
Best Pair Formulation & Accelerated Scheme for Non-convex Principal Component Pursuit

Aritra Dutta
KAUST
Thuwal, KSA
aritra.dutta@kaust.edu.sa

Filip Hanzely
KAUST
Thuwal, KSA
filip.hanzely@kaust.edu.sa

Jingwei Liang
University of Cambridge
Cambridge, UK
j1993@cam.ac.uk

Peter Richtárik
KAUST
Thuwal, KSA
peter.richtarik@kaust.edu.sa

Abstract

The *best pair* problem aims to find a pair of points that minimize the distance between two disjoint sets. In this paper, we formulate the classical robust principal component analysis (RPCA) as the best pair; which was not considered before. We design an accelerated proximal gradient scheme to solve it, for which we show global convergence, as well as the local linear rate. Our extensive numerical experiments on both real and synthetic data suggest that the algorithm outperforms relevant baseline algorithms in the literature.

1 Introduction

Let $A \in \mathbb{R}^{m \times n}$ be a given matrix, the generalized low-rank recovery model can be written as

$$\min_{L \in \mathbb{R}^{m \times n}} \mathcal{F}(A, L) + \lambda \mathcal{R}(L), \quad (1)$$

where $\mathcal{F}(A, L)$ is a loss function, $\mathcal{R}(L) \stackrel{\text{def}}{=} \sum_{i=1}^n \mathcal{R}_i(L)$ is a suitable regularizer, and $\lambda > 0$ is a balancing parameter. By an appropriate choice of the loss function and the regularizer, (1) can express a wide range of low-rank approximation problems of matrices. For example, by setting $\mathcal{F}(A, L) = \|A - L\|_F^2$, $\lambda = 1$, and $\mathcal{R}(L) = \iota_{\text{rank}(L) \leq r}(L)$ — the characteristic function (10) of the set $\{L \in \mathbb{R}^{m \times n} : \text{rank}(L) \leq r\}$, (1), specializes to:

$$\min_{L \in \mathbb{R}^{m \times n}} \|A - L\|_F^2 + \iota_{\text{rank}(L) \leq r}(L), \quad (2)$$

which is a *best approximation* formulation of the classical principal component analysis (PCA). The solution to problem (2) is given by: $\hat{L} = U \mathbf{H}_r(\Sigma) V^\top$, where $U \Sigma V^\top = A$ is a singular value decomposition (SVD) of A and $\mathbf{H}_r(\cdot)$ is the hard-thresholding operator that keeps the r largest singular values. Although PCA is vastly used and a successful designing tool in different engineering applications, it can only handle the presence of uniformly distributed noise and is rather sensitive to sparse outliers in the data matrix (Lin et al., 2010; Wright et al., 2009; Candès et al., 2011). To overcome this shortcoming and to deal with sparse errors, (Chandrasekaran et al., 2011; Candès et al., 2011) replaced the Frobenius norm in (2) by the ℓ_0 pseudo norm, and introduced the celebrated *principal component pursuit* (PCP) problem:

$$\min_{L \in \mathbb{R}^{m \times n}} \|A - L\|_{\ell_0} + \lambda \text{rank}(L). \quad (3)$$

However, the above problem is non-convex and NP-hard. One of the most commonly used, tractable surrogate reformulations of (3) is replacing the rank function with nuclear norm $\|L\|_*$ and ℓ_0 pseudo

norm with ℓ_1 -norm $\|A - L\|_{\ell_1}$ (Cai et al., 2010; Recht et al., 2010). Exploiting this idea, *Robust PCA* (RPCA) was introduced as a convex surrogate of the PCP problem (Wright et al., 2009; Lin et al., 2010; Candès et al., 2011):

$$\min_{L \in \mathbb{R}^{m \times n}} \|A - L\|_{\ell_1} + \lambda \|L\|_{\star}. \quad (4)$$

It was shown in (Chandrasekaran et al., 2011; Candès et al., 2011) that under a rank-sparsity incoherence assumption, problem (3) can be provably solved via (4), as the solutions of them lie close to each other with high probability.

Besides (4), there are other formulations of RPCA. One of the most popular way is to introduce an auxiliary variable, S , and add an additional constraint $L + S = A$, which yields:

$$\min_{L, S \in \mathbb{R}^{m \times n}} \|S\|_{\ell_1} + \lambda \|L\|_{\star} \quad \text{subject to} \quad L + S = A. \quad (5)$$

This *constrained* formulation enables several avenues to solve RPCA, such as, the exact and inexact augmented Lagrangian method of multipliers by Lin et al. (Lin et al., 2010), accelerated proximal gradient method (Wright et al., 2009), alternating direction method (Yuan and Yang, 2013), alternating projection with intermediate denoising (Netrapalli et al., 2014), dual approach (Lin et al., 2009), and SpaRCS (Waters et al., 2011), manifold optimization by Yi et al. (Yi et al., 2016) and Zhang and Yang (Zhang and Yang, 2018), are a few popular ones. We refer to (Bouwmans and Zahzah, 2014) for a comprehensive review of RPCA algorithms.

For the discussion above, A is fully observed with no data missing. One can consider that A is partially observed, that is, there exists a projection operator (or simply a Bernoulli binary mask) P_{Ω} on the set of observed data entries $\Omega \subseteq [m] \times [n]$ and is defined by

$$(P_{\Omega}[A])_{ij} = \begin{cases} A_{ij} & (i, j) \in \Omega, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

The partial observed version of (5) reads

$$\min_{L, S \in \mathbb{R}^{m \times n}} \|S\|_{\ell_1} + \lambda \|L\|_{\star} \quad \text{subject to} \quad P_{\Omega}(L + S) = P_{\Omega}(A). \quad (7)$$

Besides (5) and (7), other tractable reformulations of (3) still exist. For example, if the rank and target sparsity is user-inferred then it is common practice to relax the equality constraint in (5) and consider it in the objective function as a penalty. This, together with explicit constraints on the target rank, r , and target sparsity level, α , (*user-inferred* hyperparameters), leads to the GoDec formulation (Zhou and Tao, 2011). One can also extend the above model to the case of partially observed data that leads to a more general class of problems that is commonly known as the *robust matrix completion (RMC)* problem (Chen et al., 2011; Tao and Yang, 2011; Cherapanamjeri et al., 2017b,a) that contains the variant proposed in (Zhou and Tao, 2011) as a special case. With $S = 0$, the matrix completion (MC) problem is also a special case of the RMC problem (Candès and Plan, 2009; Jain et al., 2013; Cai et al., 2010; Jain and Netrapalli, 2015; Candès and Recht, 2009; Keshavan et al., 2010; Candès and Tao, 2010; Mareček et al., 2017; Wen et al., 2012). Lastly, when the whole matrix is observed, the RMC problem is nothing but (5).

Recently, (Dutta et al., 2018a) reformulated (3) as a non-convex feasibility problem, which does not require any objective function, convex relaxation, or surrogate convex constraints. Rather, it exploits the following idea: the solution to the PCP problem lies in the intersection of two sets—one convex and one non-convex, if one considers both the target rank r and the target sparsity α as hyperparameters. Let $X = \begin{pmatrix} S \\ L \end{pmatrix} \in \mathbb{R}^{2m \times n}$ and $K = [\text{Id}, \text{Id}]$ where Id is the identity operator of $\mathbb{R}^{m \times n}$, define

$$\mathcal{X} \stackrel{\text{def}}{=} \{X : KX = A\}, \quad \mathcal{Y} \stackrel{\text{def}}{=} \{X : \text{rank}(L) \leq r, \|S_{i,\cdot}\|_0 \leq \alpha m, \|S_{\cdot,j}\|_0 \leq \alpha n, i \in [m], j \in [n]\}.$$

Note that \mathcal{X} is convex and \mathcal{Y} is non-convex¹. Given the sets, Dutta et al. (Dutta et al., 2018a) reformulated (3) as non-convex feasibility problem:

$$\text{find } X \in \mathbb{R}^{2m \times n} \text{ such that } X \in \mathcal{X} \cap \mathcal{Y}. \quad (8)$$

Note that if we replace the Id in K with Bernoulli binary matrix, then we obtain the reformulation of PCP problem with partial observation.

¹The α -sparsity constraint on S means that for $\alpha \in (0, 1)$, each row and column of S contains no more than αn and αm number of non-zero entries, respectively. This is slightly more complicated than directly applying $\|\cdot\|_0$ constraint. However, it often works better in practice.

1.1 Formulation and Contributions

In this paper we consider reformulating the feasibility problem (8) as a *best pair* problem. Given two sets $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^{2m \times n}$, the best pair problems aims to find a pair of points $(X^*, Y^*) \in \mathcal{X} \times \mathcal{Y}$ such that they have the closest distance, that is (X^*, Y^*) a the solution of the problem below:

$$\min_{X \in \mathcal{X}, Y \in \mathcal{Y}} \frac{1}{2} \|X - Y\|^2. \quad (9)$$

When the intersection of \mathcal{X} and \mathcal{Y} is non-empty, that is $\mathcal{X} \cap \mathcal{Y} \neq \emptyset$, (9) reduces to the feasibility problem, with $X^* = Y^* \in \mathcal{X} \cap \mathcal{Y}$. Given a set \mathcal{X} , define its characteristic function by

$$\iota_{\mathcal{X}}(X) \stackrel{\text{def}}{=} \begin{cases} 0 : X \in \mathcal{X}, \\ +\infty : \text{otherwise}. \end{cases} \quad (10)$$

Then (9) can be equivalently written as

$$\min_{X, Y \in \mathbb{R}^{2m \times n}} \iota_{\mathcal{X}}(X) + \frac{1}{2} \|X - Y\|^2 + \iota_{\mathcal{Y}}(Y). \quad (11)$$

Observe that for a given Y , problem (11) becomes $\min_{X \in \mathbb{R}^{2m \times n}} \iota_{\mathcal{X}}(X) + \frac{1}{2} \|X - Y\|^2$ which is the Moreau envelope (Bauschke and Combettes, 2011) of $\iota_{\mathcal{X}}(X)$ of index 1:

$${}^1(\iota_{\mathcal{X}}(Y)) \stackrel{\text{def}}{=} \min_{X \in \mathbb{R}^{2m \times n}} \frac{1}{2} \|X - Y\|^2 + \iota_{\mathcal{X}}(X).$$

As a result, we can simplify (11) to the case of only Y ,

$$\boxed{\min_{Y \in \mathbb{R}^{2m \times n}} \iota_{\mathcal{Y}}(Y) + {}^1(\iota_{\mathcal{X}}(Y)).} \quad (12)$$

For the rest of the paper, we focus on (12) and our main contributions are summarised below:

- **New formulation and a new algorithm for non-convex PCP.** We reformulate the non-convex set feasibility formulation of RPCA to a *best pair* problem. Although our formulation was inspired by formulation (8) from (Dutta et al., 2018a), to the best of our knowledge, we are the first to formulate and solve RPCA via the best pair. To this end, we design a fast and efficient algorithm—an accelerated proximal gradient method—to solve it.
- **Theoretical convergence guarantees.** Both global and local convergence analysis of the scheme are provided. Globally, we show that our algorithm converges to a critical point. If the algorithm additionally starts sufficiently close to the optimum, we show that it converges to a global minimizer. Locally, our algorithm enjoys a fast linear rate, which we can sharply estimate. We owe this novelty to our best pair formulation. In contrast, the non-convex projection RPCA from (Dutta et al., 2018a) or GoDec (Zhou and Tao, 2011) can only guarantee a local linear convergence.
- **Numerical experiments and applications to real-world problems.** We apply the proposed method to several well-tested applications in computer vision. Our extensive experiments on both real and synthetic data suggest that our algorithm matches or outperforms relevant baseline algorithms in *fractions* of their execution time. Additionally, in the supplementary material, we provide empirical validity of the hyperparameters sensitivity of our approach.

1.2 Notations

Throughout the paper, \mathbb{N} is the set of non-negative integers. For a nonempty closed convex set $\Omega \subset \mathbb{R}^n$, denote P_Ω the orthogonal projector onto Ω . Let $\mathcal{R} : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ be a lower semi-continuous (lsc) function, its domain is defined as $\text{dom}(\mathcal{R}) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : \mathcal{R}(x) < +\infty\}$, and it is said to be proper if $\text{dom}(\mathcal{R}) \neq \emptyset$. We need the following notions from variational analysis, see e.g. (Rockafellar and Wets, 1998) for details. Given $x \in \text{dom}(\mathcal{R})$, the Fréchet subdifferential $\partial^F \mathcal{R}(x)$ of \mathcal{R} at x , is the set of vectors $v \in \mathbb{R}^n$ that satisfies $\liminf_{z \rightarrow x, z \neq x} \frac{1}{\|z-x\|} (\mathcal{R}(z) - \mathcal{R}(x) - \langle v, z-x \rangle) \geq 0$. If $x \notin \text{dom}(\mathcal{R})$, then $\partial^F \mathcal{R}(x) = \emptyset$. The limiting-subdifferential (or simply subdifferential) of \mathcal{R} at x , written as $\partial \mathcal{R}(x)$, is defined as $\partial \mathcal{R}(x) \stackrel{\text{def}}{=} \{v \in \mathbb{R}^n : \exists x_k \rightarrow x, \mathcal{R}(x_k) \rightarrow \mathcal{R}(x), v_k \in \partial^F \mathcal{R}(x_k) \rightarrow v\}$. Denote $\text{dom}(\partial \mathcal{R}) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n : \partial \mathcal{R}(x) \neq \emptyset\}$. Both $\partial^F \mathcal{R}(x)$ and $\partial \mathcal{R}(x)$ are closed, with $\partial^F \mathcal{R}(x)$ convex and $\partial^F \mathcal{R}(x) \subset \partial \mathcal{R}(x)$ (Rockafellar and Wets, 1998, Proposition 8.5). Since \mathcal{R} is lsc, it is (subdifferentially) regular at x if and only if $\partial^F \mathcal{R}(x) = \partial \mathcal{R}(x)$ (Rockafellar and Wets, 1998, Corollary 8.11). A necessary condition for x to be a minimizer of \mathcal{R} is $0 \in \partial \mathcal{R}(x)$. The set of critical points of \mathcal{R} is $\text{crit}(\mathcal{R}) = \{x \in \mathbb{R}^n : 0 \in \partial \mathcal{R}(x)\}$.

