A UNIFIED APPROACH FOR MINIMIZING COMPOSITE NORMS

N. S. AYBAT* AND G. IYENGAR†

Abstract. We propose a first-order augmented Lagrangian algorithm (FALC) to solve the composite norm minimization problem

$$\begin{array}{ll}\min_{X\in\mathbb{R}^{m\times n}} & \mu_1\|\sigma(\mathcal{F}(X)-G)\|_\alpha+\mu_2\|\mathcal{C}(X)-d\|_\beta,\\ \text{subject to} & \mathcal{A}(X)-b\in\mathcal{Q},\end{array}$$

where $\sigma(X)$ denotes the vector of singular values of $X\in\mathbb{R}^{m\times n}$, the matrix norm $\|\sigma(X)\|_\alpha$ denotes either the Frobenius, the nuclear, or the ℓ_2-operator norm of X, the vector norm $\|.\|_\beta$ denotes either the ℓ_1-norm, ℓ_2-norm or the ℓ_∞-norm; $\mathcal{Q}$ is a closed convex set and $\mathcal{A}(.)$, $\mathcal{C}(.)$, $\mathcal{F}(.)$ are linear operators from $\mathbb{R}^{m\times n}$ to vector spaces of appropriate dimensions. Basis pursuit, matrix completion, robust principal component pursuit (PCP), and stable PCP problems are all special cases of the composite norm minimization problem. Thus, FALC is able to solve all these problems in a unified manner. We show that any limit point of FALC iterate sequence is an optimal solution of the composite norm minimization problem. We also show that for *all* $\epsilon>0$, the FALC iterates are ϵ-feasible and ϵ-optimal after $\mathcal{O}(\log(\epsilon^{-1}))$ iterations, which require $\mathcal{O}(\epsilon^{-1})$ constrained shrinkage operations and Euclidean projection onto the set $\mathcal{Q}$. Surprisingly, on the problem sets we tested, FALC required only $\mathcal{O}(\log(\epsilon^{-1}))$ constrained shrinkage, instead of the $\mathcal{O}(\epsilon^{-1})$ worst case bound, to compute an ϵ-feasible and ϵ-optimal solution. To best of our knowledge, FALC is the first algorithm with a known complexity bound that solves the stable PCP problem.

1. Introduction.

In this paper we consider the class of composite norm minimization problems defined in (1.1).

$$\min_{X\in\mathbb{R}^{m\times n}} \mu_1\|\sigma(\mathcal{F}(X)-G)\|_\alpha+\mu_2\|\mathcal{C}(X)-d\|_\beta \text{ subject to } \mathcal{A}(X)-b\in\mathcal{Q}, \tag{1.1}$$

where $\mu_1,\mu_2\geq 0$, $b\in\mathbb{R}^q$, $\mathcal{A}:\mathbb{R}^{m\times n}\to\mathbb{R}^q$ denotes a linear map, $\mathcal{Q}\subset\mathbb{R}^q$ is a nonempty, closed convex set, $d\in\mathbb{R}^p$, $\mathcal{C}:\mathbb{R}^{m\times n}\to\mathbb{R}^p$ is a linear map, $G\in\mathbb{R}^{r_1\times r_2}$, $\mathcal{F}:\mathbb{R}^{m\times n}\to\mathbb{R}^{r_1\times r_2}$ is a linear map, and the function $\sigma(Z)$ denotes the singular values of the matrix Z. The parameter $\beta\in\{1,2,\infty\}$, and $\|.\|_\beta$ denotes the ℓ_β vector norm. The parameter $\alpha\in\{1,2,\infty\}$, and for $\alpha=1,2,\infty$, the norm $\|\sigma(Z)\|_\alpha$ denotes, respectively, the nuclear norm $\|X\|_*=\|\sigma(X)\|_1$, the Frobenius norm $\|X\|_F=\|\sigma(X)\|_2$, and the ℓ_2-operator norm $\|X\|_2=\|\sigma(X)\|_\infty$. Except for Section 6 we assume the following.

ASSUMPTION 1.1. *The linear map $\mathcal{A}$ is surjective, and, at least, one of the linear maps $\mathcal{C}$ and $\mathcal{F}$ in the objective function is injective.*

Since at least one of the linear maps $\mathcal{C}$ and $\mathcal{F}$ are injective, it follows that the objective function $\mu_1\|\sigma(\mathcal{F}(X)-G)\|_\alpha+\mu_2\|\mathcal{C}(X)-d\|_\beta$ of (1.1) goes to ∞ as $\|X\|_F\to\infty$, i.e. Assumption 1.1 ensures that the objective function of (1.1) is coercive, and hence, an optimal solution to (1.1) exists.

The composite norm minimization problem (1.1) appears in the context of "structured" or "sparse" optimization where desired solution is "structured" in some form – the solution matrix may be sparse, i.e. has very few non-zero components, or it may be low rank, or the indices of its non-zero coefficients may all belong to a union of few given index sets, i.e. "groups". We show in Section 1.2 that many well-known "structured" optimization problems such as basis pursuit, matrix completion, robust principal component pursuit (PCP), and stable PCP, are all special cases of (1.1). Moreover, Assumption 1.1 is satisfied in all these special cases.

Composite norm minimization problem (1.1) can be reformulated as a semidefinite programming problem (SDP); hence, it can be solved efficiently in theory. However, instances of (1.1) that arise in practice are very large and typically dense. Therefore, interior point based SDP solvers perform very poorly on these instances.

1.1. New Results.

We propose a first-order augmented Lagrangian algorithm (FALC) to solve (1.1). The main results of this paper are as follows:

(a) We establish that every limit point $\bar{X}$ of the sequence of FALC iterates $\{X^{(k)}\}_{k\in\mathbb{Z}_+}$ is an optimal solution of (1.1), i.e.,

$$\bar{X}\in\operatorname*{argmin}_{X\in\mathbb{R}^{m\times n}}\Big\{\mu_1\|\sigma(\mathcal{F}(X)-G)\|_\alpha+\mu_2\|\mathcal{C}(X)-d\|_\beta:\ \mathcal{A}(X)-b\in\mathcal{Q}\Big\}.$$

*IEOR Department, Columbia University. Email: nsa2106@columbia.edu

†IEOR Department, Columbia University. Email: gi10@columbia.edu

(b) Let P^* denote the optimal value of (1.1). For *all* $\epsilon > 0$, the FALC iterates $X^{(k)}$ are ϵ-feasible, i.e. there exists $y^{(k)} \in \mathcal{Q}$ such that

$$\|\mathcal{A}(X^{(k)}) - y^{(k)} - b\|_2 \leq \epsilon,$$

and ϵ-optimal, i.e.

$$\left| \left(\mu_1 \|\sigma(\mathcal{F}(X^{(k)}) - G)\|_\alpha + \mu_2 \|\mathcal{C}(X^{(k)}) - d\|_\beta\right) - P^* \right| \leq \epsilon,$$

after $\mathcal{O}(\log(\epsilon^{-1}))$ FALC iterations that requires $\mathcal{O}(\epsilon^{-1})$ projections on to $\mathcal{Q}$ and $\mathcal{O}(\epsilon^{-1})$ "constrained shrinkage" operations in the worst case - see (3.41) and (3.42) for the definition and complexity of each "constrained shrinkage" operation.

(c) FALC can be extended to solve the following more general optimization problem

$$\min_{X \in \mathbb{R}^{m\times n}} \quad \mu_1 \|\sigma(\mathcal{F}(X) - G)\|_\alpha + \mu_2 \|\mathcal{C}(X) - d\|_\beta + \mu_3 H(X) \ \text{ s.t. } \mathcal{A}(X) - b \in \mathcal{Q}, \tag{1.2}$$

where $H : \mathbb{R}^{m\times n} \to \mathbb{R}$ is a strongly convex function, with the same complexity guarantees.

(d) In our numerical tests we observed that FALC required only $\mathcal{O}(\log(\epsilon^{-1}))$ projection and shrinkage operations as opposed to $\mathcal{O}(\epsilon^{-1})$ the worst case theoretical bound proven in the paper to obtain an ϵ-feasible and ϵ-optimal solution.

(e) We also observed that, although, FALC is a general-purpose algorithm for the composite norm minimization problem (1.1), the numerical results show that FALC is competitive with the state-of-the-art special purpose algorithms designed for all special cases that we tested.

1.2. Special Cases. We show below that many well studied "structured" optimization problems are special cases of (1.1).

Nuclear norm-minimization. The *nuclear norm minimization problem*

$$\min_{X \in \mathbb{R}^{m\times n}} \|X\|_* \text{ subject to (s.t.) } \mathcal{A}(X) = b. \tag{1.3}$$

is a special case of (1.1) with $\mathcal{Q} = \{\mathbf{0}\}$, $\mathcal{F}(X) = X$, $G = \mathbf{0}$, $\alpha = 1$, i.e. $\|\sigma(X)\|_1 = \|X\|_*$, $\mu_1 = 1$, and $\mu_2 = 0$. The nuclear norm minimization problem is a convex approximation for the NP-hard rank minimization problem $\min_{X\in\mathbb{R}^{m\times n}}\{\text{rank}(X) : \mathcal{A}(X) = b\}$, where $\text{rank}(X)$ denotes the rank of $X \in \mathbb{R}^{m\times n}$. Rank minimization arises in many different contexts, e.g. system identification [32], optimal control [18, 19, 16], low-dimensional embedding in Euclidean space [31], and matrix completion [11].

Let $X_0 \in \mathbb{R}^{m\times n}$ be the unknown low-rank matrix such that $\mathcal{A}(X_0) = b$. Let $r = \text{rank}(X_0)$ and $\bar{n} = \max\{m, n\}$. When the linear operator $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^q$ satisfies some regularity properties, and the number of measurements $q = \mathcal{O}(r(m + n)\log(mn))$, Recht et al. [35] show that, with very high probability, (1.3) has a unique optimal solution and this solution is also optimal for the rank minimization problem.

Another related special case is the *matrix completion problem* where the operator $\mathcal{A}$ picks a subset of the matrix elements, i.e., the linear constraints are of the form: $X_{ij} = (X_0)_{ij}$ for $(i, j) \in \Omega$, where Ω is a given index set of observable entries of an unknown low rank matrix $X_0 \in \mathbb{R}^{m\times n}$. When indices (i, j) are sampled uniformly at random, and $|\Omega| = \mathcal{O}(\bar{n}^{1.2} r \log(\bar{n}))$ and the unknown matrix X_0 satisfies some regularity conditions, Candés et al. [11] show that, with high probability, X_0 is the unique solution of the matrix completion problem. The Netflix prize problem [34] is an example of the matrix completion problem.

(1.3) can be reformulated as an SDP; however, instances of (1.3) that arise in practice are so large that standard SDP solvers are unable to solve them. For existing algorithmic methodologies for solving the nuclear norm minimization problem, see [6, 22, 29, 30, 33, 36] and references therein.

Basis-pursuit problem. The *basis pursuit problem*

$$\min_{x\in\mathbb{R}^n} \|x\|_1 \text{ s.t. } Ax = b, \tag{1.4}$$

where $A \in \mathbb{R}^{q\times n}$ and $b \in \mathbb{R}^q$, is a special case of (1.1) with $\mathcal{Q} = \{\mathbf{0}\}$, $C(x) = x \in \mathbb{R}^{n\times 1}$, $d = 0$, $\beta = 1$, $\mu_1 = 0$ and $\mu_2 = 1$. The basis pursuit problem has attracted a lot of attention recently, since it appears in the context of *compressed sensing* (CS) [7, 8, 9, 15]. The goal in CS is to recover a sparse signal $x_0 \in \mathbb{R}^n$

from a small set of linear measurements or transform values $b = Ax_0$, or equivalently, to solve the NP-hard ℓ_0-minimization problem

$$\min_{x\in\mathbb{R}^n} \|x\|_0 \text{ s.t. } Ax = b, \tag{1.5}$$

where the ℓ_0-norm $\|x\|_0 = \sum_{i=1}^n \mathbf{1}(x_i \neq 0)$ and $\mathbf{1}(\cdot)$ is equal to 1 if the argument is true, 0 otherwise. Recently, Candés, Romberg and Tao [7, 8, 9] and Donoho [15] have shown that, if the target signal x_0 is s-sparse, i.e. $\|x_0\|_0 = s$, the matrix $A \in \mathbb{R}^{q\times n}$ has $q = \mathcal{O}(s\log(n))$ and is chosen randomly according to a specific set of distributions, then, with very high probability, the sparse target signal x_0 is the unique optimal solution of the basis pursuit problem (1.4). Thus, x_0 can be recovered by solving a linear program (LP), and therefore, in theory, signal recovery is very efficient. In practice, however, simplex and interior point based general purpose LP solvers are unable to solve such LPs efficiently because the matrix A in (1.4) is large, dense, and often ill-conditioned. The measurement matrix A in many CS applications has a lot of structure, in particular, the matrix-vector multiplication Ax and A^Ty can be computed efficiently in $\mathcal{O}(n\log(n))$ time. Recently, a number of different algorithms have been proposed to exploit this structural fact to efficiently solve (1.4) [3, 2, 14, 21, 23, 24, 26, 38, 39, 41] . Note that the "noisy" basis pursuit problem

$$\min_{x\in\mathbb{R}^n} \|x\|_1 \text{ s.t. } \|Ax - b\|_2 \leq \delta,$$

is a special case of (1.1) with $\mathcal{Q} = \{y \in \mathbb{R}^q : \|y\|_2 \leq \delta\}$.

Principal component pursuit. The *principal component pursuit problem*

$$\min_{X\in\mathbb{R}^{m\times n}} \|X\|_* + \mu_2\|\mathbf{vec}(X) - d\|_1, \tag{1.6}$$

is a special case of (1.1) with $\mathcal{A} = \mathbf{0}$, $b = 0$, $\mathcal{Q} = \{\mathbf{0}\}$, $\mathcal{F}(X) = X$, $G = \mathbf{0}$, $\alpha = 1$, i.e. $\|\sigma(X)\|_\alpha = \|X\|_*$, $\beta = 1$, and $\mathcal{C}(X) = \mathbf{vec}(X)$, where $\mathbf{vec}(X)$ is the vector obtained by stacking the columns of $X \in \mathbb{R}^{m\times n}$ in order, and $\beta = 1$. Suppose the data matrix $D \in \mathbb{R}^{m\times n}$ is of the form $D = X_0 + S_0$, where X_0 is a low rank matrix, S_0 is a sparse matrix, and both satisfy some regularity conditions given in [10, 29]. Then the low rank and sparse components of D can be recovered by solving (1.6) with $d = \mathbf{vec}(D)$ and $\mu_2 = 1/\sqrt{\bar{n}}$, where $\bar{n} = \max\{m, n\}$ [10, 29]. For existing algorithmic approaches for solving principal component pursuit, see [10, 22, 29, 30, 42] and references therein.

In [42], it is shown that recovery is still possible even when the data matrix D is corrupted by a dense error matrix. Suppose the data matrix D is of the form $D = X_0 + S_0 + \zeta_0$, where $(\zeta_0)_{ij}$ is independent and identically distributed (i.i.d.) for all i, j such that $\|\zeta_0\|_F \leq \delta$. Then the optimal solution (X_*, S_*) of the *stable principal component pursuit problem*

$$\begin{array}{rl} \min_{X,S\in\mathbb{R}^{m\times n}} & \|X\|_* + \frac{1}{\sqrt{\bar{n}}}\|\mathbf{vec}(S)\|_1, \\ \text{s.t.} & \|X + S - D\|_F \leq \delta, \end{array} \tag{1.7}$$

satisfies $\|X_* - X_0\|_F^2 + \|S_* - S_0\|_F^2 \leq c\, mn\delta^2$ for some constant c with high probability. The stable principle component pursuit is a special case of (1.1) with $\mathcal{A}(X, S) = \mathbf{vec}(X + S)$, $b = \mathbf{vec}(D)$, $\mathcal{Q} = \{y \in \mathbb{R}^{mn} : \|y\|_2 \leq \delta\}$.

Composite norm minimization with conic constraints. The goal in the *minimal system realization problem* is to design the lowest order discrete-time, linear time-invariant (LTI) dynamical system that is consistent with the observed data. Let x_i be the true (unknown) impulse response of the system at time i for $i = 1, \ldots, n$. Suppose that we observe noisy data $\tilde{x}_i = x_i + \varepsilon_i$, $i = 1, \ldots, n$, where $\{\varepsilon_i\}$ are i.i.d. uniform over $[-\varrho, \varrho]$. It is well-known [17, 20] that the minimum order system consistent with the observations can be computed by solving

$$\min_{x\in\mathbb{R}^{2n-1}} \text{rank}(H_n(x)) \text{ s.t. } \|x_1^n - \tilde{x}_1^n\|_\infty \leq \delta,$$

where $H_n(x)$ be the Hankel matrix formed by $x \in \mathbb{R}^{2n-1}$ and x_1^n (resp. $\tilde{x}_1^n$) denotes the vector formed by the first n components of x (resp. $\tilde{x}$). This rank-minimization problem can be approximated by the nuclear norm minimization problem

$$\min_{x\in\mathbb{R}^{2n-1}} \|H_n(x)\|_* \text{ s.t. } \|x_1^n - \tilde{x}_1^n\|_\infty \leq \delta. \tag{1.8}$$

Algorithm APG$(p, f, \mathcal{S}, x^{(0)}, \text{ITERSTOP}, \text{GRADSTOP})$

$x_1^{(0)} \leftarrow x^{(0)}$, $x_2^{(1)} \leftarrow x^{(0)}$, $t^{(1)} \leftarrow 1$, $\ell \leftarrow 0$
while ITERSTOP(ℓ) **and** GRADSTOP$(x_1^{(\ell)})$ are **false do**
  $\ell \leftarrow \ell + 1$
  $x_1^{(\ell)} \leftarrow \operatorname{argmin}\left\{p(x) + \left\langle \nabla f(x_2^{(\ell)}),\ x - x_2^{(\ell)} \right\rangle + \frac{L}{2}\|x - x_2^{(\ell)}\| : x \in \mathcal{S}\right\}$
  $t^{(\ell+1)} \leftarrow \left(1 + \sqrt{1 + 4\ \left(t^{(\ell)}\right)^2}\right)/2$
  $x_2^{(\ell+1)} \leftarrow x_1^{(\ell)} + \left(\frac{t^{(\ell)}-1}{t^{(\ell+1)}}\right)\left(x_1^{(\ell)} - x_1^{(\ell-1)}\right)$
end while
return $x_1^{(\ell)}$

Fig. 2.1: Accelerated Proximal Gradient Algorithm

Since $H_n(.)$ is injective, $\mathcal{A}(x) = [x_1, \ldots, x_n]^T$ is surjective, and $\mathcal{Q} = \{y \in \mathbb{R}^n :\ \|y\|_\infty \leq \delta\}$ is a closed convex set, (1.8) is a special case of (1.1).

In the *sparse* PCA problem

$$\min_{x \in \mathbb{R}^n} \|\Sigma - xx^T\|_F \text{ s.t. } \|x\|_0 \leq s, \tag{1.9}$$

the goal is to compute an s-sparse vector x that is "close" to the eigenvector corresponding to the largest eigenvalue of the positive semidefinite matrix Σ. Let $X = xx^T$. Then (1.9) is equivalent to

$$\min_{X \in \mathbb{R}^{m \times n}} \|X - \Sigma\|_F \text{ s.t. } \|\,\mathbf{vec}(X)\|_0 \leq s^2, \text{rank}(X) = 1, X \succeq 0.$$

Since $\|X\|_*$ is the tightest convex upper bound for $\text{rank}(X)$, and $\|X\|_* = \mathbf{Tr}(X)$ for positive semidefinite (psd) matrices, the convex relaxation

$$\begin{array}{ll} \min_{X \in \mathbb{R}^{n \times n}} & \|X - \Sigma\|_F + \mu\|\,\mathbf{vec}(X)\|_1 + \nu \langle I, X\rangle, \\ \text{s.t.} & X \succeq 0, \end{array} \tag{1.10}$$

for (1.9), where μ and ν control the sparsity on the entries and the singular values of X, respectively, is a special case of (1.1) with $\mathcal{Q}$ set as the cone of psd matrices (the linear term in the objective function over psd cone can be handled easily). See [12, 13, 25] for existing approaches for solving the sparse PCA problem.

In Section 6, we show that FALC can be easily extended to solve problems of the form given in (1.2). When $F \not\equiv 0$, we do not require that $\mathcal{C}$ or $\mathcal{F}$ be injective. Thus, FALC is able to solve regularized conic optimization problems of the form

$$\min_{X \in \mathbb{R}^{m \times n}} \langle R, X\rangle + \rho\|X\|_F^2 \text{ s.t. } B \preceq \mathcal{A}(X),$$

where $\mathcal{A}$ is a surjective map.

This paper is organized as follows. In Section 3 we prove the main convergence results for FALC and in Section 4 we discuss all the implementation details. In Section 5 we report the results from our numerical experiments comparing FALC with other algorithms to solve principle component pursuit problems. Finally, in Section 6, we briefly discuss the general problem (1.2) and conclude.

2. Preliminaries. In this section we state and briefly discuss the details of a particular implementation of Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [4] that we use as a subroutine in FALC. Let $\mathcal{E}$ be a Hilbert space and $\|.\| = \sqrt{\langle ., .\rangle}$. FISTA computes an ϵ-optimal solution of

$$\min_{x \in \mathcal{E}} p(x) + f(x), \tag{2.1}$$

in $\mathcal{O}(1/\epsilon)$ iterations, where $p : \mathcal{E} \to \mathbb{R}$ and $f : \mathcal{E} \to \mathbb{R}$ are continuous convex functions such that ∇f is Lipschitz continuous on $\mathcal{E}$ with constant L. Later, Tseng [37] showed that this rate result for FISTA also holds when $p : \mathcal{E} \to (-\infty, +\infty]$ and $f : \mathcal{E} \to (-\infty, +\infty]$ that are proper, lower semicontinuous, and convex functions such that $\mathbf{dom}\, p$ is closed and ∇f is Lipschitz continuous on $\mathcal{E}$. Moreover, for any given convex

set $\mathcal{S} \subset \mathcal{E}$ such that $\mathcal{S} \cap \operatorname{argmin}_{x\in\mathcal{E}}\{p(x)+f(x)\} \neq \emptyset$, one can ensure that the iterate sequence lies in the convex set $\mathcal{S}$ and all the iterates are ϵ-optimal after $\mathcal{O}(1/\epsilon)$ iterations. We use this property later in the paper to uniformly bound the FALC iterates.

Algorithm APG displayed in Figure 2.1 takes as input the functions f and p, the convex set $\mathcal{S} \subset \mathcal{E}$ such that $\mathcal{S} \cap \operatorname{argmin}_{x\in\mathcal{E}}\{p(x)+f(x)\} \neq \emptyset$, an initial iterate $x^{(0)} \in \mathcal{E}$ and two stopping criteria ITERstop and GRADstop. **Algorithm APG** is the same as Algorithm 2 in [37] where we have set $\mathcal{X}_\ell \equiv \mathcal{S}$ (see [37] for details). Indeed, Algorithm 2 in [37] is a modification of FISTA [4] and reduces to FISTA when $\mathcal{S} = \mathcal{E}$. FISTA and Algorithm 2 in [37] do not use ITERstop and GRADstop – we include them in the definition of **Algorithm APG** because we terminate the algorithm early when we call it as a subroutine in FALC. ITERstop(ℓ) is a stopping criterion that only depends on the current iterate counter ℓ and GRADstop(x) is a stopping criterion that only depends on its argument x. Lemma 2.1 gives the iteration complexity of **Algorithm APG**.

