# A proximal point algorithm for sequential feature extraction applications

Xuan Vinh Doan [*] Kim-Chuan Toh [†] Stephen Vavasis [‡]

August 2011

**Abstract**

We propose a proximal point algorithm to solve LAROS problem, that is the problem of finding a “large approximately rank-one submatrix”. This LAROS problem is used to sequentially extract features in data. We also develop a new stopping criterion for the proximal point algorithm, which is based on the duality conditions of $\epsilon$-optimal solutions of the LAROS problem, with a theoretical guarantee. We test our algorithm with two image databases and show that we can use the LAROS problem to extract appropriate common features from these images.

## 1 Introduction

Feature extraction is an important application in information retrieval. For example, let us consider a matrix $\boldsymbol{A} \in \mathbb{R}_+^{m\times n}$ that represents a database of pixelated and registered grayscale images which have the same size. Each column of $\boldsymbol{A}$ corresponds to one image and each row corresponds to a particular pixel position in those images. The value $A_{ij}$ is then the intensity of the $i$th pixel in the $j$th image. A common visual feature represented by the pixels in $\mathcal{J} \subset \{1,\ldots,n\}$, which occur in a subset of images in $\mathcal{I} \subset \{1,\ldots,m\}$, can be associated with the approximately rank-one submatrix $\boldsymbol{A}(\mathcal{I},\mathcal{J})$ of the matrix $\boldsymbol{A}$. We assume here the features are non-overlapping. If we want to more than one visual feature,

[*] Department of Combinatorics and Optimization, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada, vanxuan@uwaterloo.ca.

[†] Department of Mathematics, National University of Singapore, Blk S17, 10 Lower Kent Ridge Road, Singapore 119076, mattohkc@nus.edu.sg

[‡] Department of Combinatorics and Optimization, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada, vavasis@math.uwaterloo.ca.

we can iteratively find an approximately rank-one submatrix, subtract it from $\boldsymbol{A}$ (perhaps modifying the result of the subtraction to ensure that $\boldsymbol{A}$ remains nonnegative), and then repeat the procedure. Doan and Vavasis [3] proposed the LAROS problem which tries to find "large approximately rank-one submatrix". The proposed convex parametric formulation for the LAROS problem is written as follows:

$$\begin{aligned} \min \quad & \|\boldsymbol{X}\|_\theta \ := \ \|\boldsymbol{X}\|_* + \theta\|\boldsymbol{X}\|_1 \\ \text{s.t.} \quad & \langle \boldsymbol{A}, \boldsymbol{X}\rangle = 1, \end{aligned} \tag{1}$$

where $\theta > 0$. Here $\|\boldsymbol{X}\|_*$ denotes the nuclear norm of $\boldsymbol{X}$, which is defined to be the sum of the singular values of $\boldsymbol{X}$, and $\|\boldsymbol{X}\|_1$ denotes the sum the absolute values of all the entries of $\boldsymbol{X}$. Theoretical properties of LAROS problem have been developed in [3]. In this paper, we investigate algorithms to solve the problem and apply it to find features in data. We will focus on proximal point algorithmic framework, which have recently been studied for nuclear norm minimization (see Liu et al. [4] and references therein).

Throughout the paper, we use $\|\cdot\|$ to denote either the Frobenius norm of a matrix or the Euclidean norm of a vector. The spectral norm of a matrix $\boldsymbol{X}$ is denoted by $\|\boldsymbol{X}\|_2$.

## Proximal Point Algorithm

The proximal point algorithm is based on the *Moreau-Yoshida regularization* of the (non-differentiable) convex optimization problem

$$\min_{\boldsymbol{x}\in\mathcal{X}} \phi(\boldsymbol{x}), \tag{2}$$

where $\mathcal{X}$ is a finite-dimensional real Hilbert space and $\phi : \mathcal{X} \to (-\infty, \infty]$ is a proper, lower semicontinuous, convex function. For an arbitrary $\lambda > 0$, the regularization is defined as

$$\Phi_\lambda(\boldsymbol{x}) = \min_{\boldsymbol{z}\in\mathcal{X}} \left(\phi(\boldsymbol{z}) + \frac{1}{2\lambda}\|\boldsymbol{x} - \boldsymbol{z}\|^2\right), \quad \forall \boldsymbol{x} \in \mathcal{X}.$$

The above optimization problem has a unique optimal solution $p_\lambda(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{X}$, and $p_\lambda$ is called the *proximal point mapping* associated with $\phi$. One of the most important properties of $\Phi_\lambda$ and $p_\lambda$ is that the set of optimal solutions of (2) is exactly the set of optimal solutions of the following optimization problem:

$$\min_{\boldsymbol{x}\in\mathcal{X}} \Phi_\lambda(\boldsymbol{x}), \tag{3}$$

where $\Phi_\lambda$ is now a *continuously differentiable* convex function defined on $\mathcal{X}$ with a globally Lipschitz continuous gradient $\nabla\Phi_\lambda$ (with modulus $1/\lambda$). The necessary and sufficient optimality condition of (3) can then be expressed as follows:

$$\nabla\Phi_\lambda(\boldsymbol{x}) = \mathbf{0} \Leftrightarrow p_\lambda(\boldsymbol{x}) = \boldsymbol{x}, \tag{4}$$

where $p_\lambda$ is a global Lipschitz continuous function with modulus 1.

The *proximal point* algorithm is an iterative method to solve the problem (2) that uses the optimality condition written in (4). In each iteration, $\boldsymbol{x}^{k+1} \approx p_{\lambda_k}(\boldsymbol{x}^k)$ according to a sequence $\{\lambda_k\}$ of regularization parameters. The convergence of the algorithm has been studied by Rockafellar [6] in a more general setting of *inclusion* problems with *maximal monotone operators*. Note that the problem (2) is equivalent to the inclusion problem $\mathbf{0} \in \partial\phi(\boldsymbol{x})$, where $\partial\phi$ is a maximal monotone operator if $\phi$ is a proper, lower semicontinuous, and convex function. We now ready to study the proximal point mapping for our particular problem. In order to apply the framework, we reformulate Problem (1) with a redundant variable as follows:

$$\begin{aligned} \min \quad & \|\boldsymbol{X}_1\|_* + \theta\|\boldsymbol{X}_2\|_1 \\ \text{s.t.} \quad & \langle \boldsymbol{A}, \boldsymbol{X}_1\rangle = 1, \\ & \boldsymbol{X}_1 = \boldsymbol{X}_2. \end{aligned} \tag{5}$$

In addition, to introduce more flexibility into our model, we study Problem (5) under the following more general setting:

$$\begin{aligned} \min \quad & \|\boldsymbol{X}_1\|_* + \theta\|\boldsymbol{X}_2\|_1 \\ \text{s.t.} \quad & \mathcal{A}(\boldsymbol{X}) - b \in \mathcal{Q}, \quad \boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2) \in \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \end{aligned} \tag{6}$$

where $b \in \mathcal{H}$, $\mathcal{A} : \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n} \to \mathcal{H}$ is a given linear map, and $\mathcal{Q}$ is a pointed close convex cone in $\mathcal{H}$. Here $\mathcal{H}$ is a finite-dimensional Hilbert space. For the problem (5), we have $\mathcal{H} = \mathbb{R}\times\mathbb{R}^{m\times n}$, $\mathcal{Q} = \{0\}\times\{\mathbf{0}\}$, $b = (1, \mathbf{0})$, and $\mathcal{A}(\boldsymbol{X}) = (\langle \boldsymbol{A}, \boldsymbol{X}_1\rangle, \boldsymbol{X}_1 - \boldsymbol{X}_2)$. Note that the adjoint $\mathcal{A}^* : \mathcal{H} \to \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n}$ is given by $\mathcal{A}^* z = (z_1\boldsymbol{A} + \boldsymbol{Z}_2, -\boldsymbol{Z}_2)$ for any $z = (z_1, \boldsymbol{Z}_2) \in \mathcal{H}$.

## 2 Primal Proximal Point Algorithm

We define the function $\phi$ as follows:

$$\phi(\boldsymbol{X}) = \begin{cases} \|\boldsymbol{X}_1\|_* + \theta\|\boldsymbol{X}_2\|_1, & \boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2) \in \mathcal{F}, \\ +\infty, & \text{otherwise}, \end{cases} \tag{7}$$

where $\mathcal{F}$ is the feasible set of the problem (6). The problem (6) is then equivalent to the optimization problem

$$\min_{\boldsymbol{X}\in\mathcal{X}} \phi(\boldsymbol{X}),$$

where $\mathcal{X} = \mathbb{R}^{m\times n} \times \mathbb{R}^{m\times n}$.

We now introduce dual decision variables $z \in \mathbb{R}^p$ and define the Lagrangian function $L(\boldsymbol{X}, z)$,

$$L(\boldsymbol{X}, z) = \begin{cases} \|\boldsymbol{X}_1\|_* + \theta\|\boldsymbol{X}_2\|_1 + \langle z, b - \mathcal{A}(\boldsymbol{X})\rangle & \text{if } z \in \mathcal{Q}^* \\ -\infty & \text{otherwise} \end{cases} \tag{8}$$

where $\mathcal{Q}^*$ is the dual cone of $\mathcal{Q}$ defined by $\mathcal{Q}^* = \{y \in \mathcal{H} : \langle y, z\rangle \geq 0, \forall\, z \in \mathcal{Q}\}$. For our problem, $\mathcal{Q}^*$ is simply the whole space, $\mathcal{Q}^* = \mathcal{H} = \mathbb{R} \times \mathbb{R}^{m\times n}$. Clearly, $\phi(\boldsymbol{X}) = \sup\limits_{z\in\mathbb{R}^p} L(\boldsymbol{X}, z)$. We now calculate the Moreau-Yoshida regularization of $\phi$:

$$\Phi_\lambda(\boldsymbol{X}) = \min_{\boldsymbol{V}\in\mathcal{X}} \left(\phi(\boldsymbol{V}) + \frac{1}{2\lambda}\|\boldsymbol{X} - \boldsymbol{V}\|^2\right). \tag{9}$$

Applying the strong duality (or minimax theory) result in Rockafellar [5], we have:

$$\begin{aligned}
\Phi_\lambda(\boldsymbol{X}) &= \min_{\boldsymbol{V}\in\mathcal{X}} \sup_{z\in\mathcal{H}} \left(L(\boldsymbol{V}, z) + \frac{1}{2\lambda}\|\boldsymbol{X} - \boldsymbol{V}\|^2\right) \\
&= \sup_{z\in\mathcal{H}} \min_{\boldsymbol{V}\in\mathcal{X}} \left(L(\boldsymbol{V}, z) + \frac{1}{2\lambda}\|\boldsymbol{X} - \boldsymbol{V}\|^2\right) \\
&= \sup_{z\in\mathcal{Q}^*} \min_{\boldsymbol{V}\in\mathcal{X}} \left(\|\boldsymbol{V}_1\|_* + \theta\|\boldsymbol{V}_2\|_1 + \langle z, b - \mathcal{A}(\boldsymbol{V})\rangle + \frac{1}{2\lambda}\|\boldsymbol{X} - \boldsymbol{V}\|^2\right) \\
&= \sup_{z\in\mathcal{Q}^*} \langle z, b\rangle + \frac{1}{2\lambda}\|\boldsymbol{X}\|^2 - \frac{1}{2\lambda}\|\boldsymbol{X} + \lambda\mathcal{A}^* z\|^2 + \min_{\boldsymbol{V}\in\mathcal{X}} \left(\|\boldsymbol{V}_1\|_* + \theta\|\boldsymbol{V}_2\|_1 + \frac{1}{2\lambda}\|\boldsymbol{V} - (\boldsymbol{X} + \lambda\mathcal{A}^* z)\|^2\right)
\end{aligned}$$