2 An accelerated proximal gradient method

In this section, we describe a gradient-based optimization method for solving (12). Denote $P_{\mathcal{X}}, P_{\mathcal{Y}}$ the projection operators onto \mathcal{X} and \mathcal{Y} , respectively. Since \mathcal{X} is a non-empty closed convex set, its characteristic function $\iota_{\mathcal{X}}$ is proper convex and lower semi-continuous. Owing to (Bauschke and Combettes, 2011), the Moreau envelope is convex differentiable with gradient reads

$$\nabla(\iota_{\mathcal{X}}(Y)) = (\text{Id} - P_{\mathcal{X}})(Y)$$

which is 1-Lipschitz continuous. Clearly, (12) admits a “non-smooth + smooth” structure, and in literature one prevailing algorithm to apply is the proximal gradient method (Lions and Mercier, 1979), a.k.a. Forward–Backward splitting. In this paper, we consider an accelerated version of the method, see Algorithm 1, which is based on inertial technique.

Algorithm 1: An accelerated proximal gradient method

Initial: Let $\gamma \in]0, 2]$ and choose $Y_0 \in \mathbb{R}^{2m \times n}, Y_{-1} = Y_0$.

repeat

$$\begin{aligned} Z_{a,k} &= Y_k + a_k(Y_k - Y_{k-1}), \\ Z_{b,k} &= Y_k + b_k(Y_k - Y_{k-1}), \end{aligned} \tag{13}$$

$$Y_{k+1} = P_{\mathcal{Y}}(Z_{a,k} - \gamma(Z_{b,k} - P_{\mathcal{X}}(Z_{b,k}))).$$

$$k = k + 1;$$

until convergence;

Remark 2.1.

- If we choose $\gamma = 1$ and $a_k, b_k \equiv 0$, Algorithm 1 becomes the Backward–Backward splitting, which is the method of alternating projections for the considered feasibility problem (8). Therefore, we recover the method from (Dutta et al., 2018a) as a special case.
- From (8) to (12), we can also consider the Moreau envelope of the non-convex set \mathcal{Y} , that is

$$\min_{X \in \mathbb{R}^{2m \times n}} \iota_{\mathcal{X}}(X) + \iota_{\mathcal{Y}}(X),$$

which also works well in practice.

- Algorithm 1 is a special cases of the multi-step inertial proximal gradient descent method considered in (Liang et al., 2016) for general non-convex composite optimization.

Note that the two projection operators $P_{\mathcal{X}}, P_{\mathcal{Y}}$ are very easy to compute. Given $X = \begin{pmatrix} S \\ L \end{pmatrix}$, since \mathcal{X} is an affine subspace, the projection of X onto \mathcal{X} reads $P_{\mathcal{X}}(X) = \frac{1}{2} \begin{pmatrix} A + S - L \\ A - S + L \end{pmatrix}$. If $K = [P_{\Omega}, P_{\Omega}]$ where P_{Ω} is the binary mask defined in (6), then for the partial observed case, we have

$$P_{\mathcal{X}}(X) = \begin{pmatrix} S \\ L \end{pmatrix} + \frac{1}{2} \begin{pmatrix} P_{\Omega}[A - S - L] \\ P_{\Omega}[A - S + L] \end{pmatrix}.$$

For the projection $P_{\mathcal{Y}}$ which contains a low-rank projection and sparsity projection, we refer to (Dutta et al., 2018a) for more details.

2.1 Global convergence

Since set \mathcal{Y} is semi-algebraic (Bolte et al., 2010), our global convergence guarantees of Algorithm 1 is based on Kurdyka-Łojasiewicz property.

Kurdyka-Łojasiewicz property. Let $R : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper lsc function. For η_1, η_2 such that $-\infty < \eta_1 < \eta_2 < +\infty$, define the set

$$[\eta_1 < R < \eta_2] \stackrel{\text{def}}{=} \{Y \in \mathbb{R}^n : \eta_1 < R(Y) < \eta_2\}.$$

Definition 2.2. Function R is said to have the Kurdyka-Łojasiewicz property at $\bar{Y} \in \text{dom}(R)$ if there exists $\eta \in]0, +\infty]$, a neighbourhood U of \bar{Y} and a continuous concave function $\varphi : [0, \eta] \rightarrow \mathbb{R}_+$ such that

- (i) $\varphi(0) = 0$, φ is C^1 on $]0, \eta[$, and for all $s \in]0, \eta[$, $\varphi'(s) > 0$;
- (ii) for all $Y \in U \cap [R(\bar{Y}) < R < R(\bar{Y}) + \eta]$, the Kurdyka-Łojasiewicz inequality holds

$$\varphi'(R(Y) - R(\bar{Y})) \text{dist}(0, \partial R(Y)) \geq 1. \quad (14)$$

Proper lsc functions which satisfy the Kurdyka-Łojasiewicz property at each point of $\text{dom}(\partial R)$ are called KL functions.

KL functions include the class of semi-algebraic functions, see (Bolte et al., 2007, 2010). For instance, the ℓ_0 pseudo-norm and the rank function are KL.

Global convergence. To deliver the convergence result, we rewrite (12) into the following generic form

$$\min_{Y \in \mathbb{R}^{2m \times n}} \{\Phi(Y) \stackrel{\text{def}}{=} \mathcal{R}(Y) + \mathcal{F}(Y)\}, \quad (15)$$

where we assume that

- (A.1) $\mathcal{R} : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ is proper lower semi-continuous, and bounded from below;
- (A.2) $\mathcal{F} : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex differentiable and its gradient $\nabla \mathcal{F}$ is L -Lipschitz continuous.

Let $\nu > 0$ be a constant. Define the following quantities,

$$\beta_k \stackrel{\text{def}}{=} \frac{1 - \gamma L - a_k - \nu}{2\gamma}, \quad \underline{\beta} \stackrel{\text{def}}{=} \liminf_{k \in \mathbb{N}} \beta_k \quad \text{and} \quad \alpha_k \stackrel{\text{def}}{=} \frac{\gamma b_k^2 L^2 + \nu a_k}{2\nu\gamma}, \quad \bar{\alpha} \stackrel{\text{def}}{=} \limsup_{k \in \mathbb{N}} \alpha_k. \quad (16)$$

Theorem 2.3 (Global convergence). For problem (15), assume (A.1)-(A.2) hold, and that Φ is a proper lsc KL function which is bounded from below. For Algorithm 1, choose ν, γ, a_k, b_k such that

$$\delta \stackrel{\text{def}}{=} \underline{\beta} - \bar{\alpha} > 0. \quad (17)$$

Then each bounded sequence $\{Y_k\}_{k \in \mathbb{N}}$ satisfies

- (i) $\{Y_k\}_{k \in \mathbb{N}}$ has finite length, i.e. $\sum_{k \in \mathbb{N}} \|Y_k - Y_{k-1}\| < +\infty$;
- (ii) There exists a critical point $Y^* \in \text{crit}(\Phi)$ such that $\lim_{k \rightarrow \infty} Y_k = Y^*$.
- (iii) If Φ has the KL property at a global minimizer Y^* , then starting sufficiently close from Y^* , any sequence $\{Y_k\}_{k \in \mathbb{N}}$ converges to a global minimum of Φ and satisfies (i).

The proof of the above theorem can be found in the supplementary material. We also refer to (Liang et al., 2016) and the reference therein for more results on non-convex proximal gradient method.

2.2 Local linear convergence

Now we turn to the local perspective and present a local linear convergence analysis for Algorithm 1. For the constraint set \mathcal{Y} define in (8), consider the following decomposition of it

$$\mathcal{Y}_L \stackrel{\text{def}}{=} \left\{ Y = \begin{pmatrix} S \\ L \end{pmatrix} : \text{rank}(L) \leq r \right\} \quad \text{and} \quad \mathcal{Y}_S \stackrel{\text{def}}{=} \left\{ Y = \begin{pmatrix} S \\ L \end{pmatrix} : S \text{ is } \alpha\text{-sparse} \right\}.$$

For the sequence Y_k generated by (13), suppose $Y_k = \begin{pmatrix} S_k \\ L_k \end{pmatrix}$. It is immediate that $\text{rank}(L_k) \leq r$ holds for all k . For S_k , though it is always α -sparse, the locations of non-zero elements change along the course of iteration. In the following, we first show that after a finite number of iterations the locations of non-zero elements of S_k stop changing, that is S_k will have the same support as that of S^* to which S_k converges, and then Algorithm 1 enters a linear convergence regime.

Support identification of S_k . Let $Y^* = \begin{pmatrix} S^* \\ L^* \end{pmatrix}$ be a critical point of (12) to which Y_k converges. Let \mathcal{S} be the subspace extended by the support of S^* . Clearly, $S^* \in \mathcal{S}$ and we have the result below concerning the relation between S_k and \mathcal{S} .

Theorem 2.4 (Support identification). For Algorithm 1, suppose Theorem 2.3 holds. Then Y_k converges to a critical point Y^* of (12). For all k large enough, we have $S_k \in \mathcal{S}$.

Let S^* be the point that S_k converges to, the above result simply means that after finite number of iterations, $\text{supp}(S_k) = \text{supp}(S^*)$ holds for all k large enough.

Local linear convergence. Given a critical point Y^* , let $X^* = P_{\mathcal{X}}(Y^*)$, we have

$$X^* \in \mathcal{X} \text{ and } S^* \in \mathcal{S}, \quad L^* \in \mathcal{Y}_L.$$

Note that the first two sets, \mathcal{X}, \mathcal{S} are (affine) subspaces, hence smooth, and \mathcal{Y}_L is the set of fixed-rank matrices which is C^2 -smooth manifold (Lee, 2003). To derive the local linear rate, we need to utilize the smoothness of these sets. Let \mathcal{M} be a C^2 -smooth manifold and let $\mathcal{T}_{\mathcal{M}}(X)$ the tangent space of \mathcal{M} at $X \in \mathcal{M}$, we have the following lemma which is crucial for our local linear convergence analysis.

Lemma 2.5 ((Liang et al., 2014, Lemma 5.1)). *Let \mathcal{M} be a C^2 -smooth manifold around X . Then for any $X' \in \mathcal{M} \cap \mathcal{N}$, where \mathcal{N} is a neighbourhood of X , the projection operator $P_{\mathcal{M}}(X')$ is uniquely valued and C^1 around X , and thus $X' - X = P_{\mathcal{T}_{\mathcal{M}}(X)}(X' - X) + o(\|X' - X\|)$. If moreover, $\mathcal{M} = X + \mathcal{T}_{\mathcal{M}}(X)$ is an affine subspace, then $X' - X = P_{\mathcal{T}_{\mathcal{M}}(X)}(X' - X)$.*

Denote the tangent spaces of \mathcal{X}, \mathcal{Y} at X^*, Y^* as $T_{\mathcal{X}}^{X^*}$ and $T_{\mathcal{Y}}^{Y^*}$, respectively. We refer to the supplementary material for detailed expressions of these tangent spaces. Denote $P_{T_{\mathcal{X}}^{X^*}}$ and $P_{T_{\mathcal{Y}}^{Y^*}}$ the projections onto the tangent spaces. Define the matrix $\mathcal{P} \stackrel{\text{def}}{=} P_{T_{\mathcal{Y}}^{Y^*}}((1 - \gamma)\text{Id} + \gamma P_{T_{\mathcal{X}}^{X^*}})P_{T_{\mathcal{Y}}^{Y^*}}$, and

$$D_k \stackrel{\text{def}}{=} \begin{pmatrix} Y_k - Y^* \\ Y_{k-1} - Y^* \end{pmatrix} \quad \text{and} \quad \mathcal{Q} \stackrel{\text{def}}{=} \begin{bmatrix} (1+a)\mathcal{P} & -a\mathcal{P} \\ \text{Id} & 0 \end{bmatrix} \quad \text{with } a \in [0, 1].$$

Denote $\rho_{\mathcal{P}}, \rho_{\mathcal{Q}}$ the spectral radiuses of \mathcal{P}, \mathcal{Q} , respectively.

Theorem 2.6 (Local linear convergence). *For Algorithm 1, suppose Theorem 2.4 holds. Then Y_k converges to a critical point Y^* of (12). Suppose $b_k = a_k \equiv a \in [0, 1]$, there exists a $K > 0$ such that for all $k \geq K$,*

$$D_{k+1} = \mathcal{Q}D_k + o(\|D_k\|).$$

Moreover, if $\rho_{\mathcal{P}} < 1$, then so is $\rho_{\mathcal{Q}}$, and for all k large enough we have $\|Y_k - Y^*\| = O(\rho_{\mathcal{Q}}^k)$.

Remark 2.7.

- If $T_{\mathcal{X}}^{X^*} \cap T_{\mathcal{Y}}^{Y^*} = \{0\}$, then it can be shown that $\rho_{\mathcal{P}} < 1$.
- Given $\rho_{\mathcal{P}}, \rho_{\mathcal{Q}}$ can be expressed explicitly in terms of a and $\rho_{\mathcal{P}}$. For the case that $a_k \rightarrow a \in [0, 1]$ and $b_k \rightarrow b \in [0, 1]$, we refer to (Liang, 2016, Chapter 6) for detailed discussion on the local linear convergence analysis.

An numerical illustration on our theoretical rate estimation and practical observation is provided in the supplementary material Section C-Figure 13.

