Strictly Low Rank Constraint Optimization — An Asymptotically $\mathcal{O}(\frac{1}{t^2})$ Method

Mengyuan Zhang, Kai Liu
{mengyuz, kail}@clemson.edu
Clemson University

Abstract

We study a class of non-convex and non-smooth problems with *rank* regularization to promote sparsity in optimal solution. We propose to apply the proximal gradient descent method to solve the problem and accelerate the process with a novel support set projection operation on the singular values of the intermediate update. We show that our algorithms are able to achieve a convergence rate of $O(\frac{1}{t^2})$, which is exactly same as Nesterov's optimal convergence rate for first-order methods on smooth and convex problems. Strict sparsity can be expected and the support set of singular values during each update is monotonically shrinking, which to our best knowledge, is novel in momentum-based algorithms.

1 Introduction

Sparsity-induced optimization has achieved great success in many data analyses, and low-rank regularization is a powerful tool to impose sparsity. In this paper, we study the class of non-convex and non-smooth sparse learning problems in discrete space presented as follows:

$$\min F(\mathbf{X}) = g(\mathbf{X}) + h(\mathbf{X}), \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{n\times k}, h(\mathbf{X}) = \lambda \cdot rank(\mathbf{X})$ $(\lambda > 0)$. Countless problems in machine learning Shang et al. (2017); Su et al. (2020), computer vision Haeffele et al. (2014); He et al. (2015), and signal processing Dong et al. (2014); Zhou & Li (2014) naturally fall into this template, including but not limited to rank regularized matrix factorization for recommendation, image restoration and clustering, compressive sensing, and multi-task learning.

Besides the efforts trying to relax the $rank$ regularization with the nuclear norm, which is the tightest convex envelop, a class of non-convex optimization algorithms has been developed to solve rank-regularized problems with vanilla objectives. The benefit is obvious that it will bring strict sparse solution instead of low energy of singular values. For example, the cardinality constraint is studied for M-estimation problems by Iterative Hard-Thresholding (IHT) method Blumensath & Davies (2008). In addition, the alternating direction method of multipliers is also explored to apply on the rank-regularized problem Qu et al. (2023). Moreover, general iterative shrinkage and thresholding algorithm has been proposed to solve non-convex sparse regularization problems Gong et al. (2013). We refer the readers to the papers aforementioned and the references therein. In this paper, we use the proximal gradient descent method to obtain a sparse sub-optimal solution for Eq. (1). Also, with a *support set projection* operation on the singular values of the matrix $\mathbf{X}$, our proposed algorithm can be accelerated with a faster convergence rate. We explicitly list our contributions as following:

- We show that with the proximal gradient descent method, the sequence obtained converges to a critical point with *zero* gradient, at a convergence rate of $F(\mathbf{X}^{t+1}) - F(\mathbf{X}^*) \leq O(\frac{1}{t})$.
- We propose two accelerated proximal gradient descent methods (monotone and non-monotone decreasing) which can *asymptotically* achieve Nesterov Accelerated Gradient

ICML 2023 Workshop: Sampling and Optimization in Discrete Space (SODS 2023).

(NAG) convergence rate ($O(\frac{1}{t^2})$) which originally is supposed for convex problems Nesterov (2003) and we give the rigorous proof.

- The proposed algorithms can indeed admit strict sparsity, as the support set of singular values in each update is a subset of the initialized set and keeps shrinking during the update.

2 Algorithms and Convergence Analysis

Throughout this paper, we use uppercase bold letters for matrices, lowercase bold letters for vectors, and lowercase letters for scalars. We use superscript to represent the current iteration. $\sigma_i(\cdot)$ is the i-th largest singular value of a matrix. $\sigma_{min}(\cdot)$ and $\sigma_{max}(\cdot)$ represent the smallest and the largest singular value of a matrix, respectively. $\sigma(\cdot)$ indicates the vector formed by the singular values of a matrix, and $\mathrm{diag}(\sigma(\cdot))$ denotes the diagonal matrix formed by the vector. $|\cdot|$ denotes the cardinality of a set, and $\mathrm{supp}(\cdot)$ denotes the support of a vector. For a matrix $\mathbf{X} \in \mathbb{R}^{n\times k}$, let $\sigma(\mathbf{X}) = (\sigma_{max}(\mathbf{X}), \sigma_2(\mathbf{X}), \ldots, \sigma_{min}(\mathbf{X}))$, then we can see $\mathrm{rank}(\mathbf{X}) = \|\sigma(\mathbf{X})\|_0$. $S = \mathrm{supp}(\sigma(\mathbf{X}^0))$, $\mathbf{X}^0$ is the initialization. The proximal mapping associated with $h(\cdot)$ is defined as $\mathrm{prox}_h(\mathbf{U}) = \mathrm{argmin}_{\mathbf{V}}\, h(\mathbf{V}) + \frac{1}{2}\|\mathbf{U}-\mathbf{V}\|_F^2$. We make some mild assumptions regarding $g(\mathbf{X})$: 1) g$(\cdot)$ is convex, and $\nabla g(\cdot)$ has bounded singular value $\sigma_{max}(\nabla g(\mathbf{X}))) \le G$ for any $\mathbf{X}$; 2) g$(\cdot)$ has L-Lipschitz continuous gradient, $\|\nabla g(\mathbf{X}_1) - \nabla g(\mathbf{X}_2)\|_F \le L\|\mathbf{X}_1 - \mathbf{X}_2\|_F$. Regarding $g(\cdot)$, we set the squared loss $\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2$ to demonstrate the algorithm for simplicity, where $\mathbf{Y} \in \mathbb{R}^{d\times k}, \mathbf{D} \in \mathbb{R}^{d\times n}, \mathbf{X} \in \mathbb{R}^{n\times k}$. It is worth noting that the objective can be extended to other loss functions.

2.1 Proximal Gradient Descent for Low-Rank Approximation

In the t-th iteration of the proximal gradient descent (PGD) method, gradient descent is applied on the squared loss function $g(\mathbf{X})$ to get $\mathbf{X}^t - s\nabla g(\mathbf{X}^t)$, where $s \ge 0$ is the step size in gradient descent and $\frac{1}{s}$ is typically larger than the Lipschitz continuous constant L of $g(\mathbf{X})$. After applying the proximal mapping on $\mathbf{X}^t - s\nabla g(\mathbf{X}^t)$ we can get $\mathbf{X}^{t+1}$:

$$\mathbf{X}^{t+1} = \mathrm{prox}_h(\mathbf{X}^t - s\nabla g(\mathbf{X}^t)) = T_{\sqrt{2\lambda s}}(\mathbf{X}^t - s\nabla g(\mathbf{X}^t)). \tag{2}$$

$T_\theta(\cdot)$ is the singular value hard thresholding operator defined as $T_\theta(\mathbf{Q}) = \mathbf{U}\Sigma_\theta\mathbf{V}^T$, where $\mathbf{Q} = \mathbf{U}\Sigma\mathbf{V}^T$ is the singular value decomposition and $\Sigma_\theta(i,i) = \begin{cases} 0, & |\Sigma(i,i)| \le \theta, \\ \Sigma(i,i), & \text{otherwise.} \end{cases}$

The optimization algorithm to minimize problem Eq. (1) by PGD is summarized in Algorithm 1.

