI seem to attain a much lower test error during earlier iterations. To better investigate this aspect, Table 1 reports results in terms of time, test error and estimated rank attained on average across the 5 trials by the different methods *at the iteration with smallest validation error*. As it can be noticed our meta-algorithm is generally on par with its competitors in terms of test error while being relatively faster and recovering low-rank solutions. This highlights an interesting byproduct of the meta-algorithm considered in this work, namely that by exploring candidate ranks incrementally, the method allows to find potentially better factorizations than trace norm regularization both in terms of test error and estimated rank. This fact can be empirically observed also for different values of λ as we report in the Appendix.

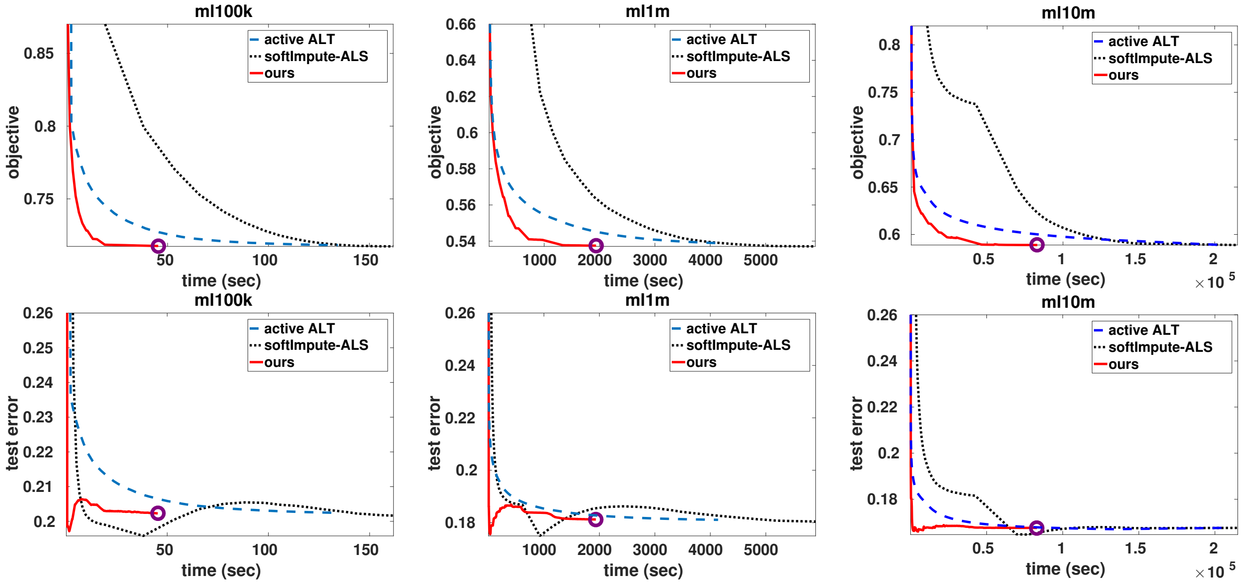

Figure 2. Convergence vs. time of the objective f_λ (Top row) and test errors (Bottom row) on three matrix completion large scale datasets for our meta-algorithm, ALT [12] and. The purple circle indicates when the global optimality criterion from Thm. 3 is satisfied.

Table 1. Average Normalized Mean Absolute Error (NMAE), convergence time and estimated rank achieved for the best validation parameters by ALT [12], ALS-SI [9] and Alg. 1, on the three Movielens datasets.

	ml100k			**ml1m**			**ml10m**		
	NMAE	time(s)	rank	NMAE	time(s)	rank	NMAE	time(s)	rank
ALT	0.2165	97	93	0.1806	4133	179	0.1670	205023	225
ALS-SI	0.1956	40	16	0.1749	832	31	0.1648	51205	36
Ours	0.1959	2	11	0.1751	39	25	0.1659	3150	41

5. Conclusions

We studied the convergence properties of low rank factorization methods for trace norm regularization. Key to our study is a necessary and sufficient condition for global optimality, which can be applied to any critical points of the non-convex problem. This condition together with a detailed analysis of the critical points lead us to propose a meta-algorithm for trace norm regularization, that incrementally expands the number of factors used by the non-convex solver. Although algorithms of this kind have been studied empirically for years, our analysis provides a fresh look and novel insights which can be used to confirm whether a global solution has been reached. Numerical experiments indicated that our optimality condition is useful in practice and the meta-algorithm is competitive with state-of-the art solvers. In the future it would be valuable to study improvements to our analysis, which would allow from one hand to derive precise rate of convergence for specific solvers used within the meta-algorithm and from another hand to study additional conditions under which our global optimality is guaranteed to activate immediately after the number of factors exceed the rank of the trace norm regularization minimizer.

Acknowledgments

We kindly thank Andreas Argyriou, Kevin Lai and Saverio Salzo for their helpful comments, corrections and suggestions. This work was supported in part by EPSRC grant EP/P009069/1.

Conflict of interest

The authors declare no conflict of interest.

References

1. J. D. M. Rennie, N. Srebro, Fast maximum margin matrix factorization for collaborative prediction, In: *ICML '05: Proceedings of the 22nd international conference on machine learning*, 2005, 713–719. http://doi.org/10.1145/1102351.1102441
2. Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, *Computer*, **42** (2009), 30–37. http://doi.org/10.1109/MC.2009.263

3. A. Argyriou, T. Evgeniou, M. Pontil, Convex multi-task feature learning, *Mach. Learn.*, **73** (2008), 243–272. http://doi.org/10.1007/s10994-007-5040-8

4. Z. Harchaoui, M. Douze, M. Paulin, M. Dudik, J. Malick, Large-scale image classification with trace-norm regularization, In: *2012 IEEE conference on computer vision and pattern recognition*, 2012, 3386–3393. http://doi.org/10.1109/CVPR.2012.6248078

5. Y. Amit, M. Fink, N. Srebro, S. Ullman, Uncovering shared structures in multiclass classification, In: *ICML '07: Proceedings of the 24th international conference on machine learning*, 2007, 17–24. http://doi.org/10.1145/1273496.1273499

6. F. R. Bach, Consistency of trace norm minimization, *The Journal of Machine Learning Research*, **9** (2008), 1019–1048.

