ALTERNATING DIRECTION METHOD OF MULTIPLIERS FOR A CLASS OF NONCONVEX AND NONSMOOTH PROBLEMS WITH APPLICATIONS TO BACKGROUND/FOREGROUND EXTRACTION

LEI YANG*, TING KEI PONG*, AND XIAOJUN CHEN*

Abstract. In this paper, we study a general optimization model, which covers a large class of existing models for many applications in imaging sciences. To solve the resulting possibly nonconvex, nonsmooth and non-Lipschitz optimization problem, we adapt the alternating direction method of multipliers (ADMM) with a general dual step-size to solve a reformulation that contains three blocks of variables, and analyze its convergence. We show that for any dual step-size less than the golden ratio, there exists a computable threshold such that if the penalty parameter is chosen above such a threshold and the sequence thus generated by our ADMM is bounded, then the cluster point of the sequence gives a stationary point of the nonconvex optimization problem. We achieve this via a potential function specifically constructed for our ADMM. Moreover, we establish the global convergence of the whole sequence if, in addition, this special potential function is a Kurdyka-Łojasiewicz function. Furthermore, we present a simple strategy for initializing the algorithm to guarantee boundedness of the sequence. Finally, we perform numerical experiments comparing our ADMM with the proximal alternating linearized minimization (PALM) proposed in [5] on the background/foreground extraction problem with real data. The numerical results show that our ADMM with a nontrivial dual step-size is efficient.

Key words. Nonsmooth and nonconvex optimization; alternating direction method of multipliers; dual step-size; background/foreground extraction

1. Introduction. In this paper, we consider the following optimization problem:

$$\min_{L,S} \ \Psi(L) + \Phi(S) + \frac{1}{2} \left\| D - \mathcal{A}\left[\mathcal{B}(L) + \mathcal{C}(S)\right] \right\|_F^2, \tag{1.1}$$

where

- $\Psi, \Phi : \mathbb{R}^{m \times n} \to \mathbb{R}_+ \cup \{\infty\}$ are proper closed nonnegative functions, and Ψ is *convex*, while Φ is possibly *nonconvex*, *nonsmooth* and *non-Lipschitz*;
- $\mathcal{A}, \mathcal{B}, \mathcal{C} : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ are linear maps and $\mathcal{B}, \mathcal{C}$ are injective.

In particular, $\Psi(L)$ and $\Phi(S)$ in (1.1) can be regularizers used for inducing the desired structures. For instance, $\Psi(L)$ can be used for inducing low rank in L. One possible choice is $\Psi(L) = \|L\|_*$ (see next section for notation and definitions). Alternatively, one may consider $\Psi(L) = \delta_\Omega(L)$, where Ω is a compact convex set such as $\Omega = \{L \in \mathbb{R}^{m \times n} \mid \|L\|_\infty \leq l, \ L_{:1} = L_{:2} = \cdots = L_{:n}\}$ with $l > 0$, or $\Omega = \{L \in \mathbb{R}^{m \times n} \mid \|L\|_* \leq r\}$ with $r > 0$; the former choice restricts L to have rank at most 1 and makes (1.1) nuclear-norm-free (see [30,33]). On the other hand, $\Phi(S)$ can be used for inducing sparsity. In the literature, $\Phi(S)$ is typically separable, i.e., taking the form

$$\Phi(S) = \mu \sum_{i=1}^{m} \sum_{j=1}^{n} \phi(s_{ij}), \tag{1.2}$$

where ϕ is a nonnegative continuous function with $\phi(0) = 0$ and $\mu > 0$ is a regularization parameter. Some concrete examples of ϕ include:

1. bridge penalty [27, 28]: $\phi(t) = |t|^p$ for $0 < p \leq 1$;
2. fraction penalty [20]: $\phi(t) = \alpha|t|/(1 + \alpha|t|)$ for $\alpha > 0$;
3. logistic penalty [39]: $\phi(t) = \log(1 + \alpha|t|)$ for $\alpha > 0$;
4. smoothly clipped absolute deviation [16]: $\phi(t) = \int_0^{|t|} \min(1, (\alpha - s/\mu)_+/(\alpha - 1)) \, \mathrm{d}s$ for $\alpha > 2$;
5. minimax concave penalty [49]: $\phi(t) = \int_0^{|t|} (1 - s/(\alpha\mu))_+ \, \mathrm{d}s$ for $\alpha > 0$;
6. hard thresholding penalty function [17]: $\phi(t) = \mu - (\mu - |t|)_+^2/\mu$.

The bridge penalty and the logistic penalty have also been considered in [13]. Finally, the linear map $\mathcal{A}$ can be suitably chosen to model different scenarios. For example, $\mathcal{A}$ can be chosen to be the identity map for extracting L and S from a noisy data D, and the blurring map for a blurred data D. The linear map $\mathcal{B}$ can be the identity map or some "dictionary" that spans the data space (see, for example, [34]), and $\mathcal{C}$ can be chosen to be the identity map or the inverse of certain sparsifying transform (see, for example, [40]). More examples of (1.1) can be found in [8–10, 13, 41, 47].

One representative application that is frequently modeled by (1.1) via a suitable choice of Φ, Ψ, $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ is the background/foreground extraction problem, which is an important problem in video processing; see [6, 7] for recent surveys. In this problem, one attempts to separate the relatively static information called "background" and the moving objects called "foreground" in a video. The problem can be modeled by (1.1), and such models are typically referred to as RPCA-based models. In these models, each image is stacked as a column of a data matrix D, the relatively static background is then modeled as a low rank matrix, while the moving foreground is modeled as sparse outliers. The data matrix D is then decomposed (approximately) as the sum of a low rank matrix $L \in \mathbb{R}^{m \times n}$ modeling the background and a sparse matrix $S \in \mathbb{R}^{m \times n}$ modeling the foreground. Various approximations are then used to induce low rank and sparsity, resulting in different RPCA-based models, most of which take the form of (1.1). One example is to set Ψ to be the nuclear norm of L, i.e., the sum of singular values of L, to promote low rank in L and Φ to be the ℓ_1 norm of S to promote sparsity in S, as in [10]. Besides convex regularizers, nonconvex models have also been widely studied recently and their performances are promising; see [13, 44] for background/foreground extraction and [4, 12, 22, 38, 39, 50] for other problems in image processing. There are also nuclear-norm-free models that do not require matrix decomposition of the matrix variable L when solving them, making the model more practical especially when the size of matrix is large. For instance, in [30], the authors set Φ to be the ℓ_1 norm of S and Ψ to be the indicator function of $\Omega = \{L \in \mathbb{R}^{m \times n} \mid L_{:1} = L_{:2} = \cdots = L_{:n}\}$. A similar approach was also adopted in [33] with promising performances. Clearly, for nuclear-norm-free models, one can also take Φ to be some nonconvex sparsity inducing regularizers, resulting in a special case of (1.1) that has not been explicitly considered in the literature before; we will consider these models in our numerical experiments in Section 5. The above discussion shows that problem (1.1) is flexible enough to cover a wide range of RPCA-based models for background/foreground extraction.

Problem (1.1), though nonconvex in general, as we will show later in Section 3, can be reformulated into an optimization problem with three blocks of variables. This kind of problems containing several blocks of variables has been widely studied in the literature; see, for example, [30, 37, 41]. Hence, it is natural to adapt the algorithm used there, namely, the alternating direction method of multipliers (ADMM), for solving (1.1). Classically, the ADMM can be applied to solving problems of the following

form that contains 2 blocks of variables:

$$\min_{x_1, x_2} \{f_1(x_1) + f_2(x_2) \mid \mathcal{A}_1(x_1) + \mathcal{A}_2(x_2) = b\}, \qquad (1.3)$$

where f_1 and f_2 are proper closed convex functions, $\mathcal{A}_1$ and $\mathcal{A}_2$ are linear operators. The iterative scheme of ADMM is

$$\begin{cases} x_1^{k+1} \in \underset{x_1}{\mathrm{Argmin}} \left\{\mathcal{L}_\beta(x_1, x_2^k, z^k)\right\}, \\ x_2^{k+1} \in \underset{x_2}{\mathrm{Argmin}} \left\{\mathcal{L}_\beta(x_1^{k+1}, x_2, z^k)\right\}, \\ z^{k+1} = z^k - \tau\beta(\mathcal{A}_1(x_1^{k+1}) + \mathcal{A}_2(x_2^{k+1}) - b), \end{cases}$$

where $\tau \in (0, \frac{\sqrt{5}+1}{2})$ is the dual step-size and $\mathcal{L}_\beta$ is the augmented Lagrangian function for (1.3) defined as

$$\begin{aligned} \mathcal{L}_\beta(x_1, x_2, z) &:= f_1(x_1) + f_2(x_2) - \langle z, \mathcal{A}_1(x_1) + \mathcal{A}_2(x_2) - b \rangle \\ &\quad + \frac{\beta}{2}\|\mathcal{A}_1(x_1) + \mathcal{A}_2(x_2) - b\|^2 \end{aligned}$$

with $\beta > 0$ being the penalty parameter. Under some mild conditions, the sequence $\{(x_1^k, x_2^k)\}$ generated by the above ADMM can be shown to converge to an optimal solution of (1.3); see for example, [3, 15, 19, 21]. However, the ADMM used in [30, 37, 41] does not have a convergence guarantee; indeed, it is shown recently in [11] that the ADMM, when applied to a convex optimization problem with 3 blocks of variables, can be divergent in general. This motivates the study of many provably convergent variants of the ADMM for convex problems with more than 2 blocks of variables; see, for example, [24, 25, 35, 36]. Recently, Hong et al. [26] established the convergence of the multi-block ADMM for certain types of nonconvex problems whose objective is a sum of a possibly *nonconvex* Lipschitz differentiable function and a bunch of convex nonsmooth functions when the penalty parameter is chosen above a computable threshold. The problem they considered covers (1.1) when Φ is convex, or smooth and possibly nonconvex. Later, Wang et al. [44] considered a more general type of nonconvex problems that contains (1.1) as a special case and allows some nonconvex nonsmooth functions in the objective. To solve this type of problems, they considered a variant of the ADMM whose subproblems are simplified by adding a Bregman proximal term. However, their results cannot be applied to the direct adaptation of the ADMM for solving (1.1).

In this paper, following the studies in [26, 44] on convergence of nonconvex AD-MM and its variant, and the recent studies in [1, 31, 45], we manage to analyze the convergence of the ADMM applied to solving the possibly nonconvex problem (1.1). In addition, we would like to point out that all the aforementioned nonconvex ADMM have a dual step-size of $\tau = 1$. While it is known that the classical ADMM converges for any $\tau \in (0, \frac{\sqrt{5}+1}{2})$ for convex problems, and that empirically $\tau \approx \frac{\sqrt{5}+1}{2}$ works best (see, for example, [18, 19, 21, 36]), to our knowledge, the algorithm with a dual step-size $\tau \neq 1$ has never been studied in the nonconvex scenarios. Thus, we also study the ADMM with a general dual step-size, which will allow more flexibility in the design of algorithms.

The contributions of this paper are as follows:
  1. We show that for any positive dual step-size τ less than the golden ratio, the cluster point of the sequence generated by our ADMM gives a stationary

point of (1.1) if the penalty parameter is chosen above a computable threshold depending on τ, whenever the sequence is bounded. We achieve this via a potential function specifically constructed for our ADMM. To the best of our knowledge, this is the first convergence result for the ADMM in the nonconvex scenario with a possibly nontrivial dual step-size ($\tau \neq 1$). This result is also new for the convex scenario for the multi-block ADMM.

2. We establish global convergence of the whole sequence generated by the AD-MM under the additional assumption that the special potential function is a Kurdyka-Łojasiewicz function. Following the discussions in [2, Section 4], one can check that this condition is satisfied for all the aforementioned ϕ.

3. Furthermore, we discuss an initialization strategy to guarantee the boundedness of the sequence generated by the ADMM.

We also conduct numerical experiments to evaluate the performance of our ADMM by using different nonconvex regularizers and real data. Our computational results illustrate the efficiency of our ADMM with a nontrivial dual step-size.

The rest of this paper is organized as follows. We present notation and preliminaries in Section 2. The ADMM for (1.1) is described in Section 3. We analyze the convergence of the method in Section 4. Numerical results are presented in Section 5, with some concluding remarks given in Section 6.

2. Notation and preliminaries. In this paper, we use $\mathbb{R}^{m \times n}$ to denote the set of all $m \times n$ matrices. For a matrix $X \in \mathbb{R}^{m \times n}$, we let x_{ij} denote its (i,j)th entry and $X_{:j}$ denote its jth column. The number of nonzero entries in X is denoted by $\|X\|_0$ and the largest entry in magnitude is denoted by $\|X\|_\infty$. Moreover, the Fröbenius norm is denoted by $\|X\|_F$, the nuclear norm is denoted by $\|X\|_*$, which is the sum of singular values of X; and ℓ_1-norm and ℓ_p-quasi-norm ($0 < p < 1$) are given by $\|X\|_1 := \sum_{i=1}^m \sum_{j=1}^n |x_{ij}|$ and $\|X\|_p := \left(\sum_{i=1}^m \sum_{j=1}^n |x_{ij}|^p \right)^{\frac{1}{p}}$, respectively. Furthermore, for two matrices X and Y of the same size, we denote their trace inner product by $\langle X, Y \rangle := \sum_{i=1}^m \sum_{j=1}^n x_{ij} y_{ij}$. Finally, for the linear map $\mathcal{A} : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ in (1.1), its adjoint is denoted by $\mathcal{A}^*$, while the largest (resp., smallest) eigenvalue of the linear map $\mathcal{A}^* \mathcal{A}$ is denoted by $\lambda_{\max}$ (resp., $\lambda_{\min}$). The identity map is denoted by $\mathcal{I}$.

For an extended-real-valued function $f : \mathbb{R}^{m \times n} \to [-\infty, \infty]$, we say that it is *proper* if $f(X) > -\infty$ for all $X \in \mathbb{R}^{m \times n}$ and its domain $\text{dom} f := \{X \in \mathbb{R}^{m \times n} \mid f(X) < \infty\}$ is nonempty. For a proper function f, we use the notation $Y \xrightarrow{f} X$ to denote $Y \to X$ and $f(Y) \to f(X)$. Our basic *(limiting-)subdifferential* [42, Definition 8.3] of f at $X \in \text{dom} f$ used in this paper, denoted by $\partial f(X)$, is defined as

$$\partial f(X) := \left\{ D \in \mathbb{R}^{m \times n} : \exists\, X^k \xrightarrow{f} X \text{ and } D^k \to D \text{ with } D^k \in \widehat{\partial} f(X^k) \text{ for all } k \right\},$$

where $\widehat{\partial} f(U)$ denotes the Fréchet subdifferential of f at $U \in \text{dom} f$, which is the set of all $D \in \mathbb{R}^{m \times n}$ satisfying

$$\liminf_{Y \neq U, Y \to U} \frac{f(Y) - f(U) - \langle D, Y - U \rangle}{\|Y - U\|_F} \geq 0.$$

From the above definition, we can easily observe that

$$\left\{ D \in \mathbb{R}^{m \times n} : \exists\, X^k \xrightarrow{f} X,\ D^k \to D,\ D^k \in \partial f(X^k) \right\} \subseteq \partial f(X). \qquad (2.1)$$

We also recall that when f is continuously differentiable or convex, the above subdifferential coincides with the classical concept of derivative or convex subdifferential of

f; see, for example, [42, Exercise 8.8] and [42, Proposition 8.12]. Moreover, from the generalized Fermat's rule [42, Theorem 10.1], we know that if $X \in \mathbb{R}^{m \times n}$ is a local minimizer of f, then $0 \in \partial f(X)$. Additionally, for a function f with several groups of variables, we write $\partial_X f$ (resp., $\nabla_X f$) for the subdifferential (resp., derivative) of f with respect to the group of variables X.

For a compact convex set $\Omega \subseteq \mathbb{R}^{m \times n}$, its indicator function δ_Ω is defined by

$$\delta_\Omega(X) = \begin{cases} 0 & \text{if } X \in \Omega, \\ +\infty & \text{otherwise.} \end{cases}$$

The normal cone of Ω at the point $X \in \Omega$ is given by $\mathcal{N}_\Omega(X) = \partial \delta_\Omega(X)$. We also use $\text{dist}(X, \Omega)$ to denote the distance from X to Ω, i.e., $\text{dist}(X, \Omega) := \inf_{Y \in \Omega} \|X - Y\|_F$, and $\mathcal{P}_\Omega(X)$ to denote the unique closest point to X in Ω.

Next, we recall the Kurdyka-Łojasiewicz (KL) property, which plays an important role in our global convergence analysis. For notational simplicity, we use Ξ_η ($\eta > 0$) to denote the class of concave functions $\varphi : [0, \eta) \to \mathbb{R}_+$ satisfying: (1) $\varphi(0) = 0$; (2) φ is continuously differentiable on $(0, \eta)$ and continuous at 0; (3) $\varphi'(x) > 0$ for all $x \in (0, \eta)$. Then the KL property can be described as follows.

DEFINITION 2.1 (**KL property and KL function**). *Let f be a proper lower semicontinuous function.*

(i) *For $\tilde{X} \in \text{dom} \, \partial f := \{X \in \mathbb{R}^{m \times n} : \partial f(X) \neq \emptyset\}$, if there exist an $\eta \in (0, +\infty]$, a neighborhood V of $\tilde{X}$ and a function $\varphi \in \Xi_\eta$ such that for all $X \in V \cap \{X \in \mathbb{R}^{m \times n} : f(\tilde{X}) < f(X) < f(\tilde{X}) + \eta\}$, it holds that*

$$\varphi'(f(X) - f(\tilde{X}))\text{dist}(0, \partial f(X)) \geq 1,$$

*then f is said to have the **Kurdyka-Łojasiewicz (KL)** property at $\tilde{X}$.*

(ii) *If f satisfies the KL property at each point of $\text{dom} \, \partial f$, then f is called a KL function.*

We refer the interested readers to [2] and references therein for examples of KL functions. We also recall the following uniformized KL property, which was established in [5, Lemma 6].

PROPOSITION 2.2 (**Uniformized KL property**). *Suppose that f is a proper lower semicontinuous function and Γ is a compact set. If $f \equiv f^*$ on Γ for some constant f^* and satisfies the KL property at each point of Γ, then there exist $\varepsilon > 0$, $\eta > 0$ and $\varphi \in \Xi_\eta$ such that*

$$\varphi'(f(X) - f^*)\text{dist}(0, \partial f(X)) \geq 1$$

for all $X \in \{X \in \mathbb{R}^{m \times n} : \text{dist}(X, \Gamma) < \varepsilon\} \cap \{X \in \mathbb{R}^{m \times n} : f^* < f(X) < f^* + \eta\}$.