LEMMA 2.1. *Let p and f be a proper, lower semicontinuous, convex functions such that* $\mathbf{dom}\, p$ *is closed and ∇f is Lipschitz continuous on $\mathcal{E}$ with constant L. Fix $\epsilon > 0$ and let $\{x_1^{(\ell)}, x_2^{(\ell)}\}_{\ell\in\mathbb{Z}_+}$ denote the* ***Algorithm APG*** *iterates when both* ITERstop *and* GRADstop *are disabled. Then $p(x_1^{(\ell)}) + f(x_1^{(\ell)}) \leq \min_{X\in\mathcal{E}}\{p(x)+f(x)\} + \epsilon$ whenever $\ell \geq \sqrt{\frac{2L}{\epsilon}}\, \|x^* - x^{(0)}\| - 1$, where $x^* \in \operatorname{argmin}_{x\in\mathcal{E}}\{p(x)+f(x)\}$.*

Proof. See Corollary 3 in [37] and Theorem 4.4 in [4] for the details of proof.□

Algorithm FALC$\left(\left\{\left(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)}\right)\right\}_{k\in\mathbb{Z}_+}, X^{(0)}\right)$

$y^{(0)} \leftarrow \mathcal{A}(X^{(0)}) - b, s^{(0)} \leftarrow \mathcal{C}(X^{(0)}) - d$
$\eta \leftarrow \mu_1\|\sigma(X^{(0)})\|_\alpha + \mu_2\|\mathcal{C}(X^{(0)}) - d\|_\beta$
$\theta_1^{(1)} \leftarrow 0$, $\theta_2^{(1)} \leftarrow 0$, $k \leftarrow 0$
while (FALCstop is **false**) **do**
  $k \leftarrow k+1$
  $p^{(k)}(X,s,y) := \lambda^{(k)}(\mu_1\|\sigma(X)\|_\alpha + \mu_2\|s\|_\beta)$
  $f^{(k)}(X,s,y) := \frac{1}{2}\|\mathcal{A}(X) - y - b - \lambda^{(k)}\theta_1^{(k)}\|_2^2 + \frac{1}{2}\|\mathcal{C}(X) - s - d - \lambda^{(k)}\theta_2^{(k)}\|_2^2$
  $\eta_1^{(k)} \leftarrow \eta + \frac{\lambda^{(k)}}{2}\left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2\right)$
  $\mathcal{S}^{(k)} := \{(X,s,y) \in \mathbb{R}^{m\times n} \times \mathbb{R}^p \times \mathbb{R}^q :\ \mu_1\|\sigma(X)\|_\alpha \leq \eta_1^{(k)},\ \mu_2\|s\|_\beta \leq \eta_1^{(k)}\}$
  ITERstop(ℓ) := $\{\ell \geq \ell_{\max}^{(k)}\}$, where $\ell_{\max}^{(k)}$ is defined in (3.12)
  GRADstop1$(X,s,y) := \left\{\exists (G,g) \in \partial_{X,s}P^{(k)}(.,.,.)|_{(X,s,y)} \text{ s.t. } \sqrt{\|G\|_F^2 + \|g\|_2^2} \leq \tau^{(k)}\right\}$
  GRADstop2$(X,s,y) := \left\{\|y - \Pi_{\mathcal{Q}}\left(y - \frac{1}{L}\nabla_y P^{(k)}(X,s,y)\right)\|_2 \leq \xi^{(k)}\right\}$
  GRADstop := GRADstop1 **and** GRADstop2
  $(X^{(k)}, s^{(k)}, y^{(k)}) \leftarrow$ **Algorithm APG**$(p^{(k)}, f^{(k)}, \mathcal{S}^{(k)}, \left(X^{(k-1)}, s^{(k-1)}, y^{(k-1)}\right), \text{ITERstop}, \text{GRADstop})$
  $\theta_1^{(k+1)} \leftarrow \theta_1^{(k)} - \frac{\mathcal{A}(X^{(k)}) - y^{(k)} - b}{\lambda^{(k)}}$
  $\theta_2^{(k+1)} \leftarrow \theta_2^{(k)} - \frac{\mathcal{C}(X^{(k)}) - s^{(k)} - d}{\lambda^{(k)}}$
end while
return $(X^{(k)}, s^{(k)}, y^{(k)})$

Fig. 3.1: Outline of First-Order Augmented Lagrangian Algorithm (FALC)

3. FALC Algorithm. For the sake of notational simplicity, we focus on the following simpler problem in this section.

$$\min_{X\in\mathbb{R}^{m\times n}} \mu_1\|\sigma(X)\|_\alpha + \mu_2\|\mathcal{C}(X) - d\|_\beta \text{ subject to } \mathcal{A}(X) - b \in \mathcal{Q}. \tag{3.1}$$

We give convergence results for FALC when $\mu_1 > 0$ and $\mu_2 > 0$. In Section 6 we briefly describe how to modify the algorithm to solve (1.2).

The linear maps $\mathcal{A}$ and $\mathcal{C}$ in (3.1) can be represented as $\mathcal{A}(X) = A\,\mathbf{vec}(X)$ and $\mathcal{C}(X) = C\,\mathbf{vec}(X)$, where $A \in \mathbb{R}^{q\times mn}$ and $C \in \mathbb{R}^{p\times mn}$. Let $\sigma_{\min}(A)$ and $\sigma_{\max}(A)$ denote the smallest and the largest singular values of A, respectively. Since we assume that $\mathcal{A}$ is surjective (see Assumption 1.1), A has full row rank; consequently, A^T has full column rank. We set $M := \begin{pmatrix} -I & 0 & C \\ 0 & -I & A \end{pmatrix}$ and $L = \sigma_{\max}^2(M)$. Let X_* denote an optimal solution of (3.1) and $\Pi_{\mathcal{Q}} : \mathbb{R}^q \to \mathcal{Q}$ denote the Euclidean projection onto $\mathcal{Q} \subset \mathbb{R}^q$.

To obtain separable and efficiently solvable subproblems, we introduce slack variables $s \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$, and reformulate (3.1) as

$$\begin{array}{rllllll} \min_{X,s,y} & \mu_1\|\sigma(X)\|_\alpha & + & \mu_2\|s\|_\beta, & & \\ \text{s.t.} & \mathcal{C}(X) & - & s & = & d, \\ & \mathcal{A}(X) & - & y & = & b, \\ & & & y & \in & \mathcal{Q}. \end{array} \tag{3.2}$$

We solve (3.2) by inexactly solving a sequence of optimization problems of the form

$$\min_{X\in\mathbb{R}^{m\times n},\ s\in\mathbb{R}^p,\ y\in\mathcal{Q}\subset\mathbb{R}^q} \left\{ \begin{array}{l} \lambda^{(k)}(\mu_1\|\sigma(X)\|_\alpha + \mu_2\|s\|_\beta) \\ -\lambda^{(k)}(\theta_1^{(k)})^T(\mathcal{A}(X)-y-b) + \frac{1}{2}\|\mathcal{A}(X)-y-b\|_2^2 \\ -\lambda^{(k)}(\theta_2^{(k)})^T(\mathcal{C}(X)-s-d) + \frac{1}{2}\|\mathcal{C}(X)-s-d\|_2^2 \end{array} \right\}, \tag{3.3}$$

for an appropriately chosen sequence $\{(\lambda^{(k)}, \theta_1^{(k)}, \theta_2^{(k)})\}_{k\in\mathbb{Z}_+}$. By completing squares, it is easy to see that (3.3) is equivalent to

$$\min_{X\in\mathbb{R}^{m\times n},\ s\in\mathbb{R}^p,\ y\in\mathcal{Q}\subset\mathbb{R}^q} P^{(k)}(X,s,y), \tag{3.4}$$

where

$$\begin{array}{rcl} P^{(k)}(X,s,y) & := & \lambda^{(k)}(\mu_1\|\sigma(X)\|_\alpha + \mu_2\|s\|_\beta) + f^{(k)}(X,s,y), \\ f^{(k)}(X,s,y) & := & \frac{1}{2}\|\mathcal{A}(X)-y-b-\lambda^{(k)}\theta_1^{(k)}\|_2^2 + \frac{1}{2}\|\mathcal{C}(X)-s-d-\lambda^{(k)}\theta_2^{(k)}\|_2^2. \end{array} \tag{3.5}$$

Algorithm FALC displayed in Figure 3.1 is the outline of Algorithm FALC. The algorithm takes as inputs the sequence $\left\{\left(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)}\right)\right\}_{k\in\mathbb{Z}_+}$ and a starting point $\left(X^{(0)}, s^{(0)}, y^{(0)}\right)$ such that $\mathcal{A}(X^{(0)})-b \in \mathcal{Q}$, $y^{(0)} := \mathcal{A}(X^{(0)})-b \in \mathcal{Q}$, and $s^{(0)} := \mathcal{C}(X^{(0)})-d$. In Section 4.3 we describe how we set the input sequence. Let $(X_*^{(k)}, s_*^{(k)}, y_*^{(k)})$ denote an optimal solution of (3.4).

3.1. Algorithm APG for subproblems. In each iteration of FALC, we call **Algorithm APG** displayed in Figure 2.1 to inexactly solve (3.4), which we call the "k-th subproblem". Let $\mathbf{1}_\mathcal{Q}$ denote the indicator function of the closed convex set $\mathcal{Q} \subset \mathbb{R}^q$, i.e., if $y \in \mathcal{Q}$, then $\mathbf{1}_\mathcal{Q}(y) = 0$; otherwise, $\mathbf{1}_\mathcal{Q}(y) = \infty$. $\mu_1\|\sigma(X)\|_\alpha + \mu_2\|s\|_\beta + \mathbf{1}_\mathcal{Q}(y)$ is a proper, lower semicontinuous (lsc), convex function of (X,s,y). Moreover, $f^{(k)}(X,s,y)$ is a proper, lsc, convex function that has a Lipschitz continuous gradient, $\nabla f^{(k)}$, defined on $\mathbb{R}^{m\times n} \times \mathbb{R}^p \times \mathbb{R}^q$ with Lipschitz constant equal to L for all $k \geq 1$. Thus, (3.4) is of the form described in (2.1). In each update step of **Algorithm APG** i.e. line 4 in Figure 2.1, we need to solve one problem of the form

$$\min_{(X,s,y)\in\mathcal{S}^{(k)}\ :\ y\in\mathcal{Q}} \left\{ \begin{array}{l} \lambda^{(k)}(\mu_1\|\sigma(X)\|_\alpha + \mu_2\|s\|_\beta) + \begin{bmatrix} \nabla_X f^{(k)}(\tilde{X},\tilde{s},\tilde{y}) \\ \nabla_s f^{(k)}(\tilde{X},\tilde{s},\tilde{y}) \\ \nabla_y f^{(k)}(\tilde{X},\tilde{s},\tilde{y}) \end{bmatrix}^T \begin{bmatrix} X-\tilde{X} \\ s-\tilde{s} \\ y-\tilde{y} \end{bmatrix} \\ \\ +\frac{L}{2}\|X-\tilde{X}\|_F^2 + \frac{L}{2}\|s-\tilde{s}\|_2^2 + \frac{L}{2}\|y-\tilde{y}\|_2^2 \end{array} \right\}, \tag{3.6}$$

for a given $(\tilde{X}, \tilde{s}, \tilde{y})$. Note that (3.6) is *separable* in X, s and y variables. Solving (3.6) reduces to one "constrained shrinkage" in $X \in \mathbb{R}^{m\times n}$, see (3.41); one "constrained shrinkage" in $s \in \mathbb{R}^p$, see (3.42); and one Euclidean projection onto $\mathcal{Q}$ in $y \in \mathbb{R}^q$.

3.2. Convex set $\mathcal{S}^{(k)}$ and the initial iterate for k-th subproblem. In the k-th FALC iteration, we solve (3.4) over the convex set $\mathcal{S}^{(k)}$ defined in Figure 3.1, using **Algorithm APG** starting from the initial iterate $(X^{(k-1)}, s^{(k-1)}, y^{(k-1)})$.

Let $\eta := \mu_1\|\sigma(X^{(0)})\|_\alpha + \mu_2\|\mathcal{C}(X^{(0)})-d\|_\beta$ and $\eta_1^{(k)} := \eta + \frac{\lambda^{(k)}}{2}\left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2\right)$. For all $k \geq 1$, since $(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) \in \operatorname{argmin}_{X\in\mathbb{R}^{m\times n},\ s\in\mathbb{R}^p,\ y\in\mathbb{R}^q}\{P^{(k)}(X,s,y) :\ y \in \mathcal{Q}\}$, we have

$$\mu_1\|\sigma(X_*^{(k)})\|_\alpha + \mu_2\|s_*^{(k)}\|_\beta \leq \frac{P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)})}{\lambda^{(k)}} \leq \frac{P^{(k)}(X^{(0)}, s^{(0)}, y^{(0)})}{\lambda^{(k)}} = \eta_1^{(k)}. \tag{3.7}$$

Above inequality ensures that $\mathcal{S}^{(k)} \cap \operatorname{argmin}\{P^{(k)}(X,s,y) : X \in \mathbb{R}^{m\times n},\ s \in \mathbb{R}^p,\ y \in \mathcal{Q}\} \neq \emptyset$.

3.3. ITERstop and GRADstop: Stopping criteria for Algorithm APG. Next, we discuss the stopping criteria set in Line 10 and Line 13 of Figure 3.1.

3.3.1. ITERstop. Let $\{(X_1^{(k,\ell)}, s_1^{(k,\ell)}, y_1^{(k,\ell)})\}_{\ell\in\mathbb{Z}_+}$ denote the sequence of x_1-iterates when **Algorithm APG** is called to solve the k-th sub-problem. For the sake of notational simplicity, let $h^{(k)}(X,s,y) := \|X - X^{(k-1)}\|_F^2 + \|s - s^{(k-1)}\|_2^2 + \|y - y^{(k-1)}\|_2^2$. Hence, Lemma 2.1 establishes that

$$P^{(k)}(X_1^{(k,\ell)}, s_1^{(k,\ell)}, y_1^{(k,\ell)}) \leq \inf_{X,s,y}\{P^{(k)}(X,s,y) :\ y\in\mathcal{Q}\} + \epsilon^{(k)} \ \text{ for } \ \ell \geq \sqrt{\frac{2Lh^{(k)}\left(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}\right)}{\epsilon^{(k)}}} - 1 \quad (3.8)$$

where $L = \sigma_{\max}^2(M)$ is the Lipschitz constant of $\nabla f^{(k)}$ for all $k \geq 1$. Triangular inequality implies that

$$\sqrt{h^{(k)}\left(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}\right)} \leq \|X_*^{(k)}\|_F + \|X^{(k-1)}\|_F + \|s_*^{(k)}\|_2 + \|s^{(k-1)}\|_2 + \|y_*^{(k)}\|_2 + \|y^{(k-1)}\|_2. \quad (3.9)$$

It is easy to show that

$$\frac{1}{I(\alpha^*)}\|\sigma(X)\|_\alpha \leq \|X\|_F \leq I(\alpha)\|\sigma(X)\|_\alpha, \qquad \frac{1}{J(\beta^*)}\|x\|_\beta \leq \|x\|_2 \leq J(\beta)\|x\|_\beta, \quad (3.10)$$

where

$$I(\alpha) = \begin{cases} \sqrt{\min\{m,n\}}, & \alpha = \infty, \\ 1, & \text{otherwise}, \end{cases} \qquad J(\beta) = \begin{cases} \sqrt{p}, & \beta = \infty, \\ 1, & \text{otherwise}, \end{cases} \quad (3.11)$$

and α^* (resp. β^*) denotes the Hölder conjugate of the α (resp. β), i.e., $\frac{1}{\alpha} + \frac{1}{\alpha^*} = 1$. In Lemma A.1 in Appendix A we show that

$$\eta_2^{(k)} := \sigma_{\max}(A)\frac{I(\alpha)}{\mu_1}\ \eta_1^{(k)} + \|b + \lambda^{(k)}\theta_1^{(k)}\|_F + 2\|\mathcal{A}(X^{(0)}) - b\|_2$$

is an upper bound on $\|y_*^{(k)}\|_2$. Note that when $\mathcal{Q}$ is a bounded set such that $\mathcal{Q} \subseteq \{y : \|y\|_2 \leq \eta_2\}$. Then, one can set $\eta_2^{(k)} := \eta_2$ for all $k \geq 1$. Let

$$\ell_{\max}^{(k)} := \sqrt{\frac{2L}{\epsilon^{(k)}}}\left[\left(\frac{I(\alpha)}{\mu_1} + \frac{J(\beta)}{\mu_2}\right)\eta_1^{(k)} + \eta_2^{(k)} + \|X^{(k-1)}\|_F + \|s^{(k-1)}\|_2 + \|y^{(k-1)}\|_2\right]. \quad (3.12)$$

Then, clearly $\ell_{\max}^{(k)}$ satisfies the following inequality

$$\sqrt{\frac{2Lh^{(k)}\left(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}\right)}{\epsilon^{(k)}}} \leq \ell_{\max}^{(k)}.$$

Thus, (3.8) implies that when **Algorithm APG** terminates due to ITERSTOP(ℓ), the iterate $(X_1^{(k,\ell)}, s_1^{(k,\ell)}, y_1^{(k,\ell)})$ is $\epsilon^{(k)}$-optimal.

3.3.2. GRADstop. The stopping condition GRADSTOP in Line 13 of Figure 3.1 is used to terminate the **Algorithm APG** when a certain set of perturbed first-order optimality conditions hold at the current iterate. Specifically, **Algorithm APG** stops according to GRADSTOP, when we have

$$\begin{array}{ll} (1) & \sqrt{\|G\|_F^2 + \|g\|_2^2} \leq \tau^{(k)}, \text{ for some } (G,g) \in \partial_{X,s}P^{(k)}(.,.,.)|_{(X^{(k)},s^{(k)},y^{(k)})} \\ (2) & \|y^{(k)} - \Pi_{\mathcal{Q}}\left(y^{(k)} - \frac{1}{L}\nabla_y P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\right)\|_2 \leq \xi^{(k)}, \end{array} \quad (3.13)$$

where

$$\begin{aligned} &\partial_{X,s}P^{(k)}(.,.,.)|_{(X^{(k)},s^{(k)},y^{(k)})} \\ &= \left\{(G,g) \in \mathbb{R}^{m\times n}\times\mathbb{R}^p : \begin{array}{ll} G & \in \lambda^{(k)}\mu_1\ \partial\|\sigma(.)\|_\alpha|_{X^{(k)}} + \nabla_X f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)}), \\ g & \in \lambda^{(k)}\mu_2\ \partial\|.\|_\beta|_{s^{(k)}} + \nabla_s f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)}), \end{array}\right\}. \end{aligned} \quad (3.14)$$

denotes the projection of $\partial P^{(k)}$ at $(X^{(k)}, s^{(k)}, y^{(k)})$ onto the X and s co-ordinates. Note that (3.13) would indeed be the first-order optimality conditions if $\tau^{(k)}$ and $\xi^{(k)}$ were both set to 0.

In our numerical experiments, we found that the calls to **Algorithm APG** were almost always terminated by the gradient-based stopping condition GRADSTOP. This suggests that relying on only ITERSTOP to terminate calls to **Algorithm APG** is likely to be very inefficient.

3.4. Convergence Results. Given $\epsilon > 0$, let $N_{\text{FALC}}(\epsilon)$ be the number of times **Algorithm APG** is called within **Algorithm FALC** until an ϵ-feasible and ϵ-optimal solution to (3.1) is found. During the k-th call, **Algorithm APG** inexactly solves the k-th subproblem (3.4). Let $N^{(k)}$ denote the number of iterations **Algorithm APG** needs until one of the stopping criteria ITERSTOP, GRADSTOP or FALCSTOP is met, where FALCSTOP is the stopping condition for **Algorithm FALC**. Finally, let N_{inner} be the total number **Algorithm APG** iterations until an ϵ-optimal and ϵ-feasible solution to (3.1) is computed, i.e. $N_{\text{inner}} = \sum_{k=1}^{N_{\text{FALC}}(\epsilon)} N^{(k)}$.

We begin by establishing bounds on the sequence of dual iterates $\{\theta_1^{(k)}\}_{k\in\mathbb{Z}_+}$ and $\{\theta_2^{(k)}\}_{k\in\mathbb{Z}_+}$. In order to establish this result, we need to bound the infeasibility of an $\epsilon^{(k)}$-optimal solution to the k-th sub-problem. In each iteration of FALC we solve a sub-problem of the form: $\min_{X,s,y}\{P(X,s,y) :\ y \in \mathcal{Q}\}$, where

$$P(X,s,y) = \lambda(\mu_1\|\sigma(X)\|_\alpha + \mu_2\|s\|_\beta) + \frac{1}{2}\|\mathcal{A}(X) - y - b - \lambda\theta_1\|_2^2 + \frac{1}{2}\|\mathcal{C}(X) - s - d - \lambda\theta_2\|_2^2.$$

Suppose $(\bar{X}, \bar{s}, \bar{y})$ is ϵ-optimal, i.e., $0 \le P(\bar{X}, \bar{s}, \bar{y}) - \min_{X,s,y}\{P(X,s,y) :\ y \in \mathcal{Q}\} \le \epsilon$. In Lemma A.2 in Appendix A we establish that

$$\begin{array}{rcl} \|\mathcal{C}(\bar{X}) - \bar{s} - d - \lambda\theta_2\|_2 & \le & J(\beta^*)\mu_2\lambda + \sigma_{\max}(M)\sqrt{2\epsilon}, \\ \|\mathcal{A}^*\left(\mathcal{A}(\bar{X}) - \bar{y} - b - \lambda\theta_1\right) + \mathcal{C}^*\left(\mathcal{C}(\bar{X}) - \bar{s} - d - \lambda\theta_2\right)\|_F & \le & I(\alpha^*)\mu_1\lambda + \sigma_{\max}(M)\sqrt{2\epsilon}, \end{array} \tag{3.15}$$

With this bound, we are now ready to show that the dual iterates are bounded.