Algorithm 1 Proximal Gradient Descent for the Rank-Regularized Problem

Input: Initialization $\mathbf{X}^0$, step size s, regularization parameter λ.
for $t = 0, \ldots$ **do**
  Update $\mathbf{X}^{t+1}$ according to Eq. (2)
end for

We present the analysis of the convergence rate of Algorithm 1. Before that, we first show the support set of the singular value vector shrinks, and then the rank of obtained solutions decreases, also the objective has a sufficient decrease during each update.

Lemma 2.1. *If* $s \le \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$, *then* $\mathrm{supp}(\sigma(\mathbf{X}^{t+1})) \subseteq \mathrm{supp}(\sigma(\mathbf{X}^t))$,*for* $t \ge 0$, *and* $\mathrm{rank}(\mathbf{X}^{t+1}) \le \mathrm{rank}(\mathbf{X}^t)$,*for* $t \ge 0$, *which means the support of the singular value vectors of the sequence* $\{\mathbf{X}^t\}_t$ *shrinks, also the rank of the sequence* $\{\mathbf{X}^t\}_t$ *decreases.*

Lemma 2.2. *With* $s \le \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$, *the sequence of the objective* $\{F(\mathbf{X}^t)\}_t$ *is monotonically non-increasing, and the following inequality holds for all* $t \ge 0$*:*

$$F(\mathbf{X}^{t+1}) \le F(\mathbf{X}^t) - (\frac{1}{2s} - \frac{L}{2})\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2. \tag{3}$$

To eliminate the concern that very small step size is required to ensure Lemma 2.1 with the prerequisite $s \le \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$, we show that $\frac{1}{L}$ is no larger than $\frac{2\lambda}{G^2}$ with high probability, thus the choice of step size s in Lemma 2.1 would not be smaller than $\frac{1}{L}$ with high probability.

Theorem 2.3. *Yang & Yu (2020) Suppose* $\mathbf{D} \in \mathbb{R}^{d\times n}(n \geq d)$ *is a random matrix with elements i.i.d. sampled from the standard Gaussian distribution N(0,1), then*

$$\textit{Probability}(\frac{1}{L} \leq \frac{2\lambda}{G^2}) \geq 1 - e^{-\frac{a^2}{2}} - ne^{-a}, \tag{4}$$

if $n \geq (\sqrt{d} + a + \sqrt{\frac{(d+2\sqrt{da}+2a)(x_0+\lambda|S|))}{\lambda}})^2$, *where* $x_0 = \|\mathbf{Y} - \mathbf{D}\mathbf{X}^0\|_F^2$, $S = \text{supp}(\sigma(\mathbf{X}^0))$, *and* a *can be chosen as* $a_0 logn$ *for* $a_0 > 0$.

The sequence $\{\mathbf{X}^t\}_t$ can be segmented into the following $|S|+1$ subsequences $\{\mathcal{X}^{k'}\}_{k'=0}^{|S|}$ with the definition as follows:

$$\mathcal{X}^{k'} = \{\mathbf{X}^t : |supp(\sigma(\mathbf{X}^t))| = k', t \geq 0\}, 0 \leq k' \leq |S|. \tag{5}$$

With the definition defined in Definition 2.4, the nonempty subsequences in $\{\mathcal{X}^{k'}\}_{k'=0}^{|S|}$ form a disjoint cover of $\{\mathcal{X}^t\}_t$ and they are in descending order of rank.

Definition 2.4. Subsequences with shrinking support: All the $K \leq |S|+1$ nonempty subsequences among $\{\mathcal{X}^{k'}\}_{k'=0}^{|S|}$ are defined to be subsequences with shrinking support, denoted by $\{\mathcal{X}^k\}_{k=1}^K$. The subsequences with shrinking support are ordered with decreasing support size of singular value vectors, i.e. $|\text{supp}(\sigma(\mathbf{X}^{t_2}))| < |\text{supp}(\sigma(\mathbf{X}^{t_1}))|$ for any $\mathbf{X}^{t_1} \in \mathcal{X}^{k_1}$ and $\mathbf{X}^{t_2} \in \mathcal{X}^{k_2}$ with any $1 \leq k_1 < k_2 \leq K$.

Based on the above definition, we have the following lemma about the property of subsequences with shrinking support:

Lemma 2.5. *(a) All the elements of each subsequence* $\mathcal{X}^k$ $(k = 1, \ldots, K)$ *in the subsequences with shrinking support have the same support. In addition, for any* $1 \leq k_1 < k_2 \leq K$ *and any* $\mathbf{X}^{t_1} \in \mathcal{X}^{k_1}$, *and* $\mathbf{X}^{t_2} \in \mathcal{X}^{k_2}$, *we have* $t_1 < t_2$ *and* $\text{supp}(\sigma(\mathbf{X}^{t_2})) \subset \text{supp}(\sigma(\mathbf{X}^{t_1}))$.

(b) All the subsequences except for the last one, $\mathcal{X}^k$ $(k = 1, \ldots, K-1)$ *have finite size, and* $\mathcal{X}^K$ *have an infinite number of elements, and there exists some* $t_0 \geq 0$ *such that* $\{\mathbf{X}^t\}_{t=t_0}^{\infty} \subseteq \mathcal{X}^K$.