Before ending this section, we discuss first-order necessary conditions for (1.1). First, recall that (1.1) is the same as

$$\min_{L,S} \ \mathcal{F}(L, S) := \Psi(L) + \Phi(S) + \frac{1}{2} \|D - \mathcal{A}[\mathcal{B}(L) + \mathcal{C}(S)]\|_F^2.$$

Hence, from [42, Theorem 10.1], we have $0 \in \partial \mathcal{F}(\bar{L}, \bar{S})$ at any local minimizer $(\bar{L}, \bar{S})$ of (1.1). On the other hand, from [42, Exercise 8.8] and [42, Proposition 10.5], we see that

$$\partial \mathcal{F}(L, S) = \begin{pmatrix} \partial \Psi(L) + \mathcal{B}^* \mathcal{A}^* \left(\mathcal{A}(\mathcal{B}(L) + \mathcal{C}(S)) - D \right) \\ \partial \Phi(S) + \mathcal{C}^* \mathcal{A}^* \left(\mathcal{A}(\mathcal{B}(L) + \mathcal{C}(S)) - D \right) \end{pmatrix}.$$

Consequently, the first-order necessary conditions of (1.1) at the local minimizer $(\bar{L}, \bar{S})$ is given by:

$$\begin{cases} 0 \in \partial\Psi(\bar{L}) + \mathcal{B}^*\mathcal{A}^* \left(\mathcal{A}(\mathcal{B}(\bar{L}) + \mathcal{C}(\bar{S})) - D\right), \\ 0 \in \partial\Phi(\bar{S}) + \mathcal{C}^*\mathcal{A}^* \left(\mathcal{A}(\mathcal{B}(\bar{L}) + \mathcal{C}(\bar{S})) - D\right). \end{cases} \tag{2.2}$$

In this paper, we say that (L^*, S^*) is a stationary point of (1.1) if (L^*, S^*) satisfies (2.2) in place of $(\bar{L}, \bar{S})$.

3. Alternating direction method of multipliers. In this section, we present an ADMM for solving (1.1), which can be equivalently written as

$$\begin{aligned} \min_{L,S,Z} \quad & \Psi(L) + \Phi(S) + \frac{1}{2}\|D - \mathcal{A}(Z)\|_F^2 \\ \text{s.t.} \quad & \mathcal{B}(L) + \mathcal{C}(S) = Z. \end{aligned} \tag{3.1}$$

To describe the iterates of the ADMM, we first introduce the augmented Lagrangian function of the above optimization problem:

$$\mathcal{L}_\beta(L, S, Z, \Lambda) = \Psi(L) + \Phi(S) + \frac{1}{2}\|D - \mathcal{A}(Z)\|_F^2$$

$$- \langle \Lambda, \ \mathcal{B}(L) + \mathcal{C}(S) - Z \rangle + \frac{\beta}{2}\|\mathcal{B}(L) + \mathcal{C}(S) - Z\|_F^2,$$

where $\Lambda \in \mathbb{R}^{m \times n}$ is the Lagrangian multiplier and $\beta > 0$ is the penalty parameter. The ADMM for solving (3.1) (equivalently (1.1)) is then presented as follows:

Algorithm 1 ADMM for solving (3.1)

Input: Initial point (S^0, Z^0, Λ^0), dual step-size parameter $\tau > 0$, penalty parameter $\beta > 0$, $k = 0$

    while a termination criterion is not met, **do**

        Step 1. Set

$$\begin{cases} L^{k+1} \in \underset{L}{\text{Argmin}} \ \mathcal{L}_\beta(L, S^k, Z^k, \Lambda^k) & (3.2a) \\[2mm] S^{k+1} \in \underset{S}{\text{Argmin}} \ \mathcal{L}_\beta(L^{k+1}, S, Z^k, \Lambda^k) & (3.2b) \\[2mm] Z^{k+1} = \underset{Z}{\text{argmin}} \ \mathcal{L}_\beta(L^{k+1}, S^{k+1}, Z, \Lambda^k) & (3.2c) \\[2mm] \Lambda^{k+1} = \Lambda^k - \tau\beta(\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^{k+1}) \end{cases}$$

        Step 2. Set $k := k + 1$

    end while

Output: (L^k, S^k)

Comparing with the ADMM considered in [26], the above algorithm has an extra dual step-size parameter $\tau > 0$ in the Λ-update. Such a dual step-size was introduced in [19, 21] for the classical ADMM (i.e., for convex problems with two separate blocks of variables), and was further studied in [18, 36, 43, 48] for other variants of the ADMM. Numerically, it was also demonstrated in [43] that a larger dual step-size ($\tau \approx \frac{\sqrt{5}+1}{2}$) results in faster convergence for the convex problems they consider. Thus, we adapt this dual step-size τ in our algorithm above. Surprisingly, in our numerical

experiments, a parameter choice of $\tau \approx \frac{\sqrt{5}+1}{2}$ leads to the worst performance for our nonconvex problems.

When $\tau = 1$, the above algorithm is a special case of the general algorithm studied in [26] when Ψ and Φ are smooth functions, or convex nonsmooth functions. The algorithm is shown to converge when β is chosen above a computable threshold. However, their convergence result cannot be directly applied when $\tau \neq 1$ or when Φ is nonsmooth and nonconvex. Nevertheless, following their analysis and the related studies [31, 44, 45], the above algorithm can be shown to be convergent under suitable assumptions. We will present the convergence analysis in Section 4.

Before ending this section, we further discuss the three subproblems in Algorithm 1. First, notice that the L-update and S-update are given by

$$\begin{cases} L^{k+1} \in \underset{L}{\text{Argmin}} \ \left\{ \Psi(L) + \frac{\beta}{2} \|\mathcal{B}(L) + \mathcal{C}(S^k) - Z^k - \frac{1}{\beta}\Lambda^k\|_F^2 \right\}, \\ S^{k+1} \in \underset{S}{\text{Argmin}} \ \left\{ \Phi(S) + \frac{\beta}{2} \|\mathcal{B}(L^{k+1}) + \mathcal{C}(S) - Z^k - \frac{1}{\beta}\Lambda^k\|_F^2 \right\}. \end{cases}$$

In general, these two subproblems are not easy to solve. However, when Ψ and Φ are chosen to be some common regularizers used in the literature, for example, $\Psi(L) = \|L\|_*$ and $\Phi(S) = \|S\|_1$, then these subproblems can be solved efficiently via the proximal gradient method. Additionally, when $\Psi(L) = \delta_\Omega(L)$ with Ω being a closed convex set and $\mathcal{B} = \mathcal{I}$, the L-update can be given explicitly by

$$L^{k+1} = \mathcal{P}_\Omega \left(-\mathcal{C}(S^k) + Z^k + \frac{1}{\beta}\Lambda^k \right),$$

which can be computed efficiently if Ω is simple, for example, when $\Omega = \{L \in \mathbb{R}^{m \times n} \mid \|L\|_\infty \leq l, \ L_{:1} = L_{:2} = \cdots = L_{:n}\}$ for some $l > 0$. For the S-update, when Φ is given by (1.2) with ϕ being one of the penalty functions presented in the introduction and $\mathcal{C} = \mathcal{I}$, it can be solved efficiently via a simple root-finding procedure. Finally, from the optimality conditions of (3.2c), the Z^{k+1} can be obtained by solving the following linear system

$$\mathcal{A}^*\mathcal{A}(Z) + \beta Z = \mathcal{A}^*(D) - \Lambda^k + \beta \left(\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) \right),$$

whose complexity would depend on the choice of $\mathcal{A}$ in our model (1.1). For example, when $\mathcal{A}$ is just the identity map, the Z^{k+1} is given explicitly by

$$Z^{k+1} = \frac{1}{1+\beta} \left[D - \Lambda^k + \beta \left(\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) \right) \right].$$

4. Convergence analysis. In this section, we discuss the convergence of Algorithm 1 for $0 < \tau < \frac{1+\sqrt{5}}{2}$. We first present the first-order optimality conditions for the subproblems in Algorithm 1 as follows, which will be used repeatedly in our convergence analysis below.

$$\begin{cases} 0 \in \partial\Psi(L^{k+1}) - \mathcal{B}^*(\Lambda^k) + \beta\mathcal{B}^* \left(\mathcal{B}(L^{k+1}) + \mathcal{C}(S^k) - Z^k \right), & \text{(4.1a)} \\ 0 \in \partial\Phi(S^{k+1}) - \mathcal{C}^*(\Lambda^k) + \beta\mathcal{C}^* \left(\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^k \right), & \text{(4.1b)} \\ 0 = \mathcal{A}^*(\mathcal{A}(Z^{k+1}) - D) + \Lambda^k - \beta(\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^{k+1}), & \text{(4.1c)} \\ \Lambda^{k+1} - \Lambda^k = -\tau\beta \left(\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^{k+1} \right). & \text{(4.1d)} \end{cases}$$

Our convergence analysis is largely based on the following potential function:

$$\Theta_{\tau,\beta}(L, S, Z, \Lambda) = \mathcal{L}_\beta(L, S, Z, \Lambda) + \theta(\tau)\beta\|\mathcal{B}(L) + \mathcal{C}(S) - Z\|_F^2,$$

where

$$\theta(\tau) := \max\left\{ 1 - \tau, \ \frac{(\tau - 1)\tau^2}{1 + \tau - \tau^2} \right\}, \quad \text{for } 0 < \tau < \frac{1 + \sqrt{5}}{2}. \tag{4.2}$$

Note that $\theta(\cdot)$ is a convex and nonnegative function on $\left(0, \frac{1+\sqrt{5}}{2} \right)$. Thus, for any (L, S, Z, Λ), we have $\Theta_{\tau,\beta}(L, S, Z, \Lambda) \geq \mathcal{L}_\beta(L, S, Z, \Lambda)$ for $0 < \tau < \frac{1+\sqrt{5}}{2}$, and the equality holds when $\tau = 1$ (so that $\theta(\tau) = 0$).

Our convergence analysis also relies on the following assumption.

ASSUMPTION 4.1. Ψ, Φ, $\mathcal{B}$, $\mathcal{C}$, β and τ satisfy

(a1) $\mathcal{B}^*\mathcal{B} \succeq \sigma\mathcal{I}$ for some $\sigma > 0$ and $\mathcal{C}^*\mathcal{C} \succeq \sigma'\mathcal{I}$ for some $\sigma' > 0$;

(a2) Ψ is continuous in its domain;

(a3) the first iterate $(L^1, S^1, Z^1, \Lambda^1)$ satisfies

$$\Theta_{\tau,\beta}(L^1, S^1, Z^1, \Lambda^1) < h_0 := \liminf_{\|L\|_F + \|S\|_F \to \infty} \Psi(L) + \Phi(S).$$

REMARK 4.1 (**Note on Assumption 4.1**). *(i) Since $\mathcal{B}$ and $\mathcal{C}$ in (1.1) are injective, (a1) holds trivially; (ii) (a2) holds for many common regularizers (for example, the nuclear norm) or the indicator function of a set; (iii) (a3) places conditions on the* first *iterate of the algorithm. It is not hard to observe that this assumption holds trivially if both Ψ and Φ are coercive, i.e., if* $\liminf_{\|L\|_F + \|S\|_F \to \infty} \Psi(L) + \Phi(S) = \infty$. *We will discuss more sufficient conditions for this assumption after our convergence results, i.e., after Theorem 4.4.*

We now start our convergence analysis by proving the following preparatory lemma, which states that the potential function is decreasing along the sequence generated from Algorithm 1 if the penalty parameter β is chosen above a computable threshold.

LEMMA 4.1. *Suppose that $0 < \tau < \frac{1+\sqrt{5}}{2}$ and $\{(L^k, S^k, Z^k, \Lambda^k)\}$ is a sequence generated by Algorithm 1. If (a1) in Assumption 4.1 holds, then for $k \geq 1$, we have*

$$\Theta_{\tau,\beta}(L^{k+1}, S^{k+1}, Z^{k+1}, \Lambda^{k+1}) - \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k)$$

$$\leq \left(\max\left\{ \frac{1}{\tau}, \frac{\tau^2}{1+\tau-\tau^2} \right\} \cdot \frac{\lambda_{\max}^2}{\beta} - \frac{\lambda_{\min} + \beta}{2} \right) \|Z^{k+1} - Z^k\|_F^2 - \frac{\sigma\beta}{2}\|L^{k+1} - L^k\|_F^2. \tag{4.3}$$

Moreover, if $\beta \geq -\frac{\lambda_{\min}}{2} + \frac{1}{2}\sqrt{\lambda_{\min}^2 + \max\left\{ \frac{1}{\tau}, \frac{\tau^2}{1+\tau-\tau^2} \right\} \cdot 8\lambda_{\max}^2}$, then the sequence $\{\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^\infty$ is decreasing.

Proof. We start our proof by noticing that

$$\Theta_{\tau,\beta}(L^{k+1}, S^{k+1}, Z^{k+1}, \Lambda^{k+1}) - \Theta_{\tau,\beta}(L^{k+1}, S^{k+1}, Z^{k+1}, \Lambda^k)$$

$$= -\langle \Lambda^{k+1} - \Lambda^k, L^{k+1} + S^{k+1} - Z^{k+1} \rangle = \frac{1}{\tau\beta}\|\Lambda^{k+1} - \Lambda^k\|_F^2, \tag{4.4}$$

where the last equality follows from (4.1d). We next derive an upper bound of $\|\Lambda^{k+1} - \Lambda^k\|_F^2$. To proceed, we first note from (4.1c) that

$$0 = \mathcal{A}^*(\mathcal{A}(Z^{k+1}) - D) + \Lambda^k - \beta(\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^{k+1})$$

$$= \mathcal{A}^*(\mathcal{A}(Z^{k+1}) - D) + \Lambda^k + \frac{1}{\tau}(\Lambda^{k+1} - \Lambda^k)$$

$$\implies \quad \Lambda^{k+1} = \tau\mathcal{A}^*(D - \mathcal{A}(Z^{k+1})) + (1 - \tau)\Lambda^k,$$

where the second equality follows from (4.1d). Hence, for $k \geq 1$,

$$
\begin{aligned}
\Lambda^{k+1} &- \Lambda^k \\
&= [\tau \mathcal{A}^*(D - \mathcal{A}(Z^{k+1})) + (1-\tau)\Lambda^k] - [\tau \mathcal{A}^*(D - \mathcal{A}(Z^k)) + (1-\tau)\Lambda^{k-1}] \\
&= \tau \mathcal{A}^* \mathcal{A}(Z^k - Z^{k+1}) + (1-\tau)(\Lambda^k - \Lambda^{k-1}).
\end{aligned} \tag{4.5}
$$

We now consider two separate cases: $0 < \tau \leq 1$ and $1 < \tau < \frac{1+\sqrt{5}}{2}$.

- For $0 < \tau \leq 1$, it follows from the convexity of $\|\cdot\|_F^2$ that

$$
\begin{aligned}
\|\Lambda^{k+1} - \Lambda^k\|_F^2 &= \left\|\tau \mathcal{A}^* \mathcal{A}(Z^k - Z^{k+1}) + (1-\tau)(\Lambda^k - \Lambda^{k-1})\right\|_F^2 \\
&\leq \tau \lambda_{\max}^2 \|Z^{k+1} - Z^k\|_F^2 + (1-\tau)\|\Lambda^k - \Lambda^{k-1}\|_F^2.
\end{aligned}
$$

We further add $-(1-\tau)\left\|\Lambda^{k+1} - \Lambda^k\right\|_F^2$ to both sides of the above inequality and simplify the resulting inequality to get

$$
\begin{aligned}
\|\Lambda^{k+1} &- \Lambda^k\|_F^2 \\
&\leq \lambda_{\max}^2 \|Z^{k+1} - Z^k\|_F^2 + \tfrac{1-\tau}{\tau}\left(\|\Lambda^k - \Lambda^{k-1}\|_F^2 - \|\Lambda^{k+1} - \Lambda^k\|_F^2\right) \\
&= (1-\tau)\tau\beta^2 \left(\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F^2 - \|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^{k+1}\|_F^2\right) \\
&\quad + \lambda_{\max}^2 \|Z^{k+1} - Z^k\|_F^2.
\end{aligned} \tag{4.6}
$$

where the last equality follows from (4.1d).

- For $1 < \tau < \frac{1+\sqrt{5}}{2}$, dividing τ from both sides of (4.5), we have

$$
\begin{aligned}
\frac{1}{\tau}\left(\Lambda^{k+1} - \Lambda^k\right) &= \mathcal{A}^* \mathcal{A}\left(Z^k - Z^{k+1}\right) + \left(\frac{1}{\tau} - 1\right)\left(\Lambda^k - \Lambda^{k-1}\right) \\
&= \frac{1}{\tau}\,\tau \mathcal{A}^* \mathcal{A}\left(Z^k - Z^{k+1}\right) + \left(1 - \frac{1}{\tau}\right)\left(\Lambda^{k-1} - \Lambda^k\right).
\end{aligned}
$$

This together with $0 < \frac{1}{\tau} < 1$ and the convexity of $\|\cdot\|_F^2$, implies that

$$
\begin{aligned}
\left\|\tfrac{1}{\tau}\left(\Lambda^{k+1} - \Lambda^k\right)\right\|_F^2 &\leq \tfrac{1}{\tau}\|\tau \mathcal{A}^* \mathcal{A}\left(Z^k - Z^{k+1}\right)\|_F^2 + \left(1 - \tfrac{1}{\tau}\right)\|\Lambda^{k-1} - \Lambda^k\|_F^2 \\
&\leq \tau \lambda_{\max}^2 \|Z^{k+1} - Z^k\|_F^2 + \left(1 - \tfrac{1}{\tau}\right)\|\Lambda^k - \Lambda^{k-1}\|_F^2 \\
\implies \quad \|\Lambda^{k+1} - \Lambda^k\|_F^2 &\leq \tau^3 \lambda_{\max}^2 \|Z^{k+1} - Z^k\|_F^2 + \left(\tau^2 - \tau\right)\|\Lambda^k - \Lambda^{k-1}\|_F^2.
\end{aligned}
$$

Then, adding $-\left(\tau^2 - \tau\right)\left\|\Lambda^{k+1} - \Lambda^k\right\|_F^2$ to both sides of the above inequality, simplifying the resulting inequality and using the fact that $1 + \tau - \tau^2 > 0$ for $1 < \tau < \frac{1+\sqrt{5}}{2}$, we see that

$$
\begin{aligned}
\|\Lambda^{k+1} &- \Lambda^k\|_F^2 \\
&\leq \tfrac{\tau^3 \lambda_{\max}^2}{1+\tau-\tau^2}\|Z^{k+1} - Z^k\|_F^2 + \tfrac{\tau^2-\tau}{1+\tau-\tau^2}\left(\|\Lambda^k - \Lambda^{k-1}\|_F^2 - \|\Lambda^{k+1} - \Lambda^k\|_F^2\right) \\
&= \tfrac{\tau^3 \lambda_{\max}^2}{1+\tau-\tau^2}\|Z^{k+1} - Z^k\|_F^2 + \tfrac{(\tau-1)\tau^3\beta^2}{1+\tau-\tau^2}\left(\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F^2\right. \\
&\quad \left. - \|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^{k+1}\|_F^2\right),
\end{aligned} \tag{4.7}
$$

where the equality follows from (4.1d).