LEMMA 3.1. *For all $k > 1$, the elements of $\{\theta_1^{(k)}\}_{k\in\mathbb{Z}_+}$ and $\{\theta_2^{(k)}\}_{k\in\mathbb{Z}_+}$ satisfy the following relation*

$$\|\theta_2^{(k)}\|_2 \le \max\left\{\sigma_{\max}(M)\sqrt{\frac{2\epsilon^{(k-1)}}{(\lambda^{(k-1)})^2}},\ \frac{\tau^{(k-1)}}{\lambda^{(k-1)}}\right\} + J(\beta^*)\mu_2, \tag{3.16a}$$

$$\|\theta_1^{(k)}\|_2 \le \frac{1}{\sigma_{\min}(A)}\left[\sigma_{\max}(C)\|\theta_2^{(k)}\|_2 + \max\left\{\sigma_{\max}(M)\sqrt{\frac{2\epsilon^{(k-1)}}{(\lambda^{(k-1)})^2}},\ \frac{\tau^{(k-1)}}{\lambda^{(k-1)}}\right\} + I(\alpha^*)\mu_1\right]. \tag{3.16b}$$

Proof. Consider the following two cases:

(a) *The k-th call to **Algorithm APG** terminates with* ITERSTOP: Since the iterate is $\epsilon^{(k)}$-optimal, the bound (3.15) implies that

$$\|\mathcal{C}(X^{(k)}) - s^{(k)} - d - \lambda^{(k)}\theta_2^{(k)}\|_2 \le J(\beta^*)\mu_2\lambda^{(k)} + \sigma_{\max}(M)\sqrt{2\epsilon^{(k)}}\ , \tag{3.17a}$$

$$\begin{aligned} &\|\mathcal{A}^*(\mathcal{A}(X^{(k)}) - y^{(k)} - b - \lambda^{(k)}\theta_1^{(k)}) + \mathcal{C}^*(\mathcal{C}(X^{(k)}) - s^{(k)} - d - \lambda^{(k)}\theta_2^{(k)})\|_F \\ &\le I(\alpha^*)\mu_1\lambda^{(k)} + \sigma_{\max}(M)\sqrt{2\epsilon^{(k)}}. \end{aligned} \tag{3.17b}$$

(b) *The k-th call to **Algorithm APG** terminates with* GRADSTOP: In this case, there exists $Q^{(k)} \in \partial\|\sigma(.)\|_\alpha|_{X^{(k)}}$ and $q^{(k)} \in \partial\|.\|_\beta|_{s^{(k)}}$ such that

$$\sqrt{\|\lambda^{(k)}\mu_1 Q^{(k)} + \nabla_X f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_F^2 + \|\lambda^{(k)}\mu_2 q^{(k)} + \nabla_s f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_2^2}\ \le\ \tau^{(k)}.$$

Since $\|q^{(k)}\|_{\beta^*} \le 1$ and $\|\sigma(Q^{(k)})\|_{\alpha^*} \le 1$, from the definition of $I(.)$ and $J(.)$ in (3.10), it follows that $\|\sigma(Q^{(k)})\|_F \le I(\alpha^*)$ and $\|q^{(k)}\|_2 \le J(\beta^*)$. Then we have

$$\|\mathcal{C}(X^{(k)}) - s^{(k)} - d - \lambda^{(k)}\theta_2^{(k)}\|_2 \le J(\beta^*)\mu_2\lambda^{(k)} + \tau^{(k)}, \tag{3.18a}$$

$$\begin{aligned} &\|\mathcal{A}^*(\mathcal{A}(X^{(k)}) - y^{(k)} - b - \lambda^{(k)}\theta_1^{(k)}) + \mathcal{C}^*(\mathcal{C}(X^{(k)}) - s^{(k)} - d - \lambda^{(k)}\theta_2^{(k)})\|_F \\ &\le\ I(\alpha^*)\mu_1\lambda^{(k)} + \tau^{(k)}. \end{aligned} \tag{3.18b}$$

Thus, combining (3.17) and (3.18), and using triangular inequality it follows that for all $k \geq 1$

$$\|\mathcal{C}(X^{(k)}) - s^{(k)} - d - \lambda^{(k)}\theta_2^{(k)}\|_2 \leq J(\beta^*)\mu_2\lambda^{(k)} + \max\left\{\sigma_{\max}(M)\sqrt{2\epsilon^{(k)}}, \tau^{(k)}\right\}, \tag{3.19a}$$

$$\begin{aligned}\|\mathcal{A}^*(\mathcal{A}(X^{(k)}) - y^{(k)} - b - \lambda^{(k)}\theta_1^{(k)})\|_F \leq\ & I(\alpha^*)\mu_1\lambda^{(k)} + \max\left\{\sigma_{\max}(M)\sqrt{2\epsilon^{(k)}}, \tau^{(k)}\right\} \\ & + \|\mathcal{C}^*(\mathcal{C}(X^{(k)}) - s^{(k)} - d - \lambda^{(k)}\theta_2^{(k)})\|_F.\end{aligned} \tag{3.19b}$$

Since $\theta_1^{(k+1)} = \theta_1^{(k)} - \frac{\mathcal{A}(X^{(k)})-y^{(k)}-b}{\lambda^{(k)}}$ and $\theta_2^{(k+1)} = \theta_2^{(k)} - \frac{\mathcal{C}(X^{(k)})-s^{(k)}-d}{\lambda^{(k)}}$, (3.19) is obtained by dividing (3.16) into $\lambda^{(k)}$ and using the fact that Assumption 1.1 implies that A has full row rank, i.e. $\sigma_{\min}(A) > 0$. Thus, $\left\{(\theta_1^{(k)}, \theta_2^{(k)})\right\}_{k\in\mathbb{Z}}$ satisfies (3.16). □

Next, we establish that the FALC iterate sequence $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k\in\mathbb{Z}_+}$ is bounded.

LEMMA 3.2. *Let $(X_*^{(k)}, s_*^{(k)}, y_*^{(k)})$ be an optimal solution to (3.4) and let $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k\in\mathbb{Z}_+}$ denote the sequence of FALC iterates corresponding to a parameter sequence $\{(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)})\}_{k\in\mathbb{Z}_+}$ such that*

(i) penalty multiplier, $\lambda^{(k)} \searrow 0$,

(ii) approximate optimality parameter, $\epsilon^{(k)} \searrow 0$ such that $\frac{\epsilon^{(k)}}{(\lambda^{(k)})^2} \leq B$ for all $k \geq 1$ for some $B > 0$,

(iii) subgradient tolerance parameters, $\tau^{(k)} \searrow 0$ and $\xi^{(k)} \searrow 0$ such that $\frac{\tau^{(k)}}{\lambda^{(k)}} \to 0$ and $\frac{\xi^{(k)}}{\lambda^{(k)}} \to 0$ as $k \to \infty$.

Then $\{(X_*^{(k)}, s_*^{(k)}, y_*^{(k)})\}_{k\in\mathbb{Z}_+}$ and $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k\in\mathbb{Z}_+}$ are bounded sequences.

Proof. In the k-th FALC iteration, the call to **Algorithm APG** terminates in at most $\ell_{\max}^{(k)}$ iterations. Since $\ell_{\max}^{(k)}$ is finite for all $k \geq 1$, the sequence $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k\in\mathbb{Z}_+}$ exists. In order to show $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k\in\mathbb{Z}_+}$ is a bounded sequence, we first establish that $\{\theta_1^{(k)}\}_{k\in\mathbb{Z}_+}$ and $\{\theta_2^{(k)}\}_{k\in\mathbb{Z}_+}$ are bounded sequences.

Because $\frac{\epsilon^{(k)}}{(\lambda^{(k)})^2} \leq B$ and $\frac{\tau^{(k)}}{\lambda^{(k)}} \to 0$, (3.16) implies that there exist constants $B_{\theta_1} > 0$ and $B_{\theta_2} > 0$ such that

$$\max_{k\geq 1}\{\|\theta_1^{(k)}\|_2\} \leq B_{\theta_1}, \text{ and } \max_{k\geq 1}\{\|\theta_2^{(k)}\|_2\} \leq B_{\theta_2}. \tag{3.20}$$

From (3.20), it follows that for $i = 1, 2$,

$$\lim_{k\to\infty} \lambda^{(k)}\theta_i^{(k)} = 0, \quad \text{and} \quad \lim_{k\to\infty} \lambda^{(k)}\|\theta_i^{(k)}\|_2^2 = 0. \tag{3.21}$$

Also, $\frac{\epsilon^{(k)}}{(\lambda^{(k)})^2} \leq B$ for all $k \geq 1$ implies that $\lim_{k\to\infty} \frac{\epsilon^{(k)}}{\lambda^{(k)}} = 0$.

Now we can prove that the iterate sequence is bounded. Trivially, the choice of $\mathcal{S}^{(k)}$ ensures that $\mu_1 \max\{\|\sigma(X^{(k)})\|_\alpha,\ \|\sigma(X_*^{(k)})\|_\alpha\} \leq \eta_1^{(k)}$ and $\mu_2 \max\{\|s^{(k)}\|_\beta, \|s_*^{(k)}\|_\beta\} \leq \eta_1^{(k)}$. From the definition of $\eta_1^{(k)}$ in Line 8 of Figure 3.1 and (3.20), it follows that for all $k \geq 1$

$$\eta_1^{(k)} \leq \eta + \lambda^{(k)}\left(\frac{B_{\theta_1}^2 + B_{\theta_2}^2}{2}\right) \leq \eta + \lambda^{(1)}\left(\frac{B_{\theta_1}^2 + B_{\theta_2}^2}{2}\right) := B_{\eta_1}. \tag{3.22}$$

Hence, for all $k \geq 1$,

$$\mu_1 \max\{\|\sigma(X^{(k)})\|_\alpha,\ \|\sigma(X_*^{(k)})\|_\alpha\} \leq B_{\eta_1} \text{ and } \ \mu_2 \max\{\|s^{(k)}\|_\beta, \|s_*^{(k)}\|_\beta\} \leq B_{\eta_1}. \tag{3.23}$$

Next we show that $\{y^{(k)}\}_{k\in\mathbb{Z}_+}$ and $\{y_*^{(k)}\}_{k\in\mathbb{Z}_+}$ are bounded. From the definition of $\theta_1^{(k+1)}$ in Line 15 of Figure 3.1, we have $y^{(k)} = \lambda^{(k)}(\theta_1^{(k+1)} - \theta_1^{(k)}) + \mathcal{A}(X^{(k)}) - b$ for all $k \geq 1$. Hence, $\|y^{(k)}\|_2 \leq \lambda^{(k)}\|(\theta_1^{(k+1)} - \theta_1^{(k)})\|_2 + \|\mathcal{A}(X^{(k)}) - b\|_2$ for all $k \geq 1$. From (3.20) and (3.23), it follows that there exists $B_y > 0$ such that

$$\|y^{(k)}\|_2 \leq B_y, \quad \forall\ k \geq 1. \tag{3.24}$$

Moreover, Lemma A.1 in Appendix A guarantees that for all $k \geq 1$, $\|y_*^{(k)}\|_2 \leq \eta_2^{(k)}$, where $\eta_2^{(k)}$ is given in (A.3) in Appendix A. Since $\eta_1^{(k)} \leq B_{\eta_1}$ for all $k \geq 1$, (3.21) and (A.3) imply that there exists a constant $B_{\eta_2} > 0$ such that

$$\|y_*^{(k)}\|_2 \leq B_{\eta_2}, \quad \forall\ k \geq 1. \tag{3.25}$$

□

THEOREM 3.3. *Let $\mathcal{X} = \{X^{(k)}\}_{k\in\mathbb{Z}_+}$ denote the FALC iterate sequence corresponding to a parameter sequence $\{(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)})\}_{k\in\mathbb{Z}}$ satisfying the conditions in Lemma 3.2. Then any limit point $\bar{X}$ of the sequence $\mathcal{X}$ is an optimal solution of the composite norm minimization problem (3.1).*

Proof. Since Lemma 3.2 guarantees that $\mathcal{X}$ is a bounded sequence, there exists a subsequence $\mathcal{K} \subset \mathbb{Z}_+$ such that $\lim_{k\in\mathcal{K}} X^{(k)} = \bar{X}$ exists. We have previously shown that $\lim_{k\to\infty} \lambda^{(k)}\theta_i^{(k)} = 0$ for $i \in \{1, 2\}$. Hence, (3.19a) guarantees that $\lim_{k\in\mathcal{K}} s^{(k)} = \bar{s}$ exists; similarly (3.19b) and the full row-rank assumption on A together guarantee that $\lim_{k\in\mathcal{K}} y^{(k)} = \bar{y}$ exists. Then, taking the limit of both sides of (3.19a) for $k \in \mathcal{K}$, we have $\|\mathcal{C}(\bar{X}) - \bar{s} - d\|_2 \leq 0$, i.e. $\bar{s} = \mathcal{C}(\bar{X}) - d$. Moreover, taking the limit of both sides of (3.19b) for $k \in \mathcal{K}$ and using the fact that $\bar{s} = \mathcal{C}(\bar{X}) - d$, we have $\|\mathcal{A}^*(\mathcal{A}(\bar{X}) - \bar{y} - b)\|_2 \leq 0$. Since A has full row rank and $y^{(k)} \in \mathcal{Q}$ for all $k \geq 1$, we have

$$\mathcal{A}(\bar{X}) - b = \bar{y}, \qquad \bar{y} \in \mathcal{Q}. \tag{3.26}$$

Therefore, we can conclude that $\bar{X}$ is feasible, i.e. $\mathcal{A}(\bar{X}) - b \in \mathcal{Q}$.

In the rest of the proof, we will show that $\bar{X} \in \operatorname{argmin}\{\mu_1\|\sigma(X)\|_\alpha + \mu_2\|\mathcal{C}(X) - d\|_\beta :\ \mathcal{A}(X) - b \in \mathcal{Q}\}$. We consider the following two cases:

(a) There exists a subsequence $\mathcal{K}_a \subset \mathcal{K}$ such that for all $k \in \mathcal{K}_a$, **Algorithm APG** terminates with ITERSTOP; hence, the iterate $(X^{(k)}, s^{(k)}, y^{(k)})$ computed in Step 14 of FALC satisfies

$$0 \leq P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)}) - P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) \leq \epsilon^{(k)} \qquad \forall k \in \mathcal{K}_a. \tag{3.27}$$

Fix $X_* \in \operatorname{argmin}_{X\in\mathbb{R}^{m\times n}}\{\mu_1\|\sigma(X)\|_\alpha + \mu_2\|\mathcal{C}(X) - d\|_\beta :\ \mathcal{A}(X) - b \in \mathcal{Q}\}$, let $s_* := \mathcal{C}(X_*) - d$ and $y_* := \mathcal{A}(X_*) - b$. Since $(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) \in \operatorname{argmin}_{X\in\mathbb{R}^{m\times n},\ s\in\mathbb{R}^p,\ y\in\mathbb{R}^q}\{P^{(k)}(X, s, y) :\ y \in \mathcal{Q}\}$, it follows that $P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) \leq P^{(k)}(X_*, s_*, y_*)$ for $k \geq 1$. Thus, (3.27) implies that $P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)}) \leq P^{(k)}(X_*, s_*, y_*) + \epsilon^{(k)}$. Hence, for all $k \in \mathcal{K}_a$,

$$\begin{aligned}\mu_1\|\sigma(X^{(k)})\|_\alpha + \mu_2\|s^{(k)}\|_\beta &\leq \frac{P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})}{\lambda^{(k)}} \ \leq \frac{P^{(k)}(X_*, s_*, y_*) + \epsilon^{(k)}}{\lambda^{(k)}},\\ &= \mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|\mathcal{C}(X_*) - d\|_\beta + \frac{\lambda^{(k)}}{2}\left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2\right) + \frac{\epsilon^{(k)}}{\lambda^{(k)}}.\end{aligned} \tag{3.28}$$

Taking the limit of both sides of (3.28) along the subsequence $\mathcal{K}_a$, and using the fact that $\bar{s} = \mathcal{C}(\bar{X}) - d$, we get

$$\begin{aligned}\mu_1\|\sigma(\bar{X})\|_\alpha + \mu_2\|\mathcal{C}(\bar{X}) - d\|_\beta &= \lim_{k\in\mathcal{K}_a} \mu_1\|\sigma(X^{(k)})\|_\alpha + \mu_2\|s^{(k)}\|_\beta,\\ &\leq \mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|\mathcal{C}(X_*) - d\|_\beta\\ &\quad + \lim_{k\in\mathcal{K}_a}\left\{\frac{\lambda^{(k)}}{2}\left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2\right) + \frac{\epsilon^{(k)}}{\lambda^{(k)}}\right\},\\ &= \mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|\mathcal{C}(X_*) - d\|_\beta,\end{aligned} \tag{3.29}$$

where (3.29) follows from the fact that $\{\theta_i^{(k)}\}$ is uniformly bounded for $i \in \{1, 2\}$, $\lambda^{(k)} \to 0$, and $\epsilon^{(k)}/\lambda^{(k)} \to 0$. Thus, from (3.26), (3.29) and the fact that X_* is optimal, it follows that $\bar{X}$ is also an optimal solution for the composite norm minimization problem (3.1).

(b) There exists $K \in \mathcal{K}$ such that, for all $k \in \mathcal{K}_b := \mathcal{K} \cap \{k \geq K\}$, **Algorithm APG** terminates with GRADSTOP; hence, $(X^{(k)}, s^{(k)}, y^{(k)})$ satisfies (3.13).

For all $k \in \mathcal{K}_b$, there exist $Q^{(k)} \in \partial\|\sigma(.)\|_\alpha|_{X^{(k)}}$ and $q^{(k)} \in \partial\|.\|_\beta|_{s^{(k)}}$ such that (3.13) holds. Hence, we have

$$\|\lambda^{(k)}\mu_2 q^{(k)} + \nabla_s f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_2 \leq \tau^{(k)}, \tag{3.30a}$$

$$\|\lambda^{(k)}\mu_1 Q^{(k)} + \nabla_X f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_F \leq \tau^{(k)}, \tag{3.30b}$$

$$\|y^{(k)} - \Pi_{\mathcal{Q}}\left(y^{(k)} - \frac{1}{L}\nabla_y f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\right)\|_2 \leq \xi^{(k)}. \tag{3.30c}$$

For all $k \in \mathcal{K}_b$, $Q^{(k)} \in \partial\|\sigma(.)\|_\alpha|_{X^{(k)}}$ and $q^{(k)} \in \partial\|.\|_\beta|_{s^{(k)}}$, therefore, $\|\sigma(Q^{(k)})\|_{\alpha^*} \leq 1$ and $\|q^{(k)}\|_{\beta^*} \leq 1$. Hence, there exists a subsequence $\mathcal{K}'_b \subset \mathcal{K}_b$ such that $\lim_{k\in\mathcal{K}'_b}(Q^{(k)}, q^{(k)}) = (\bar{Q}, \bar{q})$ exists. One can easily show that $\bar{Q} \in \partial\|\sigma(.)\|_\alpha|_{\bar{X}}$ and $\bar{q} \in \partial\|.\|_\beta|_{\bar{s}}$. Dividing both sides of (3.30a) by $\lambda^{(k)}$, we get

$$\|\mu_2 q^{(k)} + \theta_2^{(k+1)}\|_2 \leq \frac{\tau^{(k)}}{\lambda^{(k)}}, \tag{3.31}$$

for all $k \in \mathcal{K}_b \supset \mathcal{K}'_b$. Since $\lim_{k\in\mathcal{K}'_b} q^{(k)} = \bar{q}$ and $\lim_{k\in\mathbb{Z}_+} \frac{\tau^{(k)}}{\lambda^{(k)}} = 0$, it follows that $\lim_{k\in\mathcal{K}'_b} \theta_2^{(k+1)} = \bar{\theta}_2$ exists and taking the limit of both sides of (3.31) along $k \in \mathcal{K}'_b$, we have

$$\bar{\theta}_2 = -\mu_2\bar{q}.$$

Dividing both sides of (3.30b) by $\lambda^{(k)}$, we get

$$\|\mu_1 Q^{(k)} - \mathcal{A}^*(\theta_1^{(k+1)}) - \mathcal{C}^*(\theta_2^{(k+1)})\|_F \leq \frac{\tau^{(k)}}{\lambda^{(k)}}, \tag{3.32}$$

for all $k \in \mathcal{K}_b \supset \mathcal{K}'_b$. Since $\lim_{k\in\mathcal{K}'_b} Q^{(k)} = \bar{Q}$, $\lim_{k\in\mathbb{Z}_+} \frac{\tau^{(k)}}{\lambda^{(k)}} = 0$ and A has full row rank, it follows that $\lim_{k\in\mathcal{K}'_b} \theta_1^{(k+1)} = \bar{\theta}_1$ exists and taking the limit of both sides of (3.32) along $k \in \mathcal{K}'_b$, we have $\mu_1\bar{Q} + \mu_2\mathcal{C}^*(\bar{q}) = \mathcal{A}^*(\bar{\theta}_1)$. Note that $\bar{q} \in \partial\|.\|_\beta|_{\bar{s}}$ and $\bar{s} = \mathcal{C}(\bar{X}) - d$. Hence, $\mathcal{C}^*(\bar{q}) \in \partial\|\mathcal{C}(.) - d\|_\beta|_{\bar{X}}$ and we have

$$\mathcal{A}^*(\bar{\theta}_1) = G_* \quad \text{and} \quad G_* \in \partial\mu_1\|\sigma(.)\|_\alpha + \mu_2\|\mathcal{C}(.) - d\|_\beta|_{\bar{X}}, \tag{3.33}$$

where $G_* := \mu_1\bar{Q} + \mu_2\mathcal{C}^*(\bar{q})$.
Let $y_p^{(k)} := y^{(k)} - \frac{1}{L}\nabla_y f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})$ denote the gradient step. Since $\xi^{(k)} \searrow 0$, taking the limit of both sides of (3.30c) along $k \in \mathcal{K}'_b$, we get

$$\bar{y} = \lim_{k\in\mathcal{K}'_b} y^{(k)} = \lim_{k\in\mathcal{K}'_b} \Pi_{\mathcal{Q}}(y_p^{(k)}) = \Pi_{\mathcal{Q}}(\lim_{k\in\mathcal{K}'_b} y_p^{(k)}), \tag{3.34}$$

where the third equality follows from the fact that Euclidean projection $\Pi_{\mathcal{Q}}(.)$ is continuous when $\mathcal{Q}$ is a nonempty, closed, convex set; and $\lim_{k\in\mathcal{K}'_b} y_p^{(k)}$ exists since $\nabla f^{(k)}$ is Lipschitz continuous.
Dividing both sides of (3.30c) by $\lambda^{(k)}$ and taking the limit along $k \in \mathcal{K}'_b$, we get

$$\lim_{k\in\mathcal{K}'_b} \|y^{(k)}/\lambda^{(k)} - \Pi_{\mathcal{Q}/\lambda^{(k)}}(y_p^{(k)}/\lambda^{(k)})\|_2 = \lim_{k\in\mathcal{K}'_b} \|y^{(k)}/\lambda^{(k)} - \Pi_{\mathcal{Q}}(y_p^{(k)})/\lambda^{(k)}\|_2 = 0, \tag{3.35}$$

where the first equality follows from Lemma A.4.
For all $k \in \mathcal{K}_b$, (3.36) follows from the definition of Euclidean projection:

$$\left\langle \Pi_{\mathcal{Q}/\lambda^{(k)}}(y_p^{(k)}/\lambda^{(k)}) - y_p^{(k)}/\lambda^{(k)},\ y/\lambda^{(k)} - \Pi_{\mathcal{Q}/\lambda^{(k)}}(y_p^{(k)}/\lambda^{(k)}) \right\rangle \geq 0, \quad \forall y \in \mathcal{Q} \tag{3.36}$$

Since $y_p^{(k)}/\lambda^{(k)} = y^{(k)}/\lambda^{(k)} - \theta_1^{(k+1)}/L$, multiplying the second term of the inner product in (3.36) by $\lambda^{(k)}$ and using Lemma A.4, it follows that for all $k \in \mathcal{K}_b$

$$\left\langle \Pi_{\mathcal{Q}/\lambda^{(k)}}(y_p^{(k)}/\lambda^{(k)}) - y^{(k)}/\lambda^{(k)} + \theta_1^{(k+1)}/L,\ y - \Pi_{\mathcal{Q}}(y_p^{(k)}) \right\rangle \geq 0, \quad \forall y \in \mathcal{Q} \tag{3.37}$$

Since $\lim_{k\in\mathcal{K}_b'} \theta_1^{(k+1)} = \bar{\theta}_1$, taking the limit of both sides of (3.37) along $k \in \mathcal{K}_b' \subset \mathcal{K}_b$ and using (3.35), we have

$$\left\langle \bar{\theta}_1,\ y - \Pi_{\mathcal{Q}}(\lim_{k\in\mathcal{K}_b'} y_p^{(k)}) \right\rangle \geq 0, \quad \forall y \in \mathcal{Q}.$$

Thus, it follows from (3.34) and above inequality that

$$\langle \bar{\theta}_1,\ y - \bar{y} \rangle \geq 0 \qquad \forall y \in \mathcal{Q}. \tag{3.38}$$

Consequently, (3.33) and (3.38) together imply that $(\bar{X}, \bar{y})$ satisfies the first order optimality conditions of the relaxed problem (3.39).

$$\min_{X\in\mathbb{R}^{m\times n}, y\in\mathbb{R}^q} \Big\{ \mu_1\|\sigma(X)\|_\alpha + \mu_2\|\mathcal{C}(X) - d\|_\beta - (\bar{\theta}_1)^T(\mathcal{A}(X) - y - b) :\ y \in \mathcal{Q} \Big\}. \tag{3.39}$$

Since (3.39) is convex, it follows that $(\bar{X}, \bar{y})$ is an optimal solution to the relaxed problem (3.39). Moreover, from (3.26), $(\bar{X}, \bar{y})$ is feasible to the composite norm minimization problem, i.e. $\min\{\mu_1\|X\|_\alpha + \mu_2\|\mathcal{C}(X) - d\|_\beta :\ \mathcal{A}(X) - y = b, y \in \mathcal{Q}\}$. Therefore, $\bar{X} \in \operatorname{argmin}\{\mu_1\|X\|_\alpha + \mu_2\|\mathcal{C}(X) - d\|_\beta :\ \mathcal{A}(X) - b \in \mathcal{Q}\}$.