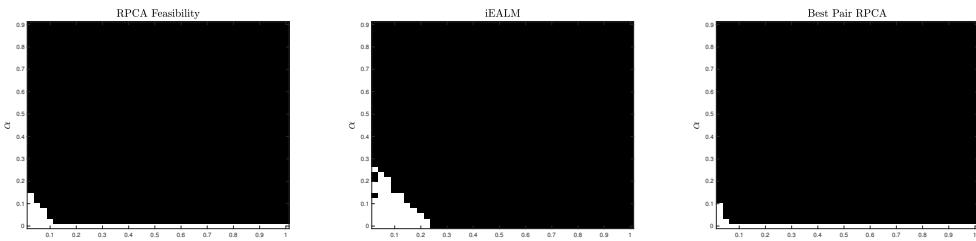


Figure 1: Phase transition diagram for RPCA F, iEALM, and APG with respect to rank and error sparsity. Here, $\rho_r = \text{rank}(L)/m$ and α is the sparsity measure. We have $(\rho_r, \alpha) \in (0.025, 1] \times (0, 1)$ with $r = 5 : 5 : 200$ and $\alpha = \text{linspace}(0, 0.99, 40)$. We perform 5 runs of each algorithm.

3 Numerical experiments

In this section, we extensively tested our best-pair formulation on both real and synthetic data against a vast genre of PCP algorithms. The first set of algorithms that we tested against, e.g. iEALM and APG, determine the target rank and sparsity *robustly* from the given set of hyperparameters. On the other hand, for the second set of algorithms, e.g. RPCA gradient descent (RPCA GD), Go decomposition (GoDec), and RPCA nonconvex feasibility (RPCA NCF), the target rank and sparsity are user-inferred. Although our accelerated proximal gradient algorithm belongs to the

second class, to show its effectiveness, we compare it with both classes of state-of-the-art robust PCP algorithms (see Table 2 in the supplementary material) on several computer vision applications—removal of shadows and specularities from face images, Background estimation or tracking from video sequences, and inlier detection from a grossly corrupted dataset (see Section A.3.1 in the supplementary material)².

Results on synthetic data. The primary goal of these set of experiments is to understand the behavior of our proposed method on some well-understood data and to test against some state-of-the-art algorithms. To construct our test matrix A , for these experiments, we used the idea proposed by Wright et al. (Wright et al., 2009). First, we generate the low-rank matrix, L , as a product of two independent full-rank matrices of size $m \times r$ with $r < m$ such that elements are independent and identically distributed (i.i.d.) and sampled from a normal distribution— $\mathcal{N}(0, 1)$. We generate the sparse matrix, S , such that its elements are chosen from the interval $[-500, 500]$. We create the sparse support set by using the operator (2). Finally, we write A as $A = L + S$. We fix $m = 200$ and define $\rho_r = \text{rank}(L)/m$, where $\text{rank}(L)$ varies. We choose the sparsity level $\alpha \in (0, 1)$.

Phase transition experiments. For each pair of (ρ_r, α) , we apply iEALM, RPCA NCF, and our algorithm to recover the pair (\hat{L}, \hat{S}) . For iEALM, we set $\lambda = 1/\sqrt{m}$ and use $\mu = 1.25/\|A\|_2$ and $\rho = 1.5$, where $\|A\|_2$ is the spectral norm (maximum singular value) of A . For a given $\varepsilon > 0$, if the recovered matrix pair (\hat{L}, \hat{S}) , satisfies the relative error $\frac{\|\hat{A} - \hat{L} - \hat{S}\|_F}{\|\hat{A}\|_F} < \varepsilon$ then we consider the construction is viable. In Figure 1, we produce the *phase transition diagrams* to show the fraction of perfect recovery of A , where white denotes *success* and black denotes *failure*. We run the experiments for 5 times and plot the results. The success of iEALM is approximately below the line $\rho_r + \alpha \approx 0.25$. On the other hand, we note that the performance of our best pair RPCA is almost similar to that of (Dutta et al., 2018a), when the sparsity level α is small and both approaches can efficiently provide a feasible reconstruction for any ρ_r in that case. We also note that for low sparsity level, iEALM can only provide a feasible reconstruction for $\rho_r \leq 0.25$. Due to their robustness to any low-rank structure when α is low, RPCA NCF and best pair RPCA can be proved to be very effective in many real-world applications. In many real-world problems, involving the video/image data can ideally have any inherent low-rank structure and are generally corrupted by very sparse outliers of arbitrary large magnitudes. In those instances, RPCA NCF and our best pair RPCA could be very useful. We show more justification in the later section.

Root mean square error measure. To validate our performance against RPCA GD of Yi et al. (Yi et al., 2016), we use a different metric—root mean square error (RMSE). Since RPCA GD does not explicitly recover a sparse matrix, S , it is unjustified to test it against the same relative error. Therefore, for the true low-rank, L , and a low-rank recovery, \hat{L} , we use the metric $\|L - \hat{L}\|_F/\sqrt{mn}$ as the measure of RMSE. From Figure 2, we can conclude that our best pair RPCA has less RMSE compare to that of RPCA GD. Moreover, the RMSE remains unaltered as the cardinality of support set, Ω increases. Also, see Figure 9 in the Appendix.

Removal of shadows and specularities. Set of images of an object under unknown pose and arbitrary lighting conditions, form a convex cone in the space of all possible images which may have *unbounded dimension* (Basri and Jacobs, 2003; Belhumeur and Kriegman, 1998). However, the images under distant, isotropic lighting can be approximated by a 9-dimensional linear subspace which is popularly referred to as the *harmonic plane*. We used three subjects B11, B12, and B13 from the Extended Yale Face Database (Georghiades et al., 2001) for our simulations. We used 63 downsampled images of resolution of 120×160 of each subject. For APG and iEALM, we set the parameters the same as in the previous section. For RPCA GD, RPCA NCF, and our method, we set target rank $r = 9$ and sparsity level to 0.1. The qualitative analysis on the recovered images from Figure 3 shows while RPCA GD recovers patchy and granular face images, our best pair reformulation provides comparable reconstruction to that of iEALM, APG, and RPCA NCF.

²In all experiments, we use the approximate projection (Dutta et al., 2018a; Yi et al., 2016; Zhang and Yang, 2018) onto \mathcal{Y} as the exact one is expensive:

$$\mathcal{T}_\alpha[S] \stackrel{\text{def}}{=} \{P_{\Omega_\alpha}(S) \in \mathbb{R}^{m \times n} : (i, j) \in \Omega_\alpha \text{ if } |S_{ij}| \geq |S_{(i,.)}^{(\alpha n)}| \text{ and } |S_{ij}| \geq |S_{(.j)}^{(\alpha m)}|\}.$$

If the sparsity constraint was defined only along rowsc (or only columns), the exact projection would be cheap. However, the approximate projection produces better results, thus we stick with it.

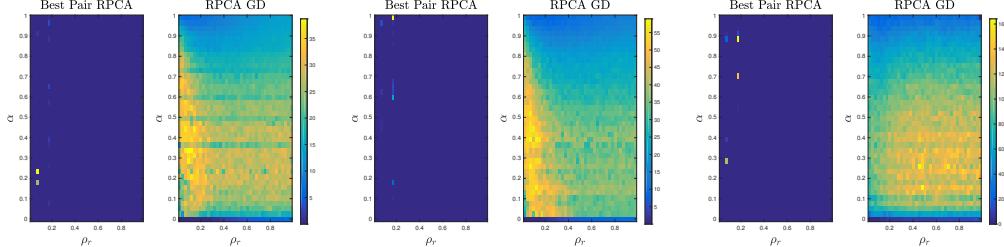


Figure 2: RMSE to compare between RPCA GD and best-pair RPCA with respect to rank and error sparsity. We set $\rho_r = \text{rank}(L)/m$ and α is the sparsity measure. We have $(\rho_r, \alpha) \in (0.025, 1] \times (0, 1)$ with $r = 5 : 5 : 200$ and $\alpha = \text{linspace}(0, 0.99, 40)$.

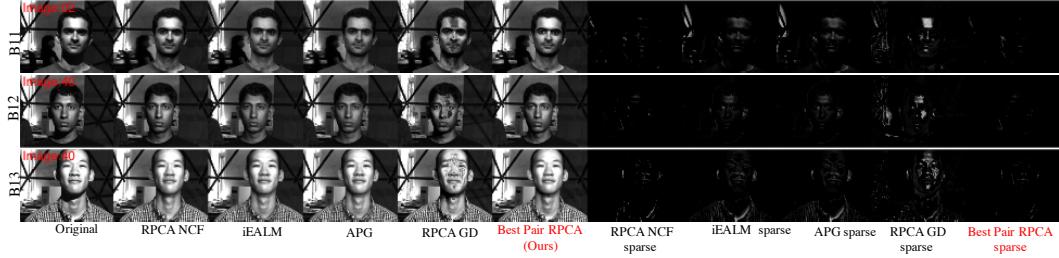


Figure 3: Shadow and specularities removal from face images captured under varying illumination and camera position. Our feasibility approach provides comparable reconstruction to that of iEALM, APG, and RPCA NCF.

Background estimation from video sequences. Background estimation or moving object tracking (Bouwmans et al., 2017; Dutta, 2016; Bouwmans et al., 2016; Dutta et al., 2017a; Bouwmans, 2014; Dutta et al., 2017b, 2018b; Dutta and Richtárik, 2019; Dutta and Li, 2017) is considered as one of the classic problems in computer vision and is used as a crucial component in human activity recognition, tracking, and video analysis from surveillance cameras. When the video is captured by a static camera, minimizing the rank of the matrix $A \in \mathbb{R}^{m \times n}$, that concatenates n video frames (after converting them into vectors) represents the structure of the linear subspace, L , that contains the background and an error, S , that emphasizes the foreground components. However, the exact desired rank is often tuned empirically, as the ideal rank-one background is often unrealistic as the changing illumination, occluded foreground/background objects, reflection, and noise are typically also a part of the video frames. Based on the above observation, we note that the problem can be cast typically as (4). However, as we explained in some cases, when the target rank and the sparsity level is user-inferred hyperparameters, one might use a different approach as in (Zhou and Tao, 2011; Dutta et al., 2018a; Yi et al., 2016) as well. Additionally, there might be missing/unobserved pixels in the video and that makes the problem more complex and only a few methods, such as RPCA NCF, GRASTA (He et al., 2012), RPCA GD remedy to that. Therefore, we tested our best pair RPCA to a wide range of methods. In our experiments, we use two different video sequences: (i) the Basic sequence from Stuttgart synthetic dataset (Brutzer et al., 2011), (ii) the waving tree sequence (Toyama et al., 1999). We extensively use the Stuttgart video sequence as it is a challenging sequence that comprises both static and dynamic foreground objects and varying illumination in

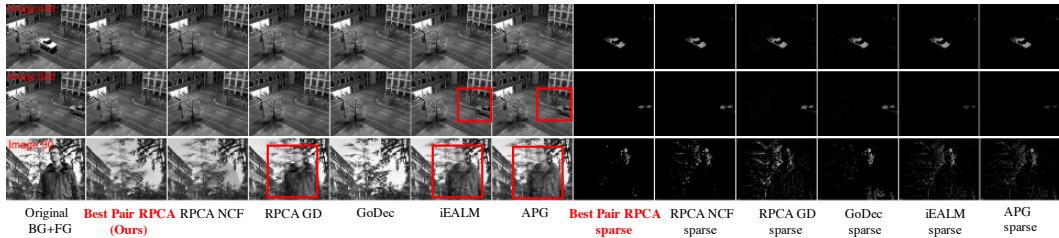


Figure 4: Background estimation from video sequences. Except Best pair RPCA, RPCA NCF, and GoDEC all other methods struggle to remove the static foreground object.

the background. Additionally, it comes with foreground ground truth for each frame. For iEALM and APG, we set the parameters the same as in the previous sections. For Best pair RPCA, RPCA GD, RPCA NCF, and GoDec, we set $r = 2$, target sparsity 10% and additionally, for GoDec, we set $q = 2$. For GRATSA, we set the parameters the same as those mentioned in the authors website (gra, 2012). The qualitative analysis on the background and foreground recovered on both, full observation (in Figure 4) and partial observation (in Figure 5), suggest that our method recovers a visually better quality background and foreground compare to the other methods. Note that, RPCA GD recovers a fragmentary foreground with more false positives compare to our method; moreover, RPCA GD, GRATSA, iEALM, and APG cannot remove the static foreground object. We provide a detailed quantitative evaluation of our best pair RPCA with respect to the ε -proximity metric- $d_\varepsilon(X, Y)$ as in (Dutta et al., 2018a) and the mean structural similarity index measure (SSIM) by (Wang et al., 2004) in recovering the foreground objects in Figures 10 and 11 in Appendix.

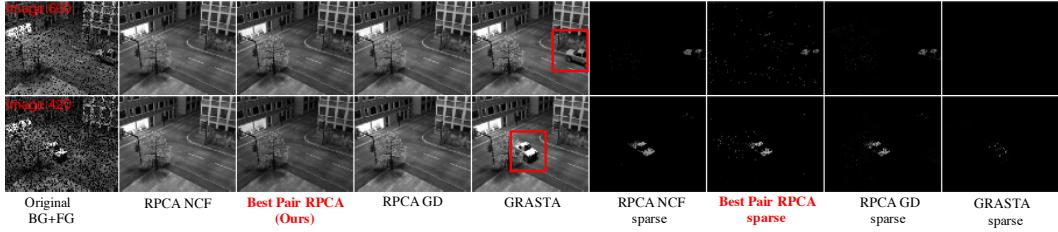


Figure 5: Background estimation on subsampled Stuttgart Basic video sequence. We use $\Omega = 0.9(m.n)$ and $\Omega = 0.8(m.n)$, respectively.