Thus, for $0 < \tau < \frac{1+\sqrt{5}}{2}$, combining (4.6), (4.7) and recalling the definition of $\theta(\tau)$ in (4.2), we have

$$
\begin{aligned}
\frac{1}{\tau\beta}\|\Lambda^{k+1} - \Lambda^k\|_F^2 &\leq \max\left\{\frac{1}{\tau}, \frac{\tau^2}{1+\tau-\tau^2}\right\} \cdot \frac{\lambda_{\max}^2}{\beta}\|Z^{k+1} - Z^k\|_F^2 \\
&+ \theta(\tau)\beta\left(\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F^2 - \|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^{k+1}\|_F^2\right).
\end{aligned}
\tag{4.8}
$$

Next, note that the function $Z \mapsto \mathcal{L}_\beta(L^{k+1}, S^{k+1}, Z, \Lambda^k)$ is strongly convex with modulus at least $\lambda_{\min} + \beta$. Using this fact and the definition of Z^{k+1} as a minimizer in (3.2c), we see that

$$
\begin{aligned}
&\Theta_{\tau,\beta}(L^{k+1}, S^{k+1}, Z^{k+1}, \Lambda^k) - \Theta_{\tau,\beta}(L^{k+1}, S^{k+1}, Z^k, \Lambda^k) \\
&= \mathcal{L}_\beta(L^{k+1}, S^{k+1}, Z^{k+1}, \Lambda^k) - \mathcal{L}_\beta(L^{k+1}, S^{k+1}, Z^k, \Lambda^k) \\
&\quad + \theta(\tau)\beta\left(\|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^{k+1}\|_F^2 - \|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^k\|_F^2\right) \\
&\leq \theta(\tau)\beta\left(\|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^{k+1}\|_F^2 - \|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^k\|_F^2\right) \\
&\quad - \frac{\lambda_{\min}+\beta}{2}\|Z^{k+1} - Z^k\|_F^2.
\end{aligned}
\tag{4.9}
$$

Moreover, using the fact that S^{k+1} is a minimizer in (3.2b), we have

$$
\begin{aligned}
&\Theta_{\tau,\beta}(L^{k+1}, S^{k+1}, Z^k, \Lambda^k) - \Theta_{\tau,\beta}(L^{k+1}, S^k, Z^k, \Lambda^k) \\
&= \mathcal{L}_\beta(L^{k+1}, S^{k+1}, Z^k, \Lambda^k) - \mathcal{L}_\beta(L^{k+1}, S^k, Z^k, \Lambda^k) \\
&\quad + \theta(\tau)\beta\left(\|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^k\|_F^2 - \|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^k) - Z^k\|_F^2\right) \\
&\leq \theta(\tau)\beta\left(\|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^k\|_F^2 - \|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^k) - Z^k\|_F^2\right).
\end{aligned}
\tag{4.10}
$$

Finally, note that $L \mapsto \mathcal{L}_\beta(L, S^k, Z^k, \Lambda^k)$ is strongly convex with modulus at least $\sigma\beta$ from (a1) in Assumption 4.1. From this, we can similarly obtain

$$
\begin{aligned}
&\Theta_{\tau,\beta}(L^{k+1}, S^k, Z^k, \Lambda^k) - \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) \\
&\leq \theta(\tau)\beta\left(\|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^k) - Z^k\|_F^2 - \|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F^2\right) \\
&\quad - \frac{\sigma\beta}{2}\|L^{k+1} - L^k\|_F^2.
\end{aligned}
\tag{4.11}
$$

Thus, summing (4.4), (4.8), (4.9), (4.10) and (4.11), we obtain (4.3).

Now, suppose in addition that $\beta \geq -\frac{\lambda_{\min}}{2} + \frac{1}{2}\sqrt{\lambda_{\min}^2 + \max\left\{\frac{1}{\tau}, \frac{\tau^2}{1+\tau-\tau^2}\right\} \cdot 8\lambda_{\max}^2}$. Then it is easy to check that

$$
\max\left\{\frac{1}{\tau}, \frac{\tau^2}{1+\tau-\tau^2}\right\} \cdot \frac{\lambda_{\max}^2}{\beta} - \frac{\lambda_{\min}+\beta}{2} \leq 0.
$$

Hence we see from (4.3) that

$$
\Theta_{\tau,\beta}(L^{k+1}, S^{k+1}, Z^{k+1}, \Lambda^{k+1}) - \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) \leq 0,
$$

which means that $\{\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^\infty$ is decreasing. This completes the proof. $\square$

We next show that the sequence generated by Algorithm 1 is bounded if β is chosen above a computable threshold, under (a1) and (a3) in Assumption 4.1. For notational simplicity, from now on, we let

$$
\bar\beta := \max\left\{\max\{1/\tau, \tau\} \cdot \lambda_{\max}, -\frac{\lambda_{\min}}{2} + \frac{1}{2}\sqrt{\lambda_{\min}^2 + \max\left\{\frac{1}{\tau}, \frac{\tau^2}{1+\tau-\tau^2}\right\} \cdot 8\lambda_{\max}^2}\right\}.
\tag{4.12}
$$

PROPOSITION 4.2 (**Boundedness of sequence generated by ADMM**). *Suppose that $0 < \tau < \frac{1+\sqrt{5}}{2}$ and $\beta > \bar{\beta}$. If (a1) and (a3) in Assumption 4.1 hold, then a sequence $\{(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^{\infty}$ generated by Algorithm 1 is bounded.*

Proof. With our choice of β and (a1) in Assumption 4.1, we see immediately from Lemma 4.1 that the sequence $\{\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^{\infty}$ is decreasing. This together with (a3) in Assumption 4.1 shows that, for $k \geq 1$,

$$
\begin{aligned}
h_0 \; > \; & \Theta_{\tau,\beta}(L^1, S^1, Z^1, \Lambda^1) \; \geq \; \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) \\
= & \Psi(L^k) + \Phi(S^k) + \frac{1}{2}\|D - \mathcal{A}(Z^k)\|_F^2 - \langle \Lambda^k, \mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k \rangle \\
& + (1 + 2\theta(\tau))\frac{\beta}{2}\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F^2 \\
= & \Psi(L^k) + \Phi(S^k) + \frac{1}{2}\|D - \mathcal{A}(Z^k)\|_F^2 + \frac{\beta}{2}\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k - \frac{1}{\beta}\Lambda^k\|_F^2 \\
& - \frac{1}{2\beta}\|\Lambda^k\|_F^2 + \theta(\tau)\beta\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F^2,
\end{aligned}
\tag{4.13}
$$

where the last equality is obtained by completing the square. We next derive an upper bound for $\|\Lambda^k\|_F^2$. We start by substituting (4.1d) into (4.1c) and rearranging terms to obtain

$$
\begin{aligned}
0 &= \mathcal{A}^*(\mathcal{A}(Z^k) - D) + \Lambda^{k-1} + \frac{1}{\tau}(\Lambda^k - \Lambda^{k-1}) \\
&\Longrightarrow -\tau\Lambda^k = \tau\mathcal{A}^*(\mathcal{A}(Z^k) - D) + (1 - \tau)(\Lambda^k - \Lambda^{k-1}).
\end{aligned}
\tag{4.14}
$$

We now consider two different cases:

• For $0 < \tau \leq 1$, it follows from the convexity of $\|\cdot\|_F^2$ and (4.14) that

$$
\begin{aligned}
\|-\tau\Lambda^k\|_F^2 &\leq \tau\|\mathcal{A}^*(\mathcal{A}(Z^k) - D)\|_F^2 + (1 - \tau)\|\Lambda^k - \Lambda^{k-1}\|_F^2 \\
&\leq \tau\lambda_{\max}\|\mathcal{A}(Z^k) - D\|_F^2 + (1 - \tau)\|\Lambda^k - \Lambda^{k-1}\|_F^2 \\
&= \tau\lambda_{\max}\|\mathcal{A}(Z^k) - D\|_F^2 + (1 - \tau)\tau^2\beta^2\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F^2,
\end{aligned}
$$

where the equality follows from (4.1d). Then, we have

$$
\|\Lambda^k\|_F^2 \leq \frac{\lambda_{\max}}{\tau}\|\mathcal{A}(Z^k) - D\|_F^2 + (1 - \tau)\beta^2\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F^2. \tag{4.15}
$$

• For $1 < \tau < \frac{1+\sqrt{5}}{2}$, by dividing $-\tau$ from both sides of (4.1c), we obtain

$$
\Lambda^k = \frac{1}{\tau}\tau\mathcal{A}^*(D - \mathcal{A}(Z^k)) + \left(1 - \frac{1}{\tau}\right)(\Lambda^k - \Lambda^{k-1}).
$$

Then, since $0 < \frac{1}{\tau} < 1$, using the convexity of $\|\cdot\|_F^2$ and (4.1d), we have

$$
\begin{aligned}
\|\Lambda^k\|_F^2 &\leq \frac{1}{\tau}\|\tau\mathcal{A}^*(D - \mathcal{A}(Z^k))\|_F^2 + \left(1 - \frac{1}{\tau}\right)\|\Lambda^k - \Lambda^{k-1}\|_F^2 \\
&\leq \tau\lambda_{\max}\|D - \mathcal{A}(Z^k)\|_F^2 + (\tau - 1)\tau\beta^2\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F^2.
\end{aligned}
\tag{4.16}
$$

Thus, combining (4.15) and (4.16), we have

$$
\|\Lambda^k\|_F^2 \leq \max\{1/\tau, \tau\} \cdot \lambda_{\max} \|D - \mathcal{A}(Z^k)\|_F^2
$$
$$
+ \max\{1 - \tau, (\tau - 1)\tau\}\beta^2 \|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F^2
$$
$$
\implies \quad -\frac{1}{2\beta}\|\Lambda^k\|_F^2 \geq -\frac{\max\{1/\tau, \tau\}\lambda_{\max}}{2\beta}\|D - \mathcal{A}(Z^k)\|_F^2 \tag{4.17}
$$
$$
-\frac{\max\{1 - \tau, (\tau - 1)\tau\}\beta}{2}\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F^2.
$$

Substituting (4.17) into (4.13), we have

$$
h_0 > \Theta_{\tau, \beta}(L^k, S^k, Z^k, \Lambda^k) \geq \Psi(L^k) + \Phi(S^k)
$$
$$
+ \frac{1}{2}\left(1 - \max\{1/\tau, \tau\} \cdot \frac{\lambda_{\max}}{\beta}\right)\|D - \mathcal{A}(Z^k)\|_F^2
$$
$$
+ \frac{\beta}{2}\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k - \frac{1}{\beta}\Lambda^k\|_F^2 \tag{4.18}
$$
$$
+ [2\theta(\tau) - \max\{1 - \tau, (\tau - 1)\tau\}] \cdot \frac{\beta}{2}\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F^2.
$$

With (4.18) established, we are now ready to prove the boundedness of the sequence. We start with the observation that for $0 < \tau < \frac{1+\sqrt{5}}{2}$ and $\beta > \bar{\beta}$, we always have

$$
1 - \max\{1/\tau, \tau\} \cdot \frac{\lambda_{\max}}{\beta} > 0 \tag{4.19}
$$

and

$$
2\theta(\tau) - \max\{1 - \tau, (\tau - 1)\tau\} = \begin{cases} 1 - \tau > 0, & \text{for } 0 < \tau < 1, \\ 0, & \text{for } \tau = 1, \\ \frac{\tau(\tau-1)(\tau^2+\tau-1)}{1+\tau-\tau^2} > 0, & \text{for } 1 < \tau < \frac{1+\sqrt{5}}{2}, \end{cases} \tag{4.20}
$$

where $\theta(\tau)$ is defined in (4.2). Then we consider two cases:

- For $\tau \in (0, 1) \cup \left(1, \frac{1+\sqrt{5}}{2}\right)$, it follows from (4.18), (4.19), (4.20), and the nonnegativity of Ψ and Φ that $\{\|D - \mathcal{A}(Z^k)\|_F\}$, $\{\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k - \frac{1}{\beta}\Lambda^k\|_F\}$ and $\{\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F\}$ are bounded; and moreover,

$$
\Psi(L^k) + \Phi(S^k) < h_0.
$$

The boundedness of $\{L^k\}$ and $\{S^k\}$ follows immediately from this last relation. Furthermore, $\{\Lambda^k\}$ is bounded since

$$
\|\Lambda^k\|_F \leq \beta\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k - \frac{1}{\beta}\Lambda^k\|_F + \beta\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k\|_F.
$$

Finally, we obtain the boundedness of $\{Z^k\}$ from

$$
\|Z^k\|_F \leq \|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k - \frac{1}{\beta}\Lambda^k\|_F + \|\mathcal{B}(L^k)\|_F
$$
$$
+ \|\mathcal{C}(S^k)\|_F + \frac{1}{\beta}\|\Lambda^k\|_F. \tag{4.21}
$$

- For $\tau = 1$, it follows from (4.18), (4.19), (4.20), and the nonnegativity of Ψ and Φ that $\{\|D - \mathcal{A}(Z^k)\|_F\}$ and $\{\|\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k - \frac{1}{\beta}\Lambda^k\|_F\}$ are bounded; and moreover $\Psi(L^k) + \Phi(S^k) < h_0$, from which we see immediately that $\{L^k\}$ and $\{S^k\}$ are bounded. The boundedness of $\{\Lambda^k\}$ now follows from (4.14) with $\tau = 1$, i.e., $\Lambda^k = \mathcal{A}^*(D - \mathcal{A}(Z^k))$. The boundedness of $\{Z^k\}$ again follows from (4.21).

This completes the proof. □

We are now ready to prove our first global convergence result for Algorithm 1, which also characterizes the cluster point of the sequence generated.

THEOREM 4.3 (**Global subsequential convergence**). *Suppose that $0 < \tau < \frac{1+\sqrt{5}}{2}$ and $\beta > \bar{\beta}$. If Assumption 4.1 holds, then*
 (i) $\lim_{k\to\infty} \|L^{k+1} - L^k\|_F + \|S^{k+1} - S^k\|_F + \|Z^{k+1} - Z^k\|_F + \|\Lambda^{k+1} - \Lambda^k\|_F = 0;$
 (ii) *Any cluster point $(L^*, S^*, Z^*, \Lambda^*)$ of a sequence $\{(L^k, S^k, Z^k, \Lambda^k)\}$ generated by Algorithm 1 is a stationary point of* (1.1).

Proof. The boundedness of the sequence $\{(L^k, S^k, Z^k, \Lambda^k)\}$ follows immediately from Proposition 4.2 and thus a cluster point exists. We now prove statement (i).

Suppose that $(L^*, S^*, Z^*, \Lambda^*)$ is a cluster point of the sequence $\{(L^k, S^k, Z^k, \Lambda^k)\}$ and let $\{(L^{k_i}, S^{k_i}, Z^{k_i}, \Lambda^{k_i})\}$ be a convergent subsequence such that

$$\lim_{i\to\infty} (L^{k_i}, S^{k_i}, Z^{k_i}, \Lambda^{k_i}) = (L^*, S^*, Z^*, \Lambda^*).$$

By summing (4.3) from $k = 1$ to $k = k_i - 1$, we have

$$
\begin{aligned}
&\Theta_{\tau,\beta}(L^{k_i}, S^{k_i}, Z^{k_i}, \Lambda^{k_i}) - \Theta_{\tau,\beta}(L^1, S^1, Z^1, \Lambda^1) \\
&\leq -C \sum_{k=1}^{k_i-1} \|Z^{k+1} - Z^k\|_F^2 - \frac{\sigma\beta}{2} \sum_{k=1}^{k_i-1} \|L^{k+1} - L^k\|_F^2,
\end{aligned}
\tag{4.22}
$$

where $C := \frac{\lambda_{\min} + \beta}{2} - \max\left\{\frac{1}{\tau}, \frac{\tau^2}{1+\tau-\tau^2}\right\} \cdot \frac{\lambda_{\max}^2}{\beta} > 0$ (since $\beta > \bar{\beta}$). Passing to the limit in (4.22) and rearranging terms in the resulting relation, we obtain

$$
\begin{aligned}
&C \sum_{k=1}^{\infty} \|Z^{k+1} - Z^k\|_F^2 + \frac{\sigma\beta}{2} \sum_{k=1}^{\infty} \|L^{k+1} - L^k\|_F^2 \\
&\leq \Theta_{\tau,\beta}(L^1, S^1, Z^1, \Lambda^1) - \Theta_{\tau,\beta}(L^*, S^*, Z^*, \Lambda^*) < \infty,
\end{aligned}
$$

where the last inequality follows from the properness of Ψ and Φ. This together with $C > 0$ and $\sigma > 0$ implies that

$$\sum_{k=1}^{\infty} \|Z^{k+1} - Z^k\|_F^2 < \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \|L^{k+1} - L^k\|_F^2 < \infty.$$

Hence, we have

$$Z^{k+1} - Z^k \to 0, \ L^{k+1} - L^k \to 0. \tag{4.23}$$

Next, by summing both sides of (4.8) from $k = 1$ to $k = k_i$ and passing to the limit, we have

$$
\begin{aligned}
\sum_{k=1}^{\infty} \|\Lambda^{k+1} - \Lambda^k\|_F^2 &\leq \max\left\{\tfrac{1}{\tau}, \tfrac{\tau^2}{1+\tau-\tau^2}\right\} \cdot \tau \lambda_{\max}^2 \sum_{k=1}^{\infty} \|Z^{k+1} - Z^k\|_F^2 \\
&\quad + \theta(\tau)\tau\beta^2 \left(\|\mathcal{B}(L^1) + \mathcal{C}(S^1) - Z^1\|_F^2 - \liminf_{k\to\infty} \|\mathcal{B}(L^{k+1}) + \mathcal{C}(S^{k+1}) - Z^{k+1}\|_F^2 \right),
\end{aligned}
$$

from which we conclude that

$$\Lambda^{k+1} - \Lambda^k \to 0. \tag{4.24}$$

Finally, we have $S^{k+1} - S^k \to 0$ from (4.23), (4.24), (4.1d) and (a1) in Assumption 4.1. This proves statement (i).