The following theorem shows the convergence property of the sequence $\{\mathbf{X}^t\}_t$:

Theorem 2.6. *Suppose* $s \leq \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$. *and* $\mathbf{X}^*$ *is a limit point of* $\{\mathbf{X}^t\}_{t=0}^{\infty}$, *and* $\sigma(\mathbf{X}^*)$ *is a limit point of* $\{\sigma(\mathbf{X}^t)\}_{t=0}^{\infty}$, *then the sequence* $\{\mathbf{X}^t\}_{t=0}^{\infty}$ *generated by Algorithm 1 converges to* $\mathbf{X}^*$, $\mathbf{X}^*$ *is a critical point of F(·), and* $\text{supp}(\sigma(\mathbf{X}^*)) = S^*$, *where* S^* *is the support of any element in* $\mathcal{X}^K$. *Moreover, there exists* $t_1 \geq 0$ *such that for all* $m \geq t_1$, *we have*

$$F(\mathbf{X}^{m+1}) - F(\mathbf{X}^*) \leq \frac{1}{2s(m-t_1+1)}\|\mathbf{X}^{t_1} - \mathbf{X}^*\|_F^2. \tag{6}$$

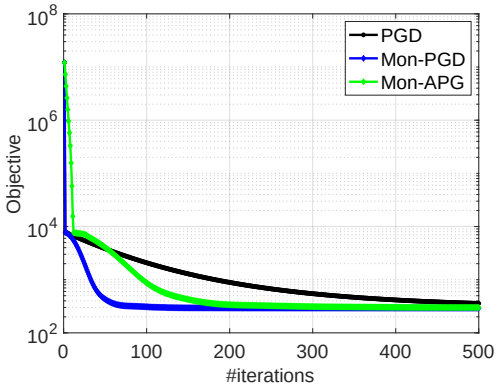


Figure 1: The convergence for Algorithm 1 PGD, Algorithm 3 Mon-PGD, and Mon-APG.

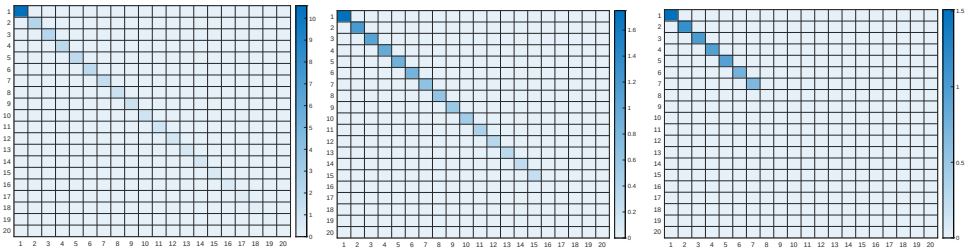
Algorithm 2 Non-monotone Accelerated Proximal Gradient Descent

Input: Initialize $\mathbf{X}^0$, $\mathbf{X}^1 = \mathbf{X}^0$, $\alpha^0 = 1, s, \lambda$.
for $t = 1, \ldots$ **do**
  Update $\mathbf{U}^t, \mathbf{V}^t, \mathbf{X}^{t+1}, \alpha^{t+1}$ by Eq. (7).
end for

Algorithm 3 Monotone Accelerated Proximal Gradient Descent

Input: Initialization $\mathbf{Z}^1 = \mathbf{X}^1 = \mathbf{X}^0$, α^0, s, λ.
for $t = 1, \ldots$ **do**
  Update $\mathbf{U}^t, \mathbf{V}^t, \mathbf{Z}^{t+1}, \alpha^{t+1}, \mathbf{X}^{t+1}$ by (9).
end for

2.2 Non-monotone Accelerated Proximal Gradient Descent with Support Projection



(a) Initialized $\mathbf{X}^0$ (b) $\mathbf{X}^t$ during update (c) Output $\mathbf{X}^T$

Figure 2: Support Shrinkage of *singular values* in Algorithm 1 PGD.

In the non-monotone accelerated proximal gradient descent with support projection, the update process in the t-th iteration is as follows:

$$\begin{aligned}\mathbf{U}^t &= \mathbf{X}^t + \frac{\alpha^{t-1}-1}{\alpha^t}(\mathbf{X}^t - \mathbf{X}^{t-1}), \mathbf{V}^t = P_{\text{supp}(\sigma(\mathbf{X}^t))}(\mathbf{U}^t), \\ \mathbf{X}^{t+1} &= \text{prox}_h(\mathbf{V}^t - s\nabla g(\mathbf{V}^t)), \alpha^{t+1} = \frac{\sqrt{1+4(\alpha^t)^2}+1}{2},\end{aligned} \tag{7}$$

where $P_{\text{supp}(\sigma(\mathbf{X}^t))}(\cdot)$ is the support projection operator which projects the singular value vector of a matrix to the support of the singular value vector of $\mathbf{X}^t$ as $P_{\text{supp}(\sigma(\mathbf{X}^t))}(\mathbf{T}) = \mathbf{A}\Sigma_{\text{projected}}\mathbf{B}^T$, with $\mathbf{T} = \mathbf{A}\Sigma\mathbf{B}^T$ being the singular value decomposition, and $\Sigma_{\text{projected}}(i,i) = \Sigma(i,i)$ if i is in the support of $\sigma(\mathbf{X}^t)$, otherwise $= 0$. Here the support projection is employed to enforce the support shrinkage property mentioned in Lemma 2.5. The algorithm for the non-monotone accelerated proximal gradient descent with support projection is summarized in Algorithm 2. Lemma 2.7 shows the support shrinkage property for Algorithm 2, and Theorem 2.8 presents the convergence rate.

Lemma 2.7. *The sequence* $\{\mathbf{X}^t\}_{t=1}^{\infty}$ *generated by Algorithm 2 satisfies* $\text{supp}(\sigma(\mathbf{X}^{t+1})) \subseteq \text{supp}(\sigma(\mathbf{X}^t)), t \geq 1$.

Theorem 2.8. *Suppose* $s \leq \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$. *and* $\mathbf{X}^*$ *is a limit point of* $\{\mathbf{X}^t\}_{t=0}^{\infty}$ *generated by Algorithm 2, then there exists* $t_0 \geq 1$ *such that for all* $m \geq t_0$, *we have*

$$F(\mathbf{X}^{m+1}) - F(\mathbf{X}^*) \leq \frac{4}{(m+1)^2}[\frac{1}{2s}\|(\alpha^{t_0-1}-1)\mathbf{X}^{t_0-1} - \alpha^{t_0-1}\mathbf{X}^{t_0} + \mathbf{X}^*\|_F^2 + (\alpha^{t_0-1})^2(F(\mathbf{X}^{t_0}) - F(\mathbf{X}^*))]. \tag{8}$$

2.3 Monotone Accelerated Proximal Gradient Descent with Support Projection

To ensure the objective is non-increasing, we introduce the following algorithm which is summarized in Algorithm 3 Beck (2017):

$$\begin{aligned}\mathbf{U}^t &= \mathbf{X}^t + \frac{\alpha^{t-1}-1}{\alpha^t}(\mathbf{X}^t - \mathbf{X}^{t-1}) + \frac{\alpha^t - 1}{\alpha^t}(\mathbf{Z}^t - \mathbf{X}^t), \mathbf{V}^t = P_{\text{supp}(\sigma(\mathbf{Z}^t))}(\mathbf{U}^t), \\ \mathbf{Z}^{t+1} &= \text{prox}_h(\mathbf{V}^t - s\nabla g(\mathbf{V}^t)), \alpha^{t+1} = \frac{\sqrt{1+4(\alpha^t)^2}+1}{2}, \mathbf{X}^{t+1} = \begin{cases}\mathbf{Z}^{t+1} & \text{if } F(\mathbf{Z}^{t+1}) \leq F(\mathbf{X}^t), \\ \mathbf{X}^t & \text{otherwise.}\end{cases}\end{aligned} \tag{9}$$

Lemma 2.9 shows the support shrinkage property and Theorem 2.10 presents the convergence rate.