References

- 2012. <https://sites.google.com/site/hejunzz/grasta>.
- R. Basri and D. Jacobs. Lambertian reflection and linear subspaces. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 25(3):218–233, 2003.
- H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer, 2011.
- P. N. Belhumeur and D. J. Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision*, 28(3):245–260, 1998.
- J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.
- J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.
- T. Bouwmans. Traditional and recent approaches in background modeforeground detection: An overview. *Computer Science Review*, 11–12:31 – 66, 2014.
- T. Bouwmans and E.-H. Zahzah. Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding*, 122:22–34, 2014.
- T. Bouwmans, A. Sobral, S. Javed, S. K. Jung, and E.-H. Zahzah. Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset. *Computer Science Review*, 2016.
- T. Bouwmans, L. Maddalena, and A. Petrosino. Scene background initialization: A taxonomy. *Pattern Recognition Letters*, 96:3–11, 2017.
- S. Brutzer, B. Hoferlin, and G. Heidemann. Evaluation of background subtraction techniques for video surveillance. In *IEEE Computer Vision and Pattern Recognition*, pages 1937–1944, 2011.
- J. F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.
- E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2009.

- E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.
- E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.
- E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the Association for Computing Machinery*, 58(3):11:1–11:37, 2011.
- V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.
- Y. Chen, H. Xu, C. Caramanis, and S. Sanghavi. Robust matrix completion and corrupted columns. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 873–880, 2011.
- Y. Cherapanamjeri, K. Gupta, and P. Jain. Nearly optimal robust matrix completion. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 797–805, 2017a.
- Y. Cherapanamjeri, P. Jain, and P. Netrapalli. Thresholding based outlier robust PCA. In *Proceedings of the 30th Conference on Learning Theory (COLT)*, pages 593–628, 2017b.
- D. Drusvyatskiy and A. S. Lewis. Optimality, identifiability, and sensitivity. *Mathematical Programming*, pages 1–32, 2013.
- A. Dutta. *Weighted Low-Rank Approximation of Matrices: Some Analytical and Numerical Aspects*. PhD thesis, University of Central Florida, 2016.
- A. Dutta and X. Li. Weighted low rank approximation for background estimation problems. In *The IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1853–1861, 2017.
- A. Dutta and P. Richtárik. Online and batch supervised background estimation via l1 regression. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 541–550, 2019.
- A. Dutta, B. Gong, X. Li, and M. Shah. Weighted singular value thresholding and its application to background estimation. *arXiv:1707.00133*, 2017a.
- A. Dutta, X. Li, and P. Richtárik. A batch-incremental video background estimation model using weighted low-rank approximation of matrices. In *The IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1835–1843, 2017b.
- A. Dutta, F. Hanzely, and P. Richtárik. A nonconvex projection method for robust PCA, 2018a. To appear in Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), arXiv:1805.07962.
- A. Dutta, X. Li, and P. Richtárik. Weighted low-rank approximation of matrices and background modeling, 2018b. *arXiv:1804.06252*.
- L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.
- A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on PAMI*, 23(6):643–660, 2001.
- J. Goes, T. Zhang, R. Arora, and G. Lerman. Robust stochastic principal component analysis. In *Proceedings of the 17th International Conference on Articial Intelligence and Statistics*, pages 266–274, 2014.
- J. He, L. Balzano, and A. Szlam. Incremental gradient on the Grassmannian for online foreground and background separation in subsampled video. *IEEE Computer Vision and Pattern Recognition*, pages 1937–1944, 2012.
- P. Jain and P. Netrapalli. Fast exact matrix completion with finite samples. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, pages 1007–1034, 2015.
- P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing*, pages 665–674, 2013.
- R. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

- J. M. Lee. *Smooth manifolds*. Springer, 2003.
- A. S. Lewis. Active sets, non-smoothness, and sensitivity. *SIAM Journal on Optimization*, 13(3):702–725, 2003.
- A. S. Lewis and S. Zhang. Partial smoothness, tilt stability, and generalized Hessians. *SIAM Journal on Optimization*, 23(1):74–94, 2013.
- J. Liang. *Convergence Rates of First-Order Operator Splitting Methods*. PhD thesis, Normandie Université; GREYC CNRS UMR 6072, 2016.
- J. Liang, J. Fadili, and G. Peyré. Local linear convergence of Forward–Backward under partial smoothness. In *Advances in Neural Information Processing Systems*, pages 1970–1978, 2014.
- J. Liang, J. Fadili, and G. Peyré. A multi-step inertial Forward–Backward splitting method for non-convex optimization. In *Advances in Neural Information Processing Systems*, 2016.
- Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Yi Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. *UIUC Technical Report UILU-ENG-09-2214*, 2009.
- Z. Lin, M. Chen, and Y. Ma. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices, 2010. arXiv1009.5055.
- P. L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.
- J. Mareček, P. Richtárik, and M. Takáč. Matrix completion under interval uncertainty. *European Journal of Operational Research*, 256(1):35 – 43, 2017.
- P. Netrapalli, U. N. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems 27*, pages 1107–1115. 2014.
- B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.
- R. T. Rockafellar and R. Wets. *Variational analysis*, volume 317. Springer Verlag, 1998.
- M. Tao and J. Yang. Recovering low-rank and sparse components of matrices from incomplete and noisy observations. *SIAM Journal on Optimization*, 21(1):57–81, 2011.
- K Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. *Seventh International Conference on Computer Vision*, pages 255–261, 1999.
- Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transaction on Image Processing*, 13(4):600–612, 2004.
- A. E. Waters, A. C. Sankaranarayanan, and R. Baraniuk. SpaRCS: Recovering low-rank and sparse matrices from compressive measurements. *Proceedings of 24nd Advances in Neural Information Processing systems*, pages 1089–1097, 2011.
- Z. Wen, W. Yin, and Y. Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4(4):333–361, 2012.
- J. Wright, Y. Peng, Y. Ma, A. Ganseh, and S. Rao. Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization. *Proceedings of 22nd Advances in Neural Information Processing systems*, pages 2080–2088, 2009.
- X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust PCA via gradient descent. *Advances in Neural Information Processing systems*, pages 361–369, 2016.
- X. Yuan and J. Yang. Sparse and low-rank matrix decomposition via alternating direction methods. *Pacific Journal of Optimization*, 9(1):167–180, 2013.
- T. Zhang and Y. Yang. Robust PCA by manifold optimization. *Journal of Machine Learning Research*, 19: 1–39, 2018.
- T. Zhou and D. Tao. Godec: Randomized low-rank and sparse matrix decomposition in noisy case. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 33–40, 2011.

Supplementary Material

The organization of this supplementary material is: extra supporting numerical experiments are reported in Section A; Proofs for the global convergence result of Algorithm 1 is provided in Section B; The proof of local linear convergence and a numerical example are provided in Section C. Lastly, we provide a comprehensive table to list all baselines we compare to in Section D.

A Extra Experiments

In this section, we empirically study convergence properties of Algorithm 1 on synthetic, well-understood data. In particular, we examine its sensitivity to user-specified parameters γ, a_k, b_k , target sparsity level α , target rank r and lastly the sensitivity to initialization. Moreover, we provide extra phase transition diagrams and both quantitative and qualitative results on the inlier detection problem.

A.1 Sensitivity to the choice of γ, a_k, b_k

In this experiment, we compare different choices of algorithm parameters γ, a_k, b_k on instances of (9) with various target sparsity level α and target rank r . In each experiment, we make sure that the solution exists; we generate random matrices \tilde{L}, \tilde{S} (with independent entries $\mathcal{N}(0, 1)$), project them onto low rank and sparse constraint set respectively to obtain \hat{L}, \hat{S} and set $A = \hat{L} + \hat{S}$. For simplicity we consider only $a_k = b_k = a$ and $m = n = 100$. Figure 6 shows the result. We see that parameter choice $\gamma = 1.1, a_k = b_k = \frac{1}{2}$ is the most reliable.

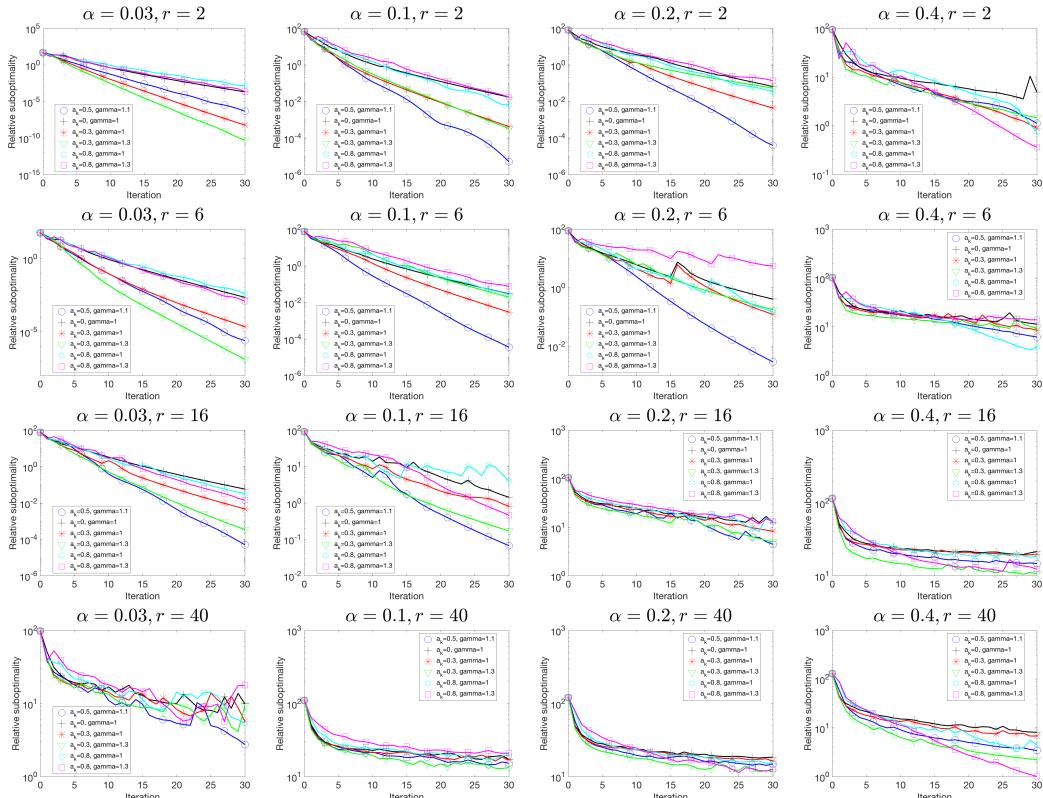


Figure 6: Sensitivity of Algorithm 1 with respect to choice of $\gamma, a_k = b_k$.

A.2 Sensitivity to the choice of r, α

In this experiment, we examine how sensitive is Algorithm 1 on the correct choice of the target sparsity level α and the target rank r .

In each experiment, we generate random matrices \tilde{L}, \tilde{S} (with independent entries $\mathcal{N}(0, 1)$), project them onto \hat{r} -low rank and $\hat{\alpha}$ -sparse constraint set respectively to obtain \hat{L}, \hat{S} and set $A = \hat{L} + \hat{S}$. Then, we run Algorithm 1 with various choices of r, α and report the results. For simplicity we consider only $\gamma = 1.1, a_k = b_k = \frac{1}{2}$ (from the previous experiment) and $m = n = 100$. Figure 7 shows the result. We can see that if sparsity level is underestimated, the method converges very slowly. Moreover, the method is more sensitive to the correct choices of target sparsity than target rank. The last take-away from this experiment is that over-estimation of target parameters usually leads to slightly slower convergence.

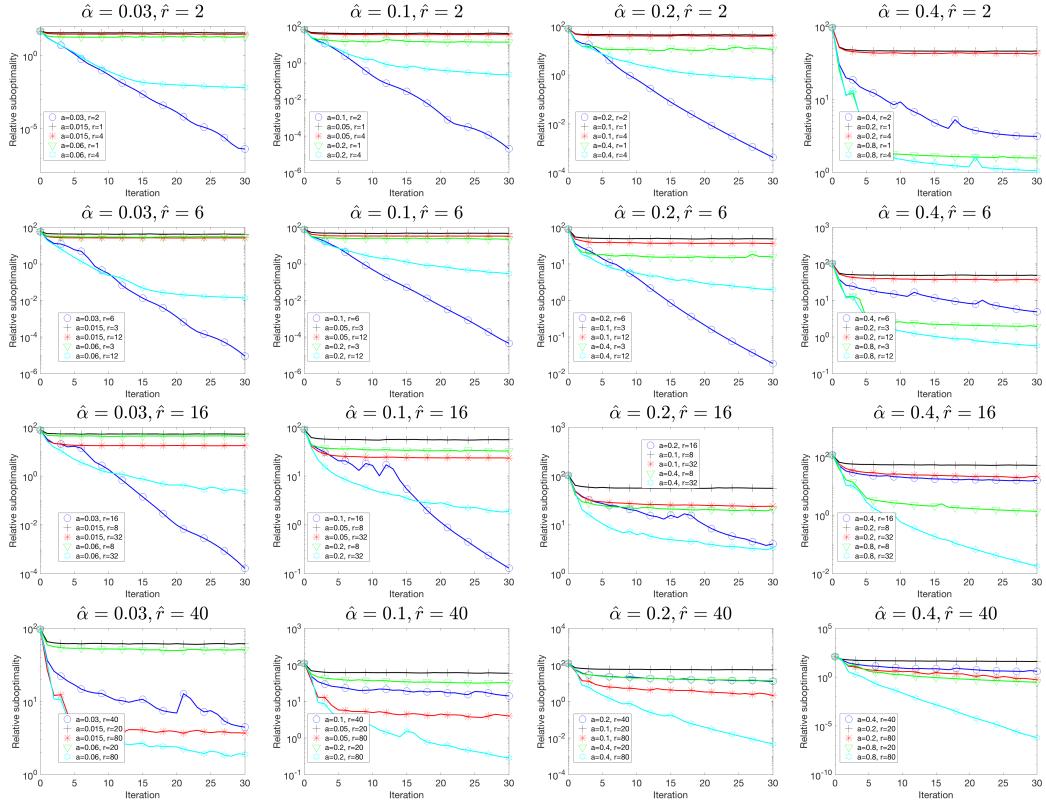


Figure 7: Sensitivity of Algorithm 1 with respect to the correct choice of target rank and target sparsity.