□

Clearly, when (3.1) has a unique solution, the FALC iterates converge to this unique solution.

COROLLARY 3.4. *Suppose the composite norm minimization problem* (3.1) *has a* unique *optimal solution* X_*. *Let* $\{X^{(k)}\}_{k\in\mathbb{Z}_+}$ *denote the sequence of FALC iterates corresponding to a parameter sequence* $\{(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)})\}_{k\in\mathbb{Z}}$ *satisfying the conditions in Lemma 3.2. Then* $\lim_{k\to\infty} X^{(k)} = X_*$.

For most sparse optimization problems such as basis pursuit and the matrix completion problems, the unknown signal can be recovered only if the corresponding convex relaxation has a unique solution. Additionally, when the set of constraints for the basis pursuit or affine rank minimization problems are defined by randomly generated Gaussian matrices, the unknown target signal is, with very high probability, the unique solution to these optimization problems.

We next establish a bound on the iteration complexity of FALC. In Lemma 3.5 we prove a uniform bound on the number of **Algorithm APG** iterations required to inexactly solve any subproblem encountered in FALC.

LEMMA 3.5. *Suppose the parameter sequence* $\{(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)})\}_{k\in\mathbb{Z}_+}$ *satisfies all the conditions in Lemma 3.2. Then there exists constant* $\mathcal{N}$ *such that, for all* $k \geq 1$, *the number of iterations* $N^{(k)}$ *required by* ***Algorithm APG*** *to compute* $(X^{(k)}, s^{(k)}, y^{(k)})$ *satisfies*

$$N^{(k)} \leq \frac{\mathcal{N}}{\sqrt{\epsilon^{(k)}}}. \tag{3.40}$$

Proof. The number of interations $N^{(k)} \leq \ell_{\max}^{(k)}$, where $\ell_{\max}^{(k)}$ denotes the number of iterations required to satisfy ITERSTOP. Since $(X^{(k)}, s^{(k)}, y^{(k)}) \in \mathcal{S}^{(k)}$, (3.12) and (3.10) imply that

$$\begin{aligned}
\ell_{\max}^{(k)} &\leq \sqrt{\frac{2L}{\epsilon^{(k)}}} \left[\left(\frac{I(\alpha)}{\mu_1} + \frac{J(\beta)}{\mu_2} \right) (\eta_1^{(k)} + \eta_1^{(k-1)}) + \eta_2^{(k)} + \|y^{(k-1)}\|_2 \right], \\
&\leq \sqrt{\frac{8L}{\epsilon^{(k)}}} \left[\left(\frac{I(\alpha)}{\mu_1} + \frac{J(\beta)}{\mu_2} \right) B_{\eta_1} + \frac{B_{\eta_2} + B_y}{2} \right] := \mathcal{N}\ \frac{1}{\sqrt{\epsilon^{(k)}}},
\end{aligned}$$

where the second inequality follows from (3.22), (3.24) and (3.25). □

In each iteration of **Algorithm APG** we need to solve one instance of each of the following problems.

(a) One *constrained matrix shrinkage problem* of the form

$$\min_{X\in\mathbb{R}^{m\times n}} \Big\{ \lambda\|\sigma(X)\|_\alpha + \frac{1}{2}\|X - \tilde{X}\|_F^2 : \|\sigma(X)\|_\alpha \leq \tilde{\eta} \Big\} \tag{3.41}$$

for a given $\tilde{X} \in \mathbb{R}^{m\times n}$ and $\tilde{\eta} > 0$. When $\alpha \in \{1, \infty\}$ the worst-case complexity of computing a solution to (3.41) is the same as that of computing a full SVD, i.e. $\mathcal{O}(\min\{nm^2, n^2m\})$, and when $\alpha = 2$, the worst-case complexity is $\mathcal{O}(mn)$. See Lemma B.2 in Appendix B for details. Exact SVD computation is not necessary – inexactly computing the SVD only adds a small additional error to (3.8).

(b) One *constrained vector shrinkage problem* of the form

$$\min_{s\in\mathbb{R}^p} \left\{ \lambda\|s\|_\beta + \frac{1}{2}\|s - \tilde{s}\|_2^2 : \|s\|_\beta \leq \tilde{\eta} \right\} \tag{3.42}$$

for a given $\tilde{s} \in \mathbb{R}^p$ and $\tilde{\eta} > 0$. The complexity of solving the vector shrinkage problem is $\mathcal{O}(p\log(p))$ when $\beta \in \{1, \infty\}$ and $\mathcal{O}(p)$ when $\beta = 2$. See Lemma B.2 in Appendix B.

(c) One *Euclidean projection problem* of the form

$$\min_{y\in\mathbb{R}^q} \{ \frac{1}{2}\|y - \tilde{y}\|_2^2 : y \in \mathcal{Q} \} \tag{3.43}$$

for a given $\tilde{y} \in \mathbb{R}^q$. The complexity of solving the Euclidean projection problem depends on $\mathcal{Q}$.

In Theorem 3.6 we establish bounds on the infeasibility and sub-optimality of the FALC iterate. This result leads to a convergence rate result in Theorem 3.7.

THEOREM 3.6. *Let $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k\in\mathbb{Z}_+}$ denote the sequence of FALC iterates corresponding to a parameter sequence $\{(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)})\}_{k\in\mathbb{Z}_+}$ satisfying the conditions in Lemma 3.2. In addition, suppose that, for all $k \geq 1$, $\tau^{(k)} = \kappa_1\epsilon^{(k)}$ and $\xi^{(k)} = \kappa_2\epsilon^{(k)}$ for some $\kappa_i \in (0, 1)$ $i = 1, 2$. Then there exist positive constants c_j, $j = 1, \ldots, 3$, such that for all $k \geq 1$,*

(i) $y^{(k)} \in \mathcal{Q}$ such that $\|\mathcal{A}(X^{(k)}) - y^{(k)} - b\|_2 \leq c_1\lambda^{(k)}$,
(ii) $\left|\left(\mu_1\|\sigma(X^{(k)})\|_\alpha + \mu_2\|\mathcal{C}(X^{(k)}) - d\|_\beta\right) - P^*\right| \leq c_2\lambda^{(k)} + c_3\sqrt{\epsilon^{(k)}}$,

where P^* denotes the optimal value of (3.1).

Proof. For all parameter sequences $\{(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)})\}_{k\in\mathbb{Z}_+}$ satisfying the conditions in Lemma 3.2, we show in (3.20) that $\|\theta_i^{(k)}\|_2 \leq B_{\theta_i}$ for $i \in \{1, 2\}$. Therefore,

$$\begin{aligned} \|\mathcal{A}(X^{(k)}) - y^{(k)} - b\|_2 &\leq \|\mathcal{A}(X^{(k)}) - y^{(k)} - b - \lambda^{(k)}\theta_1^{(k)}\|_2 + \lambda^{(k)}\|\theta_1^{(k)}\|_2, \\ &= \lambda^{(k)}\|\theta_1^{(k+1)}\|_2 + \lambda^{(k)}\|\theta_1^{(k)}\|_2, \\ &\leq 2B_{\theta_1}\lambda^{(k)}. \end{aligned}$$

This establishes (i).

In the rest of the proof, we establish (ii). Let (X_*, s_*, y_*) denote any optimal solution of (3.1), i.e. $P^* = \mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|\mathcal{C}(X_*) - d\|_\beta$. In (3.23) and (3.24) we establish that $\{(X^{(k)}, s^{(k)}, y^{(k)})\}_{k\in\mathbb{Z}_+}$ is a bounded sequence. Therefore, there exists $\Gamma < \infty$ such that for all $k \geq 1$

$$\max\left\{ \|X_*\|_F, \|s_*\|_2, \|y_*\|_2, \|X^{(k)}\|_F, \|s^{(k)}\|_2, \|y^{(k)}\|_2 \right\} \leq \Gamma. \tag{3.44}$$

Consider the following two cases:

(a) *The k-th call to* ***Algorithm APG*** *terminates with* ITERSTOP: From (3.28) it follows that

$$\mu_1\|\sigma(X^{(k)})\|_\alpha + \mu_2\|s^{(k)}\|_\beta \leq \mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|\mathcal{C}(X_*) - d\|_\beta + \frac{\lambda^{(k)}}{2}\left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2\right) + \frac{\epsilon^{(k)}}{\lambda^{(k)}}. \tag{3.45}$$

(b) *The k-th call to* ***Algorithm APG*** *terminates with* GRADSTOP: Let (G, g) belong to the set of partial subgradients $\partial_{X,s}P^{(k)}(.,.,.)|_{(X^{(k)},s^{(k)},y^{(k)})}$, defined in (3.14), and satisfy the stopping condition GRAD-

STOP. Then from the convexity of $P^{(k)}$ and Lemma A.5, it follows that

$$\begin{aligned}
&P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)}) \\
&\leq\ P^{(k)}(X_*, s_*, y_*) - \left\langle G, X_* - X^{(k)} \right\rangle - g^T(s_* - s^{(k)}) - \nabla_y P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})^T(y_* - y^{(k)}), \\
&\leq\ P^{(k)}(X_*, s_*, y_*) + \|G\|_F \|X_* - X^{(k)}\|_F + \|g\|_2 \|s_* - s^{(k)}\|_2 + L\xi^{(k)}\|y_* - y^{(k)}\|_2 \\
&\quad +\xi^{(k)}\|\nabla_y f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_2, \\
&\leq\ P^{(k)}(X_*, s_*, y_*) + \tau^{(k)}\left(\|X_* - X^{(k)}\|_F + \|s_* - s^{(k)}\|_2\right) + L\xi^{(k)}\|y_* - y^{(k)}\|_2 \\
&\quad +\xi^{(k)}\|\nabla_y f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_2. \qquad (3.46)
\end{aligned}$$

Dividing (3.46) by $\lambda^{(k)}$ and using the fact that $\theta_1^{(k+1)} = \nabla_y f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})/\lambda^{(k)}$, it follows that

$$\begin{aligned}
\mu_1\|\sigma(X^{(k)})\|_\alpha + \mu_2\|s^{(k)}\|_\beta \leq \mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|\mathcal{C}(X_*) - d\|_\beta + \frac{\lambda^{(k)}}{2}\left(\|\theta_1^{(k)}\|_2^2 + \|\theta_2^{(k)}\|_2^2\right) + \xi^{(k)}\|\theta_1^{(k+1)}\|_2 \\
+ \frac{\tau^{(k)}}{\lambda^{(k)}}\left(\|X_* - X^{(k)}\|_F + \|s_* - s^{(k)}\|_2\right) + \frac{\xi^{(k)}}{\lambda^{(k)}}L\|y_* - y^{(k)}\|_2. \qquad (3.47)
\end{aligned}$$

From (3.44), the fact that $\|\theta_i^{(k)}\|_2 \leq B_{\theta_i}$ for $i \in \{1, 2\}$, (3.45) and (3.47), it follows that

$$\begin{aligned}
\mu_1\|\sigma(X^{(k)})\|_\alpha + \mu_2\|s^{(k)}\|_\beta \leq \mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|\mathcal{C}(X_*) - d\|_\beta + \left(\frac{B_{\theta_1}^2 + B_{\theta_2}^2}{2}\right)\lambda^{(k)} \\
+ \max\left\{\frac{\epsilon^{(k)}}{\lambda^{(k)}}, \frac{4\Gamma\tau^{(k)}}{\lambda^{(k)}} + \frac{\xi^{(k)}}{\lambda^{(k)}}(2L\Gamma + \lambda^{(k)}B_{\theta_1})\right\}. \qquad (3.48)
\end{aligned}$$

The bound (3.10) relating $\|\sigma(.)\|_\alpha$ to the Frobenius norm, (3.19a), bound $\|\theta_2^{(k)}\|_2 \leq B_{\theta_2}$, together with triangle inequality imply that

$$\begin{aligned}
\|\mathcal{C}(X^{(k)}) - d\|_\beta &\leq \|s^{(k)}\|_\beta + \|\lambda^{(k)}\theta_2^{(k)}\|_\beta + J(\beta^*)\left(\max\left\{\sqrt{2\epsilon^{(k)}}\sigma_{\max}(M),\ \tau^{(k)}\right\} + J(\beta^*)\mu_2\lambda^{(k)}\right), \\
&\leq \|s^{(k)}\|_\beta + J(\beta^*)\left(B_{\theta_2} + \mu_2\ J(\beta^*)\right)\lambda^{(k)} + J(\beta^*)\max\left\{\sqrt{2}\ \sigma_{\max}(M),\ \kappa_1\sqrt{\epsilon^{(k)}}\right\}\ \sqrt{\epsilon^{(k)}},
\end{aligned}$$

where the second inequality uses the relation $\tau^{(k)} = \kappa_1\epsilon^{(k)}$. The above inequality and (3.48) imply that

$$\begin{aligned}
\mu_1\|\sigma(X^{(k)})\|_\alpha + \mu_2\|\mathcal{C}(X^{(k)}) - d\|_\beta \leq \mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|\mathcal{C}(X_*) - d\|_\beta \\
+ \left(\frac{B_{\theta_1}^2 + B_{\theta_2}^2}{2} + \mu_2 J(\beta^*)\left(B_{\theta_2} + \mu_2\ J(\beta^*)\right)\right)\lambda^{(k)} \\
+ \max\left\{\frac{\epsilon^{(k)}}{\lambda^{(k)}}, \frac{4\Gamma\tau^{(k)}}{\lambda^{(k)}} + \frac{\xi^{(k)}}{\lambda^{(k)}}(2L\Gamma + \lambda^{(k)}B_{\theta_1})\right\} \\
+\mu_2 J(\beta^*)\max\left\{\sqrt{2}\ \sigma_{\max}(M),\ \kappa_1\sqrt{\epsilon^{(k)}}\right\}\ \sqrt{\epsilon^{(k)}}. \qquad (3.49)
\end{aligned}$$

Since $\frac{\epsilon^{(k)}}{(\lambda^{(k)})^2} \leq B$, $\tau^{(k)} = \kappa_1\epsilon^{(k)}$ and $\xi^{(k)} = \kappa_2\epsilon^{(k)}$ for all $k \geq 1$, (3.49) implies one side of the bound in (ii).

Next, we establish a lower bound for $P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)})$ by comparing the following pairs of Lagrangian duals

$$\begin{aligned}
\min_{X \in \mathbb{R}^{m\times n}} \quad & \mu_1\|\sigma(X)\|_\alpha + \mu_2\|\mathcal{C}(X) - d\|_\beta, \qquad (3.50a) \\
\text{s.t.} \quad & \mathcal{A}(X) - b \in \mathcal{Q}.
\end{aligned}$$

$$\begin{aligned}
\max_{w \in \mathbb{R}^q, v \in \mathbb{R}^p} \quad & -b^T w - d^T v - \gamma_{\mathcal{Q}}(w), \qquad (3.50b) \\
\text{s.t.} \quad & \|\sigma(\mathcal{A}^*(w) + \mathcal{C}^*(v))\|_{\alpha^*} \leq \mu_1, \\
& \|v\|_{\beta^*} \leq \mu_2,
\end{aligned}$$

where $\gamma_{\mathcal{Q}}$ is the support function of $\mathcal{Q}$, i.e., $\gamma_{\mathcal{Q}}(w) := \sup_{y\in\mathcal{Q}} w^T y$, and

$$\begin{array}{rl} \min_{X\in\mathbb{R}^{m\times n}, s\in\mathbb{R}^p, y\in\mathbb{R}^q} & \lambda(\mu_1\|\sigma(X)\|_\alpha + \mu_2\|s\|_\beta) + f(X,s,y), \\ \text{s.t.} & y\in\mathcal{Q}. \end{array} \tag{3.51a}$$

$$\begin{array}{rl} \max_{w\in\mathbb{R}^q, v\in\mathbb{R}^p} & -\lambda(b+\lambda\theta_1)^T w - \lambda(d+\lambda\theta_2)^T v - \lambda\gamma_{\mathcal{Q}}(w) - \frac{\lambda^2}{2}(\|w\|_2^2 + \|v\|_2^2), \\ \text{s.t.} & \|\sigma(\mathcal{A}^*(w)+\mathcal{C}^*(v))\|_{\alpha^*} \le \mu_1, \\ & \|v\|_{\beta^*} \le \mu_2. \end{array} \tag{3.51b}$$

Above (w_*, v_*) denotes the optimal solution of the dual (3.50b) and $f(X,s,y) := \frac{1}{2}\|\mathcal{A}(X)+y-b-\lambda\theta_1\|_2^2 + \frac{1}{2}\|\mathcal{C}(X)+s-d-\lambda\theta_2\|_2^2$. Note that (w_*, v_*) is feasible for (3.51b). Therefore, by Lagrangian duality it follows that

$$\begin{aligned} & P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)}) \\ & \ge \lambda^{(k)} \left(-b^T w_* - d^T v_* - \gamma_{\mathcal{Q}}(w_*) - \frac{\lambda^{(k)}}{2}\left(\|w_*\|_2^2 + \|v_*\|_2^2 + 2(\theta_1^{(k)})^T w_* + 2(\theta_2^{(k)})^T v_* \right) \right), \\ & \ge \lambda^{(k)} \left(\mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|\mathcal{C}(X_*) - d\|_\beta \right) \\ & \quad - \frac{(\lambda^{(k)})^2}{2}\left(\|w_*\|_2^2 + \|v_*\|_2^2 + 2\|\theta_1^{(k)}\|_2\|w_*\|_2 + 2\|\theta_2^{(k)}\|_2\|v_*\|_2 \right), \end{aligned} \tag{3.52}$$

where the first inequality follows from weak duality for primal-dual pair in (3.51), and (3.52) follows from strong duality for primal-dual pair in (3.50) and the Cauchy-Schwartz inequality.

From the definition of $\{\theta_i^{(k)}\}_{k\in\mathbb{Z}_+}$ in Figure 3.1, it is clear that the FALC iterates $\{X^{(k)}\}_{k\in\mathbb{Z}}$ satisfy

$$\frac{P^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})}{\lambda^{(k)}} = \left(\mu_1\|\sigma(X^{(k)})\|_\alpha + \mu_2\|s^{(k)}\|_\beta\right) + \frac{\lambda^{(k)}}{2}\left(\|\theta_1^{(k+1)}\|_2^2 + \|\theta_2^{(k+1)}\|_2^2 \right),$$

and it follows that

$$\mu_1\|\sigma(X^{(k)})\|_\alpha + \mu_2\|s^{(k)}\|_\beta \ge \frac{P^{(k)}(X_*^{(k)}, s_*^{(k)}, y_*^{(k)})}{\lambda^{(k)}} - \frac{\lambda^{(k)}}{2}\left(\|\theta_1^{(k+1)}\|_2^2 + \|\theta_2^{(k+1)}\|_2^2 \right). \tag{3.53}$$

Thus, the bound on $\|\theta_i^{(k)}\|_2$, $i\in\{1,2\}$ established in (3.20), and the inequalities (3.52) and (3.53), together imply that

$$\begin{aligned} \mu_1\|\sigma(X^{(k)})\|_\alpha + \mu_2\|s^{(k)}\|_\beta \ge{}& \mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|\mathcal{C}(X_*) - d\|_\beta \\ & - \frac{\lambda^{(k)}}{2}\left((B_{\theta_1} + \|w_*\|_2)^2 + (B_{\theta_2} + \|v_*\|_2)^2 \right). \end{aligned} \tag{3.54}$$

The bound $\|\mathcal{A}^*(w_*) + \mathcal{C}^*(v_*)\|_F \le I(\alpha^*)\|\sigma(\mathcal{A}^*(w_*) + \mathcal{C}^*(v_*))\|_{\alpha^*} \le I(\alpha^*)\mu_1$ implies that

$$\sigma_{\min}(A)\|w_*\|_2 \le \|\mathcal{A}^*(w_*)\|_F \le I(\alpha^*)\mu_1 + \|\mathcal{C}^*(v_*)\|_F \le I(\alpha^*)\mu_1 + \sigma_{\max}(C)\|v_*\|_2,$$

and the bound $\|v_*\|_{\beta^*} \le \mu_2$ implies that $\|v_*\|_2 \le J(\beta^*)\ \mu_2$. Hence, both $\|v_*\|_2$ and $\|w_*\|_2$ in (3.54) are bounded.