Now, consider the first inner minimization problem, we have:

$$\begin{aligned}
&\min_{\boldsymbol{V}\in\mathcal{X}} \left(\|\boldsymbol{V}_1\|_* + \theta\|\boldsymbol{V}_2\|_1 + \frac{1}{2\lambda}\|\boldsymbol{V} - (\boldsymbol{X} + \lambda\mathcal{A}^* z)\|^2\right) \qquad (10) \\
&= \min_{\boldsymbol{V}_1} \left(\|\boldsymbol{V}_1\|_* + \frac{1}{2\lambda}\|\boldsymbol{V}_1 - (\boldsymbol{X}_1 + \lambda\mathcal{B}_1 z)\|^2\right) + \theta \min_{\boldsymbol{V}_2} \left(\|\boldsymbol{V}_2\|_1 + \frac{1}{2\lambda\theta}\|\boldsymbol{V}_2 - (\boldsymbol{X}_2 + \lambda\mathcal{B}_2 z)\|^2\right)
\end{aligned}$$

where we have written $\mathcal{A}^* z = (\mathcal{B}_1 z, \mathcal{B}_2 z) \in \mathcal{X}$. The first optimization problem on the right-hand side is the Moreau-Yoshida regularization of the nuclear norm function at $\boldsymbol{X}_1 + \lambda\mathcal{B}_1 z$, and the problem has an analytical solution given by

$$p_\lambda^{(1)}(\boldsymbol{X}_1 + \lambda\mathcal{B}_1 z) = \boldsymbol{U}\text{Diag}(\max\{\sigma_i - \lambda, 0\})\boldsymbol{V}^T, \tag{11}$$

which is computable from the singular value decomposition, $\boldsymbol{X}_1 + \lambda\mathcal{B}_1 z = \boldsymbol{U\Sigma V}^T$. In addition, the minimal objective value is given by

$$\frac{1}{2\lambda}\|\boldsymbol{X}_1 + \lambda\mathcal{B}_1 z\|^2 - \frac{1}{2\lambda}\|p_\lambda^{(1)}(\boldsymbol{X}_1 + \lambda\mathcal{B}_1 z)\|^2.$$

Next, we consider the second inner minimization problem on the right-hand side of (10). This optimization problem is the Moreau-Yoshida regularization of the $l_1$-norm function (with parameter $\lambda\theta$) at

$\boldsymbol{X}_2 + \lambda\mathcal{B}_2 z$, and it has the following analytical solution:

$$p^{(2)}_{\lambda\theta}(\boldsymbol{X}_2 + \lambda\mathcal{B}_2 z) = \operatorname{sgn}(\boldsymbol{X}_2 + \lambda\mathcal{B}_2 z) \circ \max\{|\boldsymbol{X}_2 + \lambda\mathcal{B}_2 z| - \theta\lambda, 0\}, \tag{12}$$

where $\circ$ is the Hadamard product (or entrywise product) and sgn is the (entrywise) sign function. The corresponding minimal objective value is given by

$$\frac{1}{2\lambda\theta}\|\boldsymbol{X}_2 + \lambda\mathcal{B}_2 z\|^2 - \frac{1}{2\lambda\theta}\|p^{(2)}_{\lambda\theta}(\boldsymbol{X}_2 + \lambda\mathcal{B}_2 z)\|^2.$$

Combining these two results, we can compute $\Phi_\lambda(\boldsymbol{X})$ as follows:

$$\Phi_\lambda(\boldsymbol{X}) = \frac{1}{2\lambda}\|\boldsymbol{X}\|^2 + \sup_{z\in\mathcal{Q}^*}\left(\langle z, b\rangle - \frac{1}{2\lambda}\|p^{(1)}_{\lambda}(\boldsymbol{X}_1 + \lambda\mathcal{B}_1 z)\|^2 - \frac{1}{2\lambda}\|p^{(2)}_{\lambda\theta}(\boldsymbol{X}_2 + \lambda\mathcal{B}_2 z)\|^2\right), \tag{13}$$

where $p^{(1)}_{\lambda}$ and $p^{(2)}_{\lambda\theta}$ are defined in (11) and (12) respectively. Now define

$$\Theta_\lambda(\boldsymbol{X}, z) = \langle z, b\rangle - \frac{1}{2\lambda}\|p^{(1)}_{\lambda}(\boldsymbol{X}_1 + \lambda\mathcal{B}_1 z)\|^2 - \frac{1}{2\lambda}\|p^{(2)}_{\lambda\theta}(\boldsymbol{X}_2 + \lambda\mathcal{B}_2 z)\|^2 \tag{14}$$

and consider

$$z_\lambda(\boldsymbol{X}) \in \arg\sup_{z\in\mathcal{Q}^*}\Theta_\lambda(\boldsymbol{X}, z).$$

Applying the saddle point theorem in Rockafellar [5], we obtain the proximal point mapping associated with $\phi$ as follows:

$$p_\lambda(\boldsymbol{X}) = \left(p^{(1)}_{\lambda}(\boldsymbol{X}_1 + \lambda\mathcal{B}_1 z_\lambda(\boldsymbol{X})),\, p^{(2)}_{\lambda\theta}(\boldsymbol{X}_2 + \lambda\mathcal{B}_2 z_\lambda(\boldsymbol{X}))\right). \tag{15}$$

The primal proximal point algorithm (primal PPA) has the following template.

**The Primal PPA.** Given $\boldsymbol{X}^0 \in \mathcal{X}$, $\lambda_0 > 0$ and $\varepsilon > 0$, perform the following loop:

**Step 1.** Find an (approximate) optimal solution

$$z^k \in \arg\sup_{z\in\mathcal{Q}^*}\Theta_{\lambda_k}(\boldsymbol{X}^k, z), \tag{16}$$

where $\Theta_{\lambda_k}$ is defined in (14).

**Step 2.** Update

$$\boldsymbol{X}_1^{k+1} = p^{(1)}_{\lambda_k}(\boldsymbol{X}_1^k + \lambda\mathcal{B}_1 z^k), \quad \boldsymbol{X}_2^{k+1} = p^{(2)}_{\lambda_k\theta}(\boldsymbol{X}_2^k + \lambda\mathcal{B}_2 z^k) \tag{17}$$

according to the proximal point mapping in (15).

**Step 3.** If $\|\boldsymbol{X}^{k+1} - \boldsymbol{X}^k\|/\lambda_k < \varepsilon$, stop; else, update $\lambda_k$, end

# 3 Dual Proximal Point Algorithm

The dual problem associated with (6) is given as follows:

$$\max_{y \in \mathcal{H}} g(y) \tag{18}$$

where $g$ is the concave function defined by

$$g(y) = \begin{cases} \inf\{\|\boldsymbol{X}_1\|_* + \theta\|\boldsymbol{X}_2\|_1 + \langle y, b - \mathcal{A}(\boldsymbol{X})\rangle \;:\; \boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2) \in \mathcal{X}\} & \text{if } y \in \mathcal{Q}^* \\ -\infty & \text{otherwise} \end{cases} \tag{19}$$

The Moreau-Yoshida regularization of $g$ is given by

$$\begin{aligned} G_\lambda(y) \;&:=\; \max_{z \in \mathcal{H}} \{g(z) - \frac{1}{2\lambda}\|z - y\|^2\} \\ &=\; \max_{z \in \mathcal{Q}^*} \inf_{\boldsymbol{X} \in \mathcal{X}} \{\|\boldsymbol{X}_1\|_* + \theta\|\boldsymbol{X}_2\|_1 + \langle z, b - \mathcal{A}(\boldsymbol{X})\rangle - \frac{1}{2\lambda}\|z - y\|^2\} \\ &=\; \inf_{\boldsymbol{X} \in \mathcal{X}} \max_{z \in \mathcal{Q}^*} \{\|\boldsymbol{X}_1\|_* + \theta\|\boldsymbol{X}_2\|_1 + \langle z, b - \mathcal{A}(\boldsymbol{X})\rangle - \frac{1}{2\lambda}\|z - y\|^2\} \\ &=\; -\frac{1}{2\lambda}\|y\|^2 + \inf_{\boldsymbol{X} \in \mathcal{X}} \Big\{\|\boldsymbol{X}_1\|_* + \theta\|\boldsymbol{X}_2\|_1 + \Psi_\lambda(\boldsymbol{X}; y)\Big\} \end{aligned} \tag{20}$$

where

$$\Psi_\lambda(\boldsymbol{X}; y) = \frac{1}{2\lambda}\|\Pi_{\mathcal{Q}^*}(y + \lambda(b - \mathcal{A}(\boldsymbol{X})))\|^2. \tag{21}$$

Note that $\nabla_{\boldsymbol{X}} \Psi_\lambda(\boldsymbol{X}; y) = -\mathcal{A}^* \Pi_{\mathcal{Q}^*}(y + \lambda(b - \mathcal{A}(\boldsymbol{X})))$. The dual algorithm can then be written as follows.

**The Dual PPA.** Given a tolerance $\varepsilon > 0$. Input $y^0 \in \mathcal{Q}^*$ and $\lambda_0 > 0$. Set $k := 0$. Iterate:

**Step 1.** Find an approximate minimizer

$$\boldsymbol{X}^k \approx \arg \inf_{X \in \mathcal{X}} \Big\{\|\boldsymbol{X}_1\|_* + \theta\|\boldsymbol{X}_2\|_1 + \Psi_{\lambda_k}(\boldsymbol{X}; y^k)\Big\}, \tag{22}$$

where $\Psi_{\lambda_k}(\boldsymbol{X}; y^k)$ is defined as in (21).

**Step 2.** Compute

$$y^{k+1} \;=\; \Pi_{\mathcal{Q}^*}\big[y^k + \lambda_k(b - \mathcal{A}(\boldsymbol{X}^k))\big]. \tag{23}$$

**Step 3.** If $\|(y^k - y^{k+1})/\lambda_k\| \le \varepsilon$; stop; else; update $\lambda_k$ ; end.

# 4 Implementation Issues

## 4.1 Primal proximal point algorithm

For the primal PPA, the most important issue we have to address first is how to solve the inner problem $\sup_{z \in \mathcal{Q}^*} \Theta_\lambda(\boldsymbol{X}, z)$. We have that $\Theta_\lambda$ is a concave function in $z$ due to the linearity in $z$ of the Lagrangian function $L$. From the general gradient formulation $\nabla \Phi_\lambda(\boldsymbol{x}) = \dfrac{1}{\lambda}(\boldsymbol{x} - p_\lambda(\boldsymbol{x}))$, we have that $\nabla \|p_\lambda^{(i)}(\boldsymbol{X}_i)\|^2 = p_\lambda^{(i)}(\boldsymbol{X}_i)$, $i = 1, 2$. Thus $\Theta_\lambda$ is continuously differentiable with

$$\nabla_z \Theta_\lambda(\boldsymbol{X}, z) \quad = \quad b - \mathcal{B}_1^* p_\lambda^{(1)}(\boldsymbol{X}_1 + \lambda \mathcal{B}_1 z) - \mathcal{B}_2^* p_{\lambda\theta}^{(2)}(\boldsymbol{X}_2 + \lambda \mathcal{B}_2 z). \tag{24}$$

Note that for the problem (5), we have $\mathcal{B}_1 z = \boldsymbol{A} z_1 + \boldsymbol{Z}_2$, $\mathcal{B}_2 z = -\boldsymbol{Z}_2$ for $z = (z_1, \boldsymbol{Z}_2) \in \mathbb{R} \times \mathbb{R}^{m \times n}$. Correspondingly, we have $\mathcal{B}_1^*(\boldsymbol{X}_1) = (\langle \boldsymbol{A}, \boldsymbol{X}_1 \rangle, \boldsymbol{X}_1)$ and $\mathcal{B}_2^*(\boldsymbol{X}_2) = (0, -\boldsymbol{X}_2)$ for any $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^{m \times n}$ and

$$\nabla_z \Theta_\lambda(\boldsymbol{X}, z) = \Big(1 - \langle \boldsymbol{A}, p_\lambda^{(1)}(\boldsymbol{X}_1 + \lambda(\boldsymbol{A} z_1 + \boldsymbol{Z}_2)) \rangle, \, p_{\lambda\theta}^{(2)}(\boldsymbol{X}_2 - \lambda \boldsymbol{Z}_2) - p_\lambda^{(1)}(\boldsymbol{X}_1 + \lambda(\boldsymbol{A} z_1 + \boldsymbol{Z}_2))\Big). \tag{25}$$

In addition, using the global Lipschitz continuity (with modulus 1) of two proximal point mappings, $p_\lambda^{(1)}$ and $p_{\lambda\theta}^{(2)}$, we can show that the gradient $\nabla_z \Theta_\lambda$ is globally Lipschitz continuous with modulus $\lambda(\|\boldsymbol{A}\|_2^2 + 2)$.