A.3 Sensitivity to the choice of the starting point

In the last experiment, we examine how the starting point influences the convergence rate. For each problem instance, we perform 50 independent runs of Algorithm 1 and report the best, worst and median performance.

For simplicity, we consider only problems with known target rank and sparsity – we generate random matrices \tilde{L}, \tilde{S} (with independent entries $\mathcal{N}(0, 1)$), project them onto low rank and sparse constraint set respectively to obtain \hat{L}, \hat{S} and set $A = \hat{L} + \hat{S}$. Further, we set $a_k = b_k = 0.5, \gamma = 1.1$ and $m = n = 100$. Figure 8 shows the result. We can see that the convergence speed of Algorithm 1 is, in most cases, not influenced significantly by the starting point. Thus, the non-convex nature of the problem is surprisingly not causing any issues. Lastly, the convergence rate of Algorithm 1 is faster for small values of α, r , which is often the most interesting case in terms of the practical application.

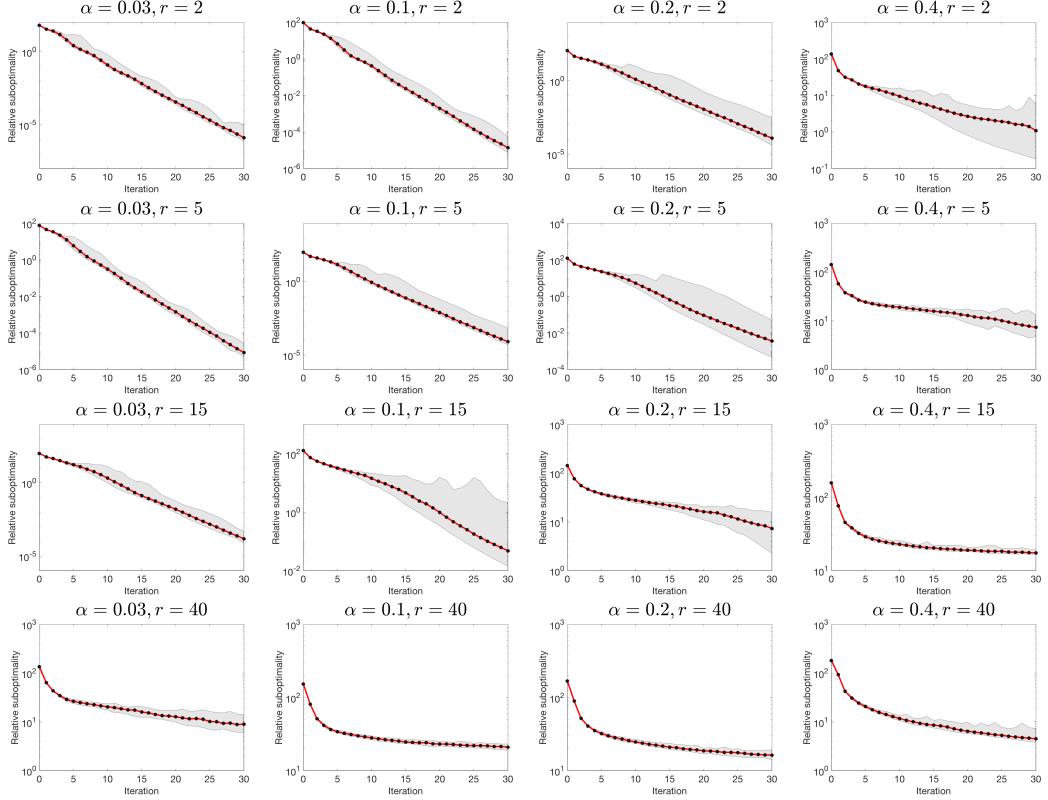


Figure 8: Sensitivity of Algorithm 1 with respect to the starting point.

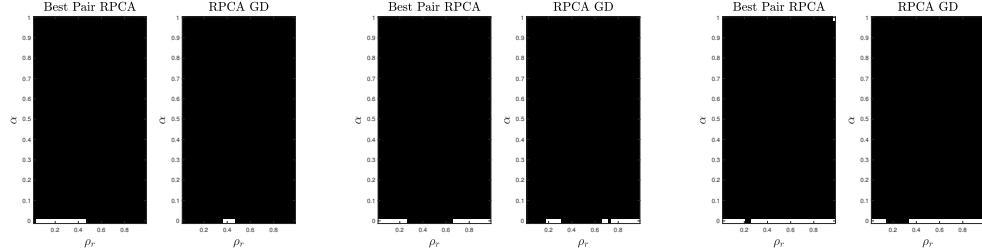


Figure 9: Phase transition diagram to compare between RPCA GD and best-pair RPCA with respect to rank and error sparsity. We set $\rho_r = \text{rank}(L)/m$ and α is the sparsity measure. We have $(\rho_r, \alpha) \in (0.025, 1] \times (0, 1)$ with $r = 5 : 5 : 200$ and $\alpha = \text{linspace}(0, 0.99, 40)$.

A.3.1 Inlier detection

Historically, PCA and RPCA are used in detecting the inliers and the outliers from a composite dataset. We infused 400 random, grayscale, downsampled (20×20 pixels) natural images from the BACKGROUND/Google folder of the Caltech101 database (Fei-Fei et al., 2007) with the Yale Extended Face Database to construct the data set. The inliers are the grayscale images of faces (of the same resolution) under different illuminations while the 400 random natural images serve as outliers. The goal is to consider a low-dimensional model and to project the inliers to a 9-dimensional linear subspace where the images of the same face lie. Goes et al. in (Goes et al., 2014) designed seven algorithms to explicitly find a low-rank subspace. To this end, Goes et al. used the classical SGD, an incremental approach, and mirror descent algorithms to find the 9-dimensional subspace. However, we split the dataset, A , into a 9-dimensional low-rank subspace L and expect

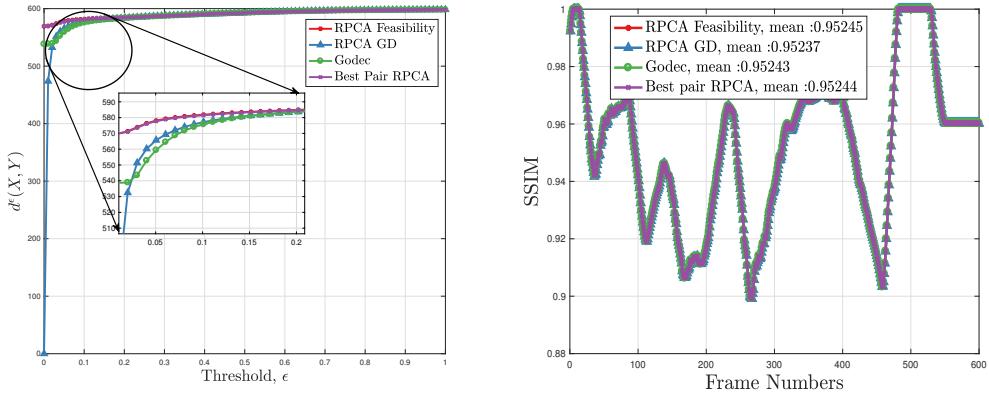


Figure 10: Quantitative comparison between different algorithms on Stuttgart Basic sequence. We compare the recovered foreground by different methods with respect to the foreground GT available for each frame on two different metrics: ε -proximity metric— $d_\varepsilon(X, Y)$ as in (Dutta et al., 2018a) and structural similarity index measure by (Wang et al., 2004). In recovering the foreground objects, our best pair RPCA is as robust as the other baseline methods with respect to both metrics.

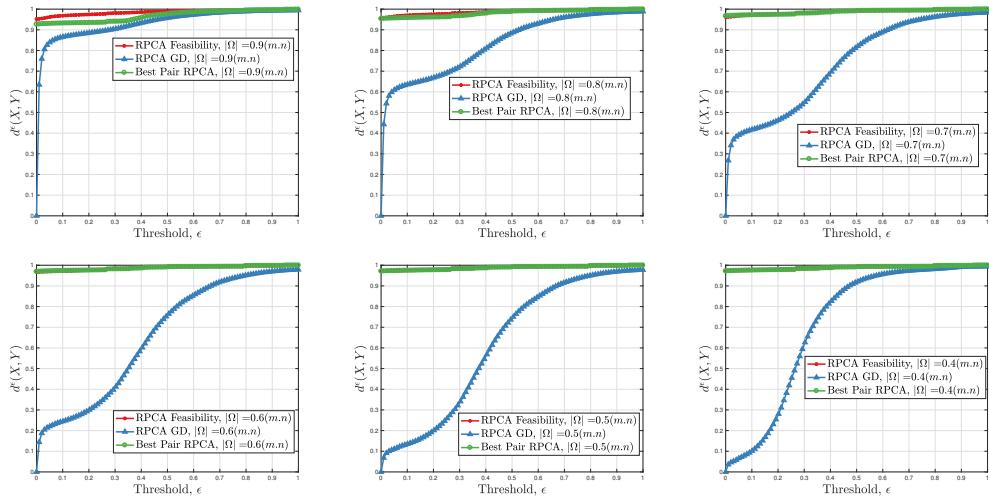


Figure 11: Quantitative comparison of foreground recovered by best pair RPCA, RPCA GD, and RPCA NCF Stuttgart Basic sequence, frame size 144×176 with observable entries: (a) $|\Omega| = 0.9(m.n)$, (b) $|\Omega| = 0.8(m.n)$, (c) $|\Omega| = 0.7(m.n)$, (d) $|\Omega| = 0.6(m.n)$, (e) $|\Omega| = 0.5(m.n)$, and (f) $|\Omega| = 0.4(m.n)$. The performance of RPCA GD drops significantly as $|\Omega|$ decreases. In contrast, the performance of our best pair RPCA and RPCA NCF stay stable irrespective of the size of $|\Omega|$.

the outliers to be in the sparse set, S . Once we find L , we find the basis of L via orthogonalization and project the faces on it. In Figure 12, we show the qualitative results of our experiments³.

As proposed in (Goes et al., 2014), we use the normalized error term $\|P_L - P_{L^*}\|_F / 3\sqrt{2}$, where L is subspace fitted by the PCA to the set of inliers and L^* be the subspace fitted by different algorithms. Note that, the metric is expected to lie between 0 and 1 where the smaller is the better. We refer to Table 1 for our quantitative results.

³The codes and datasets for experiments in Section A.3.1 are obtained from <https://github.com/jwgoes/RSPCA>

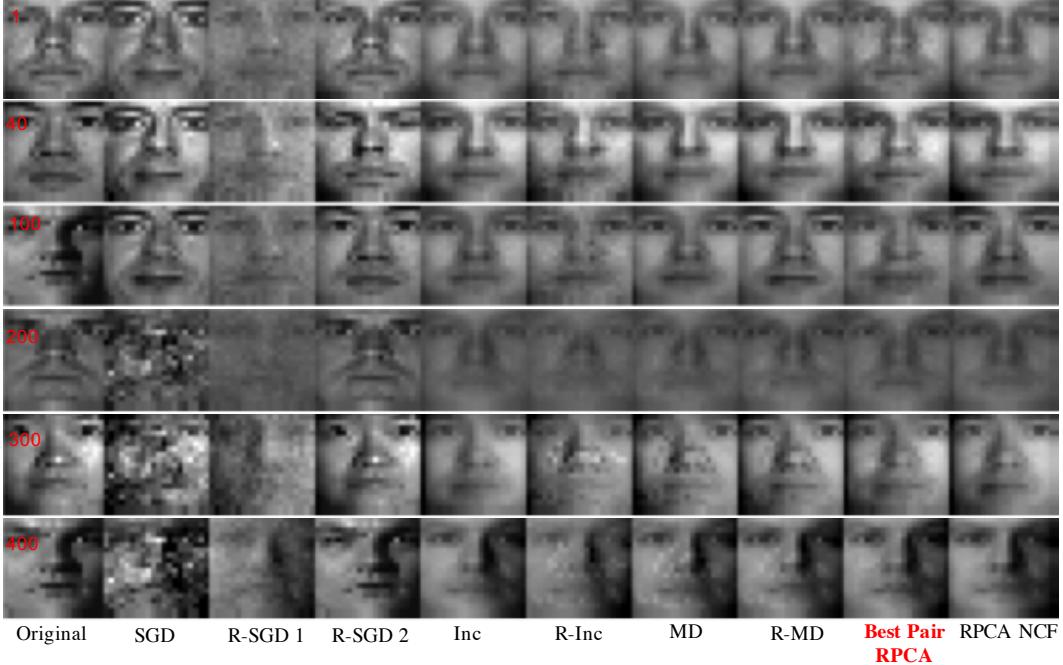


Figure 12: Inliers and outliers detection. We project the face images (inliers) to 9 dimensional subspaces found by different methods.

Metric	SGD	R-SGD1	R-SGD2	Inc	R-Inc	MD	R-MD	RPCA-F	Best pair	SVT
$\frac{\ P_L - P_{L^*}\ _F}{3\sqrt{2}}$	0.7	0.86	4.66	0.77	0.72	0.67	0.67	0.78	0.76	0.79

Table 1: Quantitative performance of different algorithms in inlier detection experiment. Except R-SGD2 all methods are competitive.