The bound (3.10), the uniform bound $\|\theta_2^{(k)}\|_2 \le B_{\theta_2}$ and triangle inequality together imply that

$$\|s^{(k)}\|_\beta \le \|\mathcal{C}(X^{(k)}) - d\|_\beta + J(\beta^*)\left(B_{\theta_2} + \mu_2\ J(\beta^*)\right)\lambda^{(k)} + J(\beta^*)\max\left\{\sqrt{2}\ \sigma_{\max}(M),\ \kappa_1\sqrt{\epsilon^{(k)}}\right\}\ \sqrt{\epsilon^{(k)}}, \tag{3.55}$$

where the second inequality uses the relation $\tau^{(k)} = \kappa_1\epsilon^{(k)}$. From (3.54) and (3.55), it follows that

$$\begin{aligned} & \mu_1\|\sigma(X^{(k)})\|_\alpha + \mu_2\|\mathcal{C}(X^{(k)}) - d\|_\beta \\ & \ge \mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|\mathcal{C}(X_*) - d\|_\beta \\ & \quad - \left(\frac{(B_{\theta_1} + \|w_*\|_2)^2 + (B_{\theta_2} + \|v_*\|_2)^2}{2} + \mu_2 J(\beta^*)\left(B_{\theta_2} + \mu_2\ J(\beta^*)\right) \right)\lambda^{(k)} \\ & \quad - \mu_2 J(\beta^*)\max\left\{\sqrt{2}\ \sigma_{\max}(M),\ \kappa_1\sqrt{\epsilon^{(k)}}\right\}\ \sqrt{\epsilon^{(k)}}. \end{aligned}$$

This establishes the result. □

Now, we have all the estimates we need to prove the main convergence rate result in this paper.

THEOREM 3.7. *Fix $\kappa_1, \kappa_2, \nu \in (0,1)$, and strictly positive parameters $(\lambda^{(1)}, \epsilon^{(1)}, \tau^{(1)}, \xi^{(1)})$. For $k \geq 1$, set parameter sequence $\{(\lambda^{(k)}, \epsilon^{(k)}, \tau^{(k)}, \xi^{(k)})\}_{k \in \mathbb{Z}_+}$ as follows:*

$$\begin{array}{rclcrcl} \lambda^{(k+1)} & = & \nu\, \lambda^{(k)}, & & \xi^{(k)} & = & \kappa_2\, \epsilon^{(k)}, \\ \epsilon^{(k+1)} & = & \nu^2\, \epsilon^{(k)}, & & \tau^{(k)} & = & \kappa_1\, \epsilon^{(k)}. \end{array} \tag{3.56}$$

*Then, for all $\epsilon > 0$, **Algorithm FALC** computes an ϵ-feasible and ϵ-optimal solution $\bar{X} \in \mathbb{R}^{m \times n}$ to the composite norm minimization problem (3.1), in $N_{\text{inner}} = \mathcal{O}\left(\frac{1}{\epsilon}\right)$ **Algorithm APG** iterations.*

Proof. For the specific choice of the parameter sequence in (3.56) we have that $\frac{\epsilon^{(k)}}{(\lambda^{(k)})^2} = \frac{\epsilon^{(1)}}{(\lambda^{(1)})^2}$, for all $k \geq 1$. Therefore, Theorem 3.6 guarantees that there exist $c_2 > 0$ and $c_3 > 0$ such that for all $k \geq 1$,

$$\left|(\mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|\mathcal{C}(X^{(k)}) - d\|_\beta) - (\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta)\right| \leq \left(c_2 \lambda^{(1)} + c_3 \sqrt{\epsilon^{(1)}}\ \right) \nu^{(k-1)}.$$

Thus, $\left|(\mu_1 \|\sigma(X^{(k)})\|_\alpha + \mu_2 \|\mathcal{C}(X^{(k)}) - d\|_\beta) - (\mu_1 \|\sigma(X_*)\|_\alpha + \mu_2 \|\mathcal{C}(X_*) - d\|_\beta)\right| \leq \epsilon$, for all

$$k > \log_{\frac{1}{\nu}} \left(\frac{c_2 \lambda^{(1)} + c_3 \sqrt{\epsilon^{(1)}}}{\epsilon} \right) + 1. \tag{3.57}$$

Moreover, Theorem 3.6 also implies that there exists $c_1 > 0$ such that $\|\mathcal{A}(X^{(k)}) - y^{(k)} - b\|_2 \leq c_1 \lambda^{(1)}\, \nu^{k-1}$, for $k \geq 1$, Thus, $\|\mathcal{A}(X^{(k)}) - y^{(k)} - b\|_2 \leq \epsilon$ for all

$$k > \log_{\frac{1}{\nu}} \left(\frac{c_1 \lambda^{(1)}}{\epsilon} \right) + 1. \tag{3.58}$$

Then (3.57) and (3.58) imply that for all $\epsilon > 0$, the number of FALC iterations required to compute an ϵ-feasible and ϵ-optimal solution

$$N_{\text{FALC}}(\epsilon) \leq \left\lceil \log_{\frac{1}{\nu}} \left(\frac{U}{\epsilon} \right) \right\rceil + 1, \tag{3.59}$$

where $U = \max\left\{ c_2 \lambda^{(1)} + c_3 \sqrt{\epsilon^{(1)}},\ c_1 \lambda^{(1)} \right\}$.

From Lemma 3.5 it follows that there exists a constant $\mathcal{N}$ such that APG iteration to solve the k-th FALC subproblem $N^{(k)} \leq \frac{\mathcal{N}}{\sqrt{\epsilon^{(k)}}}$. Therefore,

$$N_{\text{inner}} = \mathcal{N} \sum_{k=1}^{N_{\text{FALC}}(\epsilon)} \frac{1}{\sqrt{\epsilon^{(k)}}} = \frac{\mathcal{N}}{\sqrt{\epsilon^{(1)}}} \sum_{k=0}^{N_{\text{FALC}}(\epsilon)-1} \nu^{-k} = \frac{\mathcal{N}}{\sqrt{\epsilon^{(1)}}} \cdot \frac{\nu}{(1-\nu)} \cdot \left(\frac{1}{\nu}\right)^{N_{\text{FALC}}(\epsilon)} \leq \left(\frac{\mathcal{N} U}{\nu(1-\nu)\ \sqrt{\epsilon^{(1)}}} \right)\ \frac{1}{\epsilon},$$

where the last bound follows from (3.59). □

Note that we do not explicitly specify the constant hidden in the $\mathcal{O}(1/\epsilon)$ iteration complexity result of Theorem 3.7. Moreover, the bound given in the proof is very crude. The main reason is that the composite minimization problem (1.1) is very general and the constant strongly depends on the specific problem structure. However, the proof technique used to establish Theorem 3.7 can be applied as is to any special case of composite minimization problem to obtain much sharper constants. For instance, the complexity result of FALC for the basis pursuit problem in (1.4) is given by

$$N_{\text{inner}} \leq n \kappa(A)^2 \left(\frac{16 \|x_*\|_1}{\nu(1-\nu)} \cdot \frac{1}{\epsilon} + \frac{9}{\nu} \cdot \log_{\frac{1}{\nu}} \left(\frac{8 n \kappa^2(A)}{\epsilon} \right) \right) = \mathcal{O}\left(\frac{1}{\epsilon} \right),$$

where $\kappa(A) := \sigma_{\max}(A)/\sigma_{\min}(A)$ is the condition number of A. This is the same bound that was obtained for this special case in [3].

The convergence rate result in Theorem 3.7 relies on the uniform bound established in Lemma 3.5. This uniform bound in turn assumes that all calls to **Algorithm APG** are terminated by ITERSTOP. On the other hand, in our numerical experiments almost all calls to **Algorithm APG** were terminated by GRADSTOP. This suggests that the $\mathcal{O}(\frac{1}{\epsilon})$ rate result has a lot of slack. Indeed, we find that early terminating **Algorithm APG** iterations when the stopping condition GRADSTOP is satisfied reduces total number of **Algorithm APG** iterations significantly: in our numerical tests, FALC required only $\mathcal{O}(\log(\frac{1}{\epsilon}))$ inner iterations to compute an ϵ-optimal, ϵ-feasible iterate. The augmented Lagrangian algorithm FAL introduced in [3] is an implementation of FALC for the basis pursuit problem. In Figure 6.1 in [3] one can clearly observe the $\mathcal{O}(\log(1/\epsilon))$ empirical performance as opposed to the $\mathcal{O}(1/\epsilon)$ worst case complexity. We observe a similar empirical performance with FALC on the numerical problems tested in this paper.

4. Implementation details of Algorithm FALC. In this section we describe all the details of FALC. Let $\{(X_i^{(k,\ell)}, s_i^{(k,\ell)}, y_i^{(k,\ell)})\}_{\ell\in\mathbb{Z}_+}$ denote the sequence of $x_i^{(\ell)}$-iterates of **Algorithm APG** in Figure 2.1 for $i \in \{1,2\}$ when **Algorithm APG** is called in Line 14 of Figure 3.1 to solve the k-th subproblem.

4.1. Subgradient selection. We used the following slightly modified version of GRADSTOP in our implementation.

$$\text{GRADSTOP1} := \left\{\exists (G,g) \in \partial_{X,s} P^{(k)}(.,.,.)|_{(X^{(k)},s^{(k)},y^{(k)})} \text{ s.t. } \|G\|_F \le \tau_X^{(k)} \textbf{ and } \|g\|_2 \le \tau_s^{(k)}\right\} \tag{4.1}$$

for some tolerance values $\tau_X^{(k)}$ and $\tau_s^{(k)}$ such that $\{\tau_X^{(k)}\}_{k\in\mathbb{Z}_+}$ and $\{\tau_s^{(k)}\}_{k\in\mathbb{Z}_+}$ are decreasing sequences. Clearly, if (4.1) holds, the original GRADSTOP1 given in Line 11 of **Algorithm FALC** holds for $\tau^{(k)} = \tau_X^{(k)} + \tau_s^{(k)}$.

We check the stopping condition GRADSTOP1 in each **Algorithm APG** iteration. Let $\breve{Z}^{(k,\ell)} = \left(\breve{X}_1^{(k,\ell)}, \breve{s}_1^{(k,\ell)}, \breve{y}_1^{(k,\ell)}\right)$ denotes the unconstrained solution to the optimization problem in Line 4 of Figure 2.1, i.e. when the contraint $(X,s,y) \in \mathcal{S}^{(k)}$ is not enforced. A subgradient $(G,g) \in \partial_{X,s} P^{(k)}(.,.,.)|_{\breve{Z}^{(k,\ell)}}$ can be computed as follows

$$G = \lambda^{(k)}\mu_1 Q + \nabla_X f^{(k)}\left(\breve{Z}^{(k,\ell)}\right) \quad \text{and} \quad g = \lambda^{(k)}\mu_2 q + \nabla_s f^{(k)}\left(\breve{Z}^{(k,\ell)}\right),$$

where

$$Q = \frac{L}{\lambda^{(k)}\mu_1}\left(X_2^{(k,\ell)} - \frac{1}{L}\nabla_X f^{(k)}(X_2^{(k,\ell)}, s_2^{(k,\ell)}, y_2^{(k,\ell)}) - \breve{X}_1^{(k,\ell)}\right),$$
$$q = \operatorname{argmin}\left\{\|\lambda^{(k)}\mu_2 r + \nabla_s f^{(k)}\left(\breve{Z}^{(k,\ell)}\right)\|_2 : r \in \partial\|.\|_\beta|_{\breve{s}_2^{(k,\ell)}}\right\}.$$

From the first order optimality condition, it can be easily shown that $Q \in \partial\|\sigma(.)\|_\alpha|_{\breve{X}_1^{(k,\ell)}}$. Moreover, given $\nabla_s f^{(k)}$ at $\breve{Z}^{(k,\ell)}$ the complexity of computing $q \in \partial\|.\|_\beta|_{\breve{s}_1^{(k,\ell)}} \subset \mathbb{R}^p$ is $\mathcal{O}(p)$ when $\beta \in \{1,2\}$ and $\mathcal{O}(p\log(p))$ when $\beta = \infty$.

4.2. Stopping criterion for FALC. In our numerical experiments, we terminate **Algorithm FALC** either the distance between successive inner iterates are below a threshold ϱ for each component, i.e. $\|X_1^{(k,\ell)} - X_1^{(k,\ell-1)}\|_F \le \varrho$, $\|s_1^{(k,\ell)} - s_1^{(k,\ell-1)}\|_2 \le \varrho$ or there exist partial subgradients with sufficiently small norm for each component, i.e. $\|G\|_F \le \varsigma_X$, $\|g\|_2 \le \varsigma_s$ for some $(G,g) \in \partial_{X,s} P^{(k)}(.,.,.)|_{\breve{Z}^{(k,\ell)}}$ and

$$\|\breve{y}_1^{(k,\ell)} - \Pi_{\mathcal{Q}}\left(\breve{y}_1^{(k,\ell)} - \frac{1}{L}\nabla_y P^{(k)}(\breve{Z}^{(k,\ell)})\right)\|_2 \le \varsigma_y.$$

In our numerical experiments we set ϱ, ς_X, ς_s and ς_y by experimenting with a small instance of the problem.

4.3. Multiplier selection. Given $\bar{c}_\tau \in (0,1)$, $\bar{c}_\xi \in (0,1)$, $\bar{c}_\lambda > 0$, $c_\tau \in (0,1)$, $c_\xi \in (0,1)$, $c_\lambda \in (0,1)$, for all $k \ge 1$ the approximate optimality parameters $\tau_X^{(k)}$, $\tau_s^{(k)}$, $\xi^{(k)}$ and the penalty parameter $\lambda^{(k)}$ are set as

follows:

$$
\begin{aligned}
\breve{X}^{(1)} &= \operatorname{argmin}_{X\in\mathbb{R}^{m\times n}} \tfrac{1}{2}\|X-\left(X^{(0)}-\tfrac{1}{L}\nabla_X f^{(1)}(X^{(0)},s^{(0)},y^{(0)})\right)\|_F^2+\tfrac{\lambda^{(1)}\mu_1}{L}\|\sigma(X)\|_\alpha,\\
\breve{s}^{(1)} &= \operatorname{argmin}_{s\in\mathbb{R}^p} \tfrac{1}{2}\|s-\left(s^{(0)}-\tfrac{1}{L}\nabla_s f^{(1)}(X^{(0)},s^{(0)},y^{(0)})\right)\|_2^2+\tfrac{\lambda^{(1)}\mu_2}{L}\|s\|_\beta,\\
\breve{y}^{(1)} &= \operatorname{argmin}_{y\in\mathcal{Q}\subset\mathbb{R}^q} \|y-\left(y^{(0)}-\tfrac{1}{L}\nabla_y f^{(1)}(X^{(0)},s^{(0)},y^{(0)})\right)\|_2^2,\\
G^{(1)} &= L\left(X^{(0)}-\tfrac{1}{L}\nabla_X f^{(1)}(X^{(0)},s^{(0)},y^{(0)})-\breve{X}^{(1)}\right)+\nabla_X f^{(1)}(\breve{X}^{(1)},\breve{s}^{(1)},\breve{y}^{(1)}),\\
g^{(1)} &= \operatorname{argmin}_{g\in\mathbb{R}^p}\{\|g\|_2 : g=\lambda^{(1)}\mu_2 p+\nabla_s f^{(1)}(\breve{X}^{(1)},\breve{s}^{(1)},\breve{y}^{(1)}),\ p\in\partial\|.\|_\beta|_{\breve{s}^{(1)}}\},\\
\tau_X^{(1)} &= \bar{c}_\tau\|G^{(1)}\|_F,\\
\tau_s^{(1)} &= \bar{c}_\tau\|g^{(1)}\|_2,\\
\xi^{(1)} &= \bar{c}_\xi\|\breve{y}^{(1)}-\Pi_{\mathcal{Q}}\left(\breve{y}^{(1)}-\tfrac{1}{L}\nabla_y f^{(1)}(\breve{X}^{(1)},\breve{s}^{(1)},\breve{y}^{(1)})\right)\|_2,\\
\lambda^{(1)} &= \bar{c}_\lambda\|X^{(0)}\|_2,\\
\breve{X}^{(k)} &= \operatorname{argmin}_{X\in\mathbb{R}^{m\times n}} \tfrac{1}{2}\|X-\left(X^{(k-1)}-\tfrac{1}{L}\nabla_X f^{(k)}(X^{(k-1)},s^{(k-1)},y^{(k-1)})\right)\|_F^2+\tfrac{\lambda^{(k)}\mu_1}{L}\|\sigma(X)\|_\alpha,\\
\breve{s}^{(k)} &= \operatorname{argmin}_{s\in\mathbb{R}^p} \tfrac{1}{2}\|s-\left(s^{(k-1)}-\tfrac{1}{L}\nabla_s f^{(k)}(X^{(k-1)},s^{(k-1)},y^{(k-1)})\right)\|_2^2+\tfrac{\lambda^{(k)}\mu_2}{L}\|s\|_\beta,\\
\breve{y}^{(k)} &= \operatorname{argmin}_{y\in\mathcal{Q}\subset\mathbb{R}^q} \|y-\left(y^{(k-1)}-\tfrac{1}{L}\nabla_y f^{(k)}(X^{(k-1)},s^{(k-1)},y^{(k-1)})\right)\|_2^2,\\
G^{(k)} &= L\left(X^{(k-1)}-\tfrac{1}{L}\nabla_X f^{(k)}(X^{(k-1)},s^{(k-1)},y^{(k-1)})-\breve{X}^{(k)}\right)+\nabla_X f^{(k)}(\breve{X}^{(k)},\breve{s}^{(k)},\breve{y}^{(k)}),\\
g^{(k)} &= \operatorname{argmin}_{g\in\mathbb{R}^p}\{\|g\|_2 : g=\lambda^{(k)}\mu_2 p+\nabla_s f^{(k)}(\breve{X}^{(k)},\breve{s}^{(k)},\breve{y}^{(k)}),\ p\in\partial\|.\|_\beta|_{\breve{s}^{(k)}}\},\\
\tau_X^{(k)} &= \min\left\{\ c_\tau\,\tau_X^{(k-1)},\bar{c}_\tau\|G^{(k)}\|_F\ \right\},\\
\tau_s^{(k)} &= \min\left\{\ c_\tau\,\tau_s^{(k-1)},\bar{c}_\tau\|g^{(k)}\|_2\ \right\},\\
\xi^{(k)} &= \min\{c_\xi\,\xi^{(k-1)},\ \|\breve{y}^{(k)}-\Pi_{\mathcal{Q}}\left(\breve{y}^{(k)}-\tfrac{1}{L}\nabla_y f^{(k)}(\breve{X}^{(k)},\breve{s}^{(k)},\breve{y}^{(k)})\right)\|_2\}\\
\lambda^{(k)} &= c_\lambda\,\lambda^{(k-1)},
\end{aligned}
$$

for all $k\geq 2$. In all our experiments, $\bar{c}_\tau = 0.999$ and $\bar{c}_\xi = 0.999$.

We initialize FALC with $\left(X^{(0)},s^{(0)},y^{(0)}\right)$ such that $\mathcal{A}(X^{(0)})-b\in\mathcal{Q}$, $s^{(0)}=\mathcal{C}(X^{(0)})-d$ and $y^{(0)}=\mathcal{A}(X^{(0)})-b$. In first iteration of FALC, we solve the problem

$$\min_{(X,s,y)\in\mathcal{S}^{(1)},\ y\in\mathcal{Q}} P^{(1)}(X,s,y)=\min_{(X,s,y)\in\mathcal{S}^{(1)},\ y\in\mathcal{Q}}\lambda^{(1)}(\mu_1\|\sigma(X)\|_\alpha+\mu_2\|s\|_\beta)+f^{(1)}(X,s,y),$$

where $\mathcal{S}^{(1)}=\{(X,s,y):\ \mu_1\|\sigma(X)\|_\alpha+\mu_2\|s\|_\beta\leq\eta^{(1)}\}$ and $\eta^{(1)}=\mu_1\|\sigma(X^{(0)})\|_\alpha+\mu_2\|s^{(0)}\|_\beta$. Since $X^{(0)}$ is feasible, $f^{(1)}(X^{(0)},s^{(0)},y^{(0)})=0$ and $P^{(1)}(X^{(0)},s^{(0)},\,y^{(0)})=\lambda^{(1)}\eta^{(1)}$. Then $P^{(1)}(X,s,y)\geq 0$ for all $X\in\mathbb{R}^{m\times n}$ implies that the initial duality gap is less than or equal to $\lambda^{(1)}\eta^{(1)}$. Hence, we initialize $\epsilon^{(1)}=0.99\lambda^{(1)}\eta^{(1)}$ and then set $\epsilon^{(k+1)}=c_\lambda^2\epsilon^{(k)}$ for all $k\geq 1$.

5. Numerical experiments. In our numerical experiments, we focused on problems where both $\mu_1>0$ and $\mu_2>0$. Two important problems of this form are the principal component pursuit and stable principal component pursuit problems, given in (1.6) and (1.7), respectively. In the first set of experiments we solved a set of randomly generated instances of principal component pursuit problems. In this setting, we compare FALC with another augmented Lagrangian algorithm I-ALM [29], a proximal gradient algorithm APG [30] and a soft-thresholding algorithm SVT [6]. In the second set of experiments, we solved a set of randomly generated instances of stable principal component pursuit problem. Since I-ALM, APG and SVT are not able to solve this problem, we only report statistics for FALC. In Section 5.1, we describe the methodology we have used in both experimental settings for generating random problem instances. All the numerical experiments were conducted on an IBM Thinkpad laptop with a Intel Core 2 CPU T7200 @2.0 GHz processor, 3GB SDRAM running MATLAB 7.2 on Windows XP Professional operating system.

The augmented Lagrangian algorithm FAL introduced in [3] is an implementation of FALC for the basis pursuit problem. The numerical results reported in [3] show that FAL was 2-7 times faster than the specialized algorithms NESTA [5], FPC and FPC-BB [23, 24], FPC-AS [39], YALL1 [40] and SPGL1 [38]. See Tables 6.8, 6.9, 6.10, 6.11, 6.13 in [3] for details.

The numerical results in this paper and those in [3] clearly show that FALC is very competitive with the state-of-the-art algorithms for the special cases of the composite norm minimization problem.

5.1. Data generation. We tested FALC on randomly generated data matrices $D = X_0 + S_0 + \zeta_0$, where

i. $X_0 = UV^T$, such that $U \in \mathbb{R}^{n\times r}$, $V \in \mathbb{R}^{n\times r}$ for $r = 0.05n$ and $U_{ij} \sim \mathcal{N}(0,1)$, $V_{ij} \sim \mathcal{N}(0,1)$ for all i, j are independent standard Gaussian variables,

ii. $\Lambda \subset \{(i,j) : \ 1 \le i, j \le n\}$ such that cardinality of Λ, $|\Lambda| = p$ for $p = 0.05n^2$,

iii. $(S_0)_{ij} \sim \mathcal{U}[-1,1]$ for all $(i,j) \in \Lambda$ are independent uniform random variables between -1 and 1,

iv. $(\zeta_0)_{ij} \sim \delta\mathcal{U}[-1,1]$ for all i, j are independent Gaussian variables.