With all these properties of $\Theta_\gamma$, we can solve the inner problem using *first-order gradient-based* methods such as steepest descent method.

The second issue is that these inner problems are typically only solved approximately which results in inexact proximal point mappings. For inexact proximal point method, Rockafellar [6] provides two convergence criteria for global and local convergence. Based on the aforementioned convergence criteria, Liu et al. [4] have proposed some checkable stopping criteria for the inner problems to maintain global (and local) convergence of the proposed (inexact) proximal point method (for nuclear norm minimization problems). We can extend these stopping criteria for our problem.

The third issues is to calculate a partial singular value decomposition in order to compute the proximal point mapping of the nuclear norm function (the computation of the proximal point mapping of the $l_1$-norm function is straightforward). As in Liu et al. [4], we use a Lanczos bidiagonalization algorithm with partial reorthogonalization to compute a partial singular value decomposition. We also need heuristics to set the number of singular values required to be computed with this algorithm.

### 4.2 Dual proximal point algorithm

For the dual PPA, we need to look for a method to solve the inner problem $\inf\limits_{X\in\mathcal{X}}\Big\{\|\boldsymbol{X}_1\|_* + \theta\|\boldsymbol{X}_2\|_1 + \Psi_{\lambda_k}(\boldsymbol{X};y^k)\Big\}$. Similar to the approach proposed in Liu et al. [4], we will apply the accelerated proximal gradient algorithm [1] for this problem. According to Toh and Yun [8], we solve the problem $\min\limits_{\boldsymbol{X}} P(\boldsymbol{X}) + f(\boldsymbol{X})$, where $P(\boldsymbol{X}) = \|\boldsymbol{X}_1\|_* + \theta\|\boldsymbol{X}_2\|_1$ and $f(\boldsymbol{X}) = \Psi_\lambda(\boldsymbol{X};y)$. We have that the gradient $\nabla_{\boldsymbol{X}}\Psi_\lambda(\boldsymbol{X};y) = -\mathcal{A}^*\Pi_{\mathcal{Q}^*}(y+\lambda(b-\mathcal{A}(\boldsymbol{X})))$ is globally Lipschitz continuous with modulus $L = \lambda(\|\boldsymbol{A}\|_2^2 + 2)$. The proximal gradient algorithm in each iteration needs to solve the following quadratic approximation of the sum $P(\boldsymbol{X}) + f(\boldsymbol{X})$ at the current solution $\boldsymbol{Y}$:

$$\begin{aligned} Q_t(\boldsymbol{X};\boldsymbol{Y}) &= P(\boldsymbol{X}) + f(\boldsymbol{Y}) + \langle\nabla f(\boldsymbol{Y},\boldsymbol{X}-\boldsymbol{Y}\rangle + \frac{t}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2 \\ &= P(\boldsymbol{X}) + \frac{t}{2}\|\boldsymbol{X}-G_t(\boldsymbol{Y})\|_F^2 + f(\boldsymbol{Y}) - \frac{1}{2t}\|\nabla f(\boldsymbol{Y})\|_F^2, \end{aligned}$$

where $G_t(\boldsymbol{Y}) = \boldsymbol{Y} - \frac{1}{t}\nabla f(\boldsymbol{Y})$. This function is a strongly convex function in $\boldsymbol{X}$ and hence it has a unique minimizer $S_t(\boldsymbol{Y})$. We have that

$$\begin{aligned} P(\boldsymbol{X}) + \frac{t}{2}\|\boldsymbol{X}-G_t(\boldsymbol{Y})\|_F^2 &= \|\boldsymbol{X}_1\|_* + \theta\|\boldsymbol{X}_2\|_1 + \frac{t}{2}\left(\|\boldsymbol{X}_1 - G_t^1(\boldsymbol{Y})\|_F^2 + \|\boldsymbol{X}_2 - G_t^2(\boldsymbol{Y})\|_F^2\right) \\ &= \left(\|\boldsymbol{X}_1\|_* + \frac{t}{2}\|\boldsymbol{X}_1 - G_t^1(\boldsymbol{Y})\|_F^2\right) + \left(\theta\|\boldsymbol{X}_2\|_1 + \frac{t}{2}\|\boldsymbol{X}_2 - G_t^2(\boldsymbol{Y})\|_F^2\right), \end{aligned}$$

where $G_t(\boldsymbol{Y}) = (G_t^1(\boldsymbol{Y}), G_t^2(\boldsymbol{Y}))$. Thus the minimizer $S_t(\boldsymbol{Y}) = (S_t^1(\boldsymbol{Y}), S_t^2(\boldsymbol{Y}))$, where $S_t^1(\boldsymbol{Y})$ is the minimizer of the problem $\min\limits_{\boldsymbol{X}_1}\left(\|\boldsymbol{X}_1\|_* + \frac{t}{2}\|\boldsymbol{X}_1 - G_t^1(\boldsymbol{Y})\|_F^2\right)$ and $S_t^2(\boldsymbol{Y})$ is the minimizer of the optimization problem $\min\limits_{\boldsymbol{X}_2}\theta\|\boldsymbol{X}_2\|_1 + \frac{t}{2}\|\boldsymbol{X}_2 - G_t^2(\boldsymbol{Y})\|_F^2$. Similar to the previous section, the analytical solutions for these two optimization problems can be calculated and they are:

$$S_t^1(\boldsymbol{Y}) = p_{t^{-1}}^{(1)}(G_t^1(\boldsymbol{Y})), \quad S_t^2(\boldsymbol{Y}) = p_{t^{-1}\theta}^{(2)}(G_t^2(\boldsymbol{Y})). \tag{26}$$

Finally, the proximal gradient algorithm for our problem can be described as follows. Given $\tau_0 = \tau_{-1} = 1$ and $\boldsymbol{X}^0 = \boldsymbol{X}^{-1}$, each iteration includes the following steps

1. Calculate $\boldsymbol{Y}^k = \boldsymbol{X}^k + \frac{\tau_{k-1}-1}{\tau_k}\left(\boldsymbol{X}^k - \boldsymbol{X}^{k-1}\right)$
2. Update $\boldsymbol{X}^{k+1} = S_{t^k}(\boldsymbol{Y}^k)$ according the formulas in (26).
3. Update $\tau_{k+1} = \frac{1}{2}\left(\sqrt{1+4\tau_k^2}+1\right)$

The update of $\tau_k$ in the third step is to make sure that $\tau_{k+1}^2 - \tau_{k+1} \le \tau_k^2$ and $\tau_{k+1} \ge 1$, a convergence condition of the proximal gradient algorithm. We also need to have the update rule for the remaining parameter $t_k$, which affects the quadratic approximation of the function $f$ at $\boldsymbol{Y}$. Since the gradient $\nabla f(\boldsymbol{Y})$ is Lipschitz continuous with modulus $L$, for all $t \geq L$, we have:

$$P(S_t(\boldsymbol{Y})) + f(S_t(\boldsymbol{Y})) \leq Q_t(S_t(\boldsymbol{Y}); \boldsymbol{Y}).$$

In order to have a better approximation, we would like to have smaller $t$ and in the accelerated proximal gradient algorithm, we will use line search to find $t_k < L$ such that the above condition is still satisfied, starting with $t_1 = L$. More details can be found in Toh and Yun [8].

## 5 Sparse Structure of Rank-One Optimal Solutions

The proposed proximal point algorithm is a first-order iterative method, which normally does not have fast convergence. Applying duality results obtained by Doan and Vavasis [3] for Problem (1), we would like to study better stopping criteria for the proposed proximal point algorithms. We focus on the case when Problem (1) has a rank-one optimal solution $\boldsymbol{X} = \sigma \boldsymbol{u}\boldsymbol{v}^T$ since rank-one optimal solutions are what we are looking for in general. The purpose of the termination test is to obtain the correct supports of $\boldsymbol{u}$ and $\boldsymbol{v}$, that is, the positions of their nonzero entries with a guarantee certificate when we only have approximate values for $\boldsymbol{u}$ and $\boldsymbol{v}$ from the proposed first-order algorithm. Although the technique in this section is developed for Problem (1), similar ideas can be applied to other proposed formulations with the nuclear norm such as the matrix completion problem. In particular, a test like this for the matrix completion problem can be used to rigorously establish the correct rank of the optimal solution from approximate solutions obtained from a first-order method.

Now let us consider the rank-one optimal solution $\boldsymbol{X}$ of the following form

$$\boldsymbol{X} = \begin{pmatrix} \sigma_1 \boldsymbol{u}_1 \boldsymbol{v}_1^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{u}_1 \geq \boldsymbol{0}$ is a unit vector in $\mathbb{R}^M$, $M \leq m$, and $\boldsymbol{v}_1 \geq \boldsymbol{0}$ is a unit vector in $\mathbb{R}^N$, $N \leq n$. If $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$ are determined, $\sigma_1$ can be easily calculated to satisfy the optimality condition $\|\boldsymbol{X}\|_\theta = 1$ (we assume here $\boldsymbol{A} \neq \boldsymbol{0}$). Note that in general, the rank-one optimal solution $\boldsymbol{X}$ could have a different block structure. However, without loss of generality, we can assume that $\boldsymbol{u}_1 \boldsymbol{v}_1^T$ forms an upper left principal submatrix of $\boldsymbol{X}$ for ease of exposition. Under this assumption, we can set $\boldsymbol{u} = [\boldsymbol{u}_1; \boldsymbol{0}] \in \mathbb{R}^m$ and $\boldsymbol{v} = [\boldsymbol{v}_1; \boldsymbol{0}] \in \mathbb{R}^n$ with $\sigma = \sigma_1$. Similar to Theorem 5 in Doan and Vavasis [3], we can then write the optimality conditions as follows:

There exists $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{m\times n}$ such that

$$\boldsymbol{A} \;=\; \|\boldsymbol{A}\|_\theta^*(\boldsymbol{u}\boldsymbol{v}^T + \boldsymbol{W}) + \theta\|\boldsymbol{A}\|_\theta^*\boldsymbol{V} \tag{27}$$

$$\|\boldsymbol{W}\|_2 \le 1, \quad \boldsymbol{W}^T\boldsymbol{u} = \boldsymbol{0}, \quad \boldsymbol{W}\boldsymbol{v} = \boldsymbol{0}, \quad \boldsymbol{V}_{11} = \boldsymbol{E}_{M\times N}, \quad \|\boldsymbol{V}\|_\infty \le 1,$$

where $\boldsymbol{E}_{M\times N}$ is the $M \times N$ matrix of all ones.