B Proof of the global convergence

For convenience, define $\Delta_k \stackrel{\text{def}}{=} \|Y_k - Y_{k-1}\|$.

Lemma B.1. *For the update of Y_{k+1} in (13), given any $k \in \mathbb{N}$, define*

$$G_{k+1} \stackrel{\text{def}}{=} \frac{1}{\gamma}(Z_{a,k} - Y_{k+1}) - \nabla \mathcal{F}(Z_{b,k}) + \nabla \mathcal{F}(Y_{k+1}).$$

Then, we have $G_{k+1} \in \partial \Phi(Y_{k+1})$, and

$$\|G_{k+1}\| \leq \left(\frac{1}{\gamma} + L\right)\Delta_{k+1} + \left(\frac{a_k}{\gamma} + b_k L\right)\Delta_k. \quad (18)$$

Proof. From the definition of proximity operator and the update of Y_{k+1} (13), we have $Z_{a,k} - \gamma \nabla \mathcal{F}(Z_{b,k}) - Y_{k+1} \in \gamma \partial \mathcal{R}(Y_{k+1})$. Adding $\gamma \nabla \mathcal{F}(Y_{k+1})$ to both sides, we obtain

$$G_{k+1} = \frac{Z_{a,k} - \gamma \nabla \mathcal{F}(Z_{b,k}) - Y_{k+1} + \gamma \nabla \mathcal{F}(Y_{k+1})}{\gamma} \in \partial \Phi(Y_{k+1}).$$

Applying further the triangle inequality together with the Lipschitz continuity of $\nabla \mathcal{F}$, we get

$$\begin{aligned} \|G_{k+1}\| &= \left\| \frac{1}{\gamma}(Z_{a,k} - Y_{k+1}) - \nabla \mathcal{F}(Z_{b,k}) + \nabla \mathcal{F}(Y_{k+1}) \right\| \\ &\leq \frac{1}{\gamma} \|Z_{a,k} - Y_{k+1}\| + L \|Z_{b,k} - Y_{k+1}\| \leq \frac{1}{\gamma} (\Delta_{k+1} + a_k \Delta_k) + L (\Delta_{k+1} + b_k \Delta_k), \end{aligned}$$

which concludes the proof. \square

Lemma B.2. *For Algorithm 1, given the parameters γ, a_k, b_k , the following inequality holds:*

$$\Phi(Y_{k+1}) + \underline{\beta} \Delta_{k+1}^2 \leq \Phi(Y_k) + \bar{\alpha} \Delta_k^2. \quad (19)$$

Proof. Define the function

$$\mathcal{L}_k(Y) \stackrel{\text{def}}{=} \gamma \mathcal{R}(Y) + \frac{1}{2} \|Y - Z_{a,k}\|^2 + \gamma \langle Y, \nabla \mathcal{F}(Z_{b,k}) \rangle.$$

It can be shown that the update of Y_{k+1} in (13) is equivalent to

$$Y_{k+1} \in \operatorname{argmin}_{Y \in \mathbb{R}^n} \mathcal{L}_k(Y), \quad (20)$$

which means that $\mathcal{L}_k(Y_{k+1}) \leq \mathcal{L}_k(Y_k)$, which means

$$\mathcal{R}(Y_{k+1}) + \frac{1}{2\gamma} \|Y_{k+1} - Z_{a,k}\|^2 + \langle Y_{k+1}, \nabla \mathcal{F}(Z_{b,k}) \rangle \leq \mathcal{R}(Y_k) + \frac{1}{2\gamma} \|Y_k - Z_{a,k}\|^2 + \langle Y_k, \nabla \mathcal{F}(Z_{b,k}) \rangle.$$

Therefore, we get

$$\begin{aligned} \mathcal{R}(Y_k) &\geq \mathcal{R}(Y_{k+1}) + \frac{1}{2\gamma} \|Y_{k+1} - Z_{a,k}\|^2 + \langle Y_{k+1} - Y_k, \nabla \mathcal{F}(Z_{b,k}) \rangle - \frac{1}{2\gamma} \|Y_k - Z_{a,k}\|^2 \\ &= \mathcal{R}(Y_{k+1}) + \frac{1}{2\gamma} \|Y_{k+1} - Y_k + Y_k - Z_{a,k}\|^2 + \langle Y_{k+1} - Y_k, \nabla \mathcal{F}(Z_{b,k}) \rangle - \frac{1}{2\gamma} \|Y_k - Z_{a,k}\|^2 \\ &= \mathcal{R}(Y_{k+1}) + \frac{1}{2\gamma} \Delta_{k+1}^2 + \langle Y_{k+1} - Y_k, \nabla \mathcal{F}(Z_{b,k}) \rangle - \frac{a_k}{\gamma} \langle Y_k - Y_{k+1}, Y_k - Y_{k-1} \rangle \\ &= \mathcal{R}(Y_{k+1}) + \langle Y_{k+1} - Y_k, \nabla \mathcal{F}(Y_k) \rangle + \frac{1}{2\gamma} \Delta_{k+1}^2 \\ &\quad - \frac{a_k}{\gamma} \langle Y_k - Y_{k+1}, Y_k - Y_{k-1} \rangle + \langle Y_{k+1} - Y_k, \nabla \mathcal{F}(Z_{b,k}) - \nabla \mathcal{F}(Y_k) \rangle. \end{aligned} \quad (21)$$

Since $\nabla \mathcal{F}$ is L -Lipschitz, then

$$\langle \nabla \mathcal{F}(Y_k), Y_{k+1} - Y_k \rangle \geq \mathcal{F}(Y_{k+1}) - \mathcal{F}(Y_k) - \frac{L}{2} \Delta_{k+1}^2.$$

For the inner product $\langle Y_k - Y_{k+1}, Y_k - Y_{k-1} \rangle$, applying the Pythagorean relation $2\langle c_1 - c_2, c_1 - c_3 \rangle = \|c_1 - c_2\|^2 + \|c_1 - c_3\|^2 - \|c_2 - c_3\|^2$, we get

$$\begin{aligned} \langle Y_k - Y_{k+1}, Y_k - Y_{k-1} \rangle &= \frac{1}{2} (\|Y_k - Y_{k+1}\|^2 + \|Y_k - Y_{k-1}\|^2 - \|Y_{k+1} - Y_{k-1}\|^2) \\ &\leq \frac{1}{2} (\|Y_k - Y_{k+1}\|^2 + \|Y_k - Y_{k-1}\|^2). \end{aligned} \quad (22)$$

Using further Young's inequality with $\nu > 0$ we obtain

$$\begin{aligned} \langle Y_{k+1} - Y_k, \nabla \mathcal{F}(Z_{b,k}) - \nabla \mathcal{F}(Y_k) \rangle &\geq -\left(\frac{\nu}{2} \Delta_{k+1}^2 + \frac{1}{2\nu} \|\nabla \mathcal{F}(Z_{b,k}) - \nabla \mathcal{F}(Y_k)\|^2\right) \\ &\geq -\left(\frac{\nu}{2} \Delta_{k+1}^2 + \frac{b_k^2 L^2}{2\nu} \Delta_k^2\right). \end{aligned} \quad (23)$$

Combining the above 3 inequalities with (21) yields

$$\begin{aligned} \mathcal{R}(Y_k) &\geq \mathcal{R}(Y_{k+1}) + \mathcal{F}(Y_{k+1}) - \mathcal{F}(Y_k) - \frac{L}{2} \Delta_{k+1}^2 + \frac{1}{2\gamma} \Delta_{k+1}^2 \\ &\quad - \frac{a_k}{2\gamma} \|Y_k - Y_{k+1}\|^2 - \frac{a_k}{2\gamma} \|Y_k - Y_{k-1}\|^2 - \frac{\nu}{2} \Delta_{k+1}^2 - \frac{b_k^2 L^2}{2\nu} \Delta_k^2, \end{aligned} \quad (24)$$

which leads to

$$\Phi(Y_{k+1}) + \frac{1 - \gamma L - a_k - \nu}{2\gamma} \Delta_{k+1}^2 \leq \Phi(Y_k) + \frac{\gamma b_k^2 L^2 + \nu a_k}{2\nu\gamma} \Delta_k^2.$$

Owing to the definition of $\underline{\beta}$ and $\bar{\alpha}$ we conclude the proof. \square

Define \mathcal{H} the product space $\mathcal{H} \stackrel{\text{def}}{=} \mathbb{R}^n \times \mathbb{R}^n$ and $Z_k = (Y_k, Y_{k-1}) \in \mathcal{H}$. Then given Z_k , define the function

$$\Psi(Z_k) \stackrel{\text{def}}{=} \Phi(Y_k) + \bar{\alpha} \Delta_k^2,$$

which is a KL function if Φ is. Denote $\mathcal{C}_{Y_k}, \mathcal{C}_{Z_k}$ the set of cluster points of sequences $\{Y_k\}_{k \in \mathbb{N}}$ and $\{Z_k\}_{k \in \mathbb{N}}$ respectively, and $\operatorname{crit}(\Psi) = \{Z = (Y, Y) \in \mathcal{H} : Y \in \operatorname{crit}(\Phi)\}$.

Lemma B.3. *For Algorithm 1, choose ν, γ, a_k, b_k such that (17) holds. If Φ is bounded from below, then*

- (i) $\sum_{k \in \mathbb{N}} \Delta_k^2 < +\infty$;

- (ii) The sequence $\Psi(Z_k)$ is monotonically decreasing and convergent;
- (iii) The sequence $\Phi(Y_k)$ is convergent.

Proof. Define $\delta = \underline{\beta} - \bar{\alpha} > 0$, from Lemma B.2, we have

$$\delta\Delta_{k+1}^2 \leq (\Phi(Y_k) - \Phi(Y_{k+1})) + \bar{\alpha}(\Delta_k^2 - \Delta_{k+1}^2).$$

Let $Y_{-1} = Y_0$ and the above inequality over k :

$$\begin{aligned} \delta \sum_{k \in \mathbb{N}} \Delta_{k+1}^2 &\leq \sum_{k \in \mathbb{N}} (\Phi(Y_k) - \Phi(Y_{k+1})) + \sum_{k \in \mathbb{N}} \bar{\alpha}(\Delta_k^2 - \Delta_{k+1}^2) \\ &\leq \Phi(Y_0) + \bar{\alpha} \sum_{k \in \mathbb{N}} (\Delta_k^2 - \Delta_{k+1}^2) = \Phi(Y_0) + \bar{\alpha}\Delta_0^2 = \Phi(Y_0), \end{aligned}$$

which means, as $\Phi(Y_0)$ is bounded,

$$\sum_{k \in \mathbb{N}} \Delta_{k+1}^2 \leq \frac{\Phi(Y_0)}{\delta} < +\infty.$$

From Lemma B.2, by pairing terms on both sides of (19), we get

$$\Psi(Z_{k+1}) + (\underline{\beta} - \bar{\alpha})\Delta_{k+1}^2 \leq \Psi(Z_k).$$

Since we assume $\underline{\beta} - \bar{\alpha} > 0$, hence $\Psi(Z_k)$ is monotonically non-increasing. The convergence of $\Phi(Y_k)$ is straightforward. \square

Lemma B.4. *For Algorithm 1, choose ν, γ, a_k, b_k such that (17) holds. If Φ is bounded from below and $\{Y_k\}_{k \in \mathbb{N}}$ is bounded, then Y_k converges to a critical point of Φ .*

Proof. Since $\{Y_k\}_{k \in \mathbb{N}}$ is bounded, there exists a subsequence $\{Y_{k_j}\}_{j \in \mathbb{N}}$ and cluster point \bar{Y} such that $Y_{k_j} \rightarrow \bar{Y}$ as $j \rightarrow \infty$. Next we show that $\Phi(Y_{k_j}) \rightarrow \Phi(\bar{Y})$ and that \bar{Y} is a critical point of Φ .

Since \mathcal{R} is lsc, then $\liminf_{j \rightarrow \infty} \mathcal{R}(Y_{k_j}) \geq \mathcal{R}(\bar{Y})$. From (20), we have $\mathcal{L}_{k_j-1}(Y_{k_j}) \leq \mathcal{L}_{k_j-1}(\bar{Y})$ and thus

$$\begin{aligned} \mathcal{R}(\bar{Y}) &\geq \mathcal{R}(Y_{k_j}) + \frac{1}{2\gamma} \|Y_{k_j} - U_{k_j-1}\|^2 + \langle Y_{k_j} - \bar{Y}, \nabla \mathcal{F}(V_{k_j-1}) \rangle - \frac{1}{2\gamma} \|\bar{Y} - U_{k_j-1}\|^2 \\ &= \mathcal{R}(Y_{k_j}) + \frac{1}{2\gamma} (\|Y_{k_j} - \bar{Y}\|^2 + 2\langle Y_{k_j} - \bar{Y}, \bar{Y} - U_{k_j-1} \rangle) + \langle Y_{k_j} - \bar{Y}, \nabla \mathcal{F}(V_{k_j-1}) \rangle. \end{aligned}$$

Taking the limit of the above inequality and using $\Delta_{k_j}^2 \rightarrow 0$, $Y_{k_j} \rightarrow \bar{Y}$, we get $\limsup_{j \rightarrow \infty} \mathcal{R}(Y_{k_j}) \leq \mathcal{R}(\bar{Y})$. As a result, $\lim_{k \rightarrow \infty} \mathcal{R}(Y_{k_j}) = \mathcal{R}(\bar{Y})$. Since \mathcal{F} is continuous, then $\mathcal{F}(Y_{k_j}) \rightarrow \mathcal{F}(\bar{Y})$, hence $\Phi(Y_{k_j}) \rightarrow \Phi(\bar{Y})$.