5.2. Principal Component Pursuit Problem. In this section we solve the problem

$$\begin{array}{ll} \min_{X,S\in\mathbb{R}^{m\times n}} & \|X\|_* + \mu_2 \|\,\mathbf{vec}(S)\|_1, \\ \text{subject to} & X + S = D, \end{array} \tag{5.1}$$

and report the results of our numerical experiments comparing FALC with I-ALM [29], APG [30] and SVT [6]. All the codes for I-ALM, APG and SVT, can be found at `http://perception.csl.uiuc.edu/matrix-rank/home.html`. Note that SVT [6] algorithm was originally proposed for solving the matrix completion problem. The algorithm we used in our numerical study is an adaptation of the SVT algorithm by Wright and Rao at the Perception and Decision Laboratory in University of Illinois, Urbana-Champaign to solve robust PCA problem.

We created 10 random problems of size $n = 500$, i.e. $D \in \mathbb{R}^{500\times 500}$ using the procedure described in Section 5.1, where δ is set to 0, i.e. $\zeta_0 = \mathbf{0}$. We chose parameter values for each of the four algorithms so that they produce a solution X_{sol} and S_{sol} with relative-infeasibility approximately equal to 5×10^{-9}, i.e. $\frac{\|X_{sol}+S_{sol}-D\|_F}{\|D\|_F} \approx 5 \times 10^{-9}$. For each algorithm we set the parameters by solving a set of small size problems and these parameter values were fixed throughout the experiments, all other parameters are set to their default values. The termination criteria are not directly comparable due to different formulations of the problem solved by different solvers. For FALC we attempted to set the stopping parameter ϱ such that on average the stopping criterion for FALC is tighter than the stopping criteria of all the other algorithms we tested.

1. **FALC**: Problem (5.1) is a special case of problem (1.1) with $\delta = 0$, $\alpha = 1$ and $\beta = 1$. Therefore, $f^{(k)}(X, s, y)$ defined in (3.5) simplifies to $f^{(k)}(X, S) = \frac{1}{2}\|\,\mathbf{vec}(X + S) - \mathbf{vec}(D) - \lambda^{(k)}\theta_1^{(k)}\|_2^2$ (note that for all $k \ge 1$, $\theta_2^{(k)} = 0$). We set $c_\tau = 0.4$, $c_\epsilon = 0.4$, $c_\lambda = 0.4$, $\bar{c}_\tau = 0.999$, $\bar{c}_\epsilon = 0.999$, $\bar{c}_\lambda = 2$ and initialize $\theta_1^{(1)}$ as in [29], i.e.

$$\theta_1^{(1)} = \frac{1}{\max\{\|\mathrm{sign}(D)\|_2,\ \sqrt{n}\|\,\mathbf{vec}(\mathrm{sign}(D))\|_\infty\}}\,\mathbf{vec}(\mathrm{sign}(D)). \tag{5.2}$$

   Finally, we set $\varrho = 1 \times 10^{-5}$ and terminate FALC when the distance between successive inner iterates are below the threshold ϱ for each component, i.e. $\|X_1^{(k,\ell)} - X_1^{(k,\ell-1)}\|_F \le \varrho$ and $\|s_1^{(k,\ell)} - s_1^{(k,\ell-1)}\|_2 \le \varrho$ for any $k \ge 1$. We used PROPACK [27] for computing partial singular value decompositions. In order to estimate the rank of X_0, we followed the scheme proposed in Equation (17) in [29]. The code for PROPACK is available at [`http://soi.stanford.edu/~rmunk/PROPACK/`].

2. **I-ALM**: I-ALM solves $\min\{\|X\|_* + \frac{1}{\sqrt{n}}\|\,\mathbf{vec}(S)\|_1 : \ X + S = D\}$. Let $(X^{(k)}, \mathcal{S}^{(k)})$ be the k-th iterate. I-ALM terminates when $\frac{\|X^{(k)}+\mathcal{S}^{(k)}-D\|_F}{\|D\|_F} \le 1 \times 10^{-8}$.

3. **APG**: For some $\bar{\lambda} > 0$, APG solves $\min\left\{\bar{\lambda}\left(\|X\|_* + \frac{1}{\sqrt{n}}\|\,\mathbf{vec}(S)\|_1\right) + \frac{1}{2}\|X + S - D\|_F^2\right\}$. Stopping tolerance is set to 5×10^{-11} (the definition of stopping criteria is complicated, for details see partial APG code at [`http://perception.csl.uiuc.edu/matrix-rank/home.html`]. In the code, by default $\bar{\lambda}$ is set to $\sigma_{\max}(D) \times 10^{-9}$.

4. **SVT**: SVT solves a relaxation of the robust PCA problem,

$$\min\left\{\bar{\lambda}\left(\|X\|_* + \frac{1}{\sqrt{n}}\|\,\mathbf{vec}(S)\|_1\right) + \frac{1}{2}(\|X\|_F^2 + \|S\|_F^2) : \ X + S = D\right\}.$$

   Let $(X^{(k)}, \mathcal{S}^{(k)})$ be the k-th iterate when $\bar{\lambda}$ is set to 1×10^3. SVT terminates $\frac{\|X^{(k)}+\mathcal{S}^{(k)}-D\|_F}{\|D\|_F} \le 5 \times 10^{-4}$. Note that we have chosen a weaker stopping criterion for SVT.

The results of the experiments are displayed in Table 5.1. In Table 5.1, the row labeled **CPU** lists the running time of each algorithm in *seconds* and all other rows are self-explanatory. The column labeled `average` lists the average taken over the 10 random instances, the columns labeled `min` (resp. `max`) list the minimum (resp. maximum) over the 10 instances. The experimental results in Table 5.1, show that FALC is competitive with the state of the art algorithms, e.g. I-ALM, APG and SVT, specialized for solving robust PCA problem. Even though FALC is not special purpose algorithm for robust PCA, in our numerical experiments, FALC required fewer singular value decompositions when compared to APG and SVT. In addition, for all 10 randomly created problems in the test set, only FALC and I-ALM accurately identified the zero-set of the sparse component S_0, i.e. $I_0 = \{(ij) \in \{1, 2, ..., n\} \times \{1, 2, ..., n\} : (S_0)_{ij} = 0\}$ without any thresholding. This feature of FALC is very appealing in practice. For signals with a large dynamic range, almost all of the state-of-the-art efficient algorithms produce a solution with many small non zeros terms, and it is often hard to determine the threshold.

	FALC			**I-ALM**		
	Average	**Min**	**Max**	**Average**	**Min**	**Max**
svd #	40	39	44	31.6	30	33
$\|\mathbf{X_{sol}} - \mathbf{X_0}\|_\mathbf{F}/\|\mathbf{X_0}\|_\mathbf{F}$	3.5E-09	2.7E-09	4.5E-09	1.9E-09	5.9E-10	3.4E-09
$\|\mathbf{S_{sol}} - \mathbf{S_0}\|_\mathbf{F}/\|\mathbf{S_0}\|_\mathbf{F}$	1.3E-07	1.0E-07	1.8E-07	1.9E-07	4.8E-08	3.8E-07
$\mid \|\mathbf{X_{sol}}\|_* - \|\mathbf{X_0}\|_*/\|\mathbf{X_0}\|_*$	1.6E-10	2.4E-11	3.6E-10	1.1E-11	3.7E-12	2.1E-11
$\max\{\|\sigma_\mathbf{i} - \sigma_\mathbf{i}^\mathbf{0}\| : \sigma_\mathbf{i}^\mathbf{0} > \mathbf{0}\}$	2.1E-07	1.0E-07	4.0E-07	8.7E-08	2.3E-08	2.5E-07
$\max\{\|\sigma_\mathbf{i}\| : \sigma_\mathbf{i}^\mathbf{0} = \mathbf{0}\}$	1.2E-13	7.2E-14	1.9E-13	1.5E-13	5.9E-14	3.7E-13
$\mid \|\mathbf{vec}(\mathbf{S_{sol}})\|_\mathbf{1} - \|\mathbf{vec}(\mathbf{X_0})\|_\mathbf{1}/\|\mathbf{vec}(\mathbf{X_0})\|_\mathbf{1}$	1.4E-08	4.1E-09	2.6E-08	2.2E-09	4.1E-10	5.1E-09
$\max\{\|(\mathbf{S_{sol}})_\mathbf{ij} - (\mathbf{S_0})_\mathbf{ij}\| : \|(\mathbf{S_0})_\mathbf{ij}\| > \mathbf{0}\}$	8.0E-07	5.0E-07	1.4E-06	1.1E-05	2.3E-06	2.5E-05
$\max\{\|(\mathbf{S_{sol}})_\mathbf{ij}\| : (\mathbf{S_0})_\mathbf{ij} = \mathbf{0}\}$	0	0	0	0	0	0
rank	25	25	25	25	25	25
$\|\mathbf{X_{sol}} + \mathbf{S_{sol}} - \mathbf{D}\|_\mathbf{F}/\|\mathbf{D}\|_\mathbf{F}$	3.5E-09	2.6E-09	4.5E-09	4.7E-09	1.1E-09	9.6E-09
CPU	22.9	19.6	27.8	16.8	13.5	24.3

	APG			**SVT**		
	Average	**Min**	**Max**	**Average**	**Min**	**Max**
svd #	187.7	187	188	833.9	819	857
$\|\mathbf{X_{sol}} - \mathbf{X_0}\|_\mathbf{F}/\|\mathbf{X_0}\|_\mathbf{F}$	4.1E-09	4.0E-09	4.4E-09	1.8E-04	1.8E-04	1.8E-04
$\|\mathbf{S_{sol}} - \mathbf{S_0}\|_\mathbf{F}/\|\mathbf{S_0}\|_\mathbf{F}$	1.6E-07	1.6E-07	1.7E-07	2.0E-02	2.0E-02	2.1E-02
$\mid \|\mathbf{X_{sol}}\|_* - \|\mathbf{X_0}\|_*/\|\mathbf{X_0}\|_*$	4.0E-09	3.8E-09	4.2E-09	1.7E-05	1.5E-05	1.9E-05
$\max\{\|\sigma_\mathbf{i} - \sigma_\mathbf{i}^\mathbf{0}\| : \sigma_\mathbf{i}^\mathbf{0} > \mathbf{0}\}$	2.0E-06	1.9E-06	2.1E-06	1.5E-02	1.2E-02	1.7E-02
$\max\{\|\sigma_\mathbf{i}\| : \sigma_\mathbf{i}^\mathbf{0} = \mathbf{0}\}$	1.3E-13	6.8E-14	1.9E-13	2.4E-13	7.6E-14	6.8E-13
$\mid \|\mathbf{vec}(\mathbf{S_{sol}})\|_\mathbf{1} - \|\mathbf{vec}(\mathbf{X_0})\|_\mathbf{1}/\|\mathbf{vec}(\mathbf{X_0})\|_\mathbf{1}$	1.8E-07	1.8E-07	1.9E-07	5.0E-03	4.9E-03	5.1E-03
$\max\{\|(\mathbf{S_{sol}})_\mathbf{ij} - (\mathbf{S_0})_\mathbf{ij}\| : \|(\mathbf{S_0})_\mathbf{ij}\| > \mathbf{0}\}$	2.0E-07	1.8E-07	2.3E-07	1.2E-01	1.1E-01	1.3E-01
$\max\{\|(\mathbf{S_{sol}})_\mathbf{ij}\| : (\mathbf{S_0})_\mathbf{ij} = \mathbf{0}\}$	3.7E-08	2.1E-08	6.6E-08	5.5E-03	3.6E-03	8.5E-03
rank	25	25	25	25	25	25
$\|\mathbf{X_{sol}} + \mathbf{S_{sol}} - \mathbf{D}\|_\mathbf{F}/\|\mathbf{D}\|_\mathbf{F}$	5.4E-09	5.2E-09	5.8E-09	5.0E-04	5.0E-04	5.0E-04
CPU	87.7	71.6	101.6	265.2	252.0	273.1

Table 5.1: FALC vs I-ALM, APG, SVT: Numerical Test Results for PCP problem with $n = 500$, $r = 0.05n^2$, $p = 0.05n$

5.3. Stable Principal Component Pursuit Problem. In this section, we solve the problem

$$\begin{array}{rl} \min_{X,S\in\mathbb{R}^{m\times n}} & \|X\|_* + \mu_2\|\,\mathbf{vec}(S)\|_1, \\ \text{subject to} & \|\,\mathbf{vec}(X+S-D)\|_\infty \le \delta, \end{array} \tag{5.3}$$

and report the results of our numerical experiments using FALC. To best of our knowledge, there are no publicly available code specialized for solving problem in (5.3), other than general purpose SDP solvers.

We created 10 random problems of size $n = 500$, i.e. $D \in \mathbb{R}^{500\times 500}$ using the procedure described in Section 5.1, where δ is set to 1×10^4, i.e. each entry of the noise term ζ_0 is coming from a uniform distribution between $[-\delta, \delta]$. We chose the value of the stopping parameter so that FALC produces a solution X_{sol} and S_{sol} with $\frac{\|X_{sol}+S_{sol}-D\|_F}{\|D\|_F} \approx 1 \times 10^{-5}$.

Problem in (5.3) is a special case of (1.1) and $f^{(k)}(X, s, y)$ defined in (3.5) simplifies to $f^{(k)}(X, S, y) = \frac{1}{2}\|\,\mathbf{vec}(X + S) - y - \mathbf{vec}(D) - \lambda^{(k)}\theta_1^{(k)}\|_2^2$ (note that for all $k \ge 1$, $\theta_2^{(k)} = 0$). We set the parameter values

	FALC		
	Average	**Min**	**Max**
svd #	59.3	55	64
$\|\mathbf{X_{sol}} - \mathbf{X_0}\|_\mathbf{F}/\|\mathbf{X_0}\|_\mathbf{F}$	1.7E-05	1.7E-05	1.7E-05
$\|\mathbf{S_{sol}} - \mathbf{S_0}\|_\mathbf{F}/\|\mathbf{S_0}\|_\mathbf{F}$	3.7E-04	3.0E-04	4.4E-04
$\mid \|\mathbf{X_{sol}}\|_* - \|\mathbf{X_0}\|_*\mid/\|\mathbf{X_0}\|_*$	1.6E-05	1.6E-05	1.6E-05
$\max\{\|\sigma_\mathbf{i} - \sigma_\mathbf{i}^\mathbf{0}\| : \sigma_\mathbf{i}^\mathbf{0} > \mathbf{0}\}$	9.9E-03	9.7E-03	1.0E-02
$\max\{\|\sigma_\mathbf{i}\| : \sigma_\mathbf{i}^\mathbf{0} = \mathbf{0}\}$	1.6E-13	3.6E-14	3.1E-13
$\mid \|\operatorname{\mathbf{vec}}(\mathbf{S_{sol}})\|_\mathbf{1} - \|\operatorname{\mathbf{vec}}(\mathbf{X_0})\|_\mathbf{1}\mid/\|\operatorname{\mathbf{vec}}(\mathbf{X_0})\|_\mathbf{1}$	2.3E-04	2.2E-04	2.4E-04
$\max\{\|(\mathbf{S_{sol}})_\mathbf{ij} - (\mathbf{S_0})_\mathbf{ij}\| : \|(\mathbf{S_0})_\mathbf{ij}\| > \mathbf{0}\}$	3.9E-03	3.0E-03	4.6E-03
$\max\{\|(\mathbf{S_{sol}})_\mathbf{ij}\| : (\mathbf{S_0})_\mathbf{ij} = \mathbf{0}\}$	6.4E-05	0.0E+00	2.3E-04
rank	25	25	25
$\|\mathbf{X_{sol}} + \mathbf{S_{sol}} - \mathbf{D}\|_\mathbf{F}/\|\mathbf{D}\|_\mathbf{F}$	2.1E-05	2.0E-05	2.2E-05
CPU	34.6	26.1	48.3

Table 5.2: FALC: Numerical Test Results for SPCP problem with $n = 500$, $r = 0.05n^2$, $p = 0.05n$, $\delta = 1 \times 10^{-4}$

for FALC by solving a set of small size problems and these parameter values were fixed throughout the experiments, all other parameters are set to their default values, i.e. $c_\tau = 0.4$, $c_\epsilon = 0.4$, $c_\xi = 0.4$, $c_\lambda = 0.4$, $\bar{c}_\tau = 0.999$, $\bar{c}_\epsilon = 0.999$, $\bar{c}_\xi = 0.999$. We set $\bar{c}_\lambda = 1.5$ and initialize $\theta_1^{(1)}$ as in [29], i.e. as in (5.2).

Finally, We set $\varrho = 1 \times 10^{-5}$, $\varsigma = 1 \times 10^{-3}$ and terminate FALC when either the distance between successive inner iterates are below a threshold ϱ for each component, i.e. $\| \mathbf{vec}\left(X_1^{(k,\ell)}\right) - \mathbf{vec}\left(X_1^{(k,\ell-1)}\right) \|_\infty \leq \varrho$, $\| \mathbf{vec}\left(s_1^{(k,\ell)}\right) - \mathbf{vec}\left(s_1^{(k,\ell-1)}\right) \|_\infty \leq \varrho$ for any $k \geq 1$ or there exist partial subgradients with sufficiently small norm for each component, i.e.

$$\|G\|_F \leq \varsigma/2, \ \|g\|_2 \leq \varsigma \ \text{ for some } \ (G, g) \in \partial_{X,s} P^{(k)}(.,.,.)|_{\breve{Z}^{(k,\ell)}}$$

and

$$\|\breve{y}_1^{(k,\ell)} - \Pi_\mathcal{Q}\left(\breve{y}_1^{(k,\ell)} - \frac{1}{L}\nabla_y P^{(k)}(\breve{Z}^{(k,\ell)})\right)\|_2 \leq \varsigma,$$

where $\breve{Z}^{(k,\ell)} = \left(\breve{X}_1^{(k,\ell)}, \breve{s}_1^{(k,\ell)}, \breve{y}_1^{(k,\ell)}\right)$ is defined at the beginning of Section 4.

We have used PROPACK [27] for computing partial singular value decompositions. In order to estimate the rank of X_0, we followed the scheme proposed in Equation (17) in [29]. The results of the experiments are displayed in Table 5.2.

6. Extension to general composite norm problem. The algorithmic framework proposed in this paper extends to the following much more general class of problems given in (1.2). By introducing slack variables, (1.2) can be reformulated as follows.

$$\begin{array}{rl} \min & \mu_1\|\sigma(S)\|_\alpha + \mu_2\|s\|_\beta + \mu_3 H(X), \\ \text{subject to} & \mathcal{F}(X) - S = G, \\ & \mathcal{C}(X) - s = d, \\ & \mathcal{A}(X) - y = b, \quad y \in \mathcal{Q}, \end{array} \tag{6.1}$$

where the decision variables $X \in \mathbb{R}^{m\times n}$, $S \in \mathbb{R}^{r_1 \times r_2}$, $s \in \mathbb{R}^p$, and $y \in \mathbb{R}^q$. $H(.)$ is a strongly convex function with convexity parameter ς. We continue to assume that $\mathcal{A}$ is surjective; however, when $\mu_3 > 0$, we no longer require that at least one of that at least one of $\mathcal{F}$ and $\mathcal{C}$ is an injective linear map

In this more general setting, the FALC inexactly solves subproblems of the form:

$$\min_{X,\ S,\ s,\ y\in\mathcal{Q}} P^{(k)}(X, S, s, y), \tag{6.2}$$

where

$$\begin{aligned} P^{(k)}(X,S,s,y) &:= \lambda^{(k)}\big(\mu_1\|\sigma(S)\|_\alpha + \mu_2\|s\|_\beta + \mu_3 H(X)\big) + f^{(k)}(X,S,s,y), \\ f^{(k)}(X,S,s,y) &:= \ \tfrac{1}{2}\|\mathcal{F}(X) - S - G - \lambda^{(k)}\theta_3^{(k)}\|_F^2 + \tfrac{1}{2}\|\mathcal{C}(X) - s - d - \lambda^{(k)}\theta_2^{(k)}\|_2^2 \\ &\quad + \tfrac{1}{2}\|\mathcal{A}(X) - y - b - \lambda^{(k)}\theta_1^{(k)}\|_2^2. \end{aligned}$$

Let $(X_*^{(k)}, S_*^{(k)}, s_*^{(k)}, y_*^{(k)}) \in \operatorname{argmin} P^{(k)}(X, S, s, y)$. Suppose the initial iterate $X^{(0)}$ is feasible, i.e. $\mathcal{A}(X^{(0)}) - b \in \mathcal{Q}$. Let $S^{(0)} := \mathcal{F}(X^{(0)}) - G$, $s^{(0)} := \mathcal{C}(X^{(0)}) - d$, $y^{(0)} := \mathcal{A}(X^{(0)}) - b$, and $\eta := \mu_1 \|\sigma(\mathcal{F}(X^{(0)}) - G)\|_\alpha + \mu_2 \|\mathcal{C}(X^{(0)}) - d\|_\beta$.

The particular implementation of FALC depends on the nature of the objective function. In all cases we need to ensure that the iterate sequence $\{X^{(k)}\}_{k\in\mathbb{Z}_+}$ is bounded so that it has a limit point. First consider the case where $\mu_3 > 0$. Strong convexity property of $H(.)$ implies that

$$\begin{aligned}\mu_1 \|\sigma(S_*^{(k)})\|_\alpha + \mu_2 \|s_*^{(k)}\|_\beta + \frac{\varsigma}{2} \|X_*^{(k)} - \left(X^{(0)} - \frac{\nabla H(X^{(0)})}{\varsigma}\right)\|_F^2 \\ \leq \eta + \frac{1}{2\varsigma} \|\nabla H(X^{(0)})\|_F^2 + \frac{\lambda^{(k)}}{2} \sum_{i=1}^{3} \|\theta_i^{(k)}\|_F^2.\end{aligned}$$

Therefore, we can define $\mathcal{S}^{(k)}$ in line 9 in Figure 3.1 as follows:

$$\mathcal{S}^{(k)} := \left\{ (X, S, s, y) : \mu_1 \|\sigma(S)\|_\alpha + \mu_2 \|s\|_\beta \leq \eta_1^{(k)}, \ \left\| X - \left(X^{(0)} - \frac{\nabla H(X^{(0)})}{\varsigma}\right)\right\|_F \leq \sqrt{\frac{2}{\varsigma}\eta_1^{(k)}} \right\},$$

where

$$\eta_1^{(k)} := \eta + \frac{1}{2\varsigma} \|\nabla H(X^{(0)})\|_F^2 + \frac{\lambda^{(k)}}{2} \sum_{i=1}^{3} \|\theta_i^{(k)}\|_F^2. \tag{6.3}$$

The only change in the algorithm is that we need to compute ∇H at every iteration of **Algorithm APG** additional to one ∇f computation per iteration.