We next prove statement (ii). From the lower semicontinuity of $\Theta_{\tau,\beta}$ (since Ψ and Φ are lower semicontinuous), we have

$$
\begin{aligned}
\liminf_{i\to\infty} \Theta_{\tau,\beta}(L^{k_i+1}, S^{k_i+1}, Z^{k_i}, \Lambda^{k_i}) \geq{}& \Psi(L^*) + \Phi(S^*) + \frac{1}{2}\|D - \mathcal{A}(Z^*)\|_F^2 \\
&- \langle \Lambda^*,\ \mathcal{B}(L^*) + \mathcal{C}(S^*) - Z^* \rangle + (1 + 2\theta(\tau))\frac{\beta}{2}\|\mathcal{B}(L^*) + \mathcal{C}(S^*) - Z^*\|_F^2.
\end{aligned}
\tag{4.25}
$$

On the other hand, from the definition of S^{k_i+1} as a minimizer in (3.2b), we have

$$
\begin{aligned}
\Theta_{\tau,\beta}(L^{k_i+1}, S^{k_i+1}, Z^{k_i}, \Lambda^{k_i}) \leq{}& \Theta_{\tau,\beta}(L^{k_i+1}, S^*, Z^{k_i}, \Lambda^{k_i}) \\
&+ \theta(\tau)\beta \left(\|\mathcal{B}(L^{k_i+1}) + \mathcal{C}(S^{k_i+1}) - Z^{k_i}\|_F^2 - \|\mathcal{B}(L^{k_i+1}) + \mathcal{C}(S^*) - Z^{k_i}\|_F^2 \right).
\end{aligned}
$$

Taking limit in above equality, and invoking statement (i) and (a2) in Assumption 4.1, we see that

$$
\begin{aligned}
\limsup_{i\to\infty} \Theta_{\tau,\beta}(L^{k_i+1}, S^{k_i+1}, Z^{k_i}, \Lambda^{k_i}) \leq{}& \Psi(L^*) + \Phi(S^*) + \frac{1}{2}\|D - \mathcal{A}(Z^*)\|_F^2 \\
&- \langle \Lambda^*,\ \mathcal{B}(L^*) + \mathcal{C}(S^*) - Z^* \rangle + (1 + 2\theta(\tau))\frac{\beta}{2}\|\mathcal{B}(L^*) + \mathcal{C}(S^*) - Z^*\|_F^2.
\end{aligned}
\tag{4.26}
$$

Then, combining (4.25) and (4.26), we see that

$$
\begin{aligned}
\lim_{i\to\infty} \Theta_{\tau,\beta}(L^{k_i+1}, S^{k_i+1}, Z^{k_i}, \Lambda^{k_i}) ={}& \Psi(L^*) + \Phi(S^*) + \frac{1}{2}\|D - \mathcal{A}(Z^*)\|_F^2 \\
&- \langle \Lambda^*,\ \mathcal{B}(L^*) + \mathcal{C}(S^*) - Z^* \rangle + (1 + 2\theta(\tau))\frac{\beta}{2}\|\mathcal{B}(L^*) + \mathcal{C}(S^*) - Z^*\|_F^2,
\end{aligned}
$$

which, together with (a2) in Assumption 4.1, $L^{k+1} - L^k \to 0$, $S^{k+1} - S^k \to 0$ and the definition of $\Theta_{\tau,\beta}$, implies that

$$\lim_{i\to\infty} \Phi(S^{k_i+1}) = \Phi(S^*). \tag{4.27}$$

Thus, passing to the limit in (4.1a)-(4.1d) along $\{(L^{k_i}, S^{k_i}, Z^{k_i}, \Lambda^{k_i})\}$ and invoking statement (i), (4.27) and (2.1), we see that

$$
\begin{cases}
0 \in \partial\Psi(L^*) - \mathcal{B}^*(\Lambda^*) + \beta\mathcal{B}^*\left(\mathcal{B}(L^*) + \mathcal{C}(S^*) - Z^*\right), \\
0 \in \partial\Phi(S^*) - \mathcal{C}^*(\Lambda^*) + \beta\mathcal{C}^*\left(\mathcal{B}(L^*) + \mathcal{C}(S^*) - Z^*\right), \\
0 = \mathcal{A}^*(\mathcal{A}(Z^*) - D) + \Lambda^* - \beta(\mathcal{B}(L^*) + \mathcal{C}(S^*) - Z^*), \\
\mathcal{B}(L^*) + \mathcal{C}(S^*) = Z^*.
\end{cases}
\tag{4.28}
$$

Rearranging terms in (4.28), it is not hard to obtain

$$
\begin{cases}
0 \in \partial\Psi(L^*) + \mathcal{B}^*\mathcal{A}^*\left(\mathcal{A}(\mathcal{B}(L^*) + \mathcal{C}(S^*)) - D\right), \\
0 \in \partial\Phi(S^*) + \mathcal{C}^*\mathcal{A}^*\left(\mathcal{A}(\mathcal{B}(L^*) + \mathcal{C}(S^*)) - D\right).
\end{cases}
$$

This shows that $(L^*, S^*, Z^*, \Lambda^*)$ is a stationary point of (1.1). This completes the proof. □

REMARK 4.2 (**Comments on the computable threshold**). *From the above discussions, we establish under Assumption 4.1 the convergence of the ADMM with $0 < \tau < \frac{1+\sqrt{5}}{2}$ when the penalty parameter β is chosen above a computable threshold $\bar{\beta}$ which depends on τ. The existence of this kind of threshold is also obtained in the recent studies [1, 26, 31, 44, 45] on the nonconvex ADMM and its variants with $\tau = 1$. In Fig. 4.1, we plot $\bar{\beta}$ against τ with $\mathcal{A}$ being the identity map (hence, $\lambda_{\max} = \lambda_{\min} = 1$). It is not hard to see from Fig. 4.1 that for a given penalty parameter $\beta > 1$, we can always choose a dual step-size τ from an interval containing 1 so that the corresponding ADMM is convergent.*
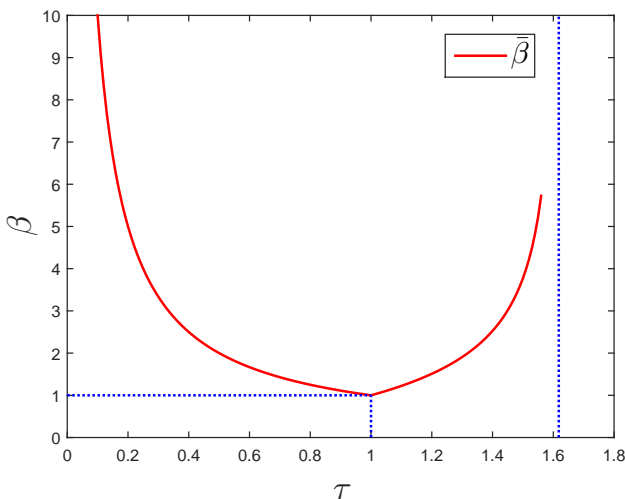

FIG. 4.1. *The computable threshold $\bar{\beta}$ for $0 < \tau < \frac{1+\sqrt{5}}{2}$.*

REMARK 4.3 (**Practical computation consideration on penalty parameter**). *In computation, for a $0 < \tau < \frac{1+\sqrt{5}}{2}$, the $\bar{\beta}$ in (4.12) may be too large and hence fixing a β close to it can lead to slow convergence. As in [32, 43], one could possibly accelerate the algorithm by initializing the algorithm with a small β (less than $\bar{\beta}$) and then increasing the β by a constant ratio until $\beta > \bar{\beta}$ if the sequence generated becomes unbounded or the successive change does not vanish sufficiently fast. Clearly, after at most finitely many increases, the penalty parameter β gets above the threshold $\bar{\beta}$ and the convergence of the resulting algorithm is guaranteed by Theorem 4.3 under Assumption 4.1. On the other hand, if β is never increased, this means that the successive change goes to zero and the sequence is bounded. Then it is routine to show that any cluster point is a stationary point if Φ is continuous in its domain.*

Under the additional assumption that the potential function $\Theta_{\tau,\beta}$ is a KL function, we show in the next theorem that the whole sequence generated by Algorithm 1 is convergent if β is greater than a computable threshold, again under Assumption 4.1. Our proof makes use of the uniformized KL property; see Proposition 2.2. This technique was previously used in [5] to prove the convergence of the proximal alternating linearized minimization algorithm for nonconvex and nonsmooth problems, and later

in [44, 45] to prove the global convergence of the Bregman ADMM with $\tau = 1$. Our analysis, though follows a similar line of arguments as in [44, 45], is much more intricate. This is because when $\tau \neq 1$, the successive change in the dual variable cannot be controlled solely by the successive changes in the primal variables.

THEOREM 4.4 (**Global convergence of the whole sequence**). *Let* $0 < \tau < \frac{1+\sqrt{5}}{2}$ *and* $\beta > \bar{\beta}$. *Suppose in addition that Assumption 4.1 holds and the potential function* $\Theta_{\tau,\beta}(\cdot)$ *is a KL function. Then, the sequence* $\{(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^{\infty}$ *generated by Algorithm 1 converges to a stationary point of* (1.1).

Proof. In view of Theorem 4.3, we only need to show that the sequence is convergent. We start by noting from (4.18), (4.19) and (4.20) that $\{\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^{\infty}$ is bounded below. Since this sequence is also decreasing from Theorem 4.1, we conclude that $\lim_{k\to\infty} \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) =: \theta^*$ exists. In the following, we will consider two cases.

Case 1) Suppose first that $\Theta_{\tau,\beta}(L^N, S^N, Z^N, \Lambda^N) = \theta^*$ for some $N \geq 1$. Since $\{\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^{\infty}$ is decreasing, we must have $\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) = \theta^*$ for all $k \geq N$. Then, it follows from (4.3) that $L^{N+t} = L^N$ and $Z^{N+t} = Z^N$ for all $t \geq 0$. Hence, $\{L^k\}$ and $\{Z^k\}$ converge finitely. Moreover, from (4.5), we have

$$\|\Lambda^{k+1} - \Lambda^k\|_F = |1 - \tau| \cdot \|\Lambda^k - \Lambda^{k-1}\|_F = \cdots = |1 - \tau|^{k+1-N} \cdot \|\Lambda^N - \Lambda^{N-1}\|_F$$

for all $k \geq N$. Since $0 < \tau < \frac{1+\sqrt{5}}{2}$, we have $0 < 1 - |1 - \tau| \leq 1$ and hence we see further that

$$\sum_{k=N}^{\infty} \|\Lambda^{k+1} - \Lambda^k\|_F \leq \frac{1}{1 - |1 - \tau|} \|\Lambda^N - \Lambda^{N-1}\|_F < \infty, \qquad (4.29)$$

which implies the convergence of $\{\Lambda^k\}$. Additionally, for all $k \geq N$, we have

$$
\begin{aligned}
\|S^{k+1} - S^k\|_F &\leq \frac{1}{\sqrt{\sigma'}} \|\mathcal{C}(S^{k+1}) - \mathcal{C}(S^k)\|_F \\
&= \frac{1}{\sqrt{\sigma'}} \left\| \frac{1}{\tau\beta}(\Lambda^k - \Lambda^{k+1}) - \frac{1}{\tau\beta}(\Lambda^{k-1} - \Lambda^k) \right\|_F \\
&\leq \frac{1}{\tau\beta\sqrt{\sigma'}} \|\Lambda^{k+1} - \Lambda^k\|_F + \frac{1}{\tau\beta\sqrt{\sigma'}} \|\Lambda^k - \Lambda^{k-1}\|_F,
\end{aligned}
$$

where the first inequality follows from (a1) in Assumption 4.1 and the equality follows from (4.1d). This together with (4.29), implies that $\sum_{k=N}^{\infty} \|S^{k+1} - S^k\|_F < \infty$. Thus, $\{S^k\}$ is also convergent. Consequently, we see that $\{(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^{\infty}$ is a convergent sequence in this case.

Case 2) From now on, we consider the case where $\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) > \theta^*$ for all $k \geq 1$. In this case, we will divide the proof into three steps: **1.** we first prove that $\Theta_{\tau,\beta}$ is constant on the set of cluster points of the sequence $\{(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^{\infty}$ and then apply the uniformized KL property; **2.** we bound the distance from 0 to $\partial\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k)$; **3.** we show that the sequence $\{(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^{\infty}$ is a Cauchy sequence and hence is convergent. The complete proof is presented as follows.

Step 1. We recall from Proposition 4.2 that the sequence $\{(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^{\infty}$ generated by Algorithm 1 is bounded and hence must have at least one cluster point. Let Γ denote the set of cluster points of $\{(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^{\infty}$. We will show that $\Theta_{\tau,\beta}$ is constant on Γ.

To this end, take any $(L^*, S^*, Z^*, \Lambda^*) \in \Gamma$ and consider a convergent subsequence $\{(L^{k_i}, S^{k_i}, Z^{k_i}, \Lambda^{k_i})\}$ with $\lim_{i \to \infty}(L^{k_i}, S^{k_i}, Z^{k_i}, \Lambda^{k_i}) = (L^*, S^*, Z^*, \Lambda^*)$. Then from the lower semicontinuity of $\Theta_{\tau,\beta}$ (since Ψ and Φ are lower semicontinuous) and the definition of θ^*, we have

$$\theta^* = \lim_{i \to \infty} \Theta_{\tau,\beta}(L^{k_i}, S^{k_i}, Z^{k_i}, \Lambda^{k_i}) \geq \Theta_{\tau,\beta}(L^*, S^*, Z^*, \Lambda^*). \qquad (4.30)$$

On the other hand, notice from the definition of S^{k+1} as a minimizer in (3.2b) that

$$\begin{aligned}
&\Theta_{\tau,\beta}(L^{k_i}, S^{k_i}, Z^{k_i-1}, \Lambda^{k_i-1}) - \Theta_{\tau,\beta}(L^{k_i}, S^*, Z^{k_i-1}, \Lambda^{k_i-1}) \\
&= \mathcal{L}_\beta(L^{k_i}, S^{k_i}, Z^{k_i-1}, \Lambda^{k_i-1}) - \mathcal{L}_\beta(L^{k_i}, S^*, Z^{k_i-1}, \Lambda^{k_i-1}) \\
&\quad + \theta(\tau)\beta\left(\|\mathcal{B}(L^{k_i}) + \mathcal{C}(S^{k_i}) - Z^{k_i-1}\|_F^2 - \|\mathcal{B}(L^{k_i}) + \mathcal{C}(S^*) - Z^{k_i-1}\|_F^2\right) \\
&\leq \theta(\tau)\beta\left(\|\mathcal{B}(L^{k_i}) + \mathcal{C}(S^{k_i}) - Z^{k_i-1}\|_F^2 - \|\mathcal{B}(L^{k_i}) + \mathcal{C}(S^*) - Z^{k_i-1}\|_F^2\right).
\end{aligned}$$

This together with Theorem 4.3(i), the continuity of $\Theta_{\tau,\beta}$ with respect to L (from (a2) in Assumption 4.1), Z and Λ; and the definition of θ^* implies that

$$\theta^* = \lim_{i \to \infty} \Theta_{\tau,\beta}(L^{k_i}, S^{k_i}, Z^{k_i}, \Lambda^{k_i}) \leq \Theta_{\tau,\beta}(L^*, S^*, Z^*, \Lambda^*). \qquad (4.31)$$

Combining (4.30) and (4.31), we conclude that $\Theta_{\tau,\beta}(L^*, S^*, Z^*, \Lambda^*) = \theta^*$. Since $(L^*, S^*, Z^*, \Lambda^*) \in \Gamma$ is arbitrary, we conclude further that the potential function $\Theta_{\tau,\beta}$ is constant on Γ.

The fact that $\Theta_{\tau,\beta} \equiv \theta^*$ on Γ together with our assumption that $\Theta_{\tau,\beta}(\cdot)$ is a KL function and Proposition 2.2 implies that there exist $\varepsilon > 0$, $\eta > 0$ and $\varphi \in \Xi_\eta$, such that

$$\varphi'\left(\Theta_{\tau,\beta}(L, S, Z, \Lambda) - \theta^*\right) \operatorname{dist}\left(0,\ \partial\Theta_{\tau,\beta}(L, S, Z, \Lambda)\right) \geq 1$$

for all (L, S, Z, Λ) satisfying $\operatorname{dist}((L, S, Z, \Lambda), \Gamma) < \varepsilon$ and $\theta^* < \Theta_{\tau,\beta}(L, S, Z, \Lambda) < \theta^* + \eta$. On the other hand, since $\lim_{k \to \infty} \operatorname{dist}((L^k, S^k, Z^k, \Lambda^k), \Gamma) = 0$ by the definition of Γ, and $\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) \to \theta^*$, then for such ε and η, there exists $k_1 \geq 3$ such that $\operatorname{dist}((L^k, S^k, Z^k, \Lambda^k), \Gamma) < \varepsilon$ and $\theta^* < \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) < \theta^* + \eta$ for all $k \geq k_1$. Thus, for $k \geq k_1$, we have

$$\varphi'\left(\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) - \theta^*\right) \operatorname{dist}\left(0,\ \partial\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k)\right) \geq 1. \qquad (4.32)$$