Lemma 2.9. *The sequence* $\{\mathbf{Z}^t\}_{t=1}^{\infty}$ *and* $\{\mathbf{X}^t\}_{t=1}^{\infty}$ *generated by Algorithm 3 satisfies* $\text{supp}(\sigma(\mathbf{Z}^{t+1})) \subseteq \text{supp}(\sigma(\mathbf{Z}^t)), t \geq 1, \text{supp}(\sigma(\mathbf{X}^{t+1})) \subseteq \text{supp}(\sigma(\mathbf{X}^t)), t \geq 1$.

Theorem 2.10. *Suppose* $s \leq \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$. *and* $\mathbf{X}^*$ *is a limit point of* $\{\mathbf{X}^t\}_{t=0}^{\infty}$ *generated by Algorithm 3, then there exists* $t_0 \geq 1$ *such that for all* $m \geq t_0$, *we have*

$$F(\mathbf{X}^{m+1}) - F(\mathbf{X}^*) \leq \frac{4}{(m+1)^2}[\frac{1}{2s}\|(\alpha^{t_0-1}-1)\mathbf{X}^{t_0-1} - \alpha^{t_0-1}\mathbf{Z}^{t_0} + \mathbf{X}^*\|_F^2 + (\alpha^{t_0-1})^2(F(\mathbf{X}^{t_0}) - F(\mathbf{X}^*))]. \tag{10}$$

We leave all the detailed proof in the Appendix due to space limit.

References

Beck, A. *First-order methods in optimization*. SIAM, 2017.

Blumensath, T. and Davies, M. E. Iterative thresholding for sparse approximations. *Journal of Fourier analysis and Applications*, 14:629–654, 2008.

Davidson, K. R. and Szarek, S. J. Local operator theory, random matrices and banach spaces. *Handbook of the geometry of Banach spaces*, 1(317-366):131, 2001.

Dong, W., Shi, G., Li, X., Ma, Y., and Huang, F. Compressive sensing via nonlocal low-rank regularization. *IEEE transactions on image processing*, 23(8):3618–3632, 2014.

Gong, Q., Liu, X., Li, G., and Qin, Y. Multiple-image encryption and authentication with sparse representation by space multiplexing. *Applied optics*, 52(31):7486–7493, 2013.

Haeffele, B., Young, E., and Vidal, R. Structured low-rank matrix factorization: Optimality, algorithm, and applications to image processing. In *International conference on machine learning*, pp. 2007–2015. PMLR, 2014.

He, W., Zhang, H., Zhang, L., and Shen, H. Total-variation-regularized low-rank matrix factorization for hyperspectral image restoration. *IEEE transactions on geoscience and remote sensing*, 54(1): 178–188, 2015.

Laurent, B. and Massart, P. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pp. 1302–1338, 2000.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Qu, W., Xiu, X., Zhang, H., and Fan, J. An efficient semi-proximal admm algorithm for low-rank and sparse regularized matrix minimization problems with real-world applications. *Journal of Computational and Applied Mathematics*, 424:115007, 2023.

Shang, R., Liu, C., Meng, Y., Jiao, L., and Stolkin, R. Nonnegative matrix factorization with rank regularization and hard constraint. *Neural Computation*, 29(9):2553–2579, 2017.

Su, Y., Hong, D., Li, Y., and Jing, P. Low-rank regularized deep collaborative matrix factorization for micro-video multi-label classification. *IEEE Signal Processing Letters*, 27:740–744, 2020.

Yang, Y. and Yu, J. Fast proximal gradient descent for a class of non-convex and non-smooth sparse learning problems. In *Uncertainty in Artificial Intelligence*, pp. 1253–1262. PMLR, 2020.

Zhou, H. and Li, L. Regularized matrix regression. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 76(2):463, 2014.

A Proofs for Subsection 2.1

Lemma A.1. *If* $s \leq \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$*, then*

$$\mathrm{supp}(\sigma(\mathbf{X}^{t+1})) \subseteq \mathrm{supp}(\sigma(\mathbf{X}^t)), \textit{for } t \geq 0, \tag{11}$$

and

$$\mathrm{rank}(\mathbf{X}^{t+1}) \leq \mathrm{rank}(\mathbf{X}^t), \textit{for } t \geq 0, \tag{12}$$

which means the support of the singular value vectors of the sequence $\{\mathbf{X}^t\}_t$ *shrinks, also the rank of the sequence* $\{\mathbf{X}^t\}_t$ *decreases.*

Proof. Let $\bar{\mathbf{X}}^{t+1} = \mathbf{X}^t - s\nabla g(\mathbf{X}^t), \mathbf{Q}^t = -s\nabla g(\mathbf{X}^t)$, thus we have $\bar{\mathbf{X}}^{t+1} = \mathbf{X}^t + \mathbf{Q}^t$. With Weyl's inequality

$$\sigma_{i+j-1}(\mathbf{A}+\mathbf{B}) \leq \sigma_i(\mathbf{A}) + \sigma_j(\mathbf{B}), \tag{13}$$

we get

$$\sigma_i(\bar{\mathbf{X}}^{t+1}) \leq \sigma_i(\mathbf{X}^t) + \sigma_1(\mathbf{Q}^t) = \sigma_1(\mathbf{Q}^t) \leq sG, \text{for all } i \text{ where } \sigma_i(\mathbf{X}^t) = 0. \tag{14}$$

With $s \leq \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$, we have $\sigma_i(\bar{\mathbf{X}}^{t+1}) \leq \sqrt{2\lambda s}$, therefore $\sigma_i(\mathbf{X}^{t+1}) = 0$. So the zero elements of $\sigma(\mathbf{X}^t)$ remain unchanged in $\sigma(\mathbf{X}^{t+1})$, and $\mathrm{supp}(\sigma(\mathbf{X}^{t+1})) \subseteq \mathrm{supp}(\sigma(\mathbf{X}^t))$, $\mathrm{rank}(\mathbf{X}^{t+1}) \leq \mathrm{rank}(\mathbf{X}^t)$. □

Lemma A.2. *The sequence of the objective* $\{F(\mathbf{X}^t)\}_t$ *is nonincreasing, and the following inequality holds for all* $t \geq 0$*:*

$$F(\mathbf{X}^{t+1}) \leq F(\mathbf{X}^t) - (\frac{1}{2s} - \frac{L}{2})\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2. \tag{15}$$