When $\mu_3 = 0$, we set

$$\mathcal{S}^{(k)} := \{(X, S, s, y) : \ \mu_1 \|\sigma(S)\|_\alpha + \mu_2 \|s\|_\beta \leq \eta_1^{(k)}\}.$$

ensuring that the iterates $\{(S^{(k)}, s^{(k)}, y^{(k)})\}_{k\in\mathbb{Z}_+}$ are bounded, see Lemma 3.2 for details. Since at least one of $\mathcal{F}$ and $\mathcal{C}$ is injective, this implies that $\{X^{(k)}\}_{k\in\mathbb{Z}_+}$ is a bounded sequence. Note that without the injectivity assumption, the sequence $\{X^{(k)}\}_{k\in\mathbb{Z}_+}$ may not have a limit point.

The general formulation (1.2) subsumes a number of different problems as special cases – see Section 1.2 for details. And, our experience with FALC leads us to believe that this algorithm is likely to be very competitive for solving all these special cases. However, the assumption that the operator $\mathcal{A}$, defining the constraints, be surjective can be restrictive. In some applications, the feasible region is the intersection of cones of the form: $\{X : \mathcal{A}_1(X) - b_1 \in \mathcal{Q}_1, \mathcal{A}_2(X) - b_2 \in \mathcal{Q}_2\}$. While, it is often the case that each $\mathcal{A}_i$, $i = 1, 2$, is surjective, the product operator $\mathcal{A}(X) = [\mathcal{A}_1(X), \mathcal{A}_2(X)]$ is not. Thus, FALC cannot be used for these problems. The extension to intersection of cones is non-trivial and one would have to design a completely new set of techniques.

The main contribution of this paper is an efficient first-order augmented Lagrangian algorithm (FALC) for the composite norm minimization problem (1.1) and for its extension (1.2). FALC solves the composite norm minimization problem by solving a sequence of augmented Lagrangian subproblems, where each subproblem is solved using **Algorithm APG** in Figure 2.1. **Algorithm APG** is essentially Algorithm 2 in [37] (see also FISTA [4]) with early termination. We show that the continuation scheme on penalty parameter λ used in FALC guarantees that the iterate sequence provably converges to the solution and we are also able to compute a convergence rate. The performance of FALC in our numerical experiments has been very promising. To best of our knowledge, for the stable PCA problem, FALC is the first algorithm with a known complexity bound.

REFERENCES

[1] N. S. Aybat and A. Chakraborty, *Fast reconstruction of CT images from parsimonious angular measurements via compressed sensing*, tech. report, Siemens Corp. Research, 2009.

[2] N. S. Aybat and G. Iyengar, *A first-order smoothed penalty method for compressed sensing*, SIAM Journal on Optimization, 21 (2011), pp. 287–313.
[3] ———, *A first-order augmented Lagrangian method for compressed sensing*, SIAM Journal on Optimization, 22 (2012), pp. 429–459.
[4] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 183–202.
[5] S. Becker, J. Bobin, and E. Candès, *Nesta: a fast and accurate first-order method for sparse recovery*, SIAM J. Imaging Sci., 4 (2011), pp. 1–39.
[6] J. Cai, E. Candès, and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM Journal on Optimization, 20 (2008), pp. 1956–1982.
[7] E. Candès and J. Romberg, *Quantitative robust uncertainty principles and optimally sparse decompositions*, Foundations of Computational Mathematics, 6 (2006), pp. 227–254.
[8] E. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. Info. Th., 52 (2006).
[9] E. Candès and T. Tao, *Near optimal signal recovery from random projections: universal encoding strategies?*, IEEE Trans. Info. Th., 52 (2006), pp. 5406–5425.
[10] E. J. Candès, X. Li, Y. Ma, and Wright J., *Robust principle component analysis?*, submitted for publication, (2009).
[11] E. J. Cands and B. Recht, *Exact matrix completion via convex optimization*, Foundations of Computational Mathematics, 9 (2008), pp. 717–772.
[12] A. d'Aspremont, F. R. Bach, and L. El. Ghaoui, *Optimal solutions for sparse principle component analysis*, Journal of Machine Learning Research, 9 (2008), pp. 1269–1294.
[13] A. d'Aspremont, L. El. Ghaoui, M. I. Jordan, and G. R. G. Lanckriet, *A direct formulation for sparse pca using semidefinite programming*, SIAM Review, 49 (2007), pp. 434–448.
[14] I. Daubechies, M. Fornasier, and I. Loris, *Accelerated projected gradient method for linear inverse problems with sparsity constraints*, Journal of Fourier Analysis and Applications, 14 (2008), pp. 764–792.
[15] D. Donoho, *Compressed sensing*, IEEE Trans. Info. Th., 52 (2006), pp. 1289–1306.
[16] L. El Ghaoui and P. Gahinet, *Rank minimization under lmi constraints: A framework for output feedback problems*, in Proceedings of the European Control Conference, 1993.
[17] M. Fazel, H. Hindi, and S. Boyd, *Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices*, in Proceedings of American Control Conference, Denver, Colorado, June 2003.
[18] ———, *A rank minimization heuristic with application to minimum order system approximation*, in Proceedings of the American Control Conference, 2003, pp. 2156–2162.
[19] ———, *Rank minimization and applications in system theory*, in American Control Conference, 2004, pp. 3273–3278.
[20] M. Fazel, T. K. Pong, D. Sun, and P. Tseng, *Hankel matrix rank minimization with applications in system identification and realization*. Submitted for publication, 2012.
[21] M. A. Figueiredo, R. Nowak, and S. J. Wright, *Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems*, IEEE Journal of Selected Topics in Signal Processing, 1 (2007), pp. 586–597.
[22] D. Goldfarb, S. Ma, and K. Scheinberg, *Fast alternating linearization methods for minimizing the sum of two convex functions*. arXiv:0912.4571v2, October 2010.
[23] E. T. Hale, W. Yin, and Y. Zhang, *A fixed-point continuation for $\ell 1$-regularized minimization with applications to compressed sensing*, tech. report, Rice University, 2007.
[24] ———, *Fixed-point continuation for $\ell 1$-minimization: Methodology and convergence*, SIAM Journal on Optimization, 19 (2008), pp. 1107–1130.
[25] M. Journée, Y. Nesterov, P. Richtárik, and Sepulchre R., *Generalized power method for sparse principle component analysis*, Journal of Machine Learning Research, 11 (2010), pp. 517–553.
[26] K. Koh, S. J. Kim, and S. Boyd, *Solver for $\ell 1$-regularized least squares problems*, tech. report, Stanford University, 2007.
[27] R.M. Larsen, *Lanczos bidiagonalization with partial reorthogonalization*, Technical report DAIMI PB-357, Department of Computer Science, Aarhus University, 1998.
[28] A. S. Lewis, *The convex analysis of unitarily invariant matrix norms*, Journal of Convex Analysis, 2 (1995), pp. 173–183.
[29] Z. Lin, M. Chen, L. Wu, and Y. Ma, *The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices*, arXiv:1009.5055v2, (2011).
[30] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma, *Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix*, tech. report, UIUC Technical Report UILU-ENG-09-2214, 2009.
[31] N. Linial, E. London, and Y. Rabinovich, *The geometry of graphs and some of its algorithmic applications*, Combinatorica, 15 (1995), pp. 215–245.
[32] Z. Liu and L. Vandenberghe, *Interior-point method for nuclear norm approximation with application to system identification*, SIAM. J. Matrix Anal. & Appl., 31 (2009), pp. 1235–1256.
[33] S. Ma, D. Goldfarb, and L. Chen, *Fixed point and bregman iterative methods for matrix rank minimization*, Mathematical Programming Series A, 128 (2011), pp. 321–353.
[34] Netflix prize. http://www.netflixprize.com/.
[35] B. Recht, M. Fazel, and P. Parrilo, *Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization*, SIAM Review, 52 (2010), pp. 471–501.
[36] K.C. Toh and S. Yun, *An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems*. preprint, 2010.
[37] P. Tseng, *On accelerated proximal gradient methods for convex-concave optimization*. 2008.

[38] E. Van den Berg and M. P. Friedlander, *Probing the pareto frontier for basis pursuit solutions*, SIAM Journal on Scientific Computing, 31 (2008), pp. 890–912.
[39] Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang, *A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization and continuation*, To appear in SIAM Journal on Scientific Computing, (2009).
[40] J. Yang and Y. Zhang, *Alternating direction algorithms for l1-problems in compressive sensing*, Tech. Report TR09-37, CAAM, Rice University, 2009.
[41] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, *Bregman iterative algorithms for ℓ_1 minimization with applications to compressed sensing*, SIAM Journal on Imaging Sciences, 1 (2008), pp. 143–168.
[42] Z. Zhou, X. Li, J. Wright, E. Candès, and Y. Ma, *Stable principle component pursuit*, in Proceedings of International Symposium on Information Theory, 2010.

Appendix A. Proofs of technical results.

Lemma A.1 and proof.

Lemma A.1. *Let $\mathcal{Q} \subset \mathbb{R}^q$ be nonempty closed convex set such that $\{X \in \mathbb{R}^{m\times n} : \mathcal{A}(X) - b \in \mathcal{Q}\} \neq \emptyset$, where $\mathcal{A}$ is surjective; and let $(X_*^{(k)}, s_*^{(k)}, y_*^{(k)})$ is an optimal solution to (3.4). Then, for all $k \geq 1$,*

$$\|y_*^{(k)}\|_2 \leq \sigma_{\max}(A)\|X_*^{(k)}\|_F + \|b + \lambda^{(k)}\theta_1^{(k)}\|_2 + 2 \min_{\tilde{y}\in\mathcal{Q}}\{\|\tilde{y}\|_2\}. \tag{A.1}$$

Proof. From the first order optimality conditions for (3.4), we have $y_*^{(k)} = \Pi_{\mathcal{Q}}(\mathcal{A}(X_*^{(k)}) - b - \lambda^{(k)}\theta_1^{(k)})$. Since Euclidean projection is nonexpansive, we have

$$\|y_*^{(k)} - \tilde{y}\|_2 \leq \|\mathcal{A}(X_*^{(k)}) - b - \lambda^{(k)}\theta_1^{(k)} - \tilde{y}\|_2 \quad \forall \tilde{y} \in \mathcal{Q}. \tag{A.2}$$

The result now follows from the triangular inequality. □

This result implies several simple bounds on $\|y_*^{(k)}\|_2$. Since the initial iterate $X^{(0)}$ is feasible, i.e. $\mathcal{A}(X^{(0)}) - b \in \mathcal{Q}$, it follows that

$$\|y_*^{(k)}\|_2 \leq \eta_2^{(k)} := \sigma_{\max}(A)\|X_*^{(k)}\|_F + \|b + \lambda^{(k)}\theta_1^{(k)}\|_2 + 2\|\mathcal{A}(X^{(0)}) - b\|_2. \tag{A.3}$$

Suppose $0 \in \mathcal{Q}$. Then $\|y_*^{(k)}\|_2 \leq \eta_2^{(k)} := \sigma_{\max}(A)\|X_*^{(k)}\|_F + \|b + \lambda^{(k)}\theta_1^{(k)}\|_2$. When $\mathcal{Q}$ is bounded with $\mathcal{Q} \subseteq \{y : \|y\|_2 \leq \eta_2\}$. Then, one can set $\eta_2^{(k)} := \eta_2$ for all $k \geq 1$.

Lemma A.2 and proof.

Lemma A.2. *Fix α, $\beta \in \{1, 2, \infty\}$. Let*

$$P(X, s, y) = \lambda(\mu_1\|\sigma(X)\|_\alpha + \mu_2\|s\|_\beta) + f(X, s, y)$$

where

$$f(X, s, y) = \frac{1}{2}\|\mathcal{A}(X) - y - b - \lambda\theta_1\|_2^2 + \frac{1}{2}\|\mathcal{C}(X) - s - d - \lambda\theta_2\|_2^2.$$

Suppose $(\bar{X}, \bar{s}, \bar{y})$ is ϵ-optimal for the problem $\min_{X,s,y}\{P(X, s, y) : \ y \in \mathcal{Q}\}$, i.e.

$$0 \leq P(\bar{X}, \bar{s}, \bar{y}) - \min_{X\in\mathbb{R}^{m\times n},\ s\in\mathbb{R}^p,\ y\in\mathcal{Q}\subset\mathbb{R}^q} P(X, s, y) \leq \epsilon.$$

Then we have

$$\|\mathcal{C}(\bar{X}) - \bar{s} - d - \lambda\theta_2\|_2 \leq J(\beta^*)\mu_2\lambda + \sigma_{max}(M)\sqrt{2\epsilon},$$
$$\|\mathcal{A}^*\left(\mathcal{A}(\bar{X}) - \bar{y} - b - \lambda\theta_1\right) + \mathcal{C}^*\left(\mathcal{C}(\bar{X}) - \bar{s} - d - \lambda\theta_2\right)\|_F \leq I(\alpha^*)\mu_1\lambda + \sigma_{max}(M)\sqrt{2\epsilon},$$

where $M = \begin{pmatrix} -I & 0 & C \\ 0 & -I & A \end{pmatrix}$, $\frac{1}{\alpha^*} + \frac{1}{\alpha} = 1$ (resp. $\frac{1}{\beta^*} + \frac{1}{\beta} = 1$) is the Hölder conjugate of α (resp. β) and the functions $I(\cdot)$ and $J(\cdot)$ are defined in (3.10).

In order to prove for Lemma A.2, we need the following result.

THEOREM A.3. *Let $f : \mathbb{R}^{m\times n} \times \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ denote a convex function with a Lipschitz continuous gradient ∇f with a Lipschitz constant L with respect to the norm $\|(X,s,y)\| = \sqrt{\|X\|_F^2 + \|s\|_2^2 + \|y\|_2^2}$. Let $(X_*, s_*, y_*) \in \operatorname{argmin}_{X,s,y}\{\lambda(\mu_1\|\sigma(X)\|_\alpha + \mu_2\|s\|_\beta) + f(X,s,y) : y \in \mathcal{Q}\}$. Suppose $(\bar{X}, \bar{s}, \bar{y}) \in \mathbb{R}^{m\times n} \times \mathbb{R}^p \times \mathbb{R}^q$ such that $\bar{y} \in \mathcal{Q}$ satisfies*

$$\lambda\big(\mu_1\|\sigma(\bar{X})\|_\alpha + \mu_2\|\bar{s}\|_\beta\big) + f(\bar{X},\bar{s},\bar{y}) \le \lambda\big(\mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|s_*\|_\beta\big) + f(X_*,s_*,y_*) + \epsilon$$

for some $\epsilon > 0$. Then

$$\|\nabla_X f(\bar{X},\bar{s},\bar{y})\|_F \le \big(\sqrt{2L\epsilon} + I(\alpha^*)\lambda\mu_1\big), \qquad \|\nabla_s f(\bar{X},\bar{s},\bar{y})\|_2 \le \big(\sqrt{2L\epsilon} + J(\beta^*)\lambda\mu_2\big)$$

where $\frac{1}{\alpha^*} + \frac{1}{\alpha} = 1$ (resp. $\frac{1}{\beta^*} + \frac{1}{\beta} = 1$) is the Hölder conjugate of α (resp. β) and the functions $I(\cdot)$ and $J(\cdot)$ are defined in (3.11).

Proof. Since ∇f is Lipschitz continuous with constant L, the triangular inequality for $\|\sigma(.)\|_\alpha$ and $\|.\|_\beta$ implies that for any $X \in \mathbb{R}^{m\times n}$, $s \in \mathbb{R}^p$ and $y \in \mathbb{R}^q$

$$\begin{aligned}
&\lambda(\mu_1\|\sigma(X)\|_\alpha + \mu_2\|s\|_\beta) + f(X,s,y)\\
&\le \lambda(\mu_1\|\sigma(\bar{X})\|_\alpha + \mu_2\|\bar{s}\|_\beta) + f(\bar{X},\bar{s},\bar{y}) + \lambda(\mu_1\|\sigma(X-\bar{X})\|_\alpha + \mu_2\|s-\bar{s}\|_\beta)\\
&\quad + \big\langle \nabla_X f(\bar{X},\bar{s},\bar{y}), (X-\bar{X})\big\rangle + \nabla_s f(\bar{X},\bar{s},\bar{y})^T(s-\bar{s}) + \nabla_y f(\bar{X},\bar{s},\bar{y})^T(y-\bar{y})\\
&\quad + \frac{L}{2}\|X-\bar{X}\|_F^2 + \frac{L}{2}\|s-\bar{s}\|_2^2 + \frac{L}{2}\|y-\bar{y}\|_2^2,
\end{aligned}$$

where $\langle X, Y\rangle = \mathbf{Tr}(X^TY) \in \mathbb{R}$ denotes the usual Euclidean inner product of $X \in \mathbb{R}^{m\times n}$ and $Y \in \mathbb{R}^{m\times n}$. Since X, s and y are arbitrary, it follows that

$$\begin{aligned}
&\lambda(\mu_1\|\sigma(X_*)\|_\alpha + \mu_2\|s_*\|_\beta) + f(X_*,s_*,y_*)\\
&\le \lambda(\mu_1\|\sigma(\bar{X})\|_\alpha + \mu_2\|\bar{s}\|_\beta) + f(\bar{X},\bar{s},\bar{y})\\
&\quad + \min_{X\in\mathbb{R}^{m\times n}}\left\{\big\langle \nabla_X f(\bar{X},\bar{s},\bar{y}), X-\bar{X}\big\rangle + \frac{L}{2}\|X-\bar{X}\|_F^2 + \lambda\mu_1\|\sigma(X-\bar{X})\|_\alpha\right\}\\
&\quad + \min_{s\in\mathbb{R}^p}\left\{\nabla_s f(\bar{X},\bar{s},\bar{y})^T(s-\bar{s}) + \frac{L}{2}\|s-\bar{s}\|_2^2 + \lambda\mu_2\|s-\bar{s}\|_\beta\right\}\\
&\quad + \min_{y\in\mathcal{Q}\subset\mathbb{R}^q}\left\{\nabla_y f(\bar{X},\bar{s},\bar{y})^T(y-\bar{y}) + \frac{L}{2}\|y-\bar{y}\|_2^2\right\}. \qquad \text{(A.4)}
\end{aligned}$$

The first minimization problem on the right hand side of (A.4) can be simplified as follows:

$$\begin{aligned}
&\min_{X\in\mathbb{R}^{m\times n}}\left\{\big\langle \nabla_X f(\bar{X},\bar{s},\bar{y}), X-\bar{X}\big\rangle + \frac{L}{2}\|X-\bar{X}\|_F^2 + \lambda\mu_1\|\sigma(X-\bar{X})\|_\alpha\right\}\\
&= \max_{W:\|\sigma(W)\|_{\alpha^*}\le\lambda\mu_1}\min_{X\in\mathbb{R}^{m\times n}}\left\{\frac{L}{2}\|X-\bar{X}\|_F^2 + \big\langle \nabla_X f(\bar{X},\bar{s},\bar{y}) + W,\ X-\bar{X}\big\rangle\right\}, \qquad \text{(A.5)}\\
&= \max_{W:\|\sigma(W)\|_{\alpha^*}\le\lambda\mu_1}\left\{\frac{L}{2}\|X^*(W)-\bar{X}\|_F^2 + \big\langle \nabla_X f(\bar{X},\bar{s},\bar{y}) + W,\ X^*(W)-\bar{X}\big\rangle\right\},\\
&= -\min_{W:\|\sigma(W)\|_{\alpha^*}\le\lambda\mu_1}\frac{\|\nabla_X f(\bar{X},\bar{s},\bar{y}) + W\|_F^2}{2L}, \qquad \text{(A.6)}
\end{aligned}$$

$X^*(W) = \bar{X} - \frac{\nabla_X f(\bar{X},\bar{s},\bar{y})+W}{L}$ is the minimizer of the inner minimization problem in (A.5).

The second minimization problem on the right hand side of (A.4) can be simplified as follows:

$$\begin{aligned}
&\min_{s\in\mathbb{R}^p}\left\{\nabla_s f(\bar{X},\bar{s},\bar{y})^T(s-\bar{s})+\frac{L}{2}\|s-\bar{s}\|_2^2+\lambda\mu_2\|s-\bar{s}\|_\beta\right\}\\
&=\max_{u:\|u\|_{\beta^*}\le\lambda\mu_2}\min_{s\in\mathbb{R}^p}\left\{\frac{L}{2}\|s-\bar{s}\|_2^2+(\nabla_s f(\bar{X},\bar{s},\bar{y})+u)^T(s-\bar{s})\right\}, \qquad (A.7)\\
&=\max_{u:\|u\|_{\beta^*}\le\lambda\mu_2}\left\{\frac{L}{2}\|s^*(u)-\bar{s}\|_2^2+(\nabla_s f(\bar{X},\bar{s},\bar{y})+u)^T(s^*(u)-\bar{s})\right\},\\
&=-\min_{u:\|u\|_{\beta^*}\le\lambda\mu_2}\frac{\|\nabla_s f(\bar{X},\bar{s},\bar{y})+u\|_2^2}{2L}, \qquad (A.8)
\end{aligned}$$

$s^*(u)=\bar{s}-\frac{\nabla_s f(\bar{X},\bar{s},\bar{y})+u}{L}$ is the minimizer of the inner minimization problem in (A.7).

Since $\bar{y}\in\mathcal{Q}$, the following is true for the third minimization problem on the right hand side of (A.4).

$$\min_{y\in\mathcal{Q}\subset\mathbb{R}^q}\left\{\nabla_y f(\bar{X},\bar{s},\bar{y})^T(y-\bar{y})+\frac{L}{2}\|y-\bar{y}\|_2^2\right\}\le 0. \qquad (A.9)$$

Thus, (A.4), (A.6), (A.8) and (A.9) together imply that

$$\begin{aligned}
\lambda(\mu_1\|\sigma(X_*)\|_\alpha+\mu_2\|s_*\|_\beta)+f(X_*,s_*,y_*)\le\ &\lambda(\mu_1\|\sigma(\bar{X})\|_\alpha+\mu_2\|\bar{s}\|_\beta)+f(\bar{X},\bar{s},\bar{y})\\
&-\min_{W:\|\sigma(W)\|_{\alpha^*}\le\lambda\mu_1}\frac{\|\nabla_X f(\bar{X},\bar{s},\bar{y})+W\|_F^2}{2L}\\
&-\min_{u:\|u\|_{\beta^*}\le\lambda\mu_2}\frac{\|\nabla_s f(\bar{X},\bar{s},\bar{y})+u\|_2^2}{2L}.
\end{aligned}$$

Since $\Big(\lambda(\mu_1\|\sigma(\bar{X})\|_\alpha+\mu_2\|\bar{s}\|_\beta)+f(\bar{X},\bar{s},\bar{y})\Big)-\Big(\lambda(\mu_1\|\sigma(X_*)\|_\alpha+\mu_2\|s_*\|_\beta)+f(X_*,s_*,y_*)\Big)\le\epsilon$, we have that

$$\min_{W:\|\sigma(W)\|_{\alpha^*}\le\lambda\mu_1}\|\nabla_X f(\bar{X},\bar{s},\bar{y})+W\|_F^2+\min_{u:\|u\|_{\beta^*}\le\lambda\mu_2}\|\nabla_s f(\bar{X},\bar{s},\bar{y})+u\|_2^2\le 2L\epsilon. \qquad (A.10)$$

From (3.10), it follows that $\|W\|_F\le I(\alpha^*)\|\sigma(W)\|_{\alpha^*}$. Thus, (A.10) implies that

$$\min_{W:\|W\|_F\le I(\alpha^*)\lambda\mu_1}\|\nabla_X f(\bar{X},\bar{s},\bar{y})+W\|_F^2\le 2L\epsilon. \qquad (A.11)$$

Suppose $\|\nabla_X f(\bar{X},\bar{s},\bar{y})\|_F>I(\alpha^*)\lambda\mu_1$. Then the optimal solution of the optimization problem in (A.11) is

$$W^*=-I(\alpha^*)\lambda\mu_1\cdot\frac{\nabla_X f(\bar{X},\bar{s},\bar{y})}{\|\nabla_X f(\bar{X},\bar{s},\bar{y})\|_F}.$$

Then (A.10) implies that $(\|\nabla_X f(\bar{X},\bar{s},\bar{y})\|_F-I(\alpha^*)\lambda\mu_1)^2\le 2L\epsilon$, i.e. $\|\nabla_X f(\bar{X},\bar{s},\bar{y})\|_F\le\sqrt{2L\epsilon}+I(\alpha^*)\lambda\mu_1$. This is trivially true when $\|\nabla_X f(\bar{X},\bar{s},\bar{y})\|_F\le I(\alpha^*)\lambda\mu_1$. Therefore, we can conclude that always

$$\|\nabla_X f(\bar{X},\bar{s},\bar{y})\|_F\le\sqrt{2L\epsilon}+I(\alpha^*)\lambda\mu_1.$$

A similar analysis establishes that $\|\nabla_s f(\bar{X},\bar{s},\bar{y})\|_2\le\sqrt{2L\epsilon}+J(\beta^*)\lambda\mu_2$. □

Now we are ready to prove Lemma A.2.