7. N. Srebro, J. D. M. Rennie, T. S. Jaakkola, Maximum-margin matrix factorization, In: *NIPS'04: Proceedings of the 17th international conference on neural information processing systems*, 2004, 1329–1336.

8. H. H. Bauschke, P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, New York, NY: Springer, 2011. http://doi.org/10.1007/978-1-4419-9467-7

9. T. Hastie, R. Mazumder, J. D. Lee, R. Zadeh, Matrix completion and low-rank SVD via fast alternating least squares, *The Journal of Machine Learning Research*, **16** (2015), 3367–3402.

10. R. Ge, J. D. Lee, T. Ma, Matrix completion has no spurious local minimum, In: *Advances in neural information processing systems 29*, 2016, 2973–2981.

11. S. Bhojanapalli, B. Neyshabur, N. Srebro, Global optimality of local search for low rank matrix recovery, In: *NIPS'16: Proceedings of the 30th international conference on neural information processing systems*, 2016, 3880–3888.

12. C.-J. Hsieh, P. Olsen, Nuclear norm minimization via active subspace selection, In: *ICML'14: Proceedings of the 31st international conference on international conference on machine learning*, 2014, 575–583.

13. J. Abernethy, F. Bach, T. Evgeniou, J.-P. Vert, A new approach to collaborative filtering: operator estimation with spectral regularization, *The Journal of Machine Learning Research*, **10** (2009), 803–826.

14. A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.*, **2** (2009), 183–202. http://doi.org/10.1137/080716542

15. M. Dudik, Z. Harchaoui, J. Malick, Lifted coordinate descent for learning with trace-norm regularization, In: *Proceedings of the fifteenth international conference on artificial intelligence and statistics*, 2012, 327–336.

16. G. J. O. Jameson, *Summing and nuclear norms in Banach space theory*, Cambridge University Press, 1987. http://doi.org/10.1017/CBO9780511569166

17. D. P. Bertsekas, *Nonlinear programming*, Athena scientific Belmont, 1999.

18. J. D. Lee, M. Simchowitz, M. I. Jordan, B. Recht, Gradient descent only converges to minimizers, In: *29th Annual conference on learning theory*, 2016, 1246–1257.

19. B. D. Haeffele, R. Vidal, Global optimality in tensor factorization, deep learning, and beyond, 2015, arXiv:1506.07540.

20. D. P. Woodruff, *Sketching as a tool for numerical linear algebra*, Now Foundations and Trends, 2014. http://doi.org/10.1561/0400000060

21. K. G. Murty, S. N. Kabadi, Some NP-complete problems in quadratic and nonlinear programming, *Mathematical Programming*, **39** (1987), 117–129. http://doi.org/10.1007/BF02592948

22. S. Boyd, L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

23. H. Attouch, J. Bolte, P. Redont, A. Soubeyran, Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality, *Math. Oper. Res.*, **35** (2010), 438–457. http://doi.org/10.1287/moor.1100.0449

24. J. Bolte, A. Daniilidis, O. Ley, L. Mazet, Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity, *Trans. Amer. Math. Soc.*, **362** (2010), 3319–3363. http://doi.org/10.1090/S0002-9947-09-05048-X

25. H. Attouch, J. Bolte, On the convergence of the proximal algorithm for nonsmooth functions involving analytic features, *Mathematical Programming*, **116** (2009), 5–16. http://doi.org/10.1007/s10107-007-0133-5

26. F. M. Harper, J. A. Konstan, The movielens datasets: history and context, *ACM Trans. Interact. Intel. Syst.*, **5** (2016), 1–19. http://doi.org/10.1145/2827872

27. A. S. Lewis, The convex analysis of unitarily invariant matrix functions, *J. Convex Anal.*, **2** (1995), 173–183.

Appendix

Here we collect some auxiliary results and we provide proofs of the results stated in the main body of the paper.

A. Auxiliary results

The first lemma establishes the variational form for the trace norm; its proof can be found in [16].

Lemma 9 (Variational form of the trace norm)**.** *For every $W \in \mathbb{R}^{n\times m}$ and $r \in \mathbb{N}$ let $\mathcal{F}_r(W) = \{(A, B) \in \mathbb{R}^{n\times k} \times \mathbb{R}^{m\times r} \; : \; AB^\top = W\}$. Let $k = \operatorname{rank}(W)$ and let $\sigma_1(W) \geq \cdots \geq \sigma_k(W) > 0$ be the k singular values of W. Then*

$$\|W\|_* = \sum_{i=1}^{k} \sigma_i(W) = \frac{1}{2} \inf\left\{ \|A\|_F^2 + \|B\|_F^2 \;\middle|\; (A, B) \in \mathcal{F}_r(W),\ r \in \mathbb{N}\right\}.$$

Furthermore, if $W = U\Sigma V^\top$ is a singular value decomposition (SVD) for W, with $\Sigma = \operatorname{diag}(\sigma_1(W), \ldots, \sigma_r(W))$, the infimum is attained for $r = \operatorname{rank}(W)$, $A = U\Sigma^{\frac{1}{2}}$, and $B = V\Sigma^{\frac{1}{2}}$.

Recall that if $\phi : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is proper convex function, its sub-differential at x is the set

$$\partial\phi(x) = \{u : \phi(x) + \langle u, y - x\rangle \leq \phi(y),\ \text{for all } y \in \operatorname{domain}(\phi)\}.$$

The elements of $\partial\phi(x)$ are called the sub-gradients of ϕ at x.