Letting $\lambda = 1/\|\boldsymbol{A}\|_\theta^*$ and splitting all matrices into four subblocks according to the sparse structure of $\boldsymbol{X}$, we obtain the following detailed optimality conditions:

$$(1, \boldsymbol{u}_1, \boldsymbol{v}_1) \text{ is a singular triple of } \lambda\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11}, \text{ and } \boldsymbol{W}_{11} = (\lambda\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11}) - \boldsymbol{u}_1\boldsymbol{v}_1^T \tag{28}$$

$$\boldsymbol{W}_{12} = \lambda\boldsymbol{A}_{12} - \theta\boldsymbol{V}_{12},\ \boldsymbol{W}_{12}^T\boldsymbol{u}_1 = \boldsymbol{0}, \text{ and } \|\boldsymbol{V}_{12}\|_\infty \le 1 \tag{29}$$

$$\boldsymbol{W}_{21} = \lambda\boldsymbol{A}_{21} - \theta\boldsymbol{V}_{21},\ \boldsymbol{W}_{21}\boldsymbol{v}_1 = \boldsymbol{0}, \text{ and } \|\boldsymbol{V}_{21}\|_\infty \le 1 \tag{30}$$

$$\boldsymbol{W}_{22} = \lambda\boldsymbol{A}_{22} - \theta\boldsymbol{V}_{22}, \text{ and } \|\boldsymbol{V}_{22}\|_\infty \le 1 \tag{31}$$

$$\|\boldsymbol{W}\|_2 \le 1. \tag{32}$$

The following lemma shows how to find $\boldsymbol{u}_1$, $\boldsymbol{v}_1$ and $\lambda$ (or $\|\boldsymbol{A}\|_\theta^*$) from the first optimality condition.

**Lemma 1.** *If $(\lambda, \boldsymbol{u}_1, \boldsymbol{v}_1)$ satisfies (28), then $\boldsymbol{x} = (\lambda, \boldsymbol{u}_1, \boldsymbol{v}_1)$ is a solution of the following system of nonlinear equations*

$$P(\boldsymbol{x}) = \begin{pmatrix} (\lambda\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11})\boldsymbol{v}_1 - \boldsymbol{u}_1 \\ (\lambda\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11})^T\boldsymbol{u}_1 - \boldsymbol{v}_1 \\ \boldsymbol{u}_1^T\boldsymbol{u}_1 - 1 \end{pmatrix} = \boldsymbol{0}. \tag{33}$$

**Proof.** It is easily to see that $\boldsymbol{v}_1^T\boldsymbol{v}_1 = \boldsymbol{v}_1^T(\lambda\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11})^T\boldsymbol{u}_1 = \boldsymbol{u}_1^T\boldsymbol{u}_1 = 1$ and the first two equations indicate that $(1, \boldsymbol{u}_1, \boldsymbol{v}_1)$ is a singular triple of $\lambda\boldsymbol{A}_{11} - \theta\boldsymbol{V}_{11}$. □

The system of equations in (33) has $M + N + 1$ variables and $M + N + 1$ equations, which can be solved using Newton method. One of the convergence results of the Newton's method is the Kantorovich theorem, which is given as follows (see Tapia [7]).

**Theorem 1** (Kantorovich)**.** *Assume that $P$ is defined and is Fréchet differentiable at each point in a given open convex set $D_0$ and for some $\boldsymbol{x}_0 \in D_0$ that $[P'(\boldsymbol{x}_0)]^{-1}$ exists and that*

*(i)* $\|[P'(\boldsymbol{x}_0)]^{-1}\| \le B$,

*(ii)* $\|[P'(\boldsymbol{x}_0)]^{-1}P(\boldsymbol{x}_0)\| \le \eta$, *and*

*(iii)* $\|P'(\boldsymbol{x}) - P'(\boldsymbol{y})\| \leq K\|\boldsymbol{x} - \boldsymbol{y}\|$*, for all* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *in* $D_0$*,*

*with* $h = BK\eta \leq \dfrac{1}{2}$*.*

*Let* $\Omega_* = \{\boldsymbol{x} \,|\, \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq t^*\}$*, where* $t^* = \left(\dfrac{1 - \sqrt{1 - 2h}}{h}\right)\eta$*. Now if* $\Omega_* \subset D_0$*, then the Newton iterates,* $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - [P'(\boldsymbol{x}_k)]^{-1}P(\boldsymbol{x}_k)$*, are well defined, remain in* $\Omega_*$*, and converge to* $\boldsymbol{x}^* \in \Omega_*$ *such that* $P(\boldsymbol{x}^*) = \boldsymbol{0}$*. In addition,*

$$\|\boldsymbol{x}^* - \boldsymbol{x}_k\| \leq \frac{\eta}{h}\left[\frac{\left(1 - \sqrt{1 - 2h}\right)^{2^k}}{2^k}\right], \quad k = 0, 1, 2, \ldots.$$

According to Theorem 1, if we can find $\boldsymbol{x}_0$ with the corresponding parameter $h \leq 1/2$, then for an arbitrary $\epsilon > 0$, an $\epsilon$-solution $\boldsymbol{x}$ such that $\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \epsilon$, can be achieved after a finite number of Newton iterations. Now assuming that we have obtained an $\epsilon$-solution $(\lambda, \boldsymbol{u}_1, \boldsymbol{v}_1)$ of the system of equations in (33), we would like to characterize the sufficient conditions which guarantee that the corresponding solution $(\lambda^*, \boldsymbol{u}_1^*, \boldsymbol{v}_1^*)$ defines the optimal solution $\boldsymbol{X}$ of (1) as described above. The following proposition shows these sufficient conditions.

**Proposition 1.** *Consider an* $\epsilon$*-solution* $(\lambda, \boldsymbol{u}_1, \boldsymbol{v}_1)$ *of the system of equations in (33),* $0 < \epsilon < 1/2$*. The corresponding solution* $(\lambda^*, \boldsymbol{u}_1^*, \boldsymbol{v}_1^*)$ *defines the rank-one optimal solution* $\boldsymbol{X}^*$*,*

$$\boldsymbol{X}^* = \begin{pmatrix} \sigma_1^* \boldsymbol{u}_1^* (\boldsymbol{v}_1^*)^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix},$$

*of (1) if there exist* $\boldsymbol{W}$ *and* $\boldsymbol{V}$ *that satisfy the following conditions:*

*(i)* $\boldsymbol{W}_{11} = (\lambda \boldsymbol{A}_{11} - \theta \boldsymbol{V}_{11}) - \boldsymbol{u}_1 \boldsymbol{v}_1^T$ *and* $\boldsymbol{V}_{11} = \boldsymbol{E}_{M \times N}$*,*

*(ii)* $\boldsymbol{W}_{12} = \lambda \boldsymbol{A}_{12} - \theta \boldsymbol{V}_{12}$*,* $\boldsymbol{W}_{12}^T \boldsymbol{u}_1 = \boldsymbol{0}$*, and* $\|\boldsymbol{V}_{12}\|_\infty \leq 1 - \theta^{-1}(\|\boldsymbol{A}_{12}\|_\infty + 5)\epsilon$*,*

*(iii)* $\boldsymbol{W}_{21} = \lambda \boldsymbol{A}_{21} - \theta \boldsymbol{V}_{21}$*,* $\boldsymbol{W}_{21} \boldsymbol{v}_1 = \boldsymbol{0}$*, and* $\|\boldsymbol{V}_{21}\|_\infty \leq 1 - \theta^{-1}(\|\boldsymbol{A}_{21}\|_\infty + 5)\epsilon$*,*

*(iv)* $\boldsymbol{W}_{22} = \lambda \boldsymbol{A}_{22} - \theta \boldsymbol{V}_{22}$*, and* $\|\boldsymbol{V}_{22}\|_\infty \leq 1$*, and*

*(v)* $\|\boldsymbol{W}\|_2 \leq 1 - (\|\boldsymbol{A}\|_2 + 7.5)\epsilon$*.*

**Remark 1.** *In order to test stopping conditions specified in Proposition 1, we need to start with an* $\epsilon$*-approximate of the optimal solution* $(\lambda^*, \boldsymbol{u}_1^*, \boldsymbol{v}_1^*)$*, where* $\boldsymbol{u}_1^*$ *and* $\boldsymbol{v}_1^*$ *are unit vectors. It is therefore better to solve the problem where* $\lambda^* = 1/\|\boldsymbol{A}\|_\theta^*$ *has the same magnitude as entries of* $\boldsymbol{u}_1^*$ *and* $\boldsymbol{v}_1^*$*. Heuristically, we could scale* $\boldsymbol{A}$ *so that* $\|\boldsymbol{A}\|_2 = 1$ *to (partially) control the magnitude of* $\lambda^*$*.*

**Proof.** Suppose we are given $\boldsymbol{W}$ and $\boldsymbol{V}$ which satisfy the conditions (i)–(v). We will construct $\boldsymbol{W}^*$ and $\boldsymbol{V}^*$ from $\boldsymbol{W}$ and $\boldsymbol{V}$ and prove that they satisfy all optimality conditions in (28)–(32) when combining with the solution $(\lambda^*, \boldsymbol{u}_1^*, \boldsymbol{v}_1^*)$ of (33). We start with the $(1,1)$ subblock . Clearly, we need $\boldsymbol{V}_{11}^* = \boldsymbol{V}_{11} = \boldsymbol{E}_{M\times N}$ and $\boldsymbol{W}_{11}^* = (\lambda^* \boldsymbol{A}_{11} - \theta \boldsymbol{V}_{11}^*) - \boldsymbol{u}_1^*(\boldsymbol{v}_1^*)^T$. We have:

$$\boldsymbol{W}_{11}^* - \boldsymbol{W}_{11} = (\lambda^* - \lambda)\boldsymbol{A}_{11} - (\boldsymbol{u}_1^*(\boldsymbol{v}_1^*)^T - \boldsymbol{u}_1\boldsymbol{v}_1^T).$$

Since $(\lambda, \boldsymbol{u}_1, \boldsymbol{v}_1)$ is an $\epsilon$-solution, we have that $\max\{|\Delta\lambda|, \|\Delta\boldsymbol{u}_1\|, \|\Delta\boldsymbol{v}_1\|\} \leq \epsilon$, where $\Delta\lambda = \lambda - \lambda^*$, $\Delta\boldsymbol{u}_1 = \boldsymbol{u}_1 - \boldsymbol{u}_1^*$, and $\Delta\boldsymbol{v}_1 = \boldsymbol{v}_1 - \boldsymbol{v}_1^*$. Hence

$$\begin{aligned}\|\boldsymbol{u}_1^*(\boldsymbol{v}_1^*)^T - \boldsymbol{u}_1\boldsymbol{v}_1^T\| &= \|\boldsymbol{u}_1^*(\boldsymbol{v}_1^*)^T - (\boldsymbol{u}_1^* + \Delta\boldsymbol{u}_1)(\boldsymbol{v}_1^* + \Delta\boldsymbol{v}_1)^T\| \\ &= \|\Delta\boldsymbol{u}_1(\boldsymbol{v}_1^*)^T + \boldsymbol{u}_1^*\Delta\boldsymbol{v}_1^T + \Delta\boldsymbol{u}_1\Delta\boldsymbol{v}_1^T\| \ \leq \ 2\epsilon + \epsilon^2,\end{aligned}$$

since $\|\boldsymbol{u}_1^*\| = \|\boldsymbol{v}_1^*\| = 1$.