Step 2. We next consider the subdifferential $\partial\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k)$. Looking at the partial subdifferential with respect to L, we have

$$\begin{aligned}
&\partial_L \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) \\
&= \partial\Psi(L^k) - \mathcal{B}^*(\Lambda^k) + (1 + 2\theta(\tau))\beta\mathcal{B}^*(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k) \\
&= \partial\Psi(L^k) - \mathcal{B}^*(\Lambda^{k-1}) + \beta\mathcal{B}^*(\mathcal{B}(L^k) + \mathcal{C}(S^{k-1}) - Z^{k-1}) + 2\theta(\tau)\beta\mathcal{B}^*(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k) \\
&\quad - \mathcal{B}^*(\Lambda^k - \Lambda^{k-1}) + \beta\mathcal{B}^*(\mathcal{C}(S^k) - Z^k - \mathcal{C}(S^{k-1}) + Z^{k-1}) \\
&\ni 2\theta(\tau)\beta\mathcal{B}^*(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k) - \mathcal{B}^*(\Lambda^k - \Lambda^{k-1}) + \beta\mathcal{B}^*(\mathcal{C}(S^k) - Z^k - \mathcal{C}(S^{k-1}) + Z^{k-1}) \\
&\overset{\text{(i)}}{=} -\left(1 + \tfrac{2\theta(\tau)}{\tau}\right)\mathcal{B}^*(\Lambda^k - \Lambda^{k-1}) + \beta\mathcal{B}^*[(\mathcal{C}(S^k) - Z^k) - (\mathcal{C}(S^{k-1}) - Z^{k-1})] \\
&\overset{\text{(ii)}}{=} -\left(1 + \tfrac{2\theta(\tau)}{\tau}\right)\mathcal{B}^*(\Lambda^k - \Lambda^{k-1}) + \beta\mathcal{B}^*\left[\left(-\mathcal{B}(L^k) - \tfrac{\Lambda^k - \Lambda^{k-1}}{\tau\beta}\right) - \left(-\mathcal{B}(L^{k-1}) - \tfrac{\Lambda^{k-1} - \Lambda^{k-2}}{\tau\beta}\right)\right] \\
&= -\left(1 + \tfrac{2\theta(\tau)+1}{\tau}\right)\mathcal{B}^*(\Lambda^k - \Lambda^{k-1}) + \tfrac{1}{\tau}\mathcal{B}^*(\Lambda^{k-1} - \Lambda^{k-2}) - \beta\mathcal{B}^*\mathcal{B}(L^k - L^{k-1}),
\end{aligned}$$

where the inclusion follows from (4.1a), and the equalities (i) and (ii) follow from (4.1d). Similarly,

$$
\begin{aligned}
&\partial_S \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) \\
&= \partial\Phi(S^k) - \mathcal{C}^*(\Lambda^k) + (1 + 2\theta(\tau))\beta\mathcal{C}^*(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k) \\
&= \partial\Phi(S^k) - \mathcal{C}^*(\Lambda^{k-1}) + \beta\mathcal{C}^*(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^{k-1}) \\
&\quad + 2\theta(\tau)\beta\mathcal{C}^*(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k) - \mathcal{C}^*(\Lambda^k - \Lambda^{k-1}) - \beta\mathcal{C}^*(Z^k - Z^{k-1}) \\
&\ni 2\theta(\tau)\beta\mathcal{C}^*(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k) - \mathcal{C}^*(\Lambda^k - \Lambda^{k-1}) - \beta\mathcal{C}^*(Z^k - Z^{k-1}) \\
&= -\left(1 + \frac{2\theta(\tau)}{\tau}\right)\mathcal{C}^*(\Lambda^k - \Lambda^{k-1}) - \beta\mathcal{C}^*(Z^k - Z^{k-1}),
\end{aligned}
$$

where the inclusion follows from (4.1b) and the last equality follows from (4.1d). Moreover,

$$
\begin{aligned}
&\nabla_Z \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) \\
&= \mathcal{A}^*(\mathcal{A}(Z^k) - D) + \Lambda^k - \beta(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k) \\
&\quad - 2\theta(\tau)\beta(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k) \\
&= \mathcal{A}^*(\mathcal{A}(Z^k) - D) + \Lambda^{k-1} - \beta(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k) \\
&\quad - 2\theta(\tau)\beta(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k) + (\Lambda^k - \Lambda^{k-1}) \\
&= -2\theta(\tau)\beta(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k) + (\Lambda^k - \Lambda^{k-1}) \\
&= \left(1 + \frac{2\theta(\tau)}{\tau}\right)(\Lambda^k - \Lambda^{k-1}),
\end{aligned}
$$

where the third equality follows from (4.1c) and the last equality follows from (4.1d). Finally,

$$
\nabla_\lambda \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) = -(\mathcal{B}(L^k) + \mathcal{C}(S^k) - Z^k) = \frac{1}{\tau\beta}(\Lambda^k - \Lambda^{k-1}),
$$

where the last equality follows from (4.1d). Thus, from the above relations, there exists $a > 0$ so that

$$
\begin{aligned}
&\mathrm{dist}\left(0, \ \partial\Theta_{\tau,\beta}\left(L^k, S^k, Z^k, \Lambda^k\right)\right) \\
&\leq a\left(\|L^k - L^{k-1}\|_F + \|Z^k - Z^{k-1}\|_F + \|\Lambda^k - \Lambda^{k-1}\|_F + \|\Lambda^{k-1} - \Lambda^{k-2}\|_F\right).
\end{aligned} \tag{4.33}
$$

Step 3. We now prove the convergence of the sequence by combining (4.33) with (4.32). For notational simplicity, define

$$
\Delta^k := \varphi\left(\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) - \theta^*\right) - \varphi\left(\Theta_{\tau,\beta}(L^{k+1}, S^{k+1}, Z^{k+1}, \Lambda^{k+1}) - \theta^*\right).
$$

Since $\Theta_{\tau,\beta}$ is decreasing and φ is monotonic, it is easy to see $\Delta^k \geq 0$ for $k \geq 1$. Then

we have for all $k \geq k_1$ that

$$
\begin{aligned}
a &\left(\|L^k - L^{k-1}\|_F + \|Z^k - Z^{k-1}\|_F + \|\Lambda^k - \Lambda^{k-1}\|_F + \|\Lambda^{k-1} - \Lambda^{k-2}\|_F \right) \cdot \Delta^k \\
&\geq \operatorname{dist}(0, \ \partial \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k)) \cdot \Delta^k \\
&\geq \operatorname{dist}(0, \ \partial \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k)) \cdot \varphi' \left(\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) - \theta^* \right) \\
&\quad \cdot \left[\Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) - \Theta_{\tau,\beta}(L^{k+1}, S^{k+1}, Z^{k+1}, \Lambda^{k+1}) \right] \\
&\geq \Theta_{\tau,\beta}(L^k, S^k, Z^k, \Lambda^k) - \Theta_{\tau,\beta}(L^{k+1}, S^{k+1}, Z^{k+1}, \Lambda^{k+1}) \\
&\geq b_1 \|L^{k+1} - L^k\|_F^2 + b_2 \|Z^{k+1} - Z^k\|_F^2 \\
&\geq \tfrac{1}{2} \min\{b_1, b_2\} \cdot \left[\|L^{k+1} - L^k\|_F + \|Z^{k+1} - Z^k\|_F \right]^2,
\end{aligned} \tag{4.34}
$$

where the first inequality follows from (4.33), the second inequality follows from the concavity of φ, the third inequality follows from (4.32), the fourth inequality follows from (4.3) with $b_1 := \frac{\sigma \beta}{2}$ and $b_2 := \frac{\lambda_{\min} + \beta}{2} - \max \left\{ \frac{1}{\tau}, \frac{\tau^2}{1 + \tau - \tau^2} \right\} \cdot \frac{\lambda_{\max}^2}{\beta}$.

Dividing both sides of (4.34) by $c := \frac{1}{2} \min\{b_1, b_2\}$, taking the square root and using the inequality $\sqrt{uv} \leq \frac{u+v}{2}$ for $u, v \geq 0$ to further upper bound the left hand side of the resulting inequality, we obtain that

$$
\begin{aligned}
&\tfrac{1}{2\gamma} \left(\|L^k - L^{k-1}\|_F + \|Z^k - Z^{k-1}\|_F + \|\Lambda^k - \Lambda^{k-1}\|_F + \|\Lambda^{k-1} - \Lambda^{k-2}\|_F \right) + \tfrac{\gamma a}{2c} \Delta^k \\
&\geq \|L^{k+1} - L^k\|_F + \|Z^{k+1} - Z^k\|_F,
\end{aligned} \tag{4.35}
$$

where γ is an arbitrary positive constant. On the other hand, it follows from (4.5) that

$$
\begin{aligned}
\|\Lambda^k - \Lambda^{k-1}\|_F &= \|\tau \mathcal{A}^* \mathcal{A}(Z^{k-1} - Z^k) + (1 - \tau)(\Lambda^{k-1} - \Lambda^{k-2})\|_F \\
&\leq \tau \lambda_{\max} \|Z^k - Z^{k-1}\|_F + |1 - \tau| \cdot \|\Lambda^{k-1} - \Lambda^{k-2}\|_F.
\end{aligned}
$$

Adding $-|1 - \tau| \cdot \|\Lambda^k - \Lambda^{k-1}\|_F$ to both sides of the above inequality and simplifying the resulting inequality, we obtain that

$$
\begin{aligned}
&\|\Lambda^k - \Lambda^{k-1}\|_F \\
&\leq \tfrac{\tau \lambda_{\max}}{1 - |1 - \tau|} \|Z^k - Z^{k-1}\|_F + \tfrac{|1-\tau|}{1 - |1 - \tau|} \left(\|\Lambda^{k-1} - \Lambda^{k-2}\|_F - \|\Lambda^k - \Lambda^{k-1}\|_F \right) \\
&= d_1 \|Z^k - Z^{k-1}\|_F + d_2 \left(\|\Lambda^{k-1} - \Lambda^{k-2}\|_F - \|\Lambda^k - \Lambda^{k-1}\|_F \right),
\end{aligned} \tag{4.36}
$$

where we write $d_1 := \frac{\tau \lambda_{\max}}{1 - |1 - \tau|}$ and $d_2 := \frac{|1 - \tau|}{1 - |1 - \tau|}$ for notational simplicity. Similarly,

$$
\begin{aligned}
\|\Lambda^{k-1} - \Lambda^{k-2}\|_F &\leq d_1 \|Z^{k-1} - Z^{k-2}\|_F \\
&\quad + d_2 \left(\|\Lambda^{k-2} - \Lambda^{k-3}\|_F - \|\Lambda^{k-1} - \Lambda^{k-2}\|_F \right).
\end{aligned} \tag{4.37}
$$

Then substituting (4.36) and (4.37) into (4.35) and rearranging terms, we have

$$
\begin{aligned}
&\left(1 - \tfrac{1}{2\gamma} \right) \|L^{k+1} - L^k\|_F + \left(1 - \tfrac{1}{2\gamma} - \tfrac{d_1}{\gamma} \right) \|Z^{k+1} - Z^k\|_F \\
&\leq \tfrac{1}{2\gamma} \left(\|L^k - L^{k-1}\|_F - \|L^{k+1} - L^k\|_F \right) \\
&\quad + \left(\tfrac{1}{2\gamma} + \tfrac{d_1}{\gamma} \right) \left(\|Z^k - Z^{k-1}\|_F - \|Z^{k+1} - Z^k\|_F \right) \\
&\quad + \tfrac{d_1}{2\gamma} \left(\|Z^{k-1} - Z^{k-2}\|_F - \|Z^k - Z^{k-1}\|_F \right) \\
&\quad + \tfrac{d_2}{2\gamma} \left(\|\Lambda^{k-1} - \Lambda^{k-2}\|_F - \|\Lambda^k - \Lambda^{k-1}\|_F \right) \\
&\quad + \tfrac{d_2}{2\gamma} \left(\|\Lambda^{k-2} - \Lambda^{k-3}\|_F - \|\Lambda^{k-1} - \Lambda^{k-2}\|_F \right) + \tfrac{\gamma a}{2c} \Delta^k.
\end{aligned} \tag{4.38}
$$

Thus, summing (4.38) from $k = k_1$ to ∞, we have

$$\left(1 - \tfrac{1}{2\gamma}\right)\sum_{k=k_1}^{\infty}\|L^{k+1} - L^k\|_F + \left(1 - \tfrac{1}{2\gamma} - \tfrac{d_1}{\gamma}\right)\sum_{k=k_1}^{\infty}\|Z^{k+1} - Z^k\|_F$$

$$\leq \tfrac{1}{2\gamma}\|L^{k_1} - L^{k_1-1}\|_F + \left(\tfrac{1}{2\gamma} + \tfrac{d_1}{\gamma}\right)\|Z^{k_1} - Z^{k_1-1}\|_F + \tfrac{d_1}{2\gamma}\|Z^{k_1-1} - Z^{k_1-2}\|_F$$

$$+ \tfrac{d_2}{2\gamma}\|\Lambda^{k_1-1} - \Lambda^{k_1-2}\|_F + \tfrac{d_2}{2\gamma}\|\Lambda^{k_1-2} - \Lambda^{k_1-3}\|_F + \tfrac{a\gamma}{2c}\,\varphi\left(\Theta_{\tau,\beta}(L^{k_1}, S^{k_1}, Z^{k_1}, \Lambda^{k_1}) - \theta^*\right)$$

$$< \infty.$$

Recall that γ introduced in (4.35) is an arbitrary positive constant. Taking $\gamma > \frac{1+2d_1}{2}$ and hence $1 - \frac{1}{2\gamma} > 1 - \frac{1}{2\gamma} - \frac{d_1}{\gamma} > 0$, we have from the above inequality that

$$\sum_{k=k_1}^{\infty}\|L^{k+1} - L^k\|_F < \infty \quad \text{and} \quad \sum_{k=k_1}^{\infty}\|Z^{k+1} - Z^k\|_F < \infty.$$

Hence $\{L^k\}$ and $\{Z^k\}$ are convergent. Additionally, summing (4.36) from $k = k_1$ to ∞, we have

$$\sum_{k=k_1}^{\infty}\|\Lambda^k - \Lambda^{k-1}\|_F \leq d_1 \sum_{k=k_1}^{\infty}\|Z^k - Z^{k-1}\|_F + d_2\|\Lambda^{k_1-1} - \Lambda^{k_1-2}\|_F < \infty,$$

which implies that $\{\Lambda^k\}$ is convergent. Finally, from (4.1d) and (a1) in Assumption 4.1, we see that $\{S^k\}$ is also convergent. Consequently, we conclude that $\{(L^k, S^k, Z^k, \Lambda^k)\}_{k=1}^{\infty}$ is a convergent sequence. This completes the proof. $\square$

Our convergence analysis relies on Assumption 4.1. While (a3) in Assumption 4.1 appears restrictive since it makes assumptions on the first iterate of Algorithm 1, we show below that this assumption would hold upon a suitable choice of initialization. Specifically, if we initialize at $(L^0, S^0, Z^0, \Lambda^0)$ satisfying

$$\begin{cases} \Theta_{\tau,\beta}(L^1, S^1, Z^1, \Lambda^1) \leq \Theta_{\tau,\beta}(L^0, S^0, Z^0, \Lambda^0), & \text{(4.39a)} \\ \Theta_{\tau,\beta}(L^0, S^0, Z^0, \Lambda^0) < h_0, & \text{(4.39b)} \end{cases}$$

then it is easy to check that (a3) in Assumption 4.1 holds. In the next proposition, we demonstrate that (4.39a) can always be satisfied with a suitable initialization. After this, we will propose a specific way to initialize Algorithm 1 for a wide range of problems so that both (4.39a) and (4.39b) are satisfied.

PROPOSITION 4.5. *Suppose that* $0 < \tau < \frac{1+\sqrt{5}}{2}$ *and* $\beta > \bar{\beta}$. *If the initialization* $(L^0, S^0, Z^0, \Lambda^0)$ *is chosen as* $(L^0, S^0) \in \operatorname{dom}\Psi \times \operatorname{dom}\Phi$ *and*

$$\Lambda^0 = \mathcal{A}^*(D - \mathcal{A}(Z^0)), \tag{4.40}$$

then we have

$$\Theta_{\tau,\beta}(L^1, S^1, Z^1, \Lambda^1) \leq \Theta_{\tau,\beta}(L^0, S^0, Z^0, \Lambda^0).$$

Proof. First, from (4.1c), we have

$$0 = \mathcal{A}^*(\mathcal{A}(Z^1) - D) + \Lambda^0 - \beta(\mathcal{B}(L^1) + \mathcal{C}(S^1) - Z^1)$$

$$\implies \mathcal{B}(L^1) + \mathcal{C}(S^1) - Z^1 = \frac{1}{\beta}\Lambda^0 + \frac{1}{\beta}\mathcal{A}^*(\mathcal{A}(Z^1) - D) = \frac{1}{\beta}\mathcal{A}^*\mathcal{A}(Z^1 - Z^0), \tag{4.41}$$

where the last equality follows from (4.40). Then,

$$
\begin{aligned}
&\Theta_{\tau,\beta}(L^1, S^1, Z^1, \Lambda^1) - \Theta_{\tau,\beta}(L^1, S^1, Z^1, \Lambda^0) \\
&= -\langle \Lambda^1 - \Lambda^0, \mathcal{B}(L^1) + \mathcal{C}(S^1) - Z^1 \rangle = \tau\beta \|\mathcal{B}(L^1) + \mathcal{C}(S^1) - Z^1\|_F^2 \\
&= (\tau + \theta(\tau))\beta\|\mathcal{B}(L^1) + \mathcal{C}(S^1) - Z^1\|_F^2 - \theta(\tau)\beta\|\mathcal{B}(L^1) + \mathcal{C}(S^1) - Z^1\|_F^2 \\
&= (\tau + \theta(\tau))\beta \left\| \tfrac{1}{\beta}\mathcal{A}^*\mathcal{A}(Z^1 - Z^0) \right\|_F^2 - \theta(\tau)\beta\|\mathcal{B}(L^1) + \mathcal{C}(S^1) - Z^1\|_F^2 \\
&\leq (\tau + \theta(\tau))\frac{\lambda_{\max}^2}{\beta}\|Z^1 - Z^0\|_F^2 - \theta(\tau)\beta\|\mathcal{B}(L^1) + \mathcal{C}(S^1) - Z^1\|_F^2,
\end{aligned}
\tag{4.42}
$$

where the second equality follows from (4.1d) and the fourth equality follows from (4.41). Additionally, using the same arguments as in the proof of Lemma 4.1 leading to (4.9), (4.10) and (4.11), it is easy to see that

$$
\begin{aligned}
\Theta_{\tau,\beta}(L^1, S^1, Z^1, \Lambda^0) - \Theta_{\tau,\beta}(L^1, S^1, Z^0, \Lambda^0) &\leq -\tfrac{\lambda_{\min}+\beta}{2}\|Z^1 - Z^0\|_F^2 \\
+ \theta(\tau)\beta \left(\|\mathcal{B}(L^1) + \mathcal{C}(S^1) - Z^1\|_F^2 - \|\mathcal{B}(L^1) + \mathcal{C}(S^1) - Z^0\|_F^2 \right), &
\end{aligned}
\tag{4.43}
$$

$$
\begin{aligned}
\Theta_{\tau,\beta}(L^1, S^1, Z^0, \Lambda^0) - \Theta_{\tau,\beta}(L^1, S^0, Z^0, \Lambda^0) & \\
\leq \theta(\tau)\beta \left(\|\mathcal{B}(L^1) + \mathcal{C}(S^1) - Z^0\|_F^2 - \|\mathcal{B}(L^1) + \mathcal{C}(S^0) - Z^0\|_F^2 \right), &
\end{aligned}
\tag{4.44}
$$

$$
\begin{aligned}
\Theta_{\tau,\beta}(L^1, S^0, Z^0, \Lambda^0) - \Theta_{\tau,\beta}(L^0, S^0, Z^0, \Lambda^0) & \\
\leq \theta(\tau)\beta \left(\|\mathcal{B}(L^1) + \mathcal{C}(S^0) - Z^0\|_F^2 - \|\mathcal{B}(L^0) + \mathcal{C}(S^0) - Z^0\|_F^2 \right). &
\end{aligned}
\tag{4.45}
$$

Summing (4.42), (4.43), (4.44) and (4.45), we obtain

$$
\begin{aligned}
&\Theta_{\tau,\beta}(L^1, S^1, Z^1, \Lambda^1) - \Theta_{\tau,\beta}(L^0, S^0, Z^0, \Lambda^0) \\
&\leq \left((\tau + \theta(\tau))\tfrac{\lambda_{\max}^2}{\beta} - \tfrac{\lambda_{\min}+\beta}{2} \right)\|Z^1 - Z^0\|_F^2 - \theta(\tau)\beta\|\mathcal{B}(L^0) + \mathcal{C}(S^0) - Z^0\|_F^2.
\end{aligned}
\tag{4.46}
$$

We now consider two cases:

- For $0 < \tau \leq 1$, it is easy to see $\theta(\tau) = 1 - \tau$ and

$$
\beta > \max\left\{ \frac{\lambda_{\max}}{\tau}, -\frac{\lambda_{\min}}{2} + \frac{1}{2}\sqrt{\lambda_{\min}^2 + \frac{8}{\tau}\lambda_{\max}^2} \right\}.
$$

Then, we have

$$
(\tau + \theta(\tau))\frac{\lambda_{\max}^2}{\beta} - \frac{\lambda_{\min} + \beta}{2} = \frac{\lambda_{\max}^2}{\beta} - \frac{\lambda_{\min} + \beta}{2} \leq \frac{\lambda_{\max}^2}{\tau\beta} - \frac{\lambda_{\min} + \beta}{2} < 0.
$$

- For $1 < \tau < \frac{1+\sqrt{5}}{2}$, it is easy to see $\theta(\tau) = \frac{(\tau-1)\tau^2}{1+\tau-\tau^2}$ and

$$
\beta > \max\left\{ \tau\lambda_{\max}, -\frac{\lambda_{\min}}{2} + \frac{1}{2}\sqrt{\lambda_{\min}^2 + \frac{8\tau^2}{1+\tau-\tau^2}\lambda_{\max}^2} \right\}.
$$

Then, we have

$$
\begin{aligned}
(\tau + \theta(\tau))\frac{\lambda_{\max}^2}{\beta} - \frac{\lambda_{\min} + \beta}{2} &= \frac{\tau\lambda_{\max}^2}{(1+\tau-\tau^2)\beta} - \frac{\lambda_{\min} + \beta}{2} \\
&< \frac{\tau^2\lambda_{\max}^2}{(1+\tau-\tau^2)\beta} - \frac{\lambda_{\min} + \beta}{2} < 0.
\end{aligned}
$$

Thus, combining the above with (4.46) and $\theta(\tau) \geq 0$, we conclude that

$$\Theta_{\tau,\beta}(L^1, S^1, Z^1, \Lambda^1) \leq \Theta_{\tau,\beta}(L^0, S^0, Z^0, \Lambda^0).$$

This completes the proof.   □

From Proposition 4.5, we see that if the initialization $(L^0, S^0, Z^0, \Lambda^0)$ is chosen to satisfy the conditions in Proposition 4.5, then (4.39a) holds. Based on this, we can now present one specific way to initialize Algorithm 1 so that both (4.39a) and (4.39b) are satisfied for a class of problems, whose objective functions $\Psi(L)$ and $\Phi(S)$ take forms $\delta_\Omega(L)$ and (1.2), respectively; here, Ω is a compact convex set.