Let $\mathbf{O}_n$ be the set of $n \times n$ orthogonal matrices. A norm $\|\cdot\| : \mathbb{R}^{m\times n} \to [0, \infty)$ is called *orthogonally invariant* if, for every $U \in \mathbf{O}_n$, $V \in \mathbf{O}_m$ and $W \in \mathbb{R}^{n\times m}$ we have that $\|UWV\| = \|W\|$ or, equivalently

$\|W\| = g(\sigma(W))$, where g is a symmetric gauge function (SGF), that is g is a norm invariant to permutations and sign changes. An important example of orthogonally invariant norms are the p-Schatten norms, $\|W\| = \|\sigma(W)\|_p$, where $\|\cdot\|$ is the ℓ_p-norm of a vector. In particular, for $p \in \{1, 2, \infty\}$ we have the trace, Frobenius, and spectral norms, respectively.

The following result is due to [27, Cor. 2.5].

Lemma 10. *If $\|\cdot\| : \mathbb{R}^{m\times n}$ is an orthogonally invariant and g is the associated SGF, then for every $W \in \mathbb{R}^{n\times m}$, it holds that*

$$\partial\|W\| = \{U\mathrm{diag}(\mu)V^\top \ : \ U \in \mathbf{O}_n,\ V \in \mathbf{O}_m,\ \mu \in \partial g(\sigma),\ W = U\mathrm{diag}(\sigma)V^\top\}.$$

B. Proofs

For convenience of the reader, we restate the results presented in the main body of the paper.

Proposition 1 (Equivalence between problems (2.1) and (2.5))**.** *Let $W_* \in \mathbb{R}^{n\times m}$ be a global minimizer of f_λ in Eq (2.1) with $r_* = rank(W_*)$. Then, for every $r \geq r_*$, every global minimizer (A_*, B_*) of $g_{\lambda,r}$ is such that*

$$g_{\lambda,r}(A_*, B_*) = f_\lambda(A_* B_*^\top) = f_\lambda(W_*). \tag{2.6}$$

Proof. Let $W_* \in \mathbb{R}^{n\times m}$ be a minimizer for f_λ of rank $r_* = \mathrm{rank}(W_*)$ and let $U\Sigma V^\top$ be a singular value decomposition of W_* with $U \in \mathbb{R}^{n\times r_*}$ and $V \in \mathbb{R}^{m\times r_*}$ with orthonormal columns and $\Sigma \in \mathbb{R}^{r_*\times r_*}$ diagonal with positive diagonal entires. Define $A_* = U\Sigma^{1/2} \in \mathbb{R}^{n\times r_*}$ and $B = V\Sigma^{1/2} \in \mathbb{R}^{m\times r_*}$. By construction $\|W_*\|_* = \frac{1}{2}(\|A_*\|_F^2 + \|B_*\|_F^2)$ and therefore

$$f_\lambda(W_*) = f_\lambda(A_* B_*^\top) = g_{\lambda,r_*}(A_*, B_*).$$

Now, we prove that $(A_*, B*)$ is a minimizer for g_{λ,r_*}. Suppose by contraddiction that there exist a couple $A_1 \in \mathbb{R}^{n\times r_*}, B_1 \in \mathbb{R}^{m\times r_*}$ such that $g_{\lambda,r_*}(A_1, B_1) < g_{\lambda,r_*}(A_*, B_*)$. Define

$$(\bar{A}_1, \bar{B}_1) = \operatorname*{argmin}_{AB^\top = A_1 B_1^\top} \|A\|_F^2 + \|B\|_F^2.$$

Then by Lemma 9 we have

$$\|\bar{A}_1 \bar{B}_1^\top\|_* = \frac{1}{2}(\|\bar{A}_1\|_F^2 + \|\bar{B}_1\|_F^2) \leq \frac{1}{2}(\|A_1\|_F^2 + \|B_1\|_F^2)$$

and therefore

$$f_\lambda(\bar{A}_1 \bar{B}_1^\top) = g_{\lambda,r_*}(\bar{A}_1, \bar{B}_1) \leq g_{\lambda,r_*}(A_1, B_1) < g_{\lambda,r_*}(A_*, B_*) = f_\lambda(W_*)$$

which is clearly not possible since W_* was a global minimizer for W_*. ■

Proposition 2 (Characterization of critical points of $g_{\lambda,r}$)**.** *Let $(A_*, B_*) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r}$ be a critical point of $g_{\lambda,r}$. Let $s \leq \min(n, m)$ and let $U \in \mathbb{R}^{n\times s}$ and $V \in \mathbb{R}^{m\times s}$ be two matrices with orthonormal columns corresponding to the left and right singular vectors of $\nabla\ell(A_* B_*^\top) \in \mathbb{R}^{n\times m}$ with singular value equal to λ. Then, there exists $C \in \mathbb{R}^{s\times r}$, such that $A_* = UC$ and $B_* = -VC$.*

Proof. Let $A_* \in \mathbb{R}^{m\times r}$, $B_* \in \mathbb{R}^{n\times r}$ be the matrices that correspond to a critical point of $g_{\lambda,r}$. We let

$$\nabla\ell(A_*{B_*}^\top) = \lambda U_1 V_1^\top + U_2\Sigma_2 V_2^\top + U_3\Sigma_3 V_3^\top \tag{B.1}$$

be the breakdown SVD of the gradient of ℓ at $A_*{B_*}^\top$, where Σ_2 is the diagonal matrix formed by the singular values strictly larger than λ and Σ_3 is the diagonal matrix formed by the singular values strictly smaller than λ (including those which are zero). For each $i = 1, 2, 3$ we denote with s_i the number of columns of $U_i \in \mathbb{R}^{n\times s_i}$ and $V_i \in \mathbb{R}^{m\times s_i}$. Recall that the matrices $[U_1\ U_2\ U_3] \in \mathbb{R}^{n\times n}$ and $[V_1\ V_2\ V_3] \in \mathbb{R}^{m\times m}$ are both orthogonal.