We continue with the $(2,2)$ subblock. Let $\boldsymbol{V}_{22}^* = \boldsymbol{V}_{22}$ and $\boldsymbol{W}_{22}^* = \lambda^*\boldsymbol{A}_{22} - \theta\boldsymbol{V}_{22}^*$, we have:

$$\boldsymbol{W}_{22}^* - \boldsymbol{W}_{22} = (\lambda^* - \lambda)\boldsymbol{A}_{22}.$$

Now consider the $(1,2)$ subblock, we would like to construct $\boldsymbol{W}_{12}^*$ that is close to $\boldsymbol{W}_{12}$ and satisfies the condition that $(\boldsymbol{W}_{12}^*)^T\boldsymbol{u}_1^* = \boldsymbol{0}$. We will use appropriate Householder matrices to construct $\boldsymbol{W}_{12}^*$ as follows. For two different unit vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, the Householder matrix $\boldsymbol{Q} = \boldsymbol{I} - 2\boldsymbol{z}\boldsymbol{z}^T$ with $\boldsymbol{z} = \pm\dfrac{\boldsymbol{x} - \boldsymbol{y}}{\|\boldsymbol{x} - \boldsymbol{y}\|}$ transforms $\boldsymbol{x}$ to $\boldsymbol{y}$ and vice versa. In other words, $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{y}$ and $\boldsymbol{Q}\boldsymbol{y} = \boldsymbol{x}$. The Householder matrix $\boldsymbol{Q}$ is symmetric and orthonormal. Now consider $\bar{\boldsymbol{u}}_1 = \boldsymbol{u}_1/\|\boldsymbol{u}_1\|$. Note that since $\|\boldsymbol{u}_1^*\| = 1$ and $\|\Delta\boldsymbol{u}_1\| \leq \epsilon$, we have that $|\|\boldsymbol{u}_1\| - 1| \leq \epsilon$, which implies $\|\Delta\bar{\boldsymbol{u}}_1\| \leq 2\epsilon$, where $\Delta\bar{\boldsymbol{u}}_1 = \bar{\boldsymbol{u}}_1 - \boldsymbol{u}_1^*$. We define $\boldsymbol{x} = -\dfrac{\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1}{\|\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1\|}$ and consider two Householder matrices, $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$, which transform $\bar{\boldsymbol{u}}_1$ to $\boldsymbol{x}$ and $\boldsymbol{x}$ to $\boldsymbol{u}_1^*$, respectively. Let us define

$$\boldsymbol{w}_1 = \bar{\boldsymbol{u}}_1 + \frac{\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1}{\|\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1\|}, \quad \boldsymbol{w}_2 = \boldsymbol{u}_1^* + \frac{\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1}{\|\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1\|},$$

then $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ can be constructed with $\boldsymbol{z}_1 = \boldsymbol{w}_1/\|\boldsymbol{w}_1\|$ and $\boldsymbol{z}_2 = \boldsymbol{w}_2/\|\boldsymbol{w}_2\|$, respectively. We have

$$\boldsymbol{w}_1^T\boldsymbol{w}_1 \ = \left(\bar{\boldsymbol{u}}_1 + \frac{\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1}{\|\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1\|}\right)^T\left(\bar{\boldsymbol{u}}_1 + \frac{\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1}{\|\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1\|}\right) \ = \ 2 + \|\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1\|,$$

or $\|\boldsymbol{w}_1\| = \sqrt{2 + \|\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1\|}$. Similarly, we can also show that $\|\boldsymbol{w}_2\| = \|\boldsymbol{w}_1\| = \sqrt{2 + \|\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1\|}$. Thus we have

$$\Delta\boldsymbol{z}_1 = \boldsymbol{z}_1 - \boldsymbol{z}_2 = \frac{1}{\sqrt{2 + \|\boldsymbol{u}_1^* + \bar{\boldsymbol{u}}_1\|}}\Delta\bar{\boldsymbol{u}}_1.$$

Hence $\|\Delta \boldsymbol{z}_1\| \leq \frac{1}{\sqrt{4-\|\Delta \bar{\boldsymbol{u}}_1\|}}\|\Delta \bar{\boldsymbol{u}}_1\| \leq \frac{2}{\sqrt{3}}\epsilon$.

Now consider $\boldsymbol{Q}_{12} = \boldsymbol{Q}_2\boldsymbol{Q}_1$, we have: $\boldsymbol{Q}\bar{\boldsymbol{u}}_1 = \boldsymbol{u}_1^*$ and $\boldsymbol{Q}$ is also an orthonormal matrix. Define $\boldsymbol{W}_{12}^* = \boldsymbol{Q}_{12}\boldsymbol{W}_{12}$, we have: $\boldsymbol{W}_{12}^*$ satisfies the condition $(\boldsymbol{W}_{12}^*)^T\boldsymbol{u}_1^* = \boldsymbol{0}$ since $\boldsymbol{W}_{12}^T\bar{\boldsymbol{u}}_1 = \boldsymbol{0}$. We can then select $\boldsymbol{V}_{12}^* = (\lambda^*\boldsymbol{A}_{12} - \boldsymbol{W}_{12}^*)/\theta$. Thus

$$\boldsymbol{W}_{12}^* - \boldsymbol{W}_{12} = (\boldsymbol{Q}_{12} - \boldsymbol{I})\boldsymbol{W}_{12}, \quad \boldsymbol{V}_{12}^* - \boldsymbol{V}_{12} = \frac{1}{\theta}\left[(\lambda^* - \lambda)\boldsymbol{A}_{12} - (\boldsymbol{Q}_{12} - \boldsymbol{I})\boldsymbol{W}_{12}\right].$$

We have

$$\boldsymbol{Q}_{12} - \boldsymbol{I} = \left(\boldsymbol{I} - 2\boldsymbol{z}_2\boldsymbol{z}_2^T\right)\left(\boldsymbol{I} - 2\boldsymbol{z}_1\boldsymbol{z}_1^T\right) - \boldsymbol{I} \; = \; 2\boldsymbol{z}_2\Delta\boldsymbol{z}_1^T - 2\Delta\boldsymbol{z}_1\boldsymbol{z}_1^T + 4(\boldsymbol{z}_2^T\Delta\boldsymbol{z}_1)\boldsymbol{z}_2\boldsymbol{z}_1^T.$$

Thus

$$\frac{1}{4}\|\boldsymbol{Q}_{12} - \boldsymbol{I}\|^2 = 2\|\Delta\boldsymbol{z}_1\|^2 + 2(\boldsymbol{z}_2^T\Delta\boldsymbol{z}_1)(\boldsymbol{z}_1^T\Delta\boldsymbol{z}_1).$$

Since $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are unit vectors, we have:

$$\|\boldsymbol{Q}_{12} - \boldsymbol{I}\| \; \leq \; 4\|\Delta\boldsymbol{z}_1\|.$$

The final $(2,1)$ subblock can be analyzed similarly. We would like to find $\boldsymbol{W}_{21}^*$ close to $\boldsymbol{W}_{21}$ such that $\boldsymbol{W}_{21}^*\boldsymbol{v}_1^* = \boldsymbol{0}$. We can define $\bar{\boldsymbol{v}}_1$, $\boldsymbol{y}_1$, and $\boldsymbol{y}_2$, and $\boldsymbol{Q}_{21}$ in a similar way to $\bar{\boldsymbol{u}}_1$, $\boldsymbol{z}_1$, and $\boldsymbol{z}_2$, and $\boldsymbol{Q}_{12}$. We then have $\boldsymbol{W}_{21}^* = \boldsymbol{W}_{21}\boldsymbol{Q}_{21}$ and $\boldsymbol{V}_{21}^* = (\lambda^*\boldsymbol{A}_{21} - \boldsymbol{W}_{21}^*)/\theta$. We also obtain

$$\Delta\boldsymbol{y}_1 = \frac{1}{\sqrt{2 + \|\boldsymbol{v}_1^* + \bar{\boldsymbol{v}}_1\|}}\Delta\bar{\boldsymbol{v}}_1,$$

and

$$\|\boldsymbol{Q}_{21} - \boldsymbol{I}\| \leq 4\|\Delta\boldsymbol{y}_1\|.$$

Finally, we need to prove $\|\boldsymbol{V}^*\|_\infty \leq 1$ and $\|\boldsymbol{W}\|_2 \leq 1$. By noting that $\|(\boldsymbol{Q}_{12} - \boldsymbol{I})\boldsymbol{W}_{12}\|_\infty \leq \|(\boldsymbol{Q}_{12} - \boldsymbol{I})\boldsymbol{W}_{12}\|_2 \leq \|\boldsymbol{Q}_{12} - \boldsymbol{I}\|_2\|\boldsymbol{W}_{12}\|_2 \leq \|\boldsymbol{Q}_{12} - \boldsymbol{I}\|$ and $\|\Delta\boldsymbol{z}_1\| \leq 2/\sqrt{3}\epsilon$, we have

$$\begin{aligned}
\|\boldsymbol{V}_{12}^*\|_\infty - \|\boldsymbol{V}_{12}\|_\infty &\leq \|\boldsymbol{V}_{12}^* - \boldsymbol{V}_{12}\|_\infty \\
&\leq \theta^{-1}\left[|\Delta\lambda|\|\boldsymbol{A}_{12}\|_\infty + \|(\boldsymbol{Q}_{12} - \boldsymbol{I})\boldsymbol{W}_{12}\|_\infty\right] \\
&\leq \theta^{-1}\left[|\Delta\lambda|\|\boldsymbol{A}_{12}\|_\infty + \|\boldsymbol{Q}_{12} - \boldsymbol{I}\|\right] \\
&\leq \theta^{-1}\left[|\Delta\lambda|\|\boldsymbol{A}_{12}\|_\infty + 4\|\Delta\boldsymbol{z}_1\|\right] \\
&\leq \theta^{-1}\left[\|\boldsymbol{A}_{12}\|_\infty + 5\right]\epsilon.
\end{aligned}$$

Thus

$$\|\boldsymbol{V}_{12}^*\|_\infty \leq \|\boldsymbol{V}_{12}\|_\infty + \theta^{-1}\left[\|\boldsymbol{A}_{12}\|_\infty + 5\right]\epsilon \leq 1.$$

Similarly, we also have

$$\|\boldsymbol{V}_{21}^*\|_\infty \le \|\boldsymbol{V}_{21}\|_\infty + \theta^{-1}\left[\|\boldsymbol{A}_{21}\|_\infty + 5\right]\epsilon \le 1.$$

Now consider $\boldsymbol{W}_2^*$. Clearly $\|\boldsymbol{W}^*\|_2 \le \|\boldsymbol{W}\|_2 + \|\boldsymbol{W}^* - \boldsymbol{W}\|_2$. We have

$$\boldsymbol{W}^* - \boldsymbol{W} = (\lambda^* - \lambda)\begin{pmatrix} \boldsymbol{A}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_{22} \end{pmatrix} + \begin{pmatrix} \boldsymbol{0} & (\boldsymbol{Q}_{12} - \boldsymbol{I})\boldsymbol{W}_{12} \\ \boldsymbol{W}_{21}(\boldsymbol{Q}_{21} - \boldsymbol{I}) & \boldsymbol{0} \end{pmatrix} - \begin{pmatrix} \boldsymbol{u}_1^*(\boldsymbol{v}_1^*)^T - \boldsymbol{u}_1\boldsymbol{v}_1^T & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}.$$

Thus we have:

$$\begin{aligned} \|\boldsymbol{W}^* - \boldsymbol{W}\|_2 &\le |\Delta\lambda| \max\{\|\boldsymbol{A}_{11}\|_2, \|\boldsymbol{A}_{22}\|_2\} + \max\{\|(\boldsymbol{Q}_{12} - \boldsymbol{I})\boldsymbol{W}_{12}\|_2, \|\boldsymbol{W}_{21}(\boldsymbol{Q}_{21} - \boldsymbol{I})\|_2\} \\ &\quad + \|\boldsymbol{u}_1^*(\boldsymbol{v}_1^*)^T - \boldsymbol{u}_1\boldsymbol{v}_1^T\|_2 \\ &\le \max\{\|\boldsymbol{A}_{11}\|_2, \|\boldsymbol{A}_{22}\|_2\}\epsilon + 5\epsilon + 2\epsilon + \epsilon^2 \\ &\le \left[\|\boldsymbol{A}\|_2 + 7.5\right]\epsilon, \end{aligned}$$

since $\max\{\|\boldsymbol{A}_{11}\|_2, \|\boldsymbol{A}_{22}\|_2\} \le \|\boldsymbol{A}\|_2$ and $0 < \epsilon < 1/2$. Thus

$$\|\boldsymbol{W}^*\|_2 \le \|\boldsymbol{W}\|_2 + \left[\|\boldsymbol{A}\|_2 + 7.5\right]\epsilon \le 1.$$

We have constructed $\boldsymbol{W}^*$ and $\boldsymbol{V}^*$ that satisfy the optimality condition for $\boldsymbol{X}^*$, which implies that $\boldsymbol{X}^*$ is indeed an optimal solution of Problem (1). □