Proof. Let $\bar{\mathbf{X}}^{t+1} = \mathbf{X}^t - s\nabla g(\mathbf{X}^t), \mathbf{Q}^t = -s\nabla g(\mathbf{X}^t)$, thus we have $\bar{\mathbf{X}}^{t+1} = \mathbf{X}^t + \mathbf{Q}^t$, and

$$\mathbf{X}^{t+1} = \operatorname*{argmin}_{\mathbf{V}} \frac{1}{2s}\|\mathbf{V} - \bar{\mathbf{X}}^{t+1}\|_F^2 + h(\mathbf{V}). \tag{16}$$

Let $\mathbf{V} = \mathbf{X}^t$, we get

$$\langle \nabla g(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{X}^t \rangle + \frac{1}{2s}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 + h(\mathbf{X}^{t+1}) \leq h(\mathbf{X}^t). \tag{17}$$

In addition, we have

$$g(\mathbf{X}^{t+1}) \leq g(\mathbf{X}^t) + \langle \nabla g(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{X}^t \rangle + \frac{L}{2}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2. \tag{18}$$

Combine Eq. (17) and Eq. (18) together, we get

$$F(\mathbf{X}^{t+1}) \leq F(\mathbf{X}^t) - (\frac{1}{2s} - \frac{L}{2})\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2, \tag{19}$$

since $s \leq \frac{1}{L}$, we have $\frac{1}{2s} \geq \frac{L}{2}$, so the sequence $\{F(\mathbf{X}^t)\}_t$ is decreasing with lower bound 0. □

Lemma A.3. *Laurent & Massart (2000) Let* $Y_1, Y_2, \ldots, Y_D$ *be i.i.d. Gaussian random variables with 0 mean and unit variance, and* $a_1, a_2, \ldots, a_D$ *be D positive numbers. Define* $Z = \sum a_i(Y_i^2 - 1)$ *and* $\mathbf{a} = [a_1, a_2, \ldots, a_D]^T$*, then for any* $t > 0$*,*

$$\textit{Probability}(Z \geq 2\|\mathbf{a}\|_2\sqrt{t} + 2\|\mathbf{a}\|_\infty t) \leq e^{-t}. \tag{20}$$

Lemma A.4. *Davidson & Szarek (2001) Suppose* $\mathbf{A} \in \mathbb{R}^{m\times n}(m \geq n)$ *is a random matrix whose entries are i.i.d. sampled from the standard Gaussian distribution N*$(0, \frac{1}{m})$*, then*

$$1 - \sqrt{\frac{n}{m}} \leq E(\sigma_n(\mathbf{A})) \leq E(\sigma_1(\mathbf{A})) \leq 1 + \sqrt{\frac{n}{m}}. \tag{21}$$

And for any $t > 0$*,*

$$\textit{Probability}(\sigma_n(\mathbf{A}) \leq 1 - \sqrt{\frac{n}{m}} - t) < e^{-\frac{mt^2}{2}}, \tag{22}$$

$$\textit{Probability}(\sigma_1(\mathbf{A}) \geq 1 + \sqrt{\frac{n}{m}} + t) < e^{-\frac{mt^2}{2}}. \tag{23}$$

Theorem A.5. *Suppose $\mathbf{D} \in \mathbb{R}^{d\times n}(n \geq d)$ is a random matrix with elements i.i.d. sampled from the standard Gaussian distribution N(0,1), then*

$$\textit{Probability}(\frac{1}{L} \leq \frac{2\lambda}{G^2}) \geq 1 - e^{-\frac{a^2}{2}} - ne^{-a}, \tag{24}$$

if

$$n \geq (\sqrt{d} + a + \sqrt{\frac{(d + 2\sqrt{da} + 2a)(x_0 + \lambda|S|))}{\lambda}})^2, \tag{25}$$

where $x_0 = \|\mathbf{Y} - \mathbf{D}\mathbf{X}^0\|_F^2$, $S = \text{supp}(\sigma(\mathbf{X}^0))$, and a can be chosen as $a_0 logn$ for $a_0 > 0$ to ensure that Eq. (25) holds with high probability.

Proof. Based on Lemma A.4, for any $a > 0$, with probability $\geq 1 - e^{-\frac{a^2}{2}}$,

$$\sigma_{max}(\mathbf{D}) > \sqrt{n} - \sqrt{d} - a, \tag{26}$$

and by Lemma A.3, for any $1 \leq i \leq n$ and $a > 0$, with probability $\geq 1 - e^{-a}$,

$$\|\mathbf{D}_i\|_2 \leq \sqrt{d + 2\sqrt{da} + 2a}, \tag{27}$$

where $\mathbf{D}_i$ denotes i-th column of $\mathbf{D}$. Then, it can be verified with the union bound that with probability $\geq 1 - e^{-\frac{a^2}{2}} - ne^{-a}$,

$$\frac{2D^2(x_0 + \lambda|S|)}{\lambda} \leq 2\sigma_{max}^2(\mathbf{D}), \tag{28}$$

where $D = \max_i \|i_{th}$ column of $\mathbf{D}\|_2$, if

$$n \geq (\sqrt{d} + a + \sqrt{\frac{(d + 2\sqrt{da} + 2a)(x_0 + \lambda|S|))}{\lambda}})^2. \tag{29}$$

□

Lemma A.6. *(a) All the elements of each subsequence $\mathcal{X}^k$ $(k = 1, \ldots, K)$ in the subsequences with shrinking support have the same support. In addition, for any $1 \leq k_1 < k_2 \leq K$ and any $\mathbf{X}^{t_1} \in \mathcal{X}^{k_1}$, and $\mathbf{X}^{t_2} \in \mathcal{X}^{k_2}$, we have $t_1 < t_2$ and $\text{supp}(\sigma(\mathbf{X}^{t_2})) \subset \text{supp}(\sigma(\mathbf{X}^{t_1}))$.*

(b) All the subsequences except for the last one, $\mathcal{X}^k$ $(k = 1, \ldots, K-1)$ have finite size, and $\mathcal{X}^K$ have an infinite number of elements, and there exists some $t_0 \geq 0$ such that $\{\mathbf{X}^t\}_{t=t_0}^{\infty} \subseteq \mathcal{X}^K$.