Taking the derivatives of Eq (2.5) w.r.t. A and B and setting them to zero gives the following optimality conditions for the critical points

$$\nabla\ell(A_*{B_*}^\top)B_* + \lambda A_* = 0 \tag{B.2}$$

$$\nabla\ell(A_*{B_*}^\top)^\top A_* + \lambda B_* = 0. \tag{B.3}$$

Solving Eq (B.3) for B_* and and replacing it in Eq (B.2) yields that

$$\left(-\frac{1}{\lambda^2}\nabla\ell(A_*{B_*}^\top)\nabla\ell(A_*{B_*}^\top)^\top + I_m\right)A_* = 0. \tag{B.4}$$

By Eq (B.1) $\ell(A_*{B_*}^\top)\nabla\ell(A_*{B_*}^\top)^\top = \lambda^2 U_1^\top + U_2\Sigma_2^2 U_2^\top + U_3\Sigma_3^2 U_3^\top$. Using this in Eq (B.4) and rearranging gives

$$\left(I_m - U_1U_1^\top - U_2\frac{\Sigma_2^2}{\lambda^2}U_2^\top - U_3\frac{\Sigma_3^2}{\lambda^2}U_3^\top\right)A_* = 0$$

which we rewrite as

$$\left(U_2(I - \frac{\Sigma_2^2}{\lambda^2})U_2^\top + U_3(I - \frac{\Sigma_3^2}{\lambda^2})U_3^\top\right)A_* = 0.$$

Therefore, the columns of A_* must be in the range of U_1 (i.e., orthogonal to U_2 and U_3), namely

$$A_* = U_1C \tag{B.5}$$

for some $C \in \mathbb{R}^{s_1\times r}$. Similarly we derive that

$$B_* = V_1D \tag{B.6}$$

for some $D \in \mathbb{R}^{s_1\times r}$. Combining Eq (B.1) with Eq (B.6) we obtain that

$$\nabla\ell(A_*{B_*}^\top)B_* = -\lambda U_1C.$$

Using this equation, Eq (B.5) and Eq (B.6) we rewrite Eq (B.2) as $\lambda U_1D + \lambda U_1C = 0$. This implies that $D = -C$ and, so, $B_* = -V_1C$. ■

Theorem 3 (A criterion for global optimality)**.** *Let (A_*, B_*) be a critical point of $g_{\lambda,r}$. Then $A_*B_*^\top$ is a minimizer for f_λ if and only if $\|\nabla\ell(A_*B_*^\top)\| \le \lambda$.*

Proof. Let $W_* = A_* B_*^{\top}$. We need to show that $0 \in \partial f_\lambda(W_*)$ or, equivalently

$$-\frac{1}{\lambda}\nabla\ell(W_*) \in \partial\|W_*\|_*. \tag{B.7}$$

Let $Z = -\frac{1}{\lambda}\nabla\ell(W_*)$. As a special case of Lemma 10 we have that $Z \in \partial\|W_*\|_*$ holds true if and only if there exists a simultaneous singular value decomposition of the form $W_* = U\mathrm{diag}(\sigma)V^{\top}$, $Z = U\mathrm{diag}(\sigma(Z))V^{\top}$ and $\sigma(Z) \in \partial\|\sigma(W_*)\|_1$, with $\sigma(Z)$ denoting the spectrum of Z (namely the vector of singular values of Z arranged in non-increasing order). Recall that

$$\partial\|\sigma\|_1 = \{z \in \mathbb{R}^m : z_i = 1 \text{ if } \sigma_i \neq 0, \text{ and } z_i \in [-1, 1] \text{ otherwise}\}.$$

Using the same notation of Prop. 2, consider the SVD

$$Z = -U_1 V_1^{\top} - U_2\frac{\Sigma_2}{\lambda}V_2^{\top} - U_3\frac{\Sigma_3}{\lambda}V_3^{\top}.$$

By Prop. 2 we have that $A_* = U_1 C$ and $B_* = -V_1 C$, with $C \in \mathbb{R}^{s_1 \times r}$. Now, let $\bar{r} = \mathrm{rank}(C) \leq r$ and let $C = P\Gamma Q^{\top}$ be the SVD of C, with $P \in \mathbb{R}^{s_1 \times \bar{r}}$ and $Q \in \mathbb{R}^{\bar{r} \times r}$ matrices with orthonormal columns and $\Gamma \in \mathbb{R}^{\bar{r} \times \bar{r}}$ diagonal with positive diagonal elements. Denote $\widetilde{P} = [P\ P^{\perp}] \in \mathbb{R}^{s_1 \times s_1}$ the orthonormal matrix obtained by completing P with a matrix $P^{\perp} \in \mathbb{R}^{s_1 \times (s_1 - \bar{r})}$ with orthonormal columns such that $P^{\top}P^{\perp} = 0$. Moreover denote $\widetilde{\Gamma} \in \mathbb{R}^{s_1 \times s_1}$ as

$$\widetilde{\Gamma} = \begin{bmatrix} \Gamma & 0_{\bar{r} \times s_1 - \bar{r}} \\ 0_{s_1 - \bar{r} \times \bar{r}} & 0_{s_1 - \bar{r} \times \bar{s}_1 - \bar{r}} \end{bmatrix}.$$

Then,

$$W_* = -U_1 C C^{\top} V_1 = (-U_1 P)\Gamma^2 (V_1 P)^{\top} = (-U_1 \widetilde{P})\widetilde{\Gamma}^2 (V_1 \widetilde{P})^{\top} = (-\widetilde{U}_1)\widetilde{\Gamma}^2(\widetilde{V}_1)$$

with $\widetilde{U}_1 = U_1\widetilde{P}$ and $\widetilde{V} = V_1\widetilde{P}$. Note that since $\widetilde{P}$ is orthonormal, $\widetilde{U}_1\widetilde{V}_1^{\top} = U_1 V_1^{\top}$. Moreover U_2, U_3 have columns orthogonal to $\widetilde{U}_1$ and V_2, V_3 have columns orthogonal to $\widetilde{V}_1$. Consequently,

$$Z = -U_1 V_1^{\top} - U_2\frac{\Sigma_2}{\lambda}V_2^{\top} - U_3\frac{\Sigma_3}{\lambda}V_3^{\top} = -\tilde{U}_1\tilde{V}_1^{\top} - U_2\frac{\Sigma_2}{\lambda}V_2 - U_3\frac{\Sigma_3}{\lambda}V_3^{\top}$$

is an alternative singular value decomposition for Z. Therefore, W_* and Z have a simultaneous singular value decomposition and we can conclude that $\sigma(Z) \in \partial\|\sigma(W_*)\|_1$ if and only if $s_2 = 0$, namely $\|\nabla\ell(W_*)\| \leq \lambda$ as desired. ■