Proposition 1 show that given an $\epsilon$-solution $(\lambda, \boldsymbol{u}_1, \boldsymbol{v}_1)$ of (33), which for example, can be obtained from the current solution $(\boldsymbol{X}_k, \boldsymbol{Y}_k)$ of the proximal point algorithm, if we could find $\boldsymbol{W}$ and $\boldsymbol{V}$ that satisfy the $\epsilon$-optimality conditions given in Proposition 1, then we can stop the algorithm with an accurate rank-one solution for Problem (1). Given $(\lambda, \boldsymbol{u}_1, \boldsymbol{v}_1)$. Let us consider the following optimization problem:

$$\begin{aligned} \min \quad & \|\boldsymbol{W}\|_2 \\ \text{s.t.} \quad & \boldsymbol{W}_{11} = (\lambda\boldsymbol{A}_{11} - \theta\boldsymbol{E}_{M\times N}) - \boldsymbol{u}_1\boldsymbol{v}_1^T, \\ & \boldsymbol{W}_{12}^T\boldsymbol{u}_1 = \boldsymbol{0}, \\ & \boldsymbol{W}_{21}\boldsymbol{v}_1 = \boldsymbol{0}, \\ & \|\boldsymbol{W}_{12} - \lambda\boldsymbol{A}_{12}\|_\infty \le \theta - (\|\boldsymbol{A}_{12}\|_\infty + 5)\epsilon, \\ & \|\boldsymbol{W}_{21} - \lambda\boldsymbol{A}_{21}\|_\infty \le \theta - (\|\boldsymbol{A}_{21}\|_\infty + 5)\epsilon, \\ & \|\boldsymbol{W}_{22} - \lambda\boldsymbol{A}_{22}\|_\infty \le \theta. \end{aligned} \tag{34}$$

Clearly, if we could find a feasible solution of Problem (34) with the objective $\|\boldsymbol{W}\|_2 \le 1 - (\|\boldsymbol{A}\|_2 + 7.5)\epsilon$, then the $\epsilon$-optimality conditions for $(\lambda, \boldsymbol{u}_1, \boldsymbol{v}_1)$ are satisfied. This problem is a non-smooth convex constrained optimization problem and our main purpose is to find a feasible solution with the objective value that is small enough. Therefore, we can simply apply the projected subgradient method to solve it. The projected subgradient method uses the iteration

$$\boldsymbol{W}_{k+1} = \Pi_{\mathcal{W}} \left[ \boldsymbol{W}_k - \alpha_k \boldsymbol{G}_k \right],$$

where $\boldsymbol{G}_k \in \partial \|\boldsymbol{W}_k\|_2$ is a subgradient of $\|.\|_2$ at $\boldsymbol{W}_k$ and $\mathcal{W}$ is the feasible set of Problem (34). The step size $\alpha_k$ can be chosen as one of the standard step sizes of the general subgradient method. For this problem, we choose $\alpha_k = O\left(1/\sqrt{k}\right)$. According to Ziętak [9], we can always select $\boldsymbol{G}_k = \boldsymbol{u}_k \boldsymbol{v}_k^T \in \partial \|\boldsymbol{W}_k\|_2$, where $(\boldsymbol{u}_k, \boldsymbol{v}_k)$ is the singular vectors corresponding to the largest singular value of $\boldsymbol{W}_k$. We now consider the projection problem $\Pi_{\mathcal{W}}(\bar{\boldsymbol{W}})$:

$$\begin{aligned} \Pi_{\mathcal{W}}(\bar{\boldsymbol{W}}) \in \arg\min \quad & \|\boldsymbol{W} - \bar{\boldsymbol{W}}\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{W}_{11} = (\lambda \boldsymbol{A}_{11} - \theta \boldsymbol{E}_{M \times N}) - \boldsymbol{u}_1 \boldsymbol{v}_1^T, \\ & \boldsymbol{W}_{12}^T \boldsymbol{u}_1 = \boldsymbol{0}, \\ & \boldsymbol{W}_{21} \boldsymbol{v}_1 = \boldsymbol{0}, \\ & \|\boldsymbol{W}_{12} - \lambda \boldsymbol{A}_{12}\|_\infty \le \theta - (\|\boldsymbol{A}_{12}\|_\infty + 5)\epsilon, \\ & \|\boldsymbol{W}_{21} - \lambda \boldsymbol{A}_{21}\|_\infty \le \theta - (\|\boldsymbol{A}_{21}\|_\infty + 5)\epsilon, \\ & \|\boldsymbol{W}_{22} - \lambda \boldsymbol{A}_{22}\|_\infty \le \theta. \end{aligned} \tag{35}$$

The objective function $\|\boldsymbol{W} - \bar{\boldsymbol{W}}\|_F^2$ is element-wise separable; therefore, Problem (35) is block-wise separable. For the $(1,1)$ subblock, we have the fixed solution $\boldsymbol{W}_{11} = (\lambda \boldsymbol{A}_{11} - \theta \boldsymbol{E}_{M \times N}) - \boldsymbol{u}_1 \boldsymbol{v}_1^T$. For the $(2,2)$ subblock, it is a simple element-wise separable optimization problem:

$$\begin{aligned} \min \quad & \|\boldsymbol{W}_{22} - \bar{\boldsymbol{W}}_{22}\|_F^2 \\ \text{s.t.} \quad & \|\boldsymbol{W}_{22} - \lambda \boldsymbol{A}_{22}\|_\infty \le \theta, \end{aligned}$$

whose optimal solution can be computed as follows:

$$\boldsymbol{W}_{22} = \max\left\{\min\left\{\bar{\boldsymbol{W}}_{22}, \lambda \boldsymbol{A}_{22} + \theta\right\}, \lambda \boldsymbol{A}_{22} - \theta\right\}.$$

For the $(1,2)$ subblock, the corresponding optimization problem is column-wise separable:

$$\begin{aligned} \min \quad & \|\boldsymbol{W}_{12} - \bar{\boldsymbol{W}}_{12}\|_F^2 \\ \text{s.t.} \quad & \boldsymbol{W}_{12}^T \boldsymbol{u}_1 = \boldsymbol{0}, \\ & \|\boldsymbol{W}_{12} - \lambda \boldsymbol{A}_{12}\|_\infty \le \theta - (\|\boldsymbol{A}_{12}\|_\infty + 5)\epsilon. \end{aligned}$$

Each subproblem is a quadratic knapsack problem which can be written as follows:

$$\begin{aligned} \min \quad & \|\boldsymbol{w} - \bar{\boldsymbol{w}}\|^2 \\ \text{s.t.} \quad & \boldsymbol{u}_1^T \boldsymbol{w} = 0, \\ \text{s.t.} \quad & \boldsymbol{l} \le \boldsymbol{w} \le \boldsymbol{u}. \end{aligned}$$

According to Brucker [2], there is an $O(n)$ algorithm for these quadratic knapsack problems. Thus we can find $\boldsymbol{W}_{12}$ efficiently. Similarly, the $(2,1)$ subblock can be found by solving a number of quadratic knapsack problems since the corresponding optimization problem for it is row-wise separable.

# 6 Numerical Examples

## 6.1 Sailboat Bitmap Image Example

In this example, we use a 80-by-50 black-and-white bitmap image of a sailboat. There are 5 distinct components or non-overlapping features in this image: left sail (Feature 1), sail mast (Feature 2), right sail (Feature 3), hull (Feature 4), and rudder (Feature 5). The bitmap image of the sailboat is shown in Figure 1. A matrix $\boldsymbol{A}$ is created with 30 columns, each of which represents a bitmap image of the sailboat with just 3 out of 5 features. The matrix $\boldsymbol{A}$ therefore has 5 rank-one submatrices composed of all ones since the bitmap image is black-and-white. The structure of matrix $\boldsymbol{A}$ and all the features (in terms of non-zero elements) are shown in Figure 2. We would like to use our proposed formulation to extract these rank-one submatrices.

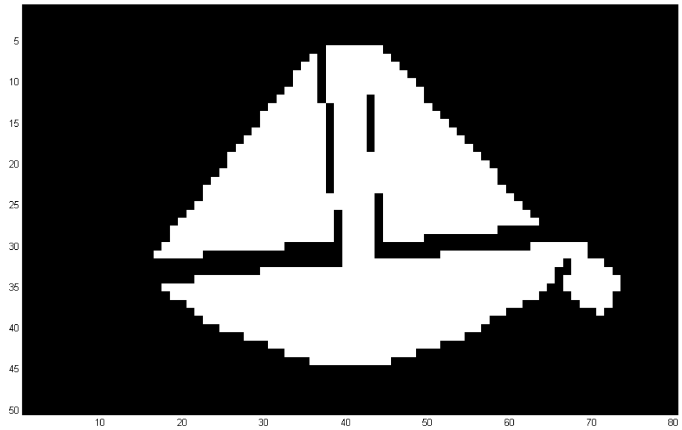

Figure 1: Bitmap image of the complete sailboat

Our main task in this example is to use our proposed formulation to extract the features from the matrix $\boldsymbol{A}$. We have developed two algorithms to solve Problem (1), the primal and dual. For these numerical examples, we will use the dual algorithm mainly due to its accuracy with respect to optimal solutions. This superior accuracy could be explained by the fact that the subproblem we solved in each iteration of the dual algorithm is similar to the original problem. The stopping criterion stated

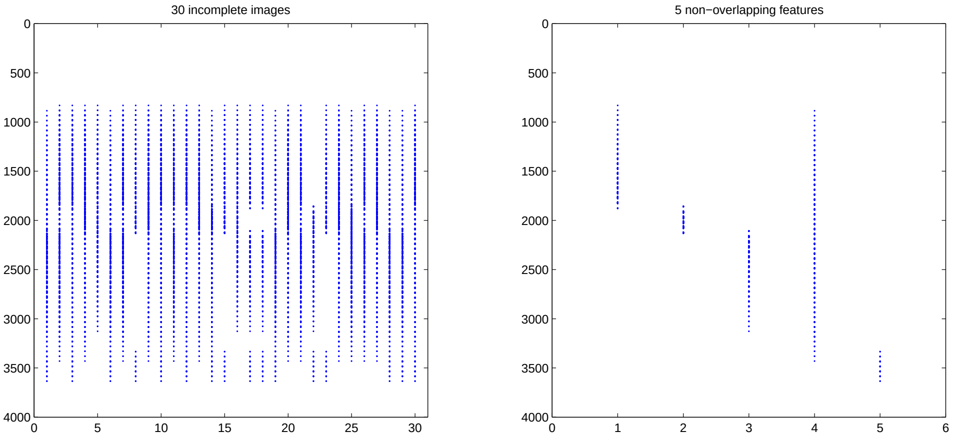

Figure 2: Collection of 30 incomplete images of the sailboat and its 5 components

in Proposition 1 is implemented. We test the conditions of Kantorovich's theorem, solve the system of equations in (33) to obtain an $\epsilon$-solution, and then use the projected subgradient method to find a feasible solution $\boldsymbol{W}$ of Problem (34). Since the projection problem with $\bar{\boldsymbol{W}} = \boldsymbol{0}$ can be considered as a relaxation of Problem (34) in which the spectral norm is replaced by the Frobenius norm, we will start the projected subgradient method with $\boldsymbol{W}_0 = \boldsymbol{0}$. The stopping criteria for this method are the condition $\|\boldsymbol{W}\|_2 \leq 1 - (\|\boldsymbol{A}\|_2 + 7.5)\epsilon$, the maximum number of iterations, and the change in objective values. As the testing process is computationally quite expensive, therefore we only use it once per a fixed number (say, 10) of outer iterations.