Proof. (a) For any $1 \leq k < K$, let $\mathbf{X}^{t_1}, \mathbf{X}^{t_2} \in \mathcal{X}^k$ and $t_1 \neq t_2$. If $t_1 < t_2$, then $\text{supp}(\sigma(\mathbf{X}^{t_2})) \subseteq \text{supp}(\sigma(\mathbf{X}^{t_1}))$ according to the support shrinkage property in Lemma A.1. If $\text{supp}(\sigma(\mathbf{X}^{t_2})) \subset \text{supp}(\sigma(\mathbf{X}^{t_1}))$ then $|\text{supp}(\sigma(\mathbf{X}^{t_2}))| < |\text{supp}(\sigma(\mathbf{X}^{t_1}))|$, which contradicts with the definition of $\mathcal{X}^k$ whose elements have the same support size. A similar argument holds for $t_2 < t_1$. Therefore, all the elements of each subsequence $\mathcal{X}^k (1 \leq k \leq K)$ have the same support.

For any $1 \leq k_1 \leq k_2 \leq K$ and any $\mathbf{X}^{t_1} \in \mathcal{X}^{k_1}$ and $\mathbf{X}^{t_2} \in \mathcal{X}^{k_2}$, note that $t_1 \neq t_2$ and $\text{supp}(\sigma(\mathbf{X}^{t_1})) \neq \text{supp}(\sigma(\mathbf{X}^{t_2}))$ since $\mathcal{X}^{k_1}$ and $\mathcal{X}^{k_2}$ have different support size. Suppose $t_1 > t_2$, we have $\text{supp}(\sigma(\mathbf{X}^{t_1})) \subset \text{supp}(\sigma(\mathbf{X}^{t_2}))$ and it follows that $|\text{supp}(\sigma(\mathbf{X}^{t_1}))| < |\text{supp}(\sigma(\mathbf{X}^{t_2}))|$, again it contradicts with the Definition 2.4. Thus, we must have $t_1 < t_2$, and it follows that $\text{supp}(\sigma(\mathbf{X}^{t_2})) \subset \text{supp}(\sigma(\mathbf{X}^{t_1}))$.

(b) Suppose $\mathcal{X}^k$ is an infinite sequence for some $1 \leq k \leq K-1$. We can get an infinite sequence from $\mathcal{X}^k$ as follows:

We have some $\mathbf{X}^{t_0} \in \mathcal{X}^k$ for some $t_0 > 0$ since $\mathcal{X}^k$ is not empty. Suppose we get $\{\mathbf{X}^{t'_j}\}_{j'=0}^{j}$ in the first $j \geq 0$ steps with increasing indices $\{t'_j\}$. Since $\mathcal{X}^k$ is an infinite sequence, $\mathcal{X}^k \backslash \{\mathbf{X}^{t'_j}\}_{j'=0}^{j}$ is still an infinite sequence. At the $(j+1)$-th step, we can find $\mathbf{X}^{t_{j+1}} \in \mathcal{X}^k \backslash \{\mathbf{X}^{t'_j}\}_{j'=0}^{j}$ with $t_{j+1} > t_j$. Therefore, we are able to get an infinite sequence $\{\mathbf{X}^{t_j}\}_{j'=0}^{\infty} \subseteq \mathcal{X}^k$ with increasing indices $\{t_j\}$. With the fact that the indices $\{t_j\}$ is increasing, we can see that $\lim_{j\to\infty} t_j = \infty$.

For any element $\mathbf{X}^q \in \mathcal{X}^{k+1}$, there must exist some $j > 0$ such that $q \leq t_j$, according to the support shrinkage property we must have $\mathrm{supp}(\sigma(\mathbf{X}^{t_j})) \subseteq \mathrm{supp}(\sigma(\mathbf{X}^q))$, and $|\mathrm{supp}(\sigma(\mathbf{X}^{t_j}))| \leq |\mathrm{supp}(\sigma(\mathbf{X}^q))|$. On the other hand, since $\mathbf{X}^{t_j} \in \mathcal{X}^k$, we have $|\mathrm{supp}(\sigma(\mathbf{X}^q))| < |\mathrm{supp}(\sigma(\mathbf{X}^{t_j}))|$. This contradiction shows that each $\mathcal{X}^k (1 \leq k \leq K-1)$ must have a finite size. Also, $\{\mathbf{X}^t\}_{t=0}^{\infty}$ is an infinite sequence and $\{\mathcal{X}^k\}_{k=1}^{K}$ form a disjoint cover of it, thus $\mathcal{X}^K$ must contain infinite number of elements.

According to the proof of (a), there exists an infinite sequence $\{\mathbf{X}^{t_j}\}_{j=0}^{\infty} \subseteq \mathcal{X}^K$, and $\lim_{j\to\infty} t_j = \infty$. For any $t > t_0$, there must be some t'_j with $j' \geq 1$ such that $t_{j'-1} \leq t \leq t_{j'}$. Then we have

$$\mathrm{supp}(\sigma(\mathbf{X}^{t_{j'}})) = S^* \subseteq \mathrm{supp}(\sigma(\mathbf{X}^t)) \subseteq \mathrm{supp}(\sigma(\mathbf{X}^{t_{j'-1}})) = S^*, \tag{30}$$

therefore we have $|\mathrm{supp}(\sigma(\mathbf{X}^t))| = |S^*|$ and $\mathbf{X}^t \in \mathcal{X}^K$ for any $t \geq t_0$, which is $\{\mathbf{X}^t\}_{t=t_0}^{\infty} \subseteq \mathcal{X}^K$. □

Theorem A.7. *Suppose $s \leq \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$. and $\mathbf{X}^*$ is a limit point of $\{\mathbf{X}^t\}_{t=0}^{\infty}$, and $\sigma(\mathbf{X}^*)$ is a limit point of $\{\sigma(\mathbf{X}^t)\}_{t=0}^{\infty}$, then the sequence $\{\mathbf{X}^t\}_{t=0}^{\infty}$ generated by Algorithm 1 converges to $\mathbf{X}^*$, $\mathbf{X}^*$ is a critical point of F(·), and $\mathrm{supp}(\sigma(\mathbf{X}^*)) = S^*$, where S^* is the support of any element in $\mathcal{X}^K$. Moreover, there exists $t_0 \geq 0$ such that for all $m \geq t_0$, we have*

$$F(\mathbf{X}^{m+1}) - F(\mathbf{X}^*) \leq \frac{1}{2s(m-t_0+1)} \|\mathbf{X}^{t_0} - \mathbf{X}^*\|_F^2. \tag{31}$$

Proof. Let S^* denote the support of any element in $\mathcal{X}^K$. First we have $\mathrm{supp}(\sigma(\mathbf{X}^*)) \subseteq S^*$, otherwise, pick an arbitrary $i \in \mathrm{supp}(\sigma(\mathbf{X}^*)) \backslash S^*$, then $\|\sigma(\mathbf{X}^{t_j}) - \sigma(\mathbf{X}^*)\|_2 \geq |\sigma_i(\mathbf{X}^*)|$ contradicts with the fact that $\sigma(\mathbf{X}^{t_j}) \to \sigma(\mathbf{X}^*)$.