Furthermore, owing to Lemma B.1, $G_{k_j} \in \partial\Phi(Y_{k_j})$, and (i) of Lemma B.3 we have $G_{k_j} \rightarrow 0$ as $k \rightarrow \infty$. Therefore, as $j \rightarrow \infty$, we have

$$G_{k_j} \in \partial\Phi(Y_{k_j}), \quad (Y_{k_j}, G_{k_j}) \rightarrow (\bar{Y}, 0) \quad \text{and} \quad \Phi(Y_{k_j}) \rightarrow \Phi(\bar{Y}).$$

Hence $0 \in \partial\Phi(\bar{Y})$, i.e. \bar{Y} is a critical point. \square

Proof of Theorem 2.3. Putting together the above lemmas, we draw the following useful conclusions:

(C.1) Denote $\delta = \underline{\beta} - \bar{\alpha}$, then $\Psi(Z_{k+1}) + \delta\Delta_{k+1}^2 \leq \Psi(Z_k)$;

(C.2) Define

$$W_k \stackrel{\text{def}}{=} \begin{pmatrix} G_k + 2\bar{\alpha}(Y_k - Y_{k-1}) \\ 2\bar{\alpha}(Y_{k-1} - Y_k) \end{pmatrix},$$

then we have $W_k \in \partial\Psi(Z_k)$. Owing to Lemma B.1, there exists a $\sigma > 0$ such that $\|W_k\| \leq \sigma(\Delta_k + \Delta_{k-1})$;

(C.3) if Y_{k_j} is a subsequence such that $Y_{k_j} \rightarrow \bar{Y}$, then $\Psi(Z_k) \rightarrow \Psi(\bar{Z})$ where $\bar{Z} = (\bar{Y}, \bar{Y})$.

(C.4) $\mathcal{C}_{Z_k} \subseteq \text{crit}(\Psi)$;

(C.5) $\lim_{k \rightarrow \infty} \text{dist}(Z_k, \mathcal{C}_{Z_k}) = 0$;

(C.6) \mathcal{C}_{Z_k} is non-empty, compact and connected;

(C.7) Ψ is finite and constant on \mathcal{C}_{Z_k} .

Next we prove the claims of Theorem 2.3.

- (i) Consider a critical point of Φ , $\bar{Y} \in \text{crit}(\Phi)$, such that $\bar{Z} = (\bar{Y}, \bar{Y}) \in \mathcal{C}_{Z_k}$. Then owing to (C.3), we have $\Psi(Z_k) \rightarrow \Psi(\bar{Z})$.

Suppose there exists K such that $\Psi(Z_K) = \Psi(\bar{Z})$. Then, the descent property (C.1) implies that $\Psi(Z_k) = \Psi(\bar{Z})$ holds for all $k \geq K$. Thus, Z_k is constant for $k \geq K$, hence has finite length.

On the other hand, suppose that $\psi_k \stackrel{\text{def}}{=} \Psi(Z_k) - \Psi(\bar{Z}) > 0$. Owing to (C.6), (C.7) and Definition 2.2, the KL property of Ψ implies that there exist ε, η and a concave function φ , and

$$\mathcal{U} \stackrel{\text{def}}{=} \{S \in \mathcal{H} : \text{dist}(S, \mathcal{C}_{Z_k}) < \varepsilon\} \cap [\Psi(\bar{Z}) < \Psi(S) < \Psi(\bar{Z}) + \eta], \quad (25)$$

such that for all $Z \in \mathcal{U}$:

$$\varphi'(\Psi(z) - \Psi(\bar{Z})) \text{dist}(0, \partial\Psi(z)) \geq 1. \quad (26)$$

Let $k_1 \in \mathbb{N}$ be such that $\Psi(Z_k) < \Psi(\bar{Z}) + \eta$ holds for all $k \geq k_1$. Owing to (C.5), there exists another $k_2 \in \mathbb{N}$ such that $\text{dist}(Z_k, \mathcal{C}_{Z_k}) < \varepsilon$ holds for all $k \geq k_2$. Let $K = \max\{k_1, k_2\}$. Then $Z_k \in \mathcal{U}$ holds for all $k \geq K$. Furthermore using (26), we have for $k \geq K$

$$\varphi'(\psi_k) \text{dist}(0, \partial\Psi(Z_k)) \geq 1.$$

Note that since φ is concave, φ' is decreasing. As $\Psi(Z_k)$ is decreasing too, we have

$$\varphi(\psi_k) - \varphi(\psi_{k+1}) \geq \varphi'(\psi_k)(\Psi(Z_k) - \Psi(Z_{k+1})) \geq \frac{\Psi(Z_k) - \Psi(Z_{k+1})}{\text{dist}(0, \partial\Psi(Z_k))}.$$

From (C.1), since $\text{dist}(0, \partial\Psi(Z_k)) \leq \|w_k\|$, we get

$$\varphi(\psi_k) - \varphi(\psi_{k+1}) \geq \frac{\Psi(Z_k) - \Psi(Z_{k+1})}{\|w_k\|} \geq \frac{\Psi(Z_k) - \Psi(Z_{k+1})}{\sigma(\Delta_k + \Delta_{k-1})}.$$

Moreover, (C.2) yields $\Psi(Z_k) - \Psi(Z_{k+1}) \geq \delta\Delta_{k+1}^2$ and thus

$$\varphi(\psi_k) - \varphi(\psi_{k+1}) \geq \frac{\delta\Delta_{k+1}^2}{\sigma(\Delta_k + \Delta_{k-1})},$$

which yields

$$\Delta_{k+1}^2 \leq \left(\frac{\sigma}{\delta}(\varphi(\psi_k) - \varphi(\psi_{k+1}))\right)(\Delta_k + \Delta_{k-1}). \quad (27)$$

Taking the square root of both sides and applying Young's inequality we further obtain

$$\Delta_{k+1} \leq \frac{1}{2}(\Delta_k + \Delta_{k-1}) + \frac{2\sigma}{\delta}(\varphi(\psi_k) - \varphi(\psi_{k+1})). \quad (28)$$

Summing up both sides over k , and using $x_0 = x_{-1}$, we get

$$\ell \stackrel{\text{def}}{=} \sum_{k \in \mathbb{N}} \Delta_k \leq \Delta_1 + \frac{2\sigma}{\delta} \varphi(\psi_1) < +\infty,$$

which concludes the finite length property of Y_k .

- (ii) Then the convergence of the sequence follows from the fact that $\{Y_k\}_{k \in \mathbb{N}}$ is a Cauchy sequence, hence convergent. Owing to Lemma B.4, there exists a critical point $Y^* \in \text{crit}(\Phi)$ such that $\lim_{k \rightarrow \infty} Y_k = Y^*$.
- (iii) We now turn to prove local convergence to a global minimizer. Note that if Y^* is a global minimizer of Φ , then Z^* is a global minimizer of Ψ . Let $r > \rho > 0$ such that $B_r(Z^*) \subset \mathcal{U}$ and $\eta < \delta(r - \rho)^2$. Suppose that the initial point Y_0 is chosen such that following conditions hold,

$$\Psi(Z^*) \leq \Psi(Z_0) < \Psi(Z^*) + \eta \quad (29)$$

$$\|Y_0 - Z^*\| + \ell(s-1) + 2\sqrt{\frac{\Psi(Z_0) - \Psi(Z^*)}{\delta}} + \frac{\sigma}{\delta} \varphi(\psi_0) < \rho. \quad (30)$$

The descent property (C.1) of Ψ together with (29) imply that for any $k \in \mathbb{N}$, $\Psi(Z^*) \leq \Psi(Z_{k+1}) \leq \Psi(Z_k) \leq \Psi(Z_0) < \Psi(Z^*) + \eta$, and

$$\|Y_{k+1} - Y_k\| \leq \sqrt{\frac{\Psi(Z_k) - \Psi(Z_{k+1})}{\delta}} \leq \sqrt{\frac{\Psi(Z_k) - \Psi(Z^*)}{\delta}}. \quad (31)$$

Therefore, given any $k \in \mathbb{N}$, if we have $Y_k \in \mathbb{B}_\rho(Y^*)$, then

$$\begin{aligned} \|Y_{k+1} - Y^*\| &\leq \|Y_k - Y^*\| + \|Y_{k+1} - Y_k\| \leq \|Y_k - Y^*\| + \sqrt{\frac{\Psi(Z_k) - \Psi(Z^*)}{\delta}} \\ &\leq \rho + (r - \rho) = r, \end{aligned} \quad (32)$$

which means that $Y_{k+1} \in \mathbb{B}_r(Y^*)$.

For any $k \in \mathbb{N}$, define the following partial sum $p_k \stackrel{\text{def}}{=} \sum_{j=k-2}^{k-1} \sum_{i=1}^j \Delta_i$. Note that $p_k = 0$ for $k = 1$, and $\lim_{k \rightarrow +\infty} p_k = \ell$. Next we prove the following claims through induction: for $k \in \mathbb{N}$

$$Y_k \in \mathbb{B}_\rho(Y^*) \quad (33)$$

$$\sum_{j=1}^k \Delta_{j+1} + \Delta_{k+1} \leq \Delta_1 + p_k + \frac{\sigma}{\delta} (\varphi(\psi_1) - \varphi(\psi_{k+1})). \quad (34)$$

From (31) we have

$$\|Y_1 - Y_0\| \leq \sqrt{\frac{\Psi(Z_0) - \Psi(Z^*)}{\delta}}. \quad (35)$$

Applying the triangle inequality we then obtain

$$\|Y_1 - Y^*\| \leq \|Y_0 - Y^*\| + \|Y_1 - Y_0\| \leq \|Y_0 - Y^*\| + \sqrt{\frac{\Psi(Z_0) - \Psi(Z^*)}{\delta}} < \rho,$$

which means $Y_1 \in \mathbb{B}_\rho(Y^*)$. Now, taking $\kappa = 1$ in (28) yields, for any $k \in \mathbb{N}$,

$$2\Delta_{k+1} \leq (\Delta_k + \Delta_{k-1}) + \frac{\sigma}{\delta} (\varphi(\psi_k) - \varphi(\psi_{k+1})). \quad (36)$$

Let $k = 1$. Since $Y_0 = Y_{-1}$, we have

$$2\Delta_2 \leq \Delta_1 + \frac{\sigma}{\delta} (\varphi(\psi_1) - \varphi(\psi_2)).$$

Therefore, (33) and (34) hold for $k = 1$.

Now assume that they hold for some $k > 1$. Using the triangle inequality and (34),

$$\begin{aligned} \|Y_{k+1} - Y^*\| &\leq \|Y_0 - Y^*\| + \Delta_1 + \sum_{j=1}^k \Delta_{j+1} \\ &\leq \|Y_0 - Y^*\| + 2\Delta_1 + p_k + \frac{\sigma}{\delta} (\varphi(\psi_1) - \varphi(\psi_{k+1})) \\ &\leq \|Y_0 - Y^*\| + 2\Delta_1 + \ell + \frac{\sigma}{\delta} (\varphi(\psi_1) - \varphi(\psi_{k+1})) \\ (35) \leq \|Y_0 - Y^*\| + 2\sqrt{\frac{\Psi(Z_0) - \Psi(Z^*)}{\delta}} + \ell + \frac{\sigma}{\delta} (\varphi(\psi_1) - \varphi(\psi_{k+1})). \end{aligned}$$

As $\varphi(\psi) \geq 0$ and $\varphi'(\psi) > 0$ for $\psi \in]0, \eta[$, and in view of (30), we arrive at

$$\|Y_{k+1} - Y^*\| \leq \|Y_0 - Y^*\| + 2\sqrt{\frac{\Psi(Z_0) - \Psi(Z^*)}{\delta}} + \ell + \frac{\sigma}{\delta} \varphi(\psi_0) < \rho$$

whence we deduce that (33) holds at $k + 1$. Now, taking (36) at $k + 1$ gives

$$\begin{aligned} 2\Delta_{k+2} &\leq (\Delta_{k+1} + \Delta_k) + \frac{\sigma}{\delta} (\varphi(\psi_{k+1}) - \varphi(\psi_{k+2})) \\ &\leq \Delta_{k+1} + (\Delta_k + \Delta_{k-1}) + \frac{\sigma}{\delta} (\varphi(\psi_{k+1}) - \varphi(\psi_{k+2})). \end{aligned} \quad (37)$$

Adding both sides of (37) and (34) we get

$$\begin{aligned} \sum_{j=1}^{k+1} \Delta_{j+1} + \Delta_{k+2} &\leq \Delta_1 + p_k + (\Delta_k + \Delta_{k-1}) + \frac{\sigma}{\delta} (\varphi(\psi_1) - \varphi(\psi_{k+2})) \\ &= \Delta_1 + p_{k+1} + \frac{\sigma}{\delta} (\varphi(\psi_1) - \varphi(\psi_{k+2})), \end{aligned}$$

meaning that (34) holds at $k + 1$. This concludes the induction proof.

In summary, the above result shows that if we start close enough from Y^* (so that (29)-(30) hold), then the sequence $\{Y_k\}_{k \in \mathbb{N}}$ will remain in the neighbourhood $\mathbb{B}_\rho(Y^*)$ and thus converges to a critical point \bar{Y} owing to Lemma B.4. Moreover, $\Psi(Z_k) \rightarrow \Psi(\bar{Z}) \geq \Psi(Z^*)$ by virtue of (C.3). Now we need to show that $\Psi(\bar{Z}) = \Psi(Z^*)$. Suppose that $\Psi(\bar{Z}) > \Psi(Z^*)$. As Ψ has the KL property at Z^* , we have

$$\varphi'(\Psi(\bar{Z}) - \Psi(Z^*)) \text{dist}(0, \partial\Psi(\bar{Z})) \geq 1.$$

But this is impossible since $\varphi'(s) > 0$ for $s \in]0, \eta[$, and $\text{dist}(0, \partial\Psi(\bar{Z})) = 0$ as \bar{Z} is a critical point. Hence we have $\Psi(\bar{Z}) = \Psi(Z^*)$, which means $\Phi(\bar{Z}) = \Phi(Y^*)$, i.e. the cluster point \bar{Y} is actually a global minimizer. This concludes the proof. \square

C Proof of local linear convergence

Before presenting the proof for local linear convergence, in Figure 13 below we provide the comparison of theoretical estimation and practical observation. The size of the problem is $\mathbb{R}^{32 \times 32}$, which is small as larger size will make the rate estimation very slow. It can be observed that our theoretical rate estimation is very tight given that the red line and the black one are parallel to each other.