For each value of the parameter $\theta$, we will obtain the optimal solution $\boldsymbol{X}_1^*$ and $\boldsymbol{X}_2^*$. The decision variable $\boldsymbol{X}_2$ corresponds to the $l_1$-norm part of the objective function; therefore, we use the sparsity structure of $\boldsymbol{X}_2^*$ to construct the final solution $\boldsymbol{X}_F^*$ with the elements of $\boldsymbol{X}_2^*$. According to Theorem 5 in [3], the optimal solution of Problem (1) indicates the exact sparsity structure of the rank-one submatrix (even under small random noise) with appropriate $\theta$. Therefore, in this experiment, we extract the rank-one approximation of $\boldsymbol{X}_F^*$ and use its sparsity structure as the sparsity structure of the extracted feature. Next, we need to select an appropriate value for $\theta$. For $\theta \approx 0$, the algorithm returns the rank-one approximation of matrix $\boldsymbol{A}$, which is an average of all features and for the purpose of extracting single features, this averaging effect is not desirable. On the other hand, we prefer large submatrices (large features) over small ones. Similar to the L-curve method used to select a regularization parameter, we construct the curve $\mathcal{L} := \{\|\boldsymbol{X}(M_\theta, N_\theta)\|_F, \|\boldsymbol{A}(M_\theta, N_\theta) - \boldsymbol{X}(M_\theta, N_\theta)\|_F, \, \theta \geq 0\}$, where $(M_\theta, N_\theta)$ is the sparsity structure obtained from the algorithm and $\boldsymbol{X}(M_\theta, N_\theta)$ is the rank-one approximation of $\boldsymbol{A}(M_\theta, N_\theta)$. We then pick $\theta$ that balances the feature largeness, $\|\boldsymbol{X}(M_\theta, N_\theta)\|_F$, and feature averaging measure $\|\boldsymbol{A}(M_\theta, N_\theta) - \boldsymbol{X}(M_\theta, N_\theta)\|_F$. After selecting $\theta$, we obtain $\boldsymbol{X}(M_\theta, N_\theta) = \boldsymbol{u}(M_\theta)\boldsymbol{v}(N_\theta)^T$, where $\max(\boldsymbol{v}(N_\theta)) = 1$. The vector $\boldsymbol{u}(M_\theta)$ represents the extracted feature and $\boldsymbol{v}(N_\theta)$ indicates how significant the feature is in each boat image. After extracting a feature, we remove the feature from the image by setting $\boldsymbol{A}(M_\theta, N_\theta) = \boldsymbol{0}$ and continue to find new (non-overlapping) features. This method for choosing $\theta$ is clearly just a heuristic and a more concrete approach for $\theta$ selection is still an important issue for future research.

We are now ready to run our algorithm on this sailboat example. We set the main tolerance to be $\epsilon = 10^{-6}$, the maximum number of iterations to be 1000, and for each subproblem, the maximum number of iterations is set to be 30. There is also the parameter $\lambda$ of the proximal point framework that we need to select. This parameter controls the convergence of the algorithm and for this example, $\lambda = O(1/\theta)$ works well most of the time. We can always adjust $\lambda$ (and number of iterations) to get better convergence if the initial setting does not achieve the tolerance required. We set $\epsilon_s = 10^{-10}$ as the tolerance used in Newton's method to test the additional stopping criterion. Finally, the values of $\theta$ are selected uniformly from three different ranges, small range $[0.01, 0.1]$, medium range $[0.1, 1]$, and large range $[1, 10]$, 10 values in each range.

We start with the matrix $\boldsymbol{A}$. Except for the first value of $\theta$ ($\theta = 0.01$), all other values result in the same rank-one submatrix of $\boldsymbol{A}$, which means $\|\boldsymbol{A}(M_\theta, N_\theta) - \boldsymbol{X}(M_\theta, N_\theta)\| = 0$. Thus we do not need to use the curve $\mathcal{L}$ and just need to pick any value of $\theta > 0.01$. The vector $\boldsymbol{v}(N_\theta)$ is a zero-one vector indicating that either the feature $\boldsymbol{u}(M_\theta)$ appears completely in an image or it does not appear at all. The feature $\boldsymbol{u}(M_\theta)$ represents the exact combination of Feature 1 and 4, which is the left sail and the hull.

We now exclude the first extracted feature from all the images and continue to find new (non-overlapping) features. Table 1 shows all the features that we obtain with the size of the features ($s_i$), number of images that share each feature ($n_i$), and their description.

The results show that our algorithm can pick out the large common features that are inherent in the structure of the image set. For example, a combination of our defined features is indeed a large common feature if there are enough images that share that combination of features.

To end this section, we would like to comment on the efficiency of the additional stopping criterion

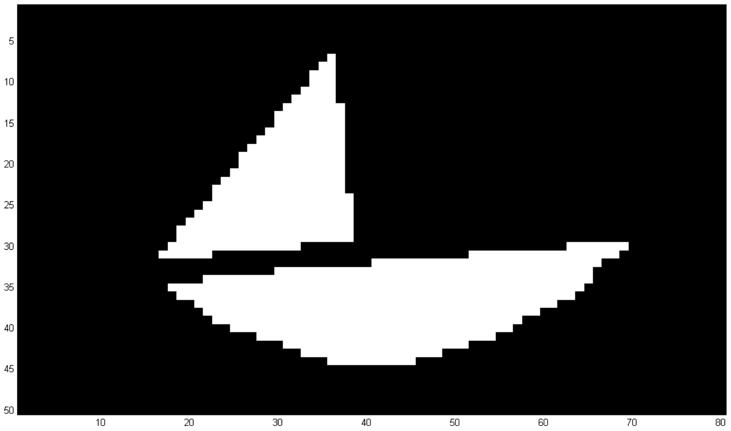

Figure 3: First extracted feature: the combination of left sail and hull

| $i$ | $s_i$ | $n_i$ | Description |
|---|---|---|---|
| 1 | 710 | 15 | Left sail and hull |
| 2 | 694 | 6 | Right sail and hull |
| 3 | 252 | 10 | Right sail |
| 4 | 156 | 5 | Sail mast and rudder |
| 5 | 268 | 7 | Left sail |
| 6 | 119 | 9 | Sail mast |
| 7 | 439 | 1 | Hull |
| 8 | 34 | 11 | Rudder |

Table 1: All extracted features obtained from the algorithm

based on Proposition 1. When the test indicates the convergence is achieved, it is guaranteed that the supports of $\boldsymbol{u}$ and $\boldsymbol{v}$ have been correctly identified for the rank-one optimal solution $\boldsymbol{X}$. On the other hand, because the test uses heuristics to find multipliers, it may occur that the optimal supports are attained and yet the test fails to indicate that. In this sailboat example, we ran the algorithm with 8 different matrices, $\boldsymbol{A}_0 = \boldsymbol{A}$, $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_7$ with subsequent extraction of features one by one. The additional stopping criterion works for 4 out of 8 matrices and we obtain a significant reduction in both computational time and number of iterations while maintaining highly accurate solutions ($\epsilon_s = 10^{-10}$). Table 2 shows these improvements with $\theta = 0.2$, where (DDPA)/(ADDPA) is the algorithm without/with the additional stopping criteria. The number of iterations with (ADDPA) is either 10 or 20 since in this example, we only test the additional stopping criterion once per 10 (outer) iterations. We can see that there are cases when the additional stopping criterion can be used very early to stop the algorithm with a guaranteed highly accurate solution.

| Matrix | (DPPA) | (ADPPA) |
|---|---|---|
| $\boldsymbol{A}_4$ | $(59, 325, 9.36s, 0.00s)$ | $(20, 123, 6.36s, 2.32s)$ |
| $\boldsymbol{A}_5$ | $(52, 267, 7.28s, 0.00s)$ | $(10, 60, 4.23s, 2.39s)$ |
| $\boldsymbol{A}_6$ | $(62, 499, 12.4s, 0.00s)$ | $(20, 187, 6.96s, 1.84s)$ |
| $\boldsymbol{A}_7$ | $(29, 148, 3.59s, 0.00s)$ | $(10, 57, 4.00s, 2.48s)$ |

Table 2: Outer iteration number, inner iteration number, total computational time, and convergence testing time for (DDPA)/(ADDPA)

### 6.2 Image Database Test Case

We conduct the experiment on the Frey face dataset, which consists of 1965 registered face images of size $28 \times 20$. The matrix $\boldsymbol{A}$ has the size of $1965 \times 560$, where each column represents a single face image. We again use the dual algorithm with the additional stopping criterion and maintain all parameters the same as in the previous example. The additional stopping criterion is less effective in this test case. However, when it works, we again have a significant improvement in computational time and number of iterations. For example, with $\boldsymbol{A}_0 = \boldsymbol{A}$ and $\theta = 0.2$, we have the following results for (DDPA) and (ADDPA) respectively: $(245, 6466, 1.21 \times 10^3 s, 0.00s)$ and $(100, 2140, 4.25 \times 10^2 s, 23.7s)$, where the tuple is explained in the caption of Figure 2.

We apply the algorithm to the matrix $\boldsymbol{A}$ and Figure 4 shows the curve $\mathcal{L}$ obtained with different

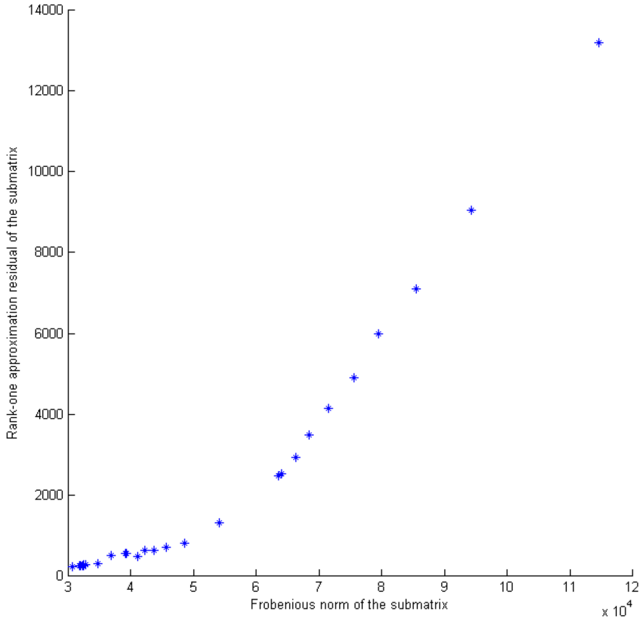

Figure 4: Feature largeness vs. feature averaging measure for different $\theta$

values of $\theta$. We select $\theta = 0.2$ at the curviest point on $\mathcal{L}$, which indicates the balance between feature largeness and feature averaging measure. We obtain the feature $\boldsymbol{u}_1$ and the significance vector $\boldsymbol{v}_1$ indicating how strong the appearance of that feature is in each image. The feature is composed of 38 pixels and there are 1557 images that are considered to have this feature with the significance factor of at least 95.14%, where the significance factor of the feature in image $j$ is defined as $v_1(j)/\|\boldsymbol{v}_1\|_\infty$. Figure 5 presents the first feature and the face image that has the significance factor of 100% for this feature. Basically, the first feature shows the right forehead, a part of right cheekbone, and the tip of the nose. This feature is common among the images (1557 of them), there are images that do not share the feature. Figure 6 shows one of such images.

We remove the first feature from all images that share that feature and continue to find new features. Table 3 shows the size of the feature $i$ ($s_i$), number of images that share the feature $i$ ($n_i$), and the minimum significance factor for each feature $i$ ($f_i^{\min}$), $i = 1, \ldots, 10$.

Figure 7 and 8 presents each feature and the ten face images that have the highest significance factors for that feature.