Proposition 4 (Escape direction from critical points)**.** *With the same notation of Prop. 2, assume rank*$(A_*) =$ *rank*$(B_*) < r$ *and* $\|\nabla\ell(A_* B_*^{\top})\| = \mu > \lambda$*. Then,* (A_*, B_*) *is a so-called strict saddle point for* $g_{\lambda,r}$*, namely the Hessian of* $g_{\lambda,r}$ *at* (A_*, B_*) *has at least one negative eigenvalue. In particular, there exists* $q \in \mathbb{R}^r$ *such that* $A_* q = 0$*,* $B_* q = 0$ *and if* $u \in \mathbb{R}^n$ *and* $v \in \mathbb{R}^m$ *are the left and right singular vectors of* $\nabla\ell(A_* B_*^{\top})$ *with singular value equal to* μ*, then* $g_{\lambda,r}$ *decreases locally at* (A_*, B_*) *along the direction* $(uq^{\top}, -vq^{\top})$*.*

Proof. By Theorem 2, $A_* = U_1 C$ and $B_* = -V_1 C$, where U_1 and V_1 are the matrices of left and right singular vectors of $\nabla\ell(A_* B_*^{\top})$ and $C \in \mathbb{R}^{s \times r}$, for $s = \mathrm{rank}(A_*) = \mathrm{rank}(B_*)$. Taking the SVD of $C = P\Gamma Q^{\top}$, we rewrite

$$A_* = U_1 P\Gamma Q^{\top}, \quad B_* = -V_1 P\Gamma Q^{\top}. \tag{B.8}$$

Since A_* and B_* are rank deficient and they have the same null space, we can choose $q \in \mathbb{R}^r$ such that $A_* q = B_* q = 0$. Let u_2 and v_2 the a left and right singular vector of $\nabla\ell(A_* B_*{}^\top)$ with singular value equal to μ.

We consider a perturbation of the objective function in the direction $(-u_2 q^\top, v_1 q^\top)$. We have that

$$L(\gamma) = \ell((A_* - \gamma u_2 q^\top)(B_* + \gamma v_1 q^\top)^\top) + \frac{\lambda}{2}(\|A_* - \gamma u_1 q^\top\|_F^2 + \|B_* + \gamma v_1 q^\top\|_F^2) \tag{B.9}$$

$$= \ell(A_* B_*{}^\top - \gamma^2 u_2 v_2^\top) + \frac{\lambda}{2}(\|A_*\|_F^2 + \|B_*\|_F^2) + \lambda\gamma^2. \tag{B.10}$$

$$\tag{B.11}$$

Thus, we have that

$$L'(\gamma) = \langle -2\gamma\nabla\ell(A_* B_*{}^\top - \gamma^2 u_2 v_2^\top), u_2 v_2^\top\rangle + 2\gamma\lambda \tag{B.12}$$

$$= -2\gamma u_2^\top \nabla\ell(A_* B_*{}^\top - \gamma^2 u_2 v_2^\top) v_2 + 2\gamma\lambda. \tag{B.13}$$

Consequently

$$L''(0) = 2(\lambda - u_2^\top \nabla\ell(A_* B_*{}^\top) v_2) = 2(\lambda - \mu) < 0$$

and the result follows. ■

Theorem 5 (Convergence to global minimizers of $g_{\lambda,r}$)**.** *Let* $r \geq \min(n,m)$. *Then the set of starting points* (A_0, B_0) *for which GD does not converge to a global minimizer of* $g_{\lambda,r}$ *has measure zero.*

Proof. We have shown in Prop. 4 that every critical point (A, B) of $g_{\lambda,r}$ for $r \geq \min(n,m)$ is either a global minimizer or a strict saddle point, namely such that the Hessian $\nabla^2 g_{\lambda,r}(A,B)$ has at least a negative eigenvalue. Therefore, since the error function ℓ is twice differentiable with gradient Lipschitz continuous also $g_{\lambda,r}$ is. Let $L_r > 0$ be the Lipschitz constant of $\nabla g_{\lambda,r}$. Then, we are in the hypotheses of Thm.4.1 in [18] which states that if $(A_k, B_k)_{k\in\mathbb{N}}$ is obtained with step $0 < \alpha < 1/L_r$ with initial point (A_0, B_0) sampled uniformly at random, then for any (A_*, B_*) *strict saddle point* of $g_{\lambda,r}$,

$$\mathrm{Prob}\left(\lim_{k\to+\infty} (A_k, B_k) = (A_*, B_*)\right) = 0$$

which implies the desired result and corresponds to Cor. 6. ■

At last we show that if the spectral norm of the gradient at a critical point is close to one from above then value of the objective function is close to the global minimum.