The initialization we consider is:

$$L^0 = \mathcal{P}_\Omega(\kappa D), \ S^0 = 0, \ Z^0 = \mathcal{B}(L^0), \ \Lambda^0 = \mathcal{A}^* \left(D - \mathcal{A}(Z^0)\right), \qquad (4.47)$$

where κ is a scaling parameter. One can easily check that this initialization satisfies (4.40). Moreover,

$$\Theta_{\tau,\beta}(L^0, S^0, Z^0, \Lambda^0) = \frac{1}{2} \left\| D - \mathcal{A}\left(Z^0\right) \right\|_F^2 = \frac{1}{2} \left\| D - \mathcal{A}\left(\mathcal{B}(\mathcal{P}_\Omega(\kappa D))\right) \right\|_F^2 .$$

Thus, the condition (4.39b) is equivalent to

$$\frac{1}{2} \left\| D - \mathcal{A}\left(\mathcal{B}(\mathcal{P}_\Omega(\kappa D))\right) \right\|_F^2 < \liminf_{\|L\|_F + \|S\|_F \to \infty} \Psi(L) + \Phi(S) = \liminf_{\|S\|_F \to \infty} \Phi(S). \quad (4.48)$$

We further discuss this inequality for some concrete examples of Φ presented in the introduction.

EXAMPLE 4.1. *Suppose that Φ is coercive. Then $\liminf_{\|S\|_F \to \infty} \Phi(S) = \infty$ and hence* (4.48) *holds trivially for any choice of κ.*

EXAMPLE 4.2. *Suppose that $\Phi(S) = \mu \sum_{i=1}^m \sum_{j=1}^n \frac{\alpha|s_{ij}|}{1+\alpha|s_{ij}|}$ for $\alpha > 0$. Then $\liminf_{\|S\|_F \to \infty} \Phi(S) = \mu$. Hence* (4.48) *holds if the parameter κ can be chosen so that $\frac{1}{2} \left\| D - \mathcal{A}\left(\mathcal{B}(\mathcal{P}_\Omega(\kappa D))\right) \right\|_F^2 < \mu$.*

EXAMPLE 4.3. *Suppose that $\Phi(S) = \mu \sum_{i=1}^m \sum_{j=1}^n \int_0^{|s_{ij}|} \min(1, (\alpha - t/\mu)_+/(\alpha - 1))$ dt for $\alpha > 2$. Then $\liminf_{\|S\|_F \to \infty} \Phi(S) = \frac{1}{2}(\alpha + 1)\mu^2$. Hence* (4.48) *holds if κ can be chosen so that $\frac{1}{2} \left\| D - \mathcal{A}\left(\mathcal{B}(\mathcal{P}_\Omega(\kappa D))\right) \right\|_F^2 < \frac{1}{2}(\alpha + 1)\mu^2$.*

EXAMPLE 4.4. *Suppose that $\Phi(S) = \mu \sum_{i=1}^m \sum_{j=1}^n \int_0^{|s_{ij}|} (1 - t/(\alpha\mu))_+$ dt for $\alpha > 0$. Then, $\liminf_{\|S\|_F \to \infty} \Phi(S) = \frac{1}{2}\alpha\mu^2$. Hence* (4.48) *holds if κ can be chosen so that $\frac{1}{2} \left\| D - \mathcal{A}\left(\mathcal{B}(\mathcal{P}_\Omega(\kappa D))\right) \right\|_F^2 < \frac{1}{2}\alpha\mu^2$.*

EXAMPLE 4.5. *Suppose that $\Phi(S) = \mu \sum_{i=1}^m \sum_{j=1}^n \mu - (\mu - |s_{ij}|)_+^2/\mu$. Then it is not hard to show that $\liminf_{\|S\|_F \to \infty} \Phi(S) = \mu^2$. Hence* (4.48) *holds if κ can be chosen so that $\frac{1}{2} \left\| D - \mathcal{A}\left(\mathcal{B}(\mathcal{P}_\Omega(\kappa D))\right) \right\|_F^2 < \mu^2$.*

5. Numerical experiments. In this section, we conduct numerical experiments to show the performances of our algorithm. All experiments are run in Matlab R2014b on a 64-bit PC with an Intel(R) Core(TM) i7-4790 CPU (3.60GHz) and 16GB of RAM equipped with Windows 8.1 OS.

5.1. Implementation details.

Testing model. We consider the problem of extracting background/foreground from a given video under different scenarios. Specifically, we consider:

$$\min_{L,S} \ \Phi(S) + \tfrac{1}{2}\|D - \mathcal{A}(L + S)\|_F^2$$
$$\text{s.t.} \ \ L \in \Omega, \tag{5.1}$$

where $\Omega = \{L \in \mathbb{R}^{m \times n} \mid \|L\|_\infty \le 1, \ L_{:1} = L_{:2} = \cdots = L_{:n}\}$ and $\mathcal{A}$ is a linear map. This model corresponds to (1.1) with $\Psi(L) = \delta_\Omega(L)$ and $\mathcal{B} = \mathcal{C} = \mathcal{I}$. We compare the performances of the ADMM with different choices of τ, as well as the proximal alternating linearized minimization (PALM) proposed in [5], on solving (5.1). For ease of future reference, we recall that the PALM for solving (5.1) is given by

$$\begin{cases} L^{k+1} = \mathcal{P}_\Omega \left(L^k - \dfrac{1}{c_k}\mathcal{A}^*(\mathcal{A}(L^k + S^k) - D) \right), \\[2ex] S^{k+1} \in \underset{S}{\mathrm{Argmin}} \left\{ \Phi(S) + \dfrac{d_k}{2} \left\| S - S^k + \dfrac{1}{d_k}\mathcal{A}^*(\mathcal{A}(L^{k+1} + S^k) - D) \right\|_F^2 \right\}, \end{cases}$$

where c_k and d_k are positive numbers.

In our experiments, we consider the following three choices of sparse regularizers $\Phi(S)$:

- bridge regularizer: $\Phi(S) = \mu\|S\|_p^p$ for $0 < p \le 1$;
- fraction regularizer: $\Phi(S) = \mu \sum_{i=1}^m \sum_{j=1}^n \frac{\alpha|s_{ij}|}{1+\alpha|s_{ij}|}$ for $\alpha > 0$;
- logistic regularizer: $\Phi(S) = \mu \sum_{i=1}^m \sum_{j=1}^n \log(1 + \alpha|s_{ij}|)$ for $\alpha > 0$;

and two choices of linear map $\mathcal{A}$:

- $\mathcal{A}(L + S) := L + S$: in this case, model (5.1) can be applied to extracting background/foreground from a surveillance video with noise.
- $\mathcal{A}(L + S) := H(L + S)$ with $H \in \mathbb{R}^{m \times m}$ being the matrix representation of a regular blurring operator (the blurring is assumed to occur frame-wise): in this case, model (5.1) can be applied to extracting background/foreground from a blurred and noisy surveillance video.

Testing videos. We choose four real videos, "Hall", "Bootstrap", "Fountain" and "ShoppingMall", from the dataset I2R[1] provided by Li et al. [29]. The details of these videos are as follows:

- **Hall** video contains 200 144×176 frames (from airport2001 to airport2200);
- **Bootstrap** video contains 200 120×160 frames (from b01801 to b02000);
- **Fountain** video contains 200 128×160 frames (from Fountain1301 to Fountain1500);
- **ShoppingMall** video contains 200 256×320 frames (from ShoppingMall1501 to ShoppingMall1700).

We show one frame of each testing video under two different scenarios (noisy and noisy blurred), and their ground-truth images of foregrounds in Fig. 5.1. Additionally, all pixel values of the testing videos are re-scaled into $[0, 1]$ in our numerical experiments.

---

[1]This dataset is available in http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html. The authors also provide 20 ground-truth images of foregrounds for each video in this dataset.
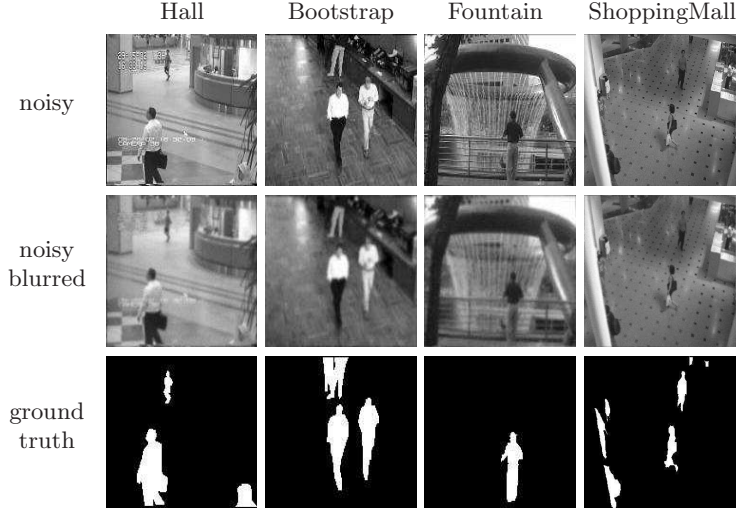
FIG. 5.1. *One frame (from left to right: airport2180, b01842, Fountain1440 and Shopping-Mall1535) of each testing video under different scenarios (the first two rows) and the ground-truth image of foreground of each testing video (the last row).*

Parameters setting. For the ADMM, we use the following heuristics[2] to update β: we initialize $n_s = 0$ and $\beta = 0.6\bar{\beta}$, where $\bar{\beta}$ is given in (4.12). In the k-th iteration, we compute

$$fnorm^k = \|L^k\|_F + \|Z^k\|_F,$$
$$succ\_chg^k = \|L^k - L^{k-1}\|_F + \|Z^k - Z^{k-1}\|_F.$$

Then, we increase n_s by 1 if $succ\_chg^k > 0.99 \cdot succ\_chg^{k-1}$. Obviously, n_s is non-decreasing in this procedure. We then update β as 1.1β whenever $\beta \leq 1.01\bar{\beta}$ and the sequence satisfies either $n_s \geq 0.3k$ or $fnorm^k > 10^{10}$. On the other hand, for PALM, we set $c_k = d_k = \frac{\lambda_{\max}}{0.99}$.

We initialize our algorithm and the PALM at the point specified in (4.47) with $\kappa = 1$. Moreover, we terminate our ADMM by the following two-stage criterion[3]: in each iteration, we check if

$$\frac{\|L^k - L^{k-1}\|_F + \|Z^k - Z^{k-1}\|_F}{\|L^k\|_F + \|Z^k\|_F + 1} < \text{Tol}_{A,1}$$

for some $\text{Tol}_{A,1} > 0$; if it holds, then we further check if

$$\frac{\|S^k - S^{k-1}\|_F + \|\Lambda^k - \Lambda^{k-1}\|_F}{\|S^k\|_F + \|\Lambda^k\|_F + 1} < \text{Tol}_{A,2}$$

---

[2]Note from Theorem 4.3(i) that the successive change of each variable goes to zero as $k \to \infty$. Thus, intuitively, it is more favorable to see a decrease in the successive change as k increases. This heuristic is designed based on this intuition.

[3]We use this two-stage criterion rather than computing the relative errors of all four variables (L, S, Z, Λ) in each iteration of our algorithm because computing matrix Frobenius norms can be expensive, especially for large scale problems. This strategy will help reduce the cost per iteration. We examine $\|L^k - L^{k-1}\|_F$ and $\|Z^k - Z^{k-1}\|_F$ in the first stage because these quantities being small intuitively implies that $\|S^k - S^{k-1}\|_F$ and $\|\Lambda^k - \Lambda^{k-1}\|_F$ are small; see the proof of Theorem 4.3, particularly (4.23), (4.24) and the discussions that follow.

for some $\text{Tol}_{A,2} > 0$. We terminate the algorithm if this latter condition is also satisfied. For the PALM, we terminate it when

$$\frac{\|L^k - L^{k-1}\|_F + \|S^k - S^{k-1}\|_F}{\|L^k\|_F + \|S^k\|_F + 1} < \text{Tol}_P$$

for some $\text{Tol}_P > 0$. The specific values of $\text{Tol}_{A,1}$, $\text{Tol}_{A,2}$ and Tol_P are given in the following experiments.

5.2. Comparisons between ADMM with different τ and PALM. In this subsection, we use the performance profile to evaluate the performances of the AD-MM with different τ and the PALM for extraction under different scenarios. The performance profile is proposed by Dolan and Moré [14] as a tool for evaluating and comparing the performance of a collection of solvers $\mathcal{K}$ on a set of test problems $\mathcal{J}$.

To describe this method, we assume that we have K solvers and J problems, and we use the iteration number as a performance measure. Then, for each problem j and solver k, we set

$$\text{iter}_{j,k} = \text{the iteration number required to solve problem } j \text{ by solver } k.$$

and compute the performance ratio

$$r_{j,k} = \frac{\text{iter}_{j,k}}{\min\{\text{iter}_{j,k} : k \in \mathcal{K}\}}. \tag{5.2}$$

The performance profile of iteration numbers is then defined as the distribution function for the performance ratio, i.e.,

$$\rho_k(\nu) = \frac{1}{J} \sharp\{j \in \mathcal{J} : r_{j,k} \leq \nu\}$$

for $\nu \geq 1$. Similarly, the performance profile of function values is obtained by using $\text{fval}_{j,k}$ in place of $\text{iter}_{j,k}$ in (5.2), where $\text{fval}_{j,k}$ denotes the function value at the solution given by solver k for solving problem j. Generally speaking, for solver $k \in \mathcal{K}$, the higher $\rho_k(\nu)$ indicates a better performance within the factor ν.

In our experiments, we evaluate the following solvers: the ADMM with $\tau = 0.8$, the ADMM with $\tau = 1$, the ADMM with $\tau = 1.6$ and the PALM.

For $\mathcal{A}(L + S) = L + S$, our test problems are described in Table 5.1, where we use the four real videos introduced above as our input data in (5.1), with 3 choices of sparse regularizers, 10 choices of μ, and 6 choices of p and α. Thus, we have 4 solvers and a total of 720 test problems, with 240 test problems for each sparse regularizer. Moreover, we set $\text{Tol}_{A,1} = 10^{-4}$, $\text{Tol}_{A,2} = 5 \times 10^{-3}$ and $\text{Tol}_P = 10^{-4}$. Fig. 5.2 shows the performance profiles of iteration numbers and function values for different regularizers under this scenario.

For $\mathcal{A}(L + S) = H(L + S)$, our test problems are described in Table 5.2, where we use 2 choices of p and α. Thus, we have 4 solvers and a total of 240 test problems, with 80 test problems for each sparse regularizer. In our experiments, we use the method described in [23] to generate the blurring matrix H, which can be represented as a Kronecker product $H = H_r \otimes H_c$ under the periodic boundary condition. The matlab codes[4] that generate H_r and H_c are shown below, where "frame_size" is the size of each frame:

---

[4]The codes are available at http://www.imm.dtu.dk/~pcha/HNO/ as a supplement to the book [23].

```
[P, center] = psfGauss(frame_size, 1);
[Hr, Hc] = kronDecomp(P, center, 'periodic');
```

Moreover, we set $\text{Tol}_{A,1} = 5 \times 10^{-3}$, $\text{Tol}_{A,2} = 10^{-2}$ and $\text{Tol}_P = 3 \times 10^{-3}$. Fig. 5.3 shows the performance profiles under this scenario.

It is not hard to see from Fig. 5.2 and Fig. 5.3 that the performance profiles of iteration numbers for the ADMM with $\tau = 0.8$ and $\tau = 1$ usually lie above those for the PALM; and their performance profiles of function values are almost the same. This shows that the ADMM with $\tau = 0.8$ or $\tau = 1$ takes less iterations for solving all the test problems while giving comparable function values. For bridge regularizer in the case where $\mathcal{A}(L + S) = L + S$ (see Fig. 5.2(a)) and in the case where $\mathcal{A}(L + S) = H(L + S)$ (see Fig. 5.3(a)), we can see that the ADMM with $\tau = 0.8$ sightly outperforms the ADMM with $\tau = 1$ in terms of the number of iterations. For other regularizers, their performances are comparable. Additionally, for the ADMM with $\tau = 1.6$, we can see from Fig. 5.2 and Fig. 5.3 that it always terminates with the worst function value, although it is always fastest in the case where $\mathcal{A}(L + S) = H(L + S)$ (see Fig. 5.3).