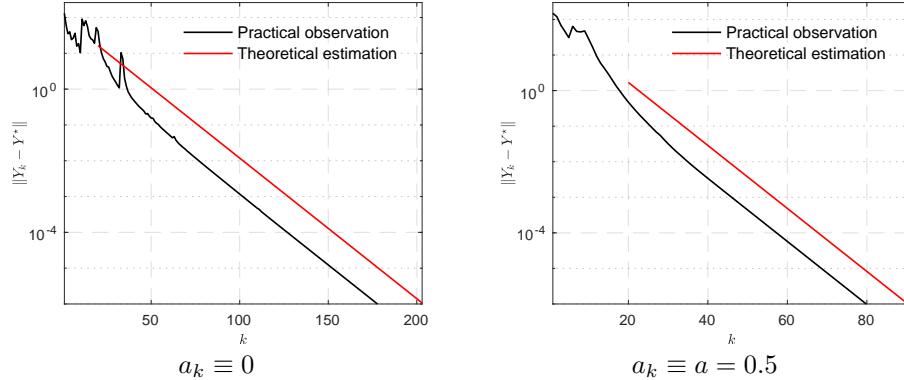


Figure 13: Local linear convergence of Algorithm 1.

Since we are in the non-convex setting, we need the prox-regularity of the non-convexity. A lower semi-continuous function \mathcal{R} is r -prox-regular at $\bar{x} \in \text{dom}(\mathcal{R})$ for $\bar{v} \in \partial\mathcal{R}(\bar{x})$ if $\exists r > 0$ such that $\mathcal{R}(x') > \mathcal{R}(x) + \langle v, x' - x \rangle - \frac{1}{2r} \|x - x'\|^2 \forall x, x' \text{ near } \bar{x}, \mathcal{R}(x) \text{ near } \mathcal{R}(\bar{x}) \text{ and } v \in \partial\mathcal{R}(x) \text{ near } \bar{v}$.

To prove Theorem 2.4, we rely on a so-called partial smoothness concept. Let $\mathcal{M} \subset \mathbb{R}^n$ be a C^2 -smooth submanifold, let $\mathcal{T}_{\mathcal{M}}(x)$ the tangent space of \mathcal{M} at any point $x \in \mathcal{M}$.

Definition C.1. The function $\mathcal{R} : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ is C^2 -partly smooth at $\bar{x} \in \mathcal{M}$ relative to \mathcal{M} for $\bar{v} \in \partial\mathcal{R}(\bar{x}) \neq \emptyset$ if \mathcal{M} is a C^2 -submanifold around \bar{x} , and

- (i) (Smoothness): \mathcal{R} restricted to \mathcal{M} is C^2 around \bar{x} ;
- (ii) (Regularity): \mathcal{R} is regular at all $x \in \mathcal{M}$ near \bar{x} and \mathcal{R} is r -prox-regular at \bar{x} for \bar{v} ;
- (iii) (Sharpness): $\mathcal{T}_{\mathcal{M}}(\bar{x}) = \text{par}(\partial\mathcal{R}(x))^\perp$;
- (iv) (Continuity): The set-valued mapping $\partial\mathcal{R}$ is continuous at \bar{x} relative to \mathcal{M} .

We denote the class of partly smooth functions at x relative to \mathcal{M} for v as $\text{PSF}_{x,v}(\mathcal{M})$. Partial smoothness was first introduced in (Lewis, 2003) and its directional version stated here is due to (Lewis and Zhang, 2013; Drusvyatskiy and Lewis, 2013). Prox-regularity is sufficient to ensure that the partly smooth submanifolds are locally unique (Lewis and Zhang, 2013, Corollary 4.12), (Drusvyatskiy and Lewis, 2013, Lemma 2.3 and Proposition 10.12).

Proof of Theorem 2.4. First we have

- \mathcal{Y}_L is a the set of fixed-rank matrices, hence it is partly smooth.
- Since \mathcal{S} is a subspace, hence it is partly smooth at S^* relative to any $W \in (\mathcal{S})^\perp$.

Under the conditions of Theorem 2.3, there exists a critical point Y^* such that $Y_k \rightarrow Y^*$ and $\Phi(Y_k) \rightarrow \Phi(Y^*)$.

Convergence properties of $\{Y_k\}_{k \in \mathbb{N}}$ (Theorem 2.3) entails $\|Z_{a,k} - Y_k\| \rightarrow 0$ and $\|Z_{b,k} - Y^*\| \rightarrow 0$. In turn,

$$\text{dist}(-\nabla\mathcal{F}(x^*), \partial\mathcal{R}(Y_{k+1})) \leq \frac{1}{\gamma} \|Z_{a,k} - Y_{k+1}\| + \|Z_{b,k} - Y^*\| \rightarrow 0.$$

Altogether, this shows that the conditions of (Lewis and Zhang, 2013, Theorem 4.10) or (Drusvyatskiy and Lewis, 2013, Proposition 10.12) are fulfilled on \mathcal{R} at Y^* for $-\nabla\mathcal{F}(Y^*)$, and the identification result follows, that is

$$(\mathcal{Y}_{r,k}, \mathcal{Y}_{\alpha,k}) \in \mathcal{Y}_L \times \mathcal{S}$$

for all k large enough, and we conclude the proof. \square

Tangent space $T_{\mathcal{X}}^{X^*}$ Given $X^* \in \mathcal{X}$, the tangent space simply reads $KX = 0$. Let E be the kernel of K , then we have the projection operator onto $KX = 0$ reads

$$P_{T_{\mathcal{X}}^{X^*}} = E(E^T E)^{-1} E^T.$$

Tangent space of \mathcal{Y}_L Let $\mathbb{M} = M_{m,n}(\mathbb{R})$ be the space of $m \times n$ matrices with the classical inner product $\langle A, B \rangle = \text{Trace}(A^T B)$. The set of matrices with fixed rank r ,

$$\mathcal{Y}_L = \{X \in \mathbb{M} : \text{rank}(X) = r\},$$

is a smooth manifold around any matrix $L \in \mathcal{Y}_L$. Given L^* , with the help of the singular value decomposition $L = U\Sigma V^T$, the tangent space at L to \mathcal{Y}_L is

$$T_{\mathcal{Y}_L}^{L^*} = \{H \in \mathbb{M} : u_i^T H v_j = 0, \text{ for all } r < i \leq m, r < j \leq n\}.$$

Let $U = [u_1, u_2, \dots, u_m]$, $V = [v_1, v_2, \dots, v_n]$ and Σ be diagonal matrix with singular value written in decreasing order.

Denote

$$\mathcal{L} = \left\{ L \in \mathbb{M} : X = u_i^T v_j, \text{ for all } \{i, j\}_{1 \leq i \leq m, 1 \leq j \leq n} \setminus \{i, j\}_{r < i \leq m, r < j \leq n} \right\},$$

then \mathcal{L} forms the basis of \mathcal{T} and $\dim(\mathcal{L}) = mn - r^2$, there for define

$$Z = [L_1(:); L_2(:); \dots; L_{mn-r^2}(:)], \quad L_i \in \mathcal{L},$$

and

$$P_{T_{\mathcal{Y}_L}^{L^*}} = Z(Z^T Z)^{-1} Z^T,$$

then $P_{T_{\mathcal{Y}_L}^{L^*}}$ is the explicit form of the projection operator of projecting onto subspace $T_{\mathcal{Y}_L}^{L^*}$.

Tangent space of \mathcal{S} Given $S^* \in \mathcal{S}$, denote the tangent space as $T_{\mathcal{S}}^{S^*}$. Let $\text{vec}(S^*)$ be the vector form of S^* , then we haves

$$P_{T_{\mathcal{S}}^{S^*}} = \text{diag}(|\text{vec}(S^*)| > 0).$$

Finally, we have

$$P_{T_{\mathcal{Y}}^{Y^*}} = \begin{bmatrix} P_{T_{\mathcal{S}}^{S^*}} & \\ & P_{T_{\mathcal{Y}_L}^{L^*}} \end{bmatrix}.$$

Proof of Theorem 2.6. From (13), when $a_k, b_k \equiv 0$, we have thats

$$Y_{k+1} = P_{\mathcal{Y}}(Y_k - \gamma(Y_k - P_{\mathcal{X}}(Y_k))).$$

Let Y^* be a critical point that Y_k converges to, then

$$Y^* = P_{\mathcal{Y}}(Y^* - \gamma(Y^* - P_{\mathcal{X}}(Y^*)).$$

Denote $X_k = P_{\mathcal{X}}(Y_k)$ and $X^* = P_{\mathcal{X}}(Y^*)$, we have

$$\begin{aligned} X_k - X^* &= P_{T_{\mathcal{X}}^{X^*}}(X_k - X^*) = P_{T_{\mathcal{X}}^{X^*}}P_{\mathcal{X}}(Y_k - Y^*) = P_{T_{\mathcal{X}}^{X^*}}(Y_k - Y^*) \\ &= P_{T_{\mathcal{X}}^{X^*}}P_{T_{\mathcal{Y}}^{Y^*}}(Y_k - Y^*) + o(\|Y_k - Y^*\|). \end{aligned}$$

Consider the difference of the above two equations, owing to Lemma 2.5, we get

$$\begin{aligned} Y_{k+1} - Y^* &= P_{\mathcal{Y}}(Y_k - \gamma(Y_k - P_{\mathcal{X}}(Y_k))) - P_{\mathcal{Y}}(Y^* - \gamma(Y^* - P_{\mathcal{X}}(Y^*))) + o(\|Y_k - Y^*\|) \\ &= P_{\mathcal{Y}}((1 - \gamma)Y_k + \gamma P_{\mathcal{X}}(Y_k)) - P_{\mathcal{Y}}((1 - \gamma)Y^* + \gamma P_{\mathcal{X}}(Y^*)) + o(\|Y_k - Y^*\|) \\ &= P_{T_{\mathcal{Y}}^{Y^*}}((1 - \gamma)Y_k + \gamma P_{\mathcal{X}}(Y_k) - (1 - \gamma)Y^* - \gamma P_{\mathcal{X}}(Y^*)) + o(\|Y_k - Y^*\|) \\ &= P_{T_{\mathcal{Y}}^{Y^*}}((1 - \gamma)(Y_k - Y^*) + \gamma(X_k - X^*)) + o(\|Y_k - Y^*\|) \\ &= P_{T_{\mathcal{Y}}^{Y^*}}((1 - \gamma)\text{Id} + \gamma P_{T_{\mathcal{X}}^{X^*}})P_{T_{\mathcal{Y}}^{Y^*}}(Y_k - Y^*) + o(\|Y_k - Y^*\|), \end{aligned}$$

which means

$$Y_{k+1} - Y^* = \mathcal{P}(Y_k - Y^*) + o(\|Y_k - Y^*\|).$$

Note that \mathcal{P} is symmetric positive semi-definite, hence all its eigenvalues are real and lie in $[0, 1]$.

Now, assume that $b_k = a_k \equiv a$, then we have from (13)

$$\begin{aligned} Z_k &= (1+a)Y_k - aY_{k-1}, \\ Y_{k+1} &= \mathbf{P}_{\mathcal{Y}}(Z_k - \gamma(Z_k - \mathbf{P}_{\mathcal{X}}(Z_k))). \end{aligned}$$

Follow the derivation of $Y_{k+1} - Y^*$ above, we get

$$\begin{aligned} Y_{k+1} - Y^* &= (1+a)\mathcal{P}(Y_k - Y^*) - a\mathcal{P}(Y_{k-1} - Y^*) + o(\|Y_k - Y^*\|) \\ &= [(1+a)\mathcal{P} - a\mathcal{P}] \begin{pmatrix} Y_k - Y^* \\ Y_{k-1} - Y^* \end{pmatrix} + o(\|Y_k - Y^*\|). \end{aligned}$$

Plus the definition of D_k and the fact that $o(\|Y_k - Y^*\|) = o(\|D_k\|)$, we obtain

$$D_{k+1} = QD_k + o(\|D_k\|).$$

Owing to (Liang, 2016, Chapter 6), if $\rho_{\mathcal{P}} < 1$, then so is $\rho_Q < 1$, and the linear convergence result follows. \square

D Table of baseline methods

Algorithm	Abbreviation	Appearing in Experiment	Reference
Inexact Augmented Lagrange Method of Multipliers	iEALM	Fig. 1, 3, 4	(Lin et al., 2010)
Accelerated Proximal Gradient Singular Value Thresholding	APG	Fig. 3, 4	(Wright et al., 2009)
SVT		Table 1	(Cai et al., 2010)
Grassmannian Robust Adaptive Subspace Tracking Algorithm	GRASTA	Fig. 5	(He et al., 2012)
Go Decomposition	GoDec	Fig. 4, 10	(Zhou and Tao, 2011)
Robust PCA Gradient Descent	RPCA GD	Fig. 2, 3, 4, 5, 9, 10, 11	(Yi et al., 2016)
Robust PCA Nonconvex Feasibility	RPCA NCF	Fig. 1, 3, 4, 5, 10, 11, 12	(Dutta et al., 2018a)
Robust stochastic PCA Algorithms	SGD, R-SGD1, R-SGD2 Inc, R-Inc, MD, R-MD	Fig. 12, Table 1	(Goes et al., 2014)

Table 2: Algorithms compared in this paper.