Proposition 11. *Let* (A_*, B_*) *be a critical point of* $g_{\lambda,r}$ *and let* $W_* = A_* B_*^\top$. *If* $\|\nabla\ell(W_*)\| \leq \lambda + \epsilon$ *with* $\epsilon \in [0, \lambda]$ *then*

$$f_\lambda(W_*) \leq \min_W f_\lambda(W) + \epsilon\left(\|W_*\|_* + \frac{\ell(0)}{\lambda}\right)$$

Proof. We write

$$\begin{aligned} \nabla\ell(W_*) &= \lambda U_1 V_1^\top + U_2\Sigma_2 V_2^\top + \lambda U_3 V_3^\top + U_3 \Sigma_3 V_3^\top \qquad (B.14)\\ &= \lambda(U_1 V_1^\top + U_3 V_3^\top) + U_2\Sigma_2 V_2^\top + \lambda U_3 V_3^\top + U_3(\Sigma_3 - \lambda U)V_3^\top. \end{aligned}$$

Let $Z = U_3(\Sigma_3 - \lambda U)V_3^\top$. By assumption $\|\Sigma_3\| \leq 2\lambda$, implying that $-Z \in \partial f_\lambda(W_*)$. That is, for every $W \in \mathbf{M}_{n,m}$,

$$f_\lambda(W_*) + \langle Z, W - W_* \rangle \leq f_\lambda(W). \tag{B.15}$$

In turn this implies that

$$f_\lambda(W_*) \leq \langle Z, W - W_* \rangle + f_\lambda(W).$$

Since the minimizer can be constrainted to be in the set $\{W : \|W\|_{\mathrm{tr}} \leq \ell(0)/\lambda\}$ we conclude from Eq (B.15) that

$$\begin{aligned} f_\lambda(W_*) &\leq f_\lambda(W) - \langle Z, W - W_* \rangle \\ &\leq \min_W f_\lambda(W) + \|Z\| \left(\|W_*\|_* + \frac{\ell(0)}{\lambda} \right) \\ &\leq \min_W f_\lambda(W) + \epsilon \left(\|W_*\|_{\mathrm{tr}} + \frac{\ell(0)}{\lambda} \right). \end{aligned}$$

■

C. Meta-algorithm with explicit escape from stationary points

We expand on the discussion in Sec. 3.3, where we formally introduced the meta-algorithm considered in this work. Specifically we observe that in the formulation of Algorithm 1 we did not exploit the result in Prop. 4, providing an explicit decreasing direction for $g_{\lambda,r+1}$ from the inflated point $([A_*\ 0], [B_*\ 0])$, where (A_*, B_*) is a stationary point for $g_{\lambda,r}$. For completeness in Algorithm 2 we report a variant of the meta-algorithm proposed in this paper that makes use of such escape direction. We care to point out that in our experiments we did not observe any statistically significant difference with the original version.

We discuss here the details of Algorithm 2. By Prop. 4 we know that there exists $\gamma > 0$ such that

$$g_{\lambda,r+1}([A_*\ 0] + \gamma[0_{n\times r},\ u], [B_*\ 0] + \gamma[0_{m\times r},\ -v]) < g_{\lambda,r}(A_*, B_*)$$

where $u \in \mathbb{R}^n, v \in \mathbb{R}^m$ are respectively the left and right singular vectors associated to the largest singular value μ of $\nabla\ell(A_* B_*^\top)$. In Algorithm 2 these are provided by the routine LargestSingularPair.

We can find a new point on the direction specified above by trying to approximate the minimizer of

$$\gamma_* = \operatorname*{argmin}_{\gamma\in[0,1]}\ g_{\lambda,r+1}([A_*\ 0] + \gamma[0_{n\times r}\ u], [B_*\ 0] + \gamma[0_{m\times r}\ -v])$$

for instance by considering a set of candidate $\gamma_1, \ldots, \gamma_M \in (0, 1]$ (e.g., $\gamma_i = \gamma_0^i$ for $i = 1, \ldots, M$) and choose the one leading to the lowest value for $g_{\lambda,r+1}$. In Algorithm 2 we refer to this procedure as LineSearch given its analogy with standard line search approaches often used in optimization. However notice that such methods can typically leverage on a criterion depending on the norm of the gradient $\|\nabla g_{\lambda,r+1}([A_*\ 0], [B_*\ 0])\|_F$ in order to guarantee that the function decreases significantly. We cannot replicate such strategy since $\nabla g_{\lambda,r+1}([A_*\ 0], [B_*\ 0]) = 0$ in our case.

As a final observation, we care to point out (as we already done in Sec. 3.3) that it is necessary to perturb the point $([A_*\ \gamma u], [B_*\ -\gamma v])$, e.g., by adding some noise, in order to guarantee that

$([A_* \ \gamma u], [B_* \ -\gamma v])$ does not belong to the set of measure zero, mentioned in Thm. 5, of initial points that converge to strict saddle points of $g_{\lambda,r}$.

Algorithm 2 Meta-algorithm with explicit escape from stationary points

```
Input: $\lambda > 0$, $\epsilon_{\text{conv}} > 0$ convergence tolerance, $\epsilon_{\text{crit}} > 0$ global criterion tolerance, $\sigma > 0$ noise parameter.
Initialize: Set $r = 1$. Sample $A'_0 \in \mathbb{R}^n$ and $B'_0 \in \mathbb{R}^n$ randomly.
For $r = 1$ to $\min(n, m)$
  $(A_r, B_r) = $ OptimizationAlgorithm$(A'_{r-1}, B'_{r-1}, g_{\lambda,r}, \epsilon_{\text{conv}})$
  If $\|\nabla \ell(A_r B_r^\top)\| \leq \lambda + \epsilon_{\text{crit}}$
    Break
  $[u, \mu, v]$ =LargestSingularPair$(\nabla \ell(A_r B_r^\top))$
  $\hat{\gamma} = $ LineSearch$([A_r \ 0], [B_r \ 0], u, v)$
  Perturb $u = u + \eta$ with $\eta \sim \mathcal{N}(0, \sigma I_{n \times n})$
  Perturb $v = v + \eta$ with $\eta \sim \mathcal{N}(0, \sigma I_{m \times m})$
  $(A'_{r+1}, B'_{r+1}) = ([A_r \ \gamma u], [B_r \ -\gamma v])$
  $r = r + 1$
End
Return $(A_r, B_r)$
```

D. On the Kurdyka-Lojasiewicz inequality

We extend here the discussion on the KL inequality (Def. 7) and corresponding convergence results reviewed in Sec. 3.4. As a special case to the problem of optimizing $g_{\lambda,r}$ considered in this work, we have recalled in Cor. 8 that if $g_{\lambda,r}$ satisfies the KL inequality, we can expect GD to exhibit polynomial rates of convergence. A natural question is when $g_{\lambda,r}$ satisfies such inequality. To provide an insight on this issue, we consider the result in [24] showing that the KL inequality is satisfied by *semi-algebraic* functions. We recall that a set $S \subseteq \mathbb{R}^d$ is said semi-algebraic if there exists a finite number of polynomials $p_{kh}, q_{kh} : \mathbb{R}^d \to \mathbb{R}$ such that

$$S = \bigcup_{k=1}^{K} \bigcap_{h=1}^{H} \left\{ x \in \mathbb{R}^d \mid p_{kh}(x) = 0, \ \ q_{kh}(x) \leq 0 \right\}.$$

A function $f : \mathbb{R}^d \to \mathbb{R}$ is said semi-algebraic if its graph

$$\operatorname{graph} f = \left\{ (x, t) \mid x \in \mathbb{R}^d, t \in \mathbb{R}, f(x) = t \right\}$$

is semi-algebraic.