To better visualize the performance of the algorithms in terms of function values, we also plot $\text{RelErr}^k := |\mathcal{F}(L^k, S^k) - \mathcal{F}_{\min}|/\mathcal{F}_{\min}$ against the number of iterations for each algorithm, where $\mathcal{F}(L^k, S^k)$ denotes the objective value obtained by each algorithm at (L^k, S^k) and $\mathcal{F}_{\min}$ denotes the minimum of the objective values obtained from all algorithms. We only consider the ADMM with $\tau = 0.8$, the ADMM with $\tau = 1$ and the PALM, and terminate them only after *at least* 500 iterations *and* the termination criteria are satisfied with $\text{Tol}_{A,1} = 10^{-5}$, $\text{Tol}_{A,2} = 5 \times 10^{-4}$ and $\text{Tol}_P = 10^{-5}$. For brevity, we focus on the scenario $\mathcal{A}(L + S) = L + S$ and use the "Hall" video. The results are presented in Fig. 5.4, from which we can see that the ADMM with $\tau = 1$ or $\tau = 0.8$ performs better than PALM for those particular instances.

TABLE 5.1
Problem setting for $\mathcal{A}(L + S) = L + S$

data	μ	regularizers
4 real videos	5e-1, 1e-1, 5e-2, 1e-2, 5e-3 1e-3, 5e-4, 1e-4, 5e-5, 1e-5	bridge: $p = 0.2, 0.4, 0.5, 0.6, 0.8, 1$ fraction/logistic: $\alpha = 0.01, 0.1, 1, 2, 5, 10$

TABLE 5.2
Problem setting for $\mathcal{A}(L + S) = H(L + S)$

data	μ	regularizers
4 real videos	5e-1, 1e-1, 5e-2, 1e-2, 5e-3 1e-3, 5e-4, 1e-4, 5e-5, 1e-5	bridge: $p = 0.5, 1$ fraction/logistic: $\alpha = 1, 2$

5.3. Simulation Results. In this subsection, we present some simulation results for the background/foreground extraction problem. In order to evaluate the performance in background/foreground extraction, we compare the support of the recovered foreground S^* with the support of the ground-truth $\tilde{S}$ by computing the following measurement:

$$\text{F-measure} := 2 \times \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}},$$

where precision and recall are defined as

$$\text{precision} := \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} := \frac{\text{TP}}{\text{TP} + \text{FN}},$$

in which,

- TP stands for true positives: the number of true foreground pixels that are recovered;
- FP stands for false positives: the number of background pixels that are misdetected as foreground;
- FN stands for false negatives: the number of true foreground pixels that are missed.

The support of the recovered foreground S^* is obtained by thresholding S^* entry-wise with a threshold value (we use 1e-3 in our numerical experiments). We would like to point out that F-measure varies between 0 and 1 according to the similarity of the support of S^* and $\tilde{S}$. The higher the F-measure value, the better the recovery accuracy of the support of $\tilde{S}$. The F-measure approaches the maximum value 1 if the supports of S^* and $\tilde{S}$ are the same, which means the foreground is recovered completely.

In our experiments below, we choose $\tau = 0.8$ for the ADMM. We also use the aforementioned four real videos as input with 3 choices of sparse regularizers and 2 choices of p and α. For each fixed p and α, we experiment with different regularization parameters μ (5e-1, 1e-1, 5e-2, 1e-2, 5e-3, 1e-3, 5e-4, 1e-4, 5e-5, 1e-5) and present only the μ corresponding to the maximal F-measure.[5]

Extraction from noisy surveillance videos. In this case, $\mathcal{A}(L+S) = L+S$, $\lambda_{\max} = \lambda_{\min} = 1$ and we set $\text{Tol}_{A,1} = 10^{-4}$, $\text{Tol}_{A,2} = 5 \times 10^{-3}$ and $\text{Tol}_P = 10^{-4}$. The computational results are reported in Table 5.3, where we report p and α, the optimal μ, the number of iterations, the CPU time (seconds) and F-measure. We also show the extracted backgrounds and foregrounds given by the ADMM in Fig. 5.5.

Extraction from noisy and blurred surveillance videos. In this case, $\mathcal{A}(L + S) = H(L + S)$, $\lambda_{\max} = \lambda_{\max}(H^*H)$, $\lambda_{\min} = \lambda_{\min}(H^*H)$ and we set $\text{Tol}_{A,1} = 5 \times 10^{-3}$, $\text{Tol}_{A,2} = 10^{-2}$ and $\text{Tol}_P = 3 \times 10^{-3}$. The blurring matrix H is generated by the same method introduced in Subsection 5.2. One frame of each corrupted video is shown in the second row in Fig. 5.1. We report the computational results in Table 5.4 and show the extracted backgrounds and foregrounds by the ADMM in Fig. 5.6.

Summary. From the results above, it can be seen that the ADMM with $\tau = 0.8$ performs better in the sense that it takes less CPU time for solving most test problems while returning comparable F-measures. The performances of our ADMM for extraction are also promising from Fig. 5.5 and Fig. 5.6.

6. Concluding remarks. In this paper, we study a general (possibly nonconvex and nonsmooth) model and adapt the ADMM with a general dual step-size τ, which can be chosen in $(0, \frac{1+\sqrt{5}}{2})$, to solve it. We establish that any cluster point of the sequence generated by our ADMM gives a stationary point under some assumptions; we also give simple sufficient conditions for these assumptions. Under an additional assumption that a potential function is a Kurdyka-Łojasiewicz function, we can further establish the global convergence of the whole sequence generated by our ADMM. Our computational results demonstrate the efficiency of our algorithm.

---

[5]If the F-measures are the same, we pick the μ that corresponds to the minimal number of iterations.

TABLE 5.3
Numerical results for extraction from noisy surveillance videos

Data	regularizer		ADMM				PALM			
		μ	iter	time	F-measure	μ	iter	time	F-measure
Hall	bri. p	1.0	5e-02	10	3.21	0.7562	5e-02	19	3.96	0.7560
		0.5	1e-02	32	11.26	0.7634	1e-02	36	9.29	0.7624
	fra. α	1.0	5e-02	23	8.26	0.7578	5e-02	33	8.53	0.7578
		2.0	5e-02	12	4.17	0.7368	5e-02	15	3.69	0.7371
	log. α	1.0	5e-02	12	21.12	0.7566	5e-02	39	68.70	0.7576
		2.0	5e-02	12	16.00	0.7368	5e-02	16	29.04	0.7368
Bootstrap	bri. p	1.0	1e-01	14	3.30	0.8180	1e-01	19	3.15	0.8180
		0.5	5e-02	23	6.77	0.8206	5e-02	22	4.93	0.8209
	fra. α	1.0	1e-01	15	4.91	0.8163	1e-01	20	5.32	0.8165
		2.0	1e-01	14	4.18	0.8264	1e-01	16	3.72	0.8261
	log. α	1.0	1e-01	16	21.92	0.8195	1e-01	22	28.62	0.8195
		2.0	1e-01	12	8.91	0.8363	1e-01	10	6.50	0.8363
Fountain	bri. p	1.0	1e-01	9	2.19	0.7749	1e-01	7	1.10	0.7749
		0.5	5e-02	13	3.54	0.7000	5e-02	11	2.13	0.6922
	fra. α	1.0	1e-01	9	2.39	0.7717	1e-01	8	1.63	0.7717
		2.0	5e-02	10	2.82	0.7717	5e-02	9	1.87	0.7717
	log. α	1.0	1e-01	9	13.41	0.7738	1e-01	7	9.65	0.7738
		2.0	5e-02	9	12.46	0.7717	5e-02	8	11.51	0.7717
ShoppingMall	bri. p	1.0	1e-01	10	9.66	0.7046	1e-01	13	8.73	0.7043
		0.5	1e-02	39	52.39	0.7087	1e-02	79	83.62	0.7078
	fra. α	1.0	1e-01	12	14.33	0.7055	1e-01	18	16.95	0.7055
		2.0	5e-02	15	18.46	0.7062	5e-02	26	25.34	0.7064
	log. α	1.0	1e-01	11	66.96	0.7055	1e-01	16	94.06	0.7055
		2.0	5e-02	12	40.23	0.7057	5e-02	18	74.83	0.7057

Note that our ADMM may not be beneficial when $\mathcal{B}$ or $\mathcal{C}$ has no special structure, because the corresponding subproblems of ADMM may not have closed-form solutions. Nonetheless, as in [31, 44, 45], it may be possible to add "proximal terms" to simplify the subproblems of our ADMM. In addition, in view of the recent work [46], it may also be possible to study the convergence of our ADMM for some specially structured nonconvex Ψ. These are possible future research directions.

REFERENCES

[1] B.P.W. Ames and M. Hong. Alternating direction method of multipliers for penalized zero-variance discriminant analysis. *arXiv preprint arXiv:1401.5492*, 2014.
[2] H. Attouch, J. Bolte, P. Redont and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2): 438–457, 2010.
[3] D. Bertsekas and J.N. Tsitsiklis. Parallel and Distributed Computation: Numerical Methods. Prentice Hall, 1989.
[4] W. Bian and X. Chen. Linearly constrained non-Lipschitz optimization for image restoration. *SIAM Journal on Imaging Sciences*, 8(4): 2294–2322, 2015.

TABLE 5.4
Numerical results for extraction from noisy and blurred surveillance videos

| Data	regularizer		ADMM				PALM		
		μ	iter	time	F-measure	μ	iter	time	F-measure
Hall	bri. p 1.0	5e-02	24	15.37	0.6801	5e-02	36	26.72	0.6626
	0.5	1e-02	44	31.96	0.6358	1e-02	45	35.11	0.6357
	fra. α 1.0	5e-02	57	40.40	0.5265	5e-02	49	37.02	0.5616
	2.0	1e-02	66	52.25	0.5381	1e-02	61	50.51	0.5445
	log. α 1.0	5e-02	42	93.40	0.5970	5e-02	44	100.84	0.6033
	2.0	1e-02	54	118.58	0.5188	1e-02	52	120.68	0.5211
Bootstrap	bri. p 1.0	1e-01	22	10.56	0.7651	1e-01	50	26.20	0.7364
	0.5	1e-02	56	36.50	0.6692	1e-02	86	55.18	0.6589
	fra. α 1.0	5e-02	66	36.75	0.5705	5e-02	99	60.04	0.5270
	2.0	5e-02	73	41.08	0.5265	5e-02	111	61.89	0.4674
	log. α 1.0	5e-02	47	83.24	0.5651	5e-02	87	158.16	0.5666
	2.0	5e-02	73	131.17	0.4891	5e-02	115	213.62	0.4179
Fountain	bri. p 1.0	5e-02	28	12.72	0.7229	5e-02	64	36.25	0.6970
	0.5	1e-02	51	28.54	0.6881	1e-02	78	45.68	0.6606
	fra. α 1.0	5e-02	64	34.55	0.5155	5e-02	84	49.98	0.5000
	2.0	1e-02	62	37.87	0.4482	1e-02	87	56.47	0.4341
	log. α 1.0	5e-02	50	98.81	0.6095	5e-02	77	162.12	0.5760
	2.0	1e-02	53	105.63	0.4438	1e-02	80	169.25	0.4525
ShoppingMall	bri. p 1.0	5e-02	22	51.99	0.6431	5e-02	25	62.71	0.6411
	0.5	5e-03	54	157.53	0.6271	5e-03	40	119.91	0.6328
	fra. α 1.0	1e-02	33	109.64	0.5045	5e-02	60	188.48	0.5106
	2.0	1e-02	51	153.19	0.5810	1e-02	38	127.47	0.5935
	log. α 1.0	5e-02	58	263.00	0.5453	5e-02	39	178.60	0.5967
	2.0	1e-02	38	174.88	0.5856	1e-02	33	149.36	0.5913

[5] J. Bolte, S. Sabach and M. Teboublle. Proximal alternating linearized minimization for non-convex and nonsmooth problems. *Mathematical Programming*, 146(1–2): 459–494, 2014.

[6] T. Bouwmans. Recent advanced statistical background modelling for foreground detection: A systematic survey. *Recent Patents on Computer Science*, 4(3): 147–176, 2011.

[7] T. Bouwmans. Traditional and recent aproaches in background modeling for foreground detection: An overview. *Computer Science Review*, 11–12: 31–66, 2014.

[8] 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. *arXiv preprint arXiv:1511.01245*, 2015.

[9] 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.

[10] E.J. Candès, X. Li, Y. Ma and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3), Article 11, 2011.

[11] C. Chen, B. He, Y. Ye and X. Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155: 57–79, 2016.

[12] X. Chen and W. Zhou. Smoothing nonlinear conjugate gradient method for image restoration using nonsmooth nonconvex minimization. *SIAM Journal on Imaging Sciences*, 3(4): 765–790, 2010.

[13] Y. Deng, Q. Dai, R. Liu, Z. Zhang and S. Hu. Low-rank structure learning via nonconvex heuristic recovery. *IEEE Transactions on Neural Networks and Learning Systems*, 24(3): 383–396, 2013.

[14] E.D. Dolan and J.J. Moré. Benchmarking optimization software with performance profiles.

Mathematical Programming, 91(2): 201–213, 2012.

[15] J. Eckstein and D. Bertsekas. On the DouglasRachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55, 293–318, 1992.

[16] J. Fan. Comments on "wavelets in statistics: A review" by A.Antoniadis. *Journal of the Italian Statistical Society*, 6(2): 131–138, 1997.

[17] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360, 2001.

[18] M. Fazel, T.K. Pong, D. Sun and P. Tseng. Hankel matrix rank minimization with applications to system identification and realization. *SIAM Journal on Matrix Analysis and Applications*, 34(3): 946–977, 2013.

[19] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximations. *Computers & Mathematics with Applications*, 2(1): 17–40, 1976.

[20] D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(3): 367–383, 1992.

[21] R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une classe de problèmes de Dirichlet non linéares. *Revue Francaise d'Automatique, Informatique, Recherche Opérationelle.*, 9(R-2): 41–76, 1975.

[22] S. Gu, L. Zhang, W. Zuo and X. Feng. Weighted nuclear norm minimization with application to image denoising. in *CVPR*, 2862–2869, 2014.

[23] P.C. Hansen, J.G. Nagy and D.P. O'leary. Deblurring Images: Matrices, Spectra, and Filtering. Fundamentals of Algorithms 3, SIAM, Philadelphia, 2006.

[24] B. He, M. Tao and X. Yuan. Alternating direction method with Gaussian back substitution for separable convex programming. *SIAM Journal on Optimization*, 22(2): 313–340, 2012.

[25] B. He and X. Yuan. Linearized alternating direction method with Gaussian back substitution for separable convex programming. *Numerical Algebra, Control and Optimization*, 3(2), 247–260, 2013.

[26] M. Hong, Z.-Q. Luo and M. Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1): 337–364, 2016.

[27] J. Huang, J. L. Horowitz and S. Ma. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, 36(2): 587–613, 2008.

[28] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5): 1356–1378, 2000.

[29] L. Li, W. Huang, I. Y.-H. Gu and Q. Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing*, 13(11): 1459–1472, 2004.

[30] X. Li, M. K. Ng and X. Yuan. Median filtering–based methods for static background extraction from surveillance video. *Numerical Linear Algebra with Applications*, 22: 845–865, 2015.

[31] G. Li and T.K. Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4): 2434–2460, 2015.

[32] G. Li and T.K. Pong. Douglas-Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems. *Mathematical Programming*, 159(1): 371–401, 2016.

[33] L. Li, P. Wang, Q. Hu and S. Cai. Efficient background modeling based on sparse representation and outlier iterative removal. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(2): 278–289, 2014.

[34] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1): 171–184, 2013.

[35] Z. Lin, R. Liu and H. Li. Linearized alternating direction method with parallel splitting and adaptive penalty for separable convex programs in machine learning. *Machine Learning*, 99(2): 287–325, 2015.

[36] M. Lin, D. Sun and K.-C. Toh. A convergent 3-block semi-proximal ADMM for convex minimization problems with one strongly convex block. *Asia-Pacific Journal of Operational Research*, 32(3): 1550024(19p), 2015.

[37] K. Mohan, P. London, M. Fazel, D. Witten and S.-I. Lee. Node-based learning of multiple gaussian graphical models. *Journal of Machine Learning Research*, 15(1): 445–488, 2014.

[38] M. Nikolova. Energy Minimization Methods. in *Handbook of Mathematical Methods in Imaging*, O. Scherzer, Ed. Springer, pp. 139–185, 2011.

[39] M. Nikolova, M.K. Ng, S. Zhang and W.-K. Ching. Efficient reconstruction of piecewise constant

images using nonsmooth nonconvex minimization. *SIAM Journal on Imaging Sciences*, 1(1): 2–25, 2008.

[40] R. Otazo, E.J. Candès and D. K. Sodickson. Low–rank plus sparse matrix decomposition for accelerated dynamic MRI with separation of background and dynamic components. *Magnetic Resonance in Medicine*, 73(3): 1125–1136, 2015.

[41] Y. Peng, A. Ganesh, J. Wright, W. Xu and Y. Ma. Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11): 2233–2246, 2012.

[42] R.T. Rockafellar and R.J-B. Wets. *Variational Analysis*. Springer, 1998.

[43] D. Sun, K.-C. Toh and L. Yang. A convergent 3–block semi-proximal alternating direction method of multipliers for conic programming with 4–type constraints. *SIAM Journal on Optimization*, 25(2): 882–915, 2015.

[44] F. Wang, W. Cao and Z. Xu. Convergence of multi-block Bregman ADMM for nonconvex composite problems. *arXiv preprint arXiv:1505.03063*, 2015.

[45] F. Wang, Z. Xu and H.-K. Xu. Convergence of Bregman alternating direction method with multipliers for nonconvex composite problems. *arXiv preprint arXiv:1410.8625*, 2014.

[46] Y. Wang, W. Yin and J. Zeng. Global convergence of ADMM in nonconvex nonsmooth optimization. *arXiv preprint arXiv:1511.06324*, 2015.

[47] H. Xu, C. Caramanis and S. Sanghavi. Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5): 3047–3064, 2012.

[48] M.H. Xu and T. Wu. A class of linearized proximal alternating direction methods. *Journal of Optimization Theory and Applications*, 151(2): 321–337, 2011.

[49] C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2): 894–942, 2010.