Moreover, suppose $\mathrm{supp}(\sigma(\mathbf{X}^*)) \subset S^*$, we can pick an arbitrary $i \in S^* \backslash \mathrm{supp}(\sigma(\mathbf{X}^*))$. And it can be shown that $\sigma_i(\mathbf{X}^{t_j}) \to 0$. Otherwise there exists $\epsilon > 0$, for any j, there exists $j' \geq j$ such that $|\sigma_i(\mathbf{X}^{t_{j'}})| \geq \epsilon$. It follows that $\|\sigma(\mathbf{X}^{t_{j'}}) - \sigma(\mathbf{X}^*)\|_2 \geq |\sigma_i(\mathbf{X}^{t_{j'}}| \geq \epsilon$, contradicting with the fact that $\sigma(\mathbf{X}^{t_j}) \to \sigma(\mathbf{X}^*)$.

Let $\epsilon > 0$ be a sufficiently small positive number such that $sG + \epsilon < \sqrt{2\lambda s}$. Since $\sigma_i(\mathbf{X}^{t_j}) \to 0$, there exists sufficiently large j such that $|\sigma_i(\mathbf{X}^{t_j})| < \epsilon$. Let $\bar{\mathbf{X}}^{t_j+1} = \mathbf{X}^{t_j} - s\nabla g(\mathbf{X}^{t_j})$, then

$$|\sigma_i(\bar{\mathbf{X}}^{t_j+1})| \leq |\sigma_i(\mathbf{X}^{t_j})| + sG < \epsilon + sG \leq \sqrt{2\lambda s}. \tag{32}$$

Then according to the update rule we have $\sigma_i(\mathbf{X}^{t_j+1}) = 0$, so $\mathrm{supp}(\sigma(\mathbf{X}^{t_j+1})) \subseteq \mathrm{supp}(\sigma(\mathbf{X}^{t_j})) \backslash \{i\}$. On the other hand, $\mathbf{X}^{t_j+1} \in \mathcal{X}^k$, so we have $\mathrm{supp}(\sigma(\mathbf{X}^{t_j+1})) = \mathrm{supp}(\sigma(\mathbf{X}^{t_j}))$. Such contradict shows that $\mathrm{supp}(\sigma(\mathbf{X}^*)) \subset S^*$ cannot be true. So $\mathrm{supp}(\sigma(\mathbf{X}^*)) = S^*$.

Now we will show that $\{\mathbf{X}^t\}_{t=t_0}^{\infty}$ converges to $\mathbf{X}^*$.

For any $\mathbf{V}, \mathbf{U}$, we have

$$g(\mathbf{V}) \leq g(\mathbf{U}) + \langle \nabla g(\mathbf{U}), \mathbf{V} - \mathbf{U} \rangle + \frac{L}{2}\|\mathbf{V} - \mathbf{U}\|_F^2, \tag{33}$$

also since $g(\cdot)$ is convex, for any $\mathbf{V}$ and $t \geq 0$:

$$g(\mathbf{X}^{t+1}) + \langle \nabla g(\mathbf{X}^{t+1}), \mathbf{V} - \mathbf{X}^{t+1} \rangle \leq g(\mathbf{V}). \tag{34}$$

Also, since

$$\mathbf{X}^{t+1} = \operatorname*{argmin}_{\mathbf{V}} \frac{1}{2s}\|\mathbf{V} - (\mathbf{X}^t - s\nabla g(\mathbf{X}^t))\|_F^2 + h(\mathbf{V}), \tag{35}$$

we have

$$-\nabla g(\mathbf{X}^t) - \frac{1}{s}(\mathbf{X}^{t+1} - \mathbf{X}^t) \in \partial h(\mathbf{X}^{t+1}), \tag{36}$$

and

$$\begin{aligned} &\frac{1}{s}(\mathbf{X}^{t+1} - (\mathbf{X}^t - s\nabla g(\mathbf{X}^t))) + \partial h(\mathbf{X}^{t+1}) = 0, \\ &\frac{1}{s}(\sum_{i \in \mathrm{supp}(\sigma(\mathbf{X}^{t+1}))} \sigma_i \mathbf{u}_i \mathbf{v}_i^T - \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^T) + \partial h(\mathbf{X}^{t+1}) = 0, \end{aligned} \tag{37}$$

where $\mathbf{X}^t - s\nabla g(\mathbf{X}^t) = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ is the singular value decomposition, then it follows

$$\partial h(\mathbf{X}^{t+1}) = \frac{1}{s} \sum_{i \notin \mathrm{supp}(\sigma(\mathbf{X}^{t+1}))} \sigma_i \mathbf{u}_i \mathbf{v}_i^T, \tag{38}$$

therefore

$$\langle \partial h(\mathbf{X}^{t+1}), \mathbf{X}^{t+1} \rangle = 0. \tag{39}$$

For any matrix $\mathbf{V}$ such that $\mathrm{supp}(\sigma(\mathbf{V})) = \mathrm{supp}(\sigma(\mathbf{X}^{t+1}))$, we have $h(\mathbf{V}) = h(\mathbf{X}^{t+1}) + \langle \partial h(\mathbf{X}^{t+1}), \mathbf{X}^{t+1} \rangle$.

For $t \geq t_0$, we have

$$\begin{aligned}
F(\mathbf{X}^{t+1}) &\leq g(\mathbf{X}^t) + \langle \nabla g(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{X}^t \rangle + \frac{L}{2}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 + h(\mathbf{X}^{t+1}) \\
&\leq g(\mathbf{V}) + \langle \nabla g(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{V} \rangle + \langle \nabla g(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{X}^t \rangle + \frac{L}{2}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 + h(\mathbf{X}^{t+1}) \\
&= g(\mathbf{V}) + \langle \nabla g(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{V} \rangle + \frac{L}{2}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 + h(\mathbf{X}^{t+1}) \\
&= g(\mathbf{V}) + \langle \nabla g(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{V} \rangle + \frac{L}{2}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 + h(\mathbf{V}) + \langle \nabla g(\mathbf{X}^t) + \frac{1}{s}(\mathbf{X}^{t+1} - \mathbf{X}^t), \mathbf{V} - \mathbf{X}^{t+1} \rangle \\
&= F(\mathbf{V}) + \frac{1}{s}\langle \mathbf{X}^{t+1} - \mathbf{X}^t, \mathbf{V} - \mathbf{X}^{t+1} \rangle + \frac{L}{2}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 \\
&= F(\mathbf{V}) + \frac{1}{s}\langle \mathbf{X}^{t+1} - \mathbf{X}^t, \mathbf{V} - \mathbf{X}^t \rangle - \frac{1}{s}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 + \frac{L}{2}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 \\
&= F(\mathbf{V}) + \frac{1}{s}\langle \mathbf{X}^{t+1} - \mathbf{X}^t, \mathbf{V} - \mathbf{X}^t \rangle - (\frac{1}{s} - \frac{L}{2})\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 \\
&\leq F(\mathbf{V}) + \frac{1}{s}\langle \mathbf{X}^{t+1} - \mathbf{X}^t, \mathbf{V} - \mathbf{X}^t \rangle - \frac{1}{2s}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2.
\end{aligned} \tag{40}$$