Note that a variety of error functions $\ell : \mathbb{R}^{n \times m} \to \mathbb{R}$ typically used in machine learning and matrix factorization problems are semi-algebraic (e.g., the square loss). Interestingly, we have that if ℓ is semi-algebraic, then $\ell_r : \mathbb{R}^{n \times r} \times \mathbb{R}^{m \times r} \to \mathbb{R}$ such that $\ell_r(A, B) = \ell(AB^\top)$, is semi-algebraic as well. Indeed, by definition of semi-algebraic function we have that

$$\operatorname{graph} \ell = \{ (X, t) \mid X \in \mathbb{R}^{n \times m}, t \in \mathbb{R}, \ell(X) = t \}$$

is semi algebraic, therefore there exist $p_{kh}, q_{kh} : \mathbb{R}^{n\times m} \times \mathbb{R} \to \mathbb{R}$ such that

$$\text{graph } \ell = \bigcup_{k=1}^{K} \bigcap_{h=1}^{H} \{(X,t) \mid p_{kh}(X,t) = 0, q_{kh}(X,t) \leq 0\}\,.$$

Now, denote X_{ij} the (i,j)-th entry of X. For $X = AB^\top$ with $A \in \mathbb{R}^{n\times r}$ and $B \in \mathbb{R}^{m\times r}$ we have $X_{ij} = \sum_{s=1}^{r} A_{is}B_{js}$ namely $X_{ij} = m_{ij}(A,B)$ with $m_{ij} : \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r} \to \mathbb{R}$ a real polynomial. Let us denote $m : \mathbb{R}^{n\times r} \times \mathbb{R}^{m\times r} \to \mathbb{R}^{n\times m}$ the matrix valued function with (i,j)-th entry $m(A,B)_{ij} = m_{ij}(A,B)$. Since every p_{kh} and q_{kh} are polynomial in the variables of X, we have that $p_{kh} \circ m$ and $h_{kh} \circ m$ are still polynomials and therefore the graph of ℓ_r

$$\begin{aligned}\text{graph } \ell_r &= \{(A,B,t) \mid A \in \mathbb{R}^{n\times r}, B \in \mathbb{R}^{m\times r}, t \in \mathbb{R}, \ell(AB^t op) = t\}\\ &= \bigcup_{k=1}^{K} \bigcap_{h=1}^{H} \{(A,B,t) \mid p_{kh}(m(A,B)),t) = 0,\ \ q_{kh}(m(A,B),t) \leq 0\}\end{aligned}$$

is semi-algebraic, which implies that ℓ_r is a semi-algebraic function.

Going back to the question of whether $g_{\lambda,r}(A,B) = \ell(AB^\top) + \frac{\lambda}{2}(\|A\|_F^2 + \|B\|_F^2)$ is semi-algebraic, we now can conclude that it is sufficient to assume the error function ℓ to be semi-algebric. Indeed, it is well-known (see e.g., [24]) that the squared Frobenius norm of a matrix is semi-algebraic and that the finite sum of semi-algebraic functions is still semi-algebraic. Therefore we can re-formulate Cor. 8 in terms only of the error ℓ, namely

Corollary 12 (Convergence rate of gradient descent). *Let $(A_k, B_k)_{k\in\mathbb{N}}$ a sequence produced by GD method applied to $g_{\lambda,r}$. If ℓ is semi-algebraic, then $g_{\lambda,r}$ satisfies the KL inequality for some constant $\alpha \in [0,1)$, and there exists a critical point (A_*, B_*) of $g_{\lambda,r}$ and constants $C > 0$, $b \in (0,1)$ such that*

$$\|(A_k,B_k) - (A_*,B_*)\|_F^2 \leq \begin{cases} Cb^k & \text{if } \alpha \in (0,1/2], \\ Ck^{-\frac{1-\alpha}{2\alpha-1}} & \text{if } \alpha \in (1/2,1). \end{cases} \tag{D.1}$$

Furthermore, if $\alpha = 0$ convergence is achieved in a finite number of steps.

E. Further experiments

This last section provides more comparative experiments between the meta-algorithm and the two state-of-the art solvers, as well as a comparison to the algorithm in [19].

E.1. Large scale matrix completion

In Sec. 4 we reported on the performance of the three methods considered for λ chosen by validation as the parameter leading to the lowest Normalized Mean Average Error (NMAE). For completeness here we report the same experiments for a range of candidate values of λ.

Figures 3 and 4 report respectively the value of the objective function f_λ and the test error (NMAE) for the three methods considered in our experiments. Interestingly we observe a similar pattern to the one of the optimal λ, with our method exhibiting comparable performance in terms of both time and test error to the state-of-the-art competitors for most of the λ considered.