Proof. Let $f(X,s,y)=\frac{1}{2}\|\mathcal{A}(X)-y-b-\lambda\theta_1\|_2^2+\frac{1}{2}\|\mathcal{C}(X)-s-d-\lambda\theta_2\|_2^2$ and let $\|(X,s,y)\|=$

$\sqrt{\|X\|_F^2 + \|s\|_2^2 + \|y\|_2^2}$, then for any $X_1, X_2 \in \mathbb{R}^{m\times n}$, $s_1, s_2 \in \mathbb{R}^p$ and $y_1, y_2 \in \mathbb{R}^q$, we have

$$\begin{aligned}
&\|\nabla f(X_1, s_1, y_1) - \nabla f(X_2, s_2, y_2)\|^2 \\
&= \left\| \begin{pmatrix} \nabla_X f(X_1, s_1, y_1) - \nabla_X f(X_2, s_2, y_2) \\ \nabla_s f(X_1, s_1, y_1) - \nabla_s f(X_2, s_2, y_2) \\ \nabla_y f(X_1, s_1, y_1) - \nabla_y f(X_2, s_2, y_2) \end{pmatrix} \right\|^2, \\
&= \|\nabla_X f(X_1, s_1, y_1) - \nabla_X f(X_2, s_2, y_2)\|_F^2 + \|\nabla_s f(X_1, s_1, y_1) - \nabla_s f(X_2, s_2, y_2)\|_2^2 \\
&\quad + \|\nabla_y f(X_1, s_1, y_1) - \nabla_y f(X_2, s_2, y_2)\|_2^2, \\
&= \|\mathcal{A}^*(\mathcal{A}(X_1 - X_2) - y_1 + y_2) + \mathcal{C}^*(\mathcal{C}(X_1 - X_2) - s_1 + s_2)\|_F^2 \\
&\quad + \|\mathcal{C}(X_1 - X_2) - s_1 + s_2\|_2^2 + \|\mathcal{A}(X_1 - X_2) - y_1 + y_2\|_2^2, \\
&= \|A^T(A\ \mathbf{vec}(X_1 - X_2) - y_1 + y_2) + C^T(C\ \mathbf{vec}(X_1 - X_2) - s_1 + s_2)\|_2^2 \\
&\quad + \|C\ \mathbf{vec}(X_1 - X_2) - s_1 + s_2\|_2^2 + \|A\ \mathbf{vec}(X_1 - X_2) - y_1 + y_2\|_2^2, \\
&= \left\| M^T M \begin{pmatrix} s_1 - s_2 \\ y_1 - y_2 \\ \mathbf{vec}(X_1 - X_2) \end{pmatrix} \right\|_2^2.
\end{aligned}$$

Hence,

$$\begin{aligned}
\|\nabla f(X_1, s_1, y_1) - \nabla f(X_2, s_2, y_2)\| &\le \sigma_{\max}^2(M) \left\| \begin{pmatrix} s_1 - s_2 \\ y_1 - y_2 \\ \mathbf{vec}(X_1 - X_2) \end{pmatrix} \right\|_2, \\
&= \sigma_{\max}^2(M)\ \sqrt{\|X_1 - X_2\|_F^2 + \|s_1 - s_2\|_2^2 + \|y_1 - y_2\|_2^2}, \\
&= \sigma_{\max}^2(M)\ \|(X_1, s_1, y_1) - (X_2, s_2, y_2)\|,
\end{aligned}$$

where $\sigma_{\max}(M)$ is the maximum singular-value of M. Thus, $f : \mathbb{R}^{m\times n} \times \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ is a convex function and ∇f is Lipschitz continuous with respect to $\|.\|$ with Lipschitz constant $L = \sigma_{\max}^2(M)$.

Since $(\bar{X}, \bar{s}, \bar{y})$ is an ϵ-optimal solution to the problem $\min\{P(X, s, y) : X \in \mathbb{R}^{m\times n}, s \in \mathbb{R}^p, y \in \mathcal{Q} \subset \mathbb{R}^q\}$, Theorem A.3 guarantees that

$$\begin{aligned}
\|\nabla_X f(\bar{X}, \bar{s}, \bar{y})\|_F &= \|\mathcal{A}^*(\mathcal{A}(\bar{X}) - \bar{y} - b - \lambda\theta_1) + \mathcal{C}^*(\mathcal{C}(\bar{X}) - \bar{s} - d - \lambda\theta_2)\|_F \\
&\le \sqrt{2\epsilon}\ \sigma_{\max}(M) + I(\alpha^*)\lambda\mu_1, \qquad \text{(A.12)} \\
\|\nabla_s f(\bar{X}, \bar{s}, \bar{y})\|_2 &= \|\mathcal{C}(\bar{X}) - \bar{s} - d - \lambda\theta_2\|_2 \le \sqrt{2\epsilon}\ \sigma_{\max}(M) + J(\beta^*)\lambda\mu_2. \qquad \text{(A.13)}
\end{aligned}$$

□

Lemma A.4 and proof.

LEMMA A.4. *Let $\mathcal{Q} \subset \mathbb{R}^q$ be a nonempty, closed, and convex set. Then for all $\tilde{y} \in \mathbb{R}^q$ and $\lambda > 0$, we have $\Pi_{\mathcal{Q}}(\lambda\tilde{y}) = \lambda\ \Pi_{\mathcal{Q}/\lambda}(\tilde{y})$, or equivalently, $\Pi_{\mathcal{Q}}(\tilde{y}) = \lambda\ \Pi_{\mathcal{Q}/\lambda}(\tilde{y}/\lambda)$, where $\mathcal{Q}/\lambda = \{x : \lambda x \in \mathcal{Q}\}$.*

Proof. Fix $\tilde{y} \in \mathbb{R}^q$ and $\lambda > 0$. Then

$$\Pi_{\mathcal{Q}}(\lambda\tilde{y}) = \operatorname*{argmin}_{x\in\mathcal{Q}} \|x - \lambda\tilde{y}\|_2 = \lambda \operatorname*{argmin}_{y\in\mathcal{Q}/\lambda} \|y - \tilde{y}\|_2 = \lambda\ \Pi_{\mathcal{Q}/\lambda}(\tilde{y}). \qquad \text{(A.14)}$$

□

Lemma A.5 and proof.

LEMMA A.5. *Let (X_*, s_*, y_*) be an optimal solution to (3.2) and suppose that $\|\Pi_{\mathcal{Q}}\left(y_p^{(k)}\right) - y^{(k)}\|_2 \le \xi^{(k)}$ for some $k \ge 1$, where $y_p^{(k)} := y^{(k)} - \frac{1}{L}\nabla_y f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})$. Then we have*

$$-\left\langle \nabla_y f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)}),\ y_* - y^{(k)} \right\rangle \le L\xi^{(k)}\|y_* - y^{(k)}\|_2 + \xi^{(k)}\|\nabla_y f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})\|_2. \qquad \text{(A.15)}$$

Proof. From the definition of $\Pi_{\mathcal{Q}}(.)$, we have

$$\begin{aligned}
&\left\langle \Pi_{\mathcal{Q}}(y_p^{(k)}) - y_p^{(k)},\ y - \Pi_{\mathcal{Q}}(y_p^{(k)}) \right\rangle \geq 0, \quad \forall\ y \in \mathcal{Q}, \\
\Rightarrow &\left\langle \Pi_{\mathcal{Q}}(y_p^{(k)}) - y^{(k)},\ y - y^{(k)} \right\rangle + \left\langle \Pi_{\mathcal{Q}}(y_p^{(k)}) - y^{(k)},\ y^{(k)} - \Pi_{\mathcal{Q}}(y_p^{(k)}) \right\rangle \\
&+ \left\langle y^{(k)} - y_p^{(k)},\ y - y^{(k)} \right\rangle + \left\langle y^{(k)} - y_p^{(k)},\ y^{(k)} - \Pi_{\mathcal{Q}}(y_p^{(k)}) \right\rangle \geq 0, \quad \forall\ y \in \mathcal{Q}. \qquad \text{(A.16)}
\end{aligned}$$

Since $y_* \in \mathcal{Q}$, $y^{(k)} - y_p^{(k)} = \frac{1}{L}\nabla_y f^{(k)}(X^{(k)}, s^{(k)}, y^{(k)})$ and $\|\Pi_{\mathcal{Q}}\left(y_p^{(k)}\right) - y^{(k)}\|_2 \leq \xi^{(k)}$, (A.15) follows from (A.16). □

Appendix B. Auxiliary results for simple optimization problems.

LEMMA B.1. *Let $(\mathcal{E}, \|.\|)$ be a normed vector space, $f : \mathcal{E} \to \mathbb{R}$ be a strictly convex function and $\chi \subset \mathcal{E}$ be a closed, convex set with a non-empty interior. Let $\bar{x} = \operatorname{argmin}_{x\in\chi} f(x)$ and $x^* = \operatorname{argmin}_{x\in\mathcal{E}} f(x)$. If $x^* \notin \chi$, then $\bar{x} \in \mathbf{bd}\,\chi$, where $\mathbf{bd}\,\chi$ denotes the boundary of χ.*

Proof. We will establish the result by contradiction. Assume $\bar{x}$ is in the interior of χ, i.e. $\bar{x} \in \mathbf{int}(\chi)$. Then $\exists\ \epsilon > 0$ such that $B(\bar{x}, \epsilon) = \{x \in \mathcal{E} \ :\ \|x - \bar{x}\| < \epsilon\} \subset \chi$. Since f is strictly convex and $x^* \neq \bar{x}$, $f(x^*) < f(\bar{x})$. Choose $0 < \lambda < \frac{\epsilon}{\|\bar{x}-x^*\|} < 1$ so that $\lambda x^* + (1-\lambda)\bar{x} \in B(\bar{x}, \epsilon) \subset \chi$. Since f is strictly convex,

$$f(\lambda x^* + (1-\lambda)\bar{x}) < \lambda f(x^*) + (1-\lambda) f(\bar{x}) < f(\bar{x}). \qquad \text{(B.1)}$$

However, $\lambda x^* + (1-\lambda)\bar{x} \in B(\bar{x}, \epsilon) \subset \chi$ and $f(\lambda x^* + (1-\lambda)\bar{x}) < f(\bar{x})$ contradicts the fact that $f(\bar{x}) < f(x)$ for all $x \in \chi$. Therefore, $\bar{x} \notin \mathbf{int}(\chi)$. Since $\bar{x} \in \chi$, it follows that $\bar{x} \in \mathbf{bd}\,\chi$. □

Next, we collect together complexity results for optimization problems of the form

$$\begin{aligned}
&\min_{X\in\mathbb{R}^{m\times n}} \{\lambda\|\sigma(X)\|_\alpha + \frac{1}{2}\|X - \tilde{X}\|_F^2 : \|\sigma(X)\|_\alpha \leq \eta\} \\
&\min_{s\in\mathbb{R}^p} \quad \{\lambda\|s\|_\beta + \frac{1}{2}\|s - \tilde{s}\|_2^2 : \|s\|_\beta \leq \eta\}
\end{aligned}$$

that need to be solved in each **Algorithm APG** update step, displayed in Figure 2.1.

LEMMA B.2. *Let $\bar{X} = \operatorname{argmin}_{X\in\mathbb{R}^{m\times n}} \left\{\lambda\|\sigma(X)\|_\alpha + \frac{1}{2}\|X - \tilde{X}\|_F^2 : \|\sigma(X)\|_\alpha \leq \eta\right\}$ of the constrained matrix shrinkage problem. Then*

$$\bar{X} = U\,\mathrm{diag}(\bar{s})V^T,$$

where $U\,\mathrm{diag}(\sigma)V^T$ denotes the SVD of $\tilde{X}$ such that $\sigma \in \mathbb{R}^r_+$ and $r = \mathbf{rank}(\tilde{X})$; and $\bar{s}$ denotes the optimal solution of the constrained vector shrinkage problem

$$\min_{s\in\mathbb{R}^r} \left\{\lambda\|s\|_\alpha + \frac{1}{2}\|s - \sigma\|_2^2 :\ \|s\|_\alpha \leq \eta\right\}.$$

Since the worst case complexity of computing the SVD of $\tilde{X}$ is $\mathcal{O}(\min\{n^2m, m^2n\})$ the complexity of the computing $\bar{X}$ is $\mathcal{O}(\min\{n^2m, m^2n\} + T_v(r,\alpha))$, where $T_v(r,\alpha)$ denotes the complexity of computing the solution of an r-dimensional constrained vector shrinkage problem with norm $\|.\|_\alpha$. The function

$$T_v(p,\alpha) = \begin{cases} \mathcal{O}(p\log(p)) & \alpha = 1, \infty, \\ \mathcal{O}(p), & \alpha = 2, \end{cases} \qquad \text{(B.2)}$$

Proof. The standard results in non-linear convex optimization over matrices implies that $\bar{X}$ is of the form $\bar{X} = U\,\mathrm{diag}(\bar{s})V^T$ (see Corollary 2.5 in [28]).

Now, consider the vector constrained shrinkage problem

$$\min_{s\in\mathbb{R}^p} \left\{\lambda\|s\|_\beta + \frac{1}{2}\|s - \tilde{s}\|_2^2 :\ \|s\|_\beta \leq \eta\right\}.$$

(i) $\beta = 1$: First considered the unconstrained case, i.e. $\eta = \infty$. The unconstrained solution s^* has a closed form $s^* = \text{sign}(\tilde{s}) \odot \max\{|\tilde{s}| - \lambda \mathbf{1}, \mathbf{0}\}$ and can be computed with $\mathcal{O}(p)$ complexity, where $\odot$ denotes componentwise multiplication and $\mathbf{1}$ is a vector of ones.
When $\eta < \infty$, the constrained optimal solution, $\bar{s}$, can be computed with $\mathcal{O}(p\log(p))$ complexity. See Lemma A.4 in [1].

(ii) $\beta = 2$: First considered the unconstrained case, i.e. $\eta = \infty$. Since ℓ_2-norm is self dual, $\lambda\|s\|_2 = \max\{u^T s : \|u\|_2 \leq 1\}$. Thus,

$$\begin{aligned}
\min_{s\in\mathbb{R}^p}\left\{\lambda\|s\|_2 + \frac{1}{2}\|s-\tilde{s}\|_2^2\right\} &= \min_{s\in\mathbb{R}^p}\ \max_{u:\ \|u\|_2\leq\lambda}\left\{u^T s + \frac{1}{2}\|s-\tilde{s}\|_2^2\right\},\\
&= \max_{u:\ \|u\|_2\leq\lambda}\ \min_{s\in\mathbb{R}^p}\left\{u^T s + \frac{1}{2}\|s-\tilde{s}\|_2^2\right\},\\
&= \max_{u:\ \|u\|_2\leq\lambda}\left\{u^T(\tilde{s}-u) + \frac{1}{2}\|u\|_2^2\right\}, \qquad \text{(B.3)}\\
&= \frac{1}{2}\|\tilde{s}\|_2^2 - \min_{u:\ \|u\|_2\leq\lambda}\frac{1}{2}\|u-\tilde{s}\|_2^2,
\end{aligned}$$

where (B.3) follows from the fact that $s^*(u) := \text{argmin}_{s\in\mathbb{R}^p}\{u^T s + \frac{1}{2}\|s-\tilde{s}\|_2^2\} = \tilde{s} - u$.
Define

$$u^* := \operatorname*{argmin}_{u:\ \|u\|_2\leq\lambda}\frac{1}{2}\|u-\tilde{s}\|_2^2 = \tilde{s}\ \min\left\{\frac{\lambda}{\|\tilde{s}\|_2},\ \mathbf{1}\right\}.$$

Then the unconstrained optimal solution $s^* = s^*(u^*) = \tilde{s}\max\left\{1 - \frac{\lambda}{\|\tilde{s}\|_2},\ 0\right\}$ and the complexity of computing $\bar{s}$ is $\mathcal{O}(p)$.
Next, consider the constrained optimization problem, i.e. $\eta < \infty$. The constrained optimum $\bar{s} = s^*$, whenever s^* is feasible, i.e. $\|s^*\|_2 \leq \eta$. Since $f(s) := \lambda\|s\|_2 + \frac{1}{2}\|s-\tilde{s}\|_2^2$ is strongly convex, Lemma B.1 implies that $\|\bar{s}\|_2 = \eta$ whenever $\|s^*\|_2 > \eta$. Thus,

$$\min\left\{\lambda\|s\|_2 + \frac{1}{2}\|s-\tilde{s}\|_2^2 :\ \|s\|_2 \leq \eta\right\} = \lambda\eta + \min\left\{\frac{1}{2}\|s-\tilde{s}\|_2^2 : \|s\|_2^2 = \eta^2\right\}.$$

The unique KKT point for the optimization problem $\min\left\{\frac{1}{2}\|s-\tilde{s}\|_2^2 : \frac{1}{2}\|s\|_2^2 = \frac{\eta^2}{2}\right\}$, is given by $\bar{s} = \eta\frac{\tilde{s}}{\|\tilde{s}\|}$ and KKT multiplier for the constraint $\frac{1}{2}\|s\|_2^2 = \frac{\eta^2}{2}$ is $\vartheta = \frac{\|\tilde{s}\|_2}{\eta} - 1$. It is easy to check that $\vartheta > 0$ whenever $\|s^*\|_2 > \eta$. Thus, $\bar{s}$ is optimal for the convex optimization problem $\min\left\{\frac{1}{2}\|s-\tilde{s}\|_2^2 : \|s\|_2^2 \leq \eta^2\right\}$; consequently, optimal for equality constrained optimization problem $\min\left\{\frac{1}{2}\|s-\tilde{s}\|_2^2 : \|s\|_2 = \eta\right\}$. Hence, the complexity of computing $\bar{s}$ is $\mathcal{O}(p)$

(iii) $\beta = \infty$: First consider the unconstrained problem. Since ℓ_1-norm is the dual norm of the ℓ_∞-norm, we have that

$$\begin{aligned}
\min_{s\in\mathbb{R}^p}\left\{\lambda\|s\|_\infty + \frac{1}{2}\|s-\tilde{s}\|_2^2\right\} &= \min_{s\in\mathbb{R}^p}\ \max_{u:\ \|u\|_1\leq\lambda}\left\{u^T s + \frac{1}{2}\|s-\tilde{s}\|_2^2\right\},\\
&= \max_{u:\ \|u\|_1\leq\lambda}\ \min_{s\in\mathbb{R}^p}\left\{u^T s + \frac{1}{2}\|s-\tilde{s}\|_2^2\right\},\\
&= \max_{u:\ \|u\|_1\leq\lambda}\left\{u^T(\tilde{s}-u) + \frac{1}{2}\|u\|_2^2\right\}, \qquad \text{(B.4)}\\
&= \frac{1}{2}\|\tilde{s}\|_2^2 - \min_{u:\ \|u\|_1\leq\lambda}\frac{1}{2}\|u-\tilde{s}\|_2^2,
\end{aligned}$$

where (B.4) follows from the fact that $s^*(u) := \text{argmin}_{s\in\mathbb{R}^p}\{u^T s + \frac{1}{2}\|s-\tilde{s}\|_2^2\} = \tilde{s} - u$. The result in (i) implies that complexity of computing $u^* = \min_{u:\ \|u\|_1\leq\lambda}\frac{1}{2}\|u-\tilde{s}\|_2^2$ is $\mathcal{O}(p\log(p))$. Thus, the unconstrained optimal solution $s^* = s^*(u^*) = \tilde{s} - u^*$ can be computed in $\mathcal{O}(p\log(p))$ operations.

Next, consider the constrained optimization problem. The constrained optimum, $\bar{s} = s^*$ whenever s^* is feasible, i.e. $\|s^*\|_\infty \leq \eta$. Since $f(s) = \lambda\|s\|_\infty + \frac{1}{2}\|s - \tilde{s}\|_2^2$ is strictly convex, Lemma B.1 implies that $\|\bar{s}\|_\infty = \eta$, whenever $\|s^*\|_\infty > \eta$. Therefore,

$$\min\Big\{\lambda\|s\|_\infty + \frac{1}{2}\|s - \tilde{s}\|_2^2 : \|s\|_\infty \leq \eta\Big\} = \lambda\eta + \min\Big\{\frac{1}{2}\|s - \tilde{s}\|_2^2 : \|s\|_\infty = \eta\Big\}.$$

Then, it is easy to check $\text{sign}(\bar{s}_i) = \text{sign}(\tilde{s}_i)$ for all $i = 1, \ldots, p$. Moreover, $\|s^*\|_\infty > \eta$ implies that $\|\tilde{s}\|_\infty > \eta$. These two facts imply that

$$\min\Big\{\frac{1}{2}\|s - \tilde{s}\|_2^2 : \|s\|_\infty = \eta\Big\} = \min\Big\{\frac{1}{2}\|s - |\tilde{s}|\,\|_2^2 : 0 \leq s_i \leq \eta\Big\}.$$

For $1 \leq i \leq p$, we have $\min\{|\tilde{s}_i|\,, \eta\} = \text{argmin}_{s_i \in \mathbb{R}}\big\{\frac{1}{2}(s_i - |\tilde{s}_i|)^2 : 0 \leq s_i \leq \eta\big\}$. Thus, it follows that $\bar{s} = \text{sign}(\tilde{s}) \odot \min\{|\tilde{s}|, \eta\mathbf{1}\}$. Hence the complexity of computing $\bar{s}$ is $\mathcal{O}(p\log(p))$.

□