The features are not easy to observe or distinguish. However, with images that have high significance factors, we can see that some features could be associated with a certain orientation of the face or lighting

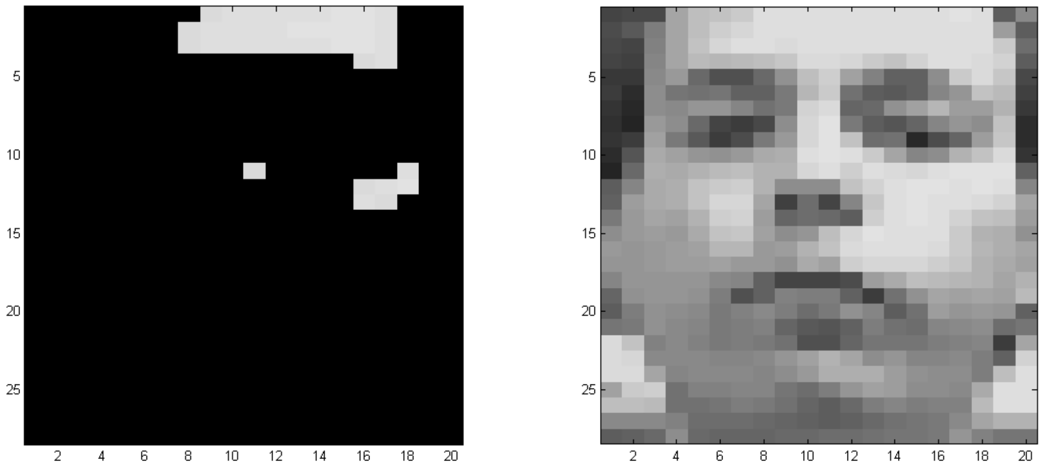

Figure 5: First feature and the image that has the highest significance factor

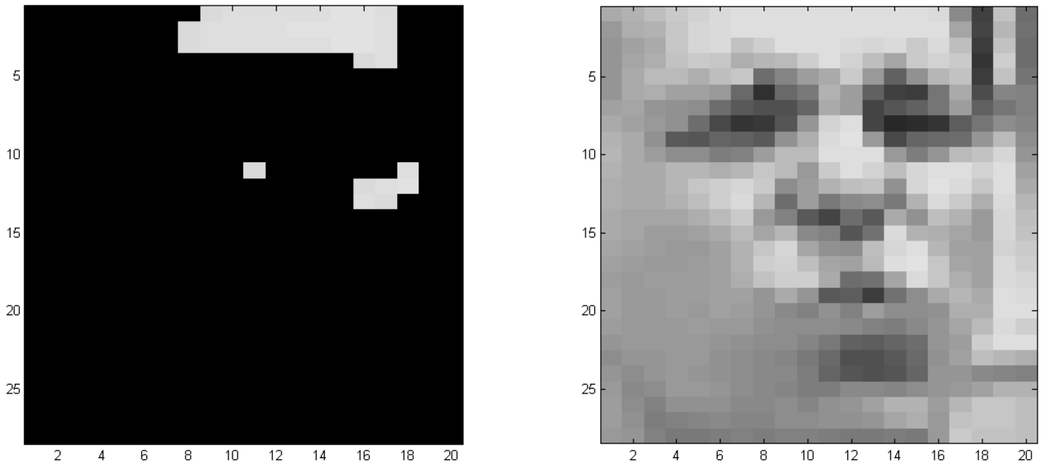

Figure 6: An image without the first feature

| $i$ | $s_i$ | $n_i$ | $f_i^{\min}$ | $i$ | $s_i$ | $n_i$ | $f_i^{\min}$ |
|---|---|---|---|---|---|---|---|
| 1 | 38 | 1557 | 95.14% | 6 | 28 | 673 | 83.27% |
| 2 | 27 | 896 | 92.19% | 7 | 21 | 578 | 80.38% |
| 3 | 29 | 1096 | 87.61% | 8 | 20 | 555 | 87.59% |
| 4 | 24 | 847 | 83.12% | 9 | 35 | 291 | 80.73% |
| 5 | 25 | 791 | 83.67% | 10 | 13 | 598 | 71.31% |

Table 3: Information of the first ten extracted features

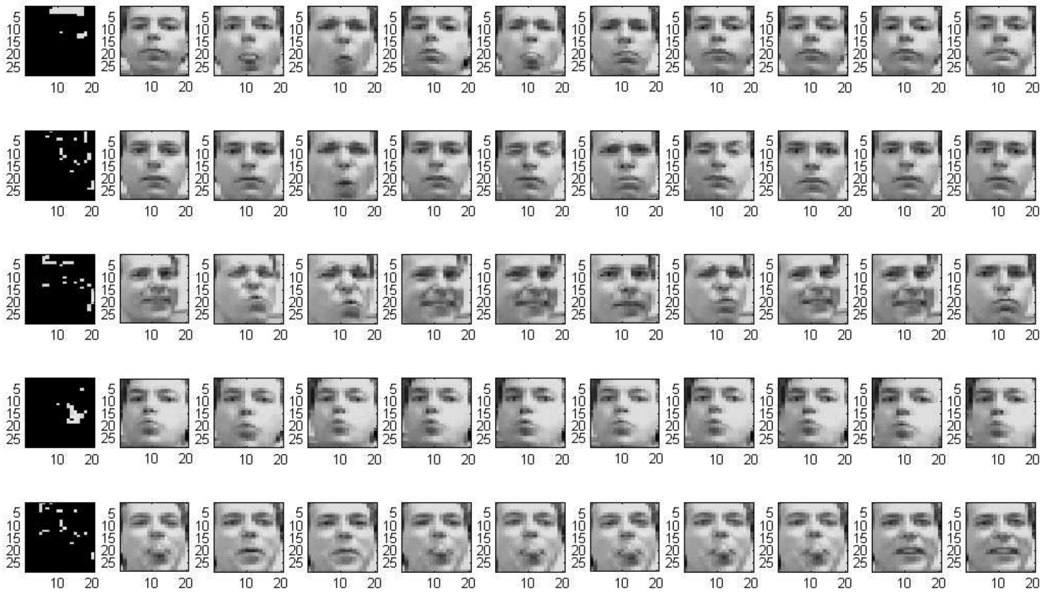

Figure 7: First five features and images with highest significance factors

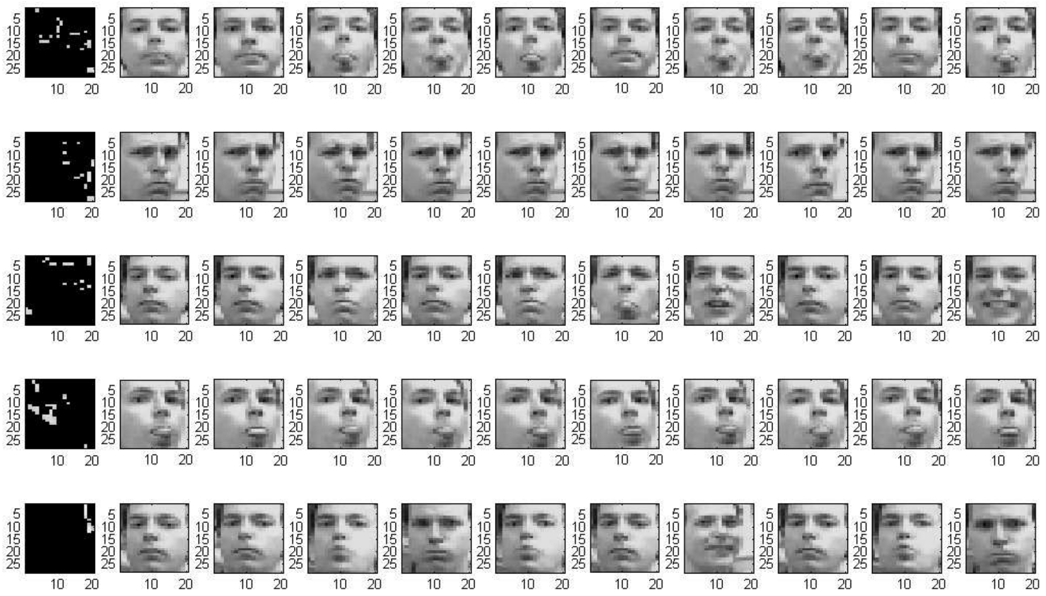

Figure 8: Feature 6 to Feature 10 and images with highest significant factors

of images. For example, Feature 4 and Feature 9 clearly show the right (or left) cheek when Frey faces left (or right). Certain lighting of the background can also define features, which is the case of Feature 3 and Feature 7. Another observation is that since this approach favors large submatrices, which means other features defined by small entries (dark pixels), for examples, eyes or mouth, will not be picked up as major features. In this particular application of visual features, we can define *negative* features, which correspond to the features of the negative images. In order to find these negative features, we construct the negative images and apply the algorithm to this set of images. In this example, the algorithm is applied to $\boldsymbol{B} = 255\boldsymbol{E} - \boldsymbol{A}$, where $\boldsymbol{E}$ is the matrix of all ones. The coefficient 255 appears due to the range of pixel intensities in these images. For each feature $\boldsymbol{u}$ extracted from $\boldsymbol{B}$, we define $\boldsymbol{u}_n = 255\boldsymbol{e} - \boldsymbol{u}$, where $\boldsymbol{e}$ is the vector of all ones, as the negative feature of the original set of images. The first three extracted features are presented in Figure 9. The first negative feature has both straight dark eyes with two dark background columns at both sides on top. The second one focuses on the darker right eye, the left nostril, and also the chin. And the third one is a long dark background column on the top left.

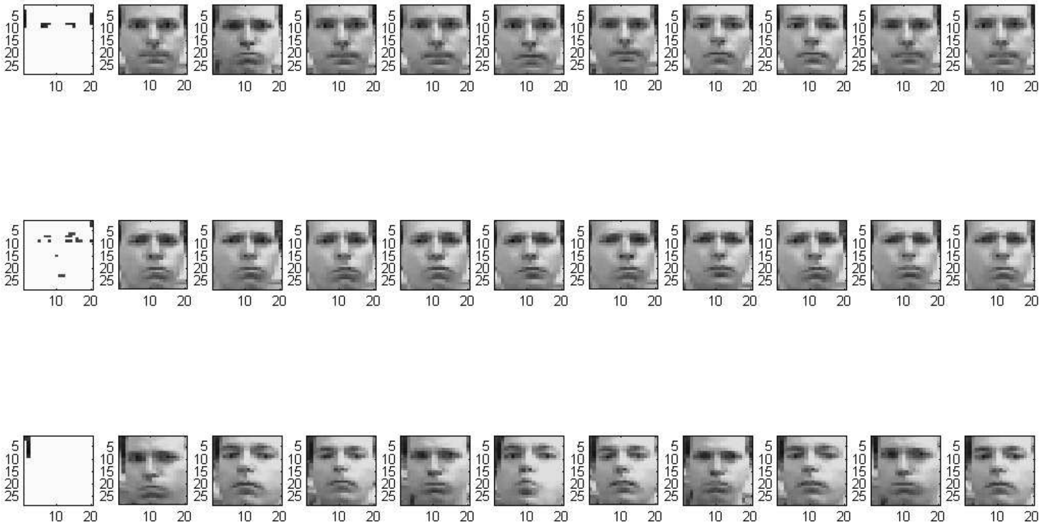

Figure 9: First three negative features and images with highest significance factors

# References

[1] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 1:183–202, 2009.

[2] P. Brucker. An $O(n)$ algorithm for quadratic knapsack problems. *Operations Research Letters*, 3:163–166, 1984.

[3] X. V. Doan and S. Vavasis. Finding approximately rank-one submatrix. Under review, SIAM Journal of Optimization, URL: `http://arxiv.org/abs/1011.1839`, 2010.

[4] Y. J. Liu, D. Sun, and K. C. Toh. An implementable proximal point algorithmic framework for nuclear norm minimization, 2009. Preprint.

[5] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, NJ, 1970.

[6] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal of Control and Optimization*, 14:877–898, 1976.

[7] R. A. Tapia. The Kantorovich theorem for Newton method. *The American Mathematical Monthly*, 78:389–392, 1971.

[8] K. C. Toh and S. W. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems, 2009. To appear in Pacific Journal of Optimization.

[9] K. Ziętak. Properties of linear approximations of matrices in the spectral norm. *Linear Algebra Applications*, 183:41–60, 1993.