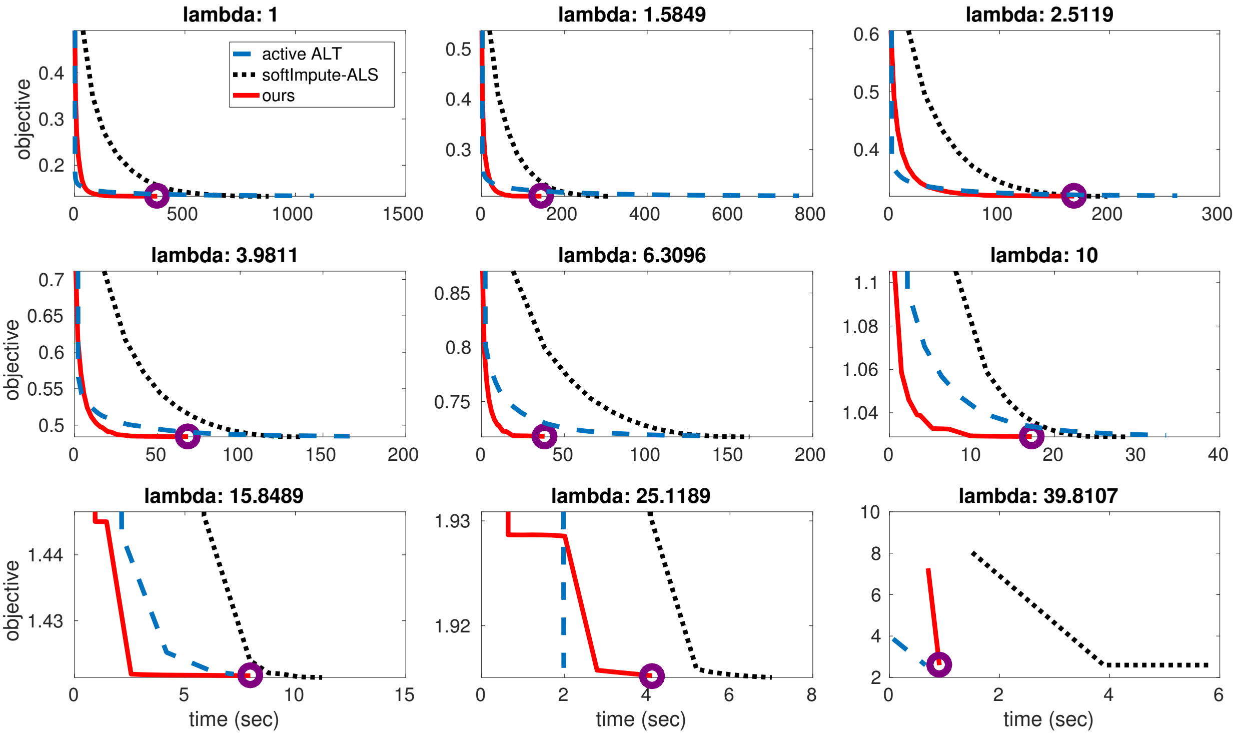

Figure 3. Convergence of the objective function on *ml*100*k* for various lambda values. The center plot is for the optimal lambda value based on the validation errors. The circle indicates that our proposed global optimality criterion has been satisfied.

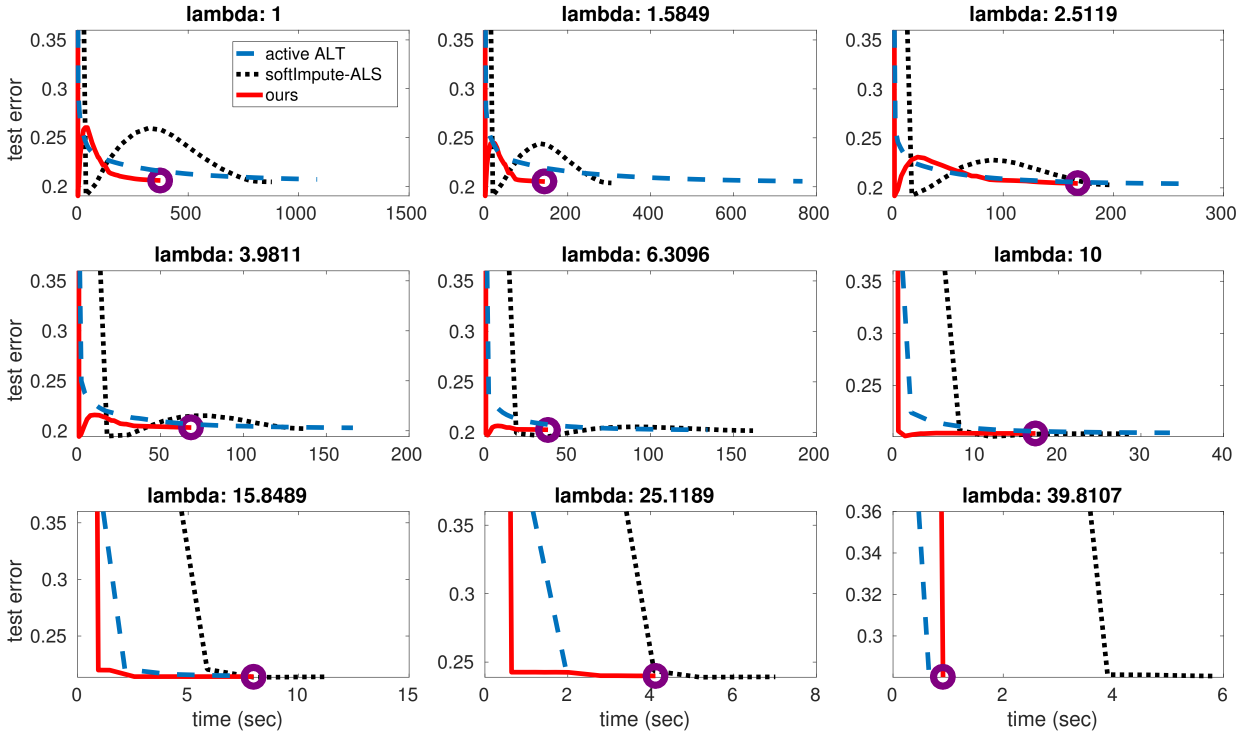

Figure 4. Test error on *ml*100*k* for various lambda values. The center plot is for the optimal lambda value based on the validation errors. The circle indicates that our proposed global optimality criterion has been satisfied on the original problem.