[50] W. Zuo, D. Ren, S. Gu, L. Lin and L. Zhang. Discriminative learning of iteration-wise priors for blind deconvolution. in *CVPR*, 2015.
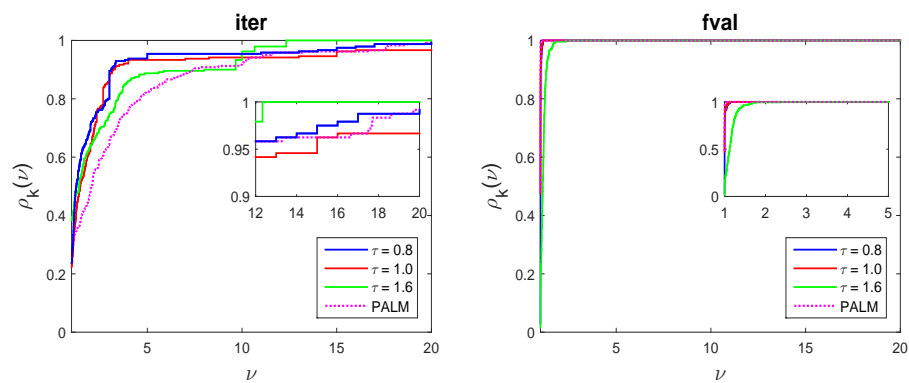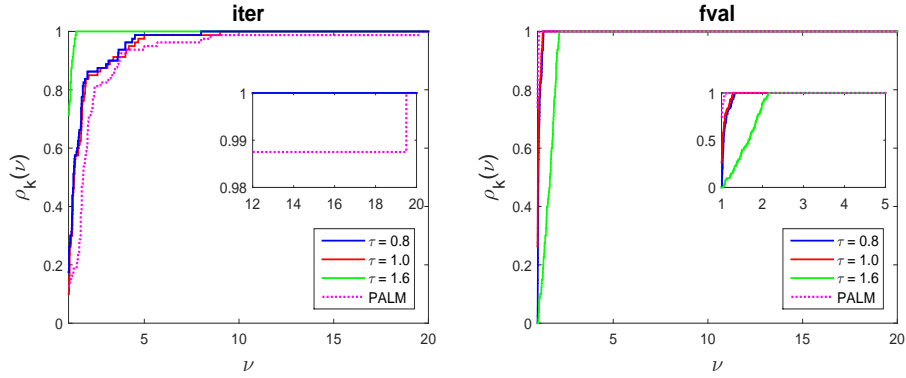
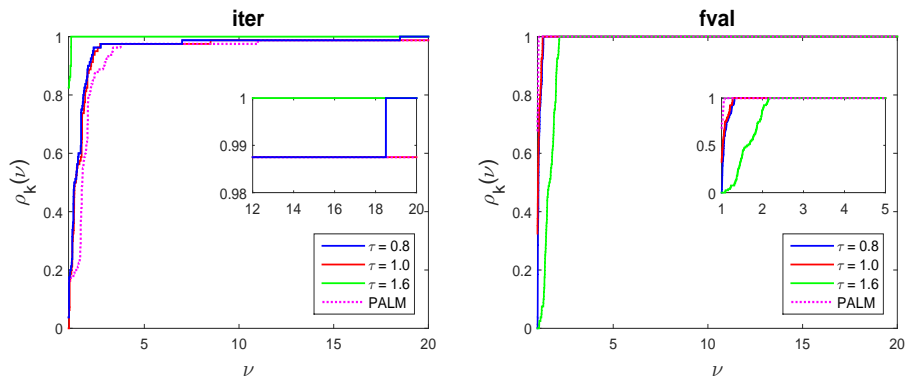(a) bridge regularizer
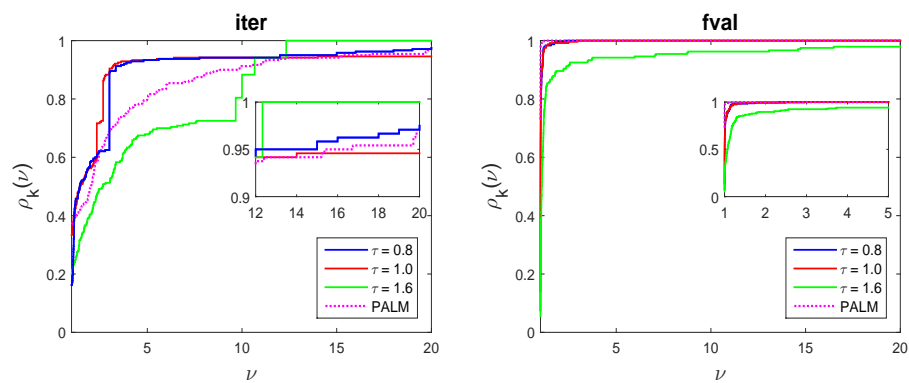
(b) fraction regularizer
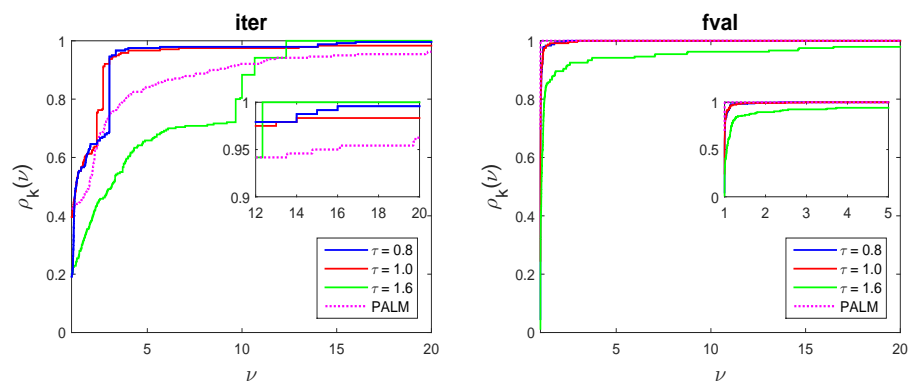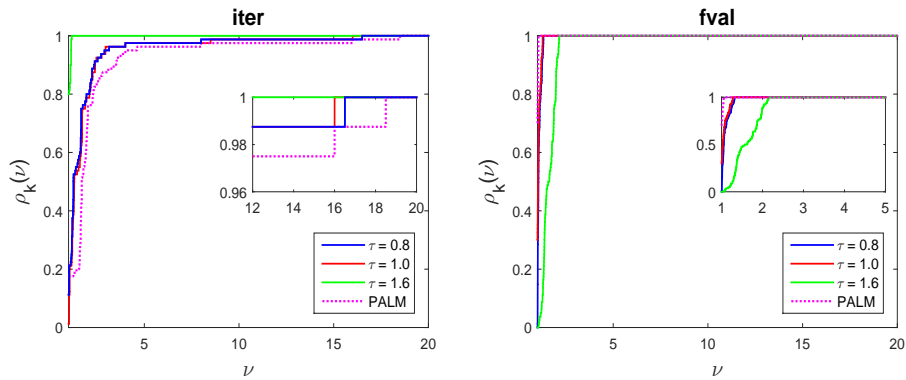

(c) logistic regularizer

Fig. 5.2. *Performance profiles of iteration numbers (denoted by "**iter**" on the left) and function values (denoted by "**fval**" on the right) for each sparse regularizer with $\mathcal{A}(L + S) = L + S$. The blown-up subfigures are used to highlight the differences in a specific range of ν.*

(a) bridge regularizer

(b) fraction regularizer

(c) logistic regularizer

FIG. 5.3. *Performance profiles of iteration numbers (denoted by "**iter**" on the left) and function values (denoted by "**fval**" on the right) for each sparse regularizer with $\mathcal{A}(L+S) = H(L+S)$. The blown-up subfigures are used to highlight the differences in a specific range of ν.*
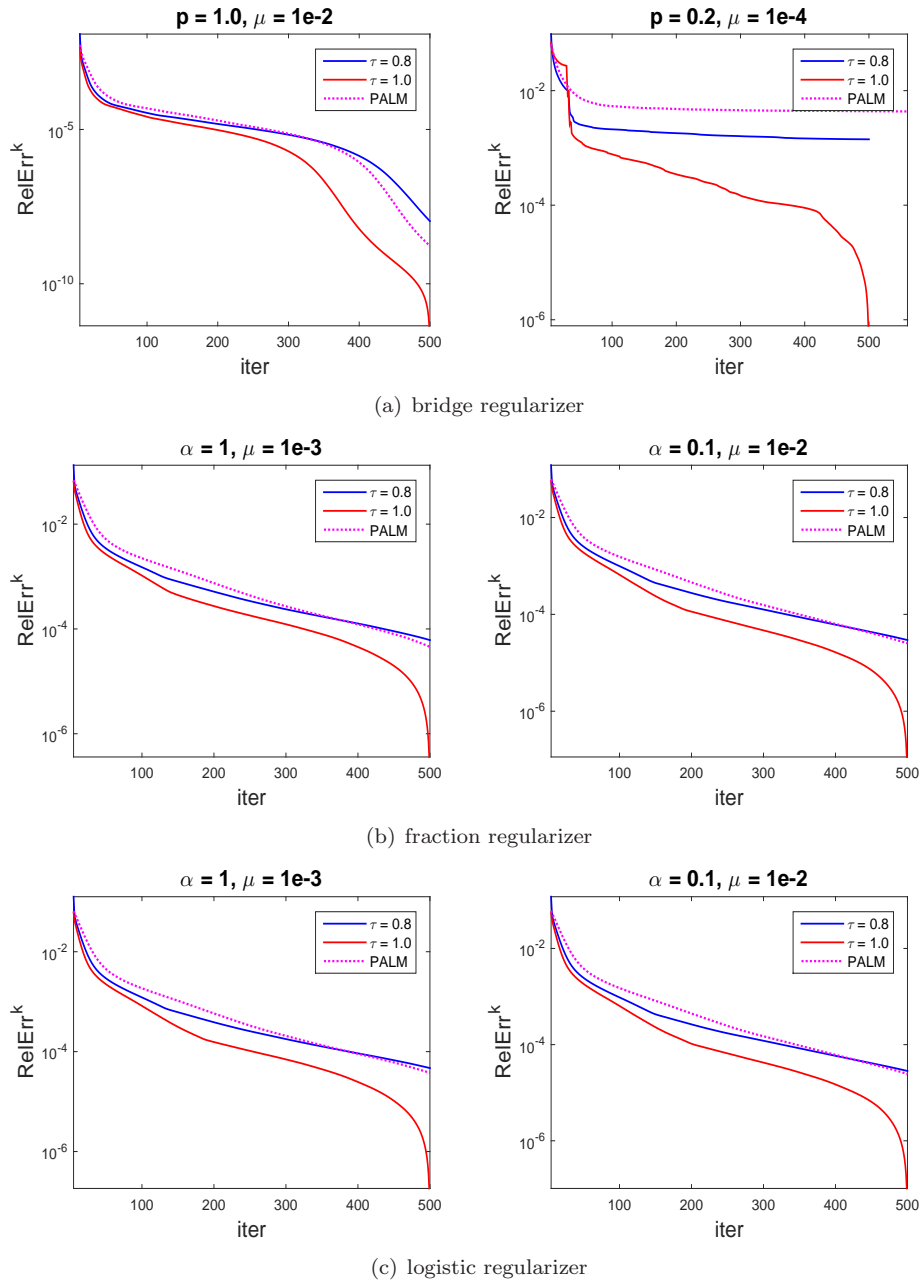
(a) bridge regularizer

(b) fraction regularizer

(c) logistic regularizer

Fig. 5.4. *The* RelErr^k *vs the number of iterations for each sparse regularizer*

bri.$p = 1.0$ bri.$p = 0.5$ fra.$\alpha = 1$

fra.$\alpha = 2$ log.$\alpha = 1$ log.$\alpha = 2$
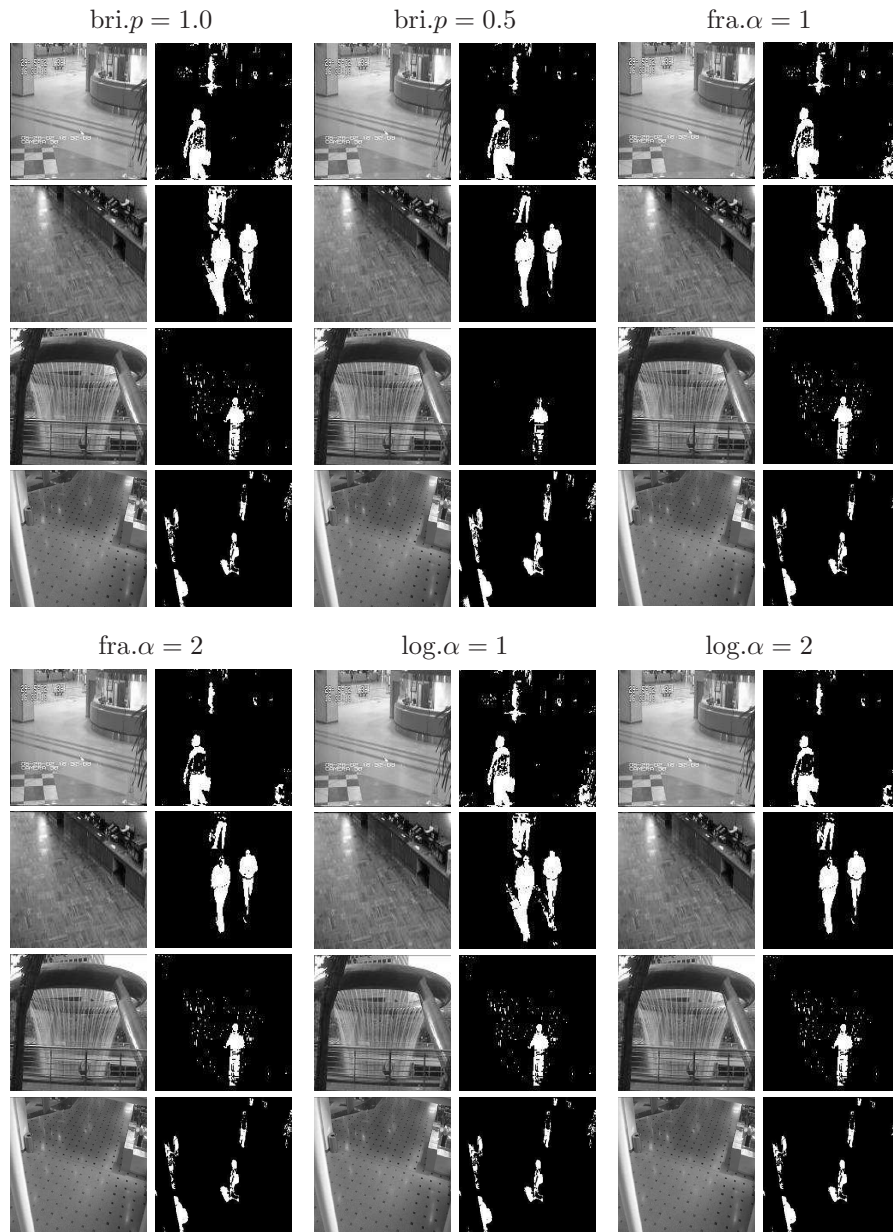
FIG. 5.5. *Extracted backgrounds and foregrounds given by the ADMM for noisy surveillance videos.*

bri.$p = 1.0$ bri.$p = 0.5$ fra.$\alpha = 1$

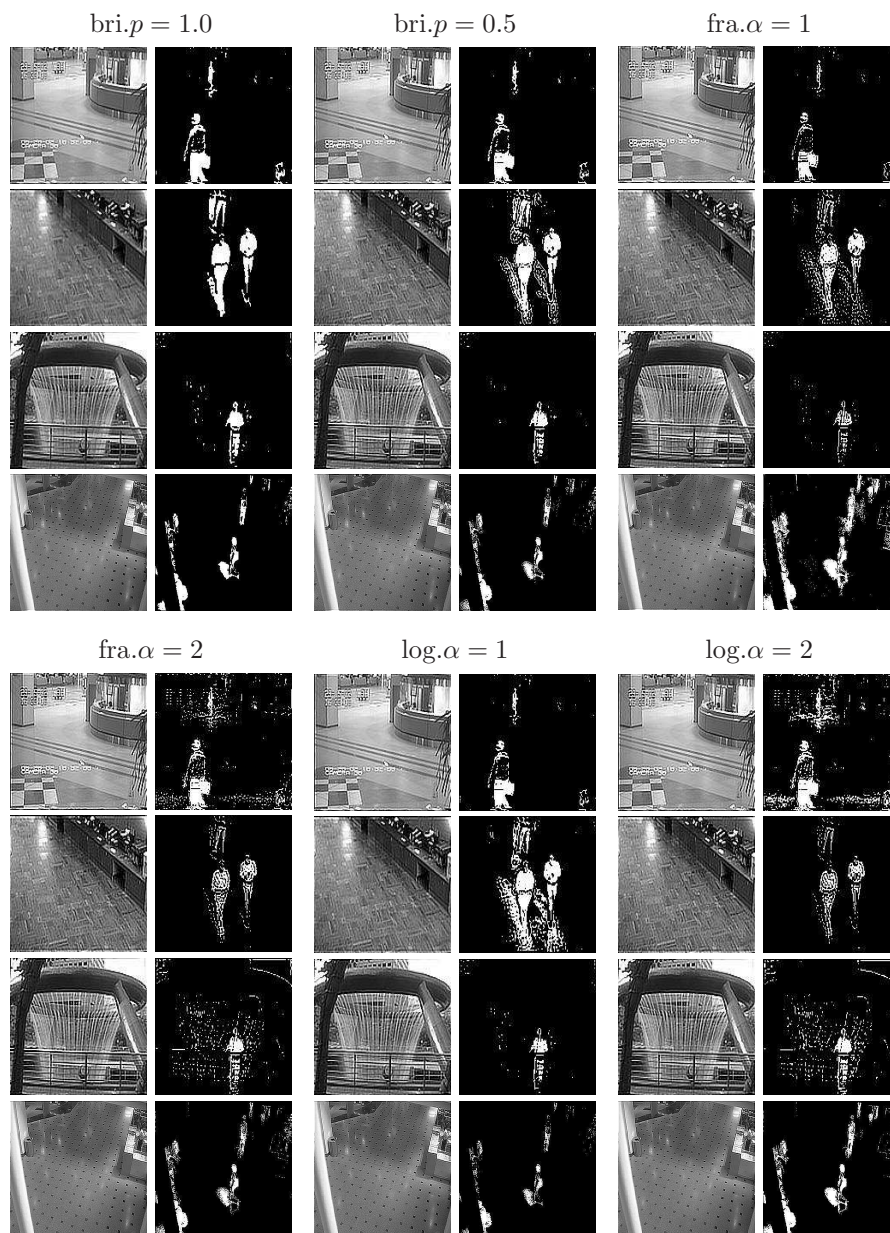fra.$\alpha = 2$ log.$\alpha = 1$ log.$\alpha = 2$

Fig. 5.6. *Extracted backgrounds and foregrounds given by the ADMM for noisy and blurred surveillance videos.*