Now suppose $\mathrm{supp}(\sigma(\mathbf{X}^*)) = \mathrm{supp}(\sigma(\mathbf{X}^{t+1})) = S^*$, let $\mathbf{V} = \mathbf{X}^*$, we have

$$F(\mathbf{X}^{t+1}) - F(\mathbf{X}^*) \leq \frac{1}{s}\langle \mathbf{X}^{t+1} - \mathbf{X}^t, \mathbf{X}^* - \mathbf{X}^t \rangle - \frac{1}{2s}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 = \frac{1}{2s}(\|\mathbf{X}^t - \mathbf{X}^*\|_F^2 - \|\mathbf{X}^{t+1} - \mathbf{X}^*\|_F^2). \tag{41}$$

Now, sum the above equation over $t = t_0, \ldots, m$ with $m \geq t_0$, we get

$$\sum_{t=t_0}^{m} F(\mathbf{X}^{t+1}) - F(\mathbf{X}^*) \leq \sum_{t=t_0}^{m} \frac{1}{2s}(\|\mathbf{X}^t - \mathbf{X}^*\|_F^2 - \|\mathbf{X}^{t+1} - \mathbf{X}^*\|_F^2) = \frac{1}{2s}(\|\mathbf{X}^{t_0} - \mathbf{X}^*\|_F^2 - \|\mathbf{X}^{m+1} - \mathbf{X}^*\|_F^2). \tag{42}$$

Since $\{F(\mathbf{X}^t)\}_t$ is non-increasing, $\sum_{t=t_0}^{m} F(\mathbf{X}^{t+1}) - F(\mathbf{X}^*) > (m - t_0 + 1)F(\mathbf{X}^{m+1}) - F(\mathbf{X}^*)$, therefore,

$$F(\mathbf{X}^{m+1}) - F(\mathbf{X}^*) \leq \frac{1}{2s(m - t_0 + 1)}(\|\mathbf{X}^{t_0} - \mathbf{X}^*\|_F^2 - \|\mathbf{X}^{m+1} - \mathbf{X}^*\|_F^2) \leq \frac{1}{2s(m - t_0 + 1)}(\|\mathbf{X}^{t_0} - \mathbf{X}^*\|_F^2). \tag{43}$$

Again, since $\mathbf{X}^{t+1} = \mathrm{argmin}_{\mathbf{V}} \langle \nabla g(\mathbf{X}^t), \mathbf{V} - \mathbf{X}^t \rangle + \frac{1}{2s}\|\mathbf{V} - \mathbf{X}^t\|_F^2 + h(\mathbf{V})$, then

$$\langle \nabla g(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{X}^t \rangle + \frac{1}{2s}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 + h(\mathbf{X}^{t+1}) \leq \langle \nabla g(\mathbf{X}^t), \mathbf{X}^t - \mathbf{X}^t \rangle + \frac{1}{2s}\|\mathbf{X}^t - \mathbf{X}^t\|_F^2 + h(\mathbf{X}^t) = h(\mathbf{X}^t). \tag{44}$$

Therefore,

$$\begin{aligned}
F(\mathbf{X}^{t+1}) &\leq g(\mathbf{V}) + \langle \nabla g(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{V} \rangle + \frac{L}{2}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 + h(\mathbf{X}^{t+1}) \\
&\leq g(\mathbf{V}) + \langle \nabla g(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{V} \rangle + \frac{L}{2}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 + h(\mathbf{X}^t) - \langle \nabla g(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{X}^t \rangle - \frac{1}{2s}\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2.
\end{aligned} \tag{45}$$

Let $\mathbf{V} = \mathbf{X}^t$, we get $F(\mathbf{X}^{t+1}) \leq F(\mathbf{X}^t) - (\frac{1}{2s} - \frac{L}{2})\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2$. Thus, we have

$$(\frac{1}{2s} - \frac{L}{2}) \sum_{t=0}^{\infty} \|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 \leq F(\mathbf{X}^0) - F(\mathbf{X}^*) < \infty, \tag{46}$$

then

$$\|\mathbf{X}^{t+1} - \mathbf{X}^t\|_F^2 \to 0, \text{ as } t \to \infty. \tag{47}$$

Now we show $\mathbf{X}^*$ is a critical point of $F(\cdot)$. For $t_j \geq 1$, we have

$$\nabla g(\mathbf{X}^{t_j}) - \nabla g(\mathbf{X}^{t_j-1}) - \frac{1}{s}(\mathbf{X}^{t_j} - \mathbf{X}^{t_j-1}) \in \partial F(\mathbf{X}^{t_j}), \tag{48}$$

when $j \to \infty$ we have

$$\begin{aligned} \|\partial F(\mathbf{X}^{t_j})\|_F &= \|\nabla g(\mathbf{X}^{t_j}) - \nabla g(\mathbf{X}^{t_j-1}) - \frac{1}{s}(\mathbf{X}^{t_j} - \mathbf{X}^{t_j-1})\|_F \\ &\leq L\|\mathbf{X}^{t_j} - \mathbf{X}^{t_j-1}\|_F + \frac{1}{s}\|\mathbf{X}^{t_j} - \mathbf{X}^{t_j-1}\|_F \to 0. \end{aligned} \tag{49}$$

Also, when $j \to \infty$,

$$F(\mathbf{X}^{t_j}) = g(\mathbf{X}^{t_j}) + h(\mathbf{X}^{t_j}) = g(\mathbf{X}^{t_j}) + \lambda|S^*| \to g(\mathbf{X}^*) + \lambda|S^*| = F(\mathbf{X}^*). \tag{50}$$

Therefore, $0 \in \partial F(\mathbf{X}^*)$ and $\mathbf{X}^*$ is a critical point. □

B Proofs for Subsection 2.2

Lemma B.1. *The sequence $\{\mathbf{X}^t\}_{t=1}^{\infty}$ generated by Algorithm 2 satisfies*

$$\text{supp}(\sigma(\mathbf{X}^{t+1})) \subseteq \text{supp}(\sigma(\mathbf{X}^t)), t \geq 1. \tag{51}$$

Proof. We will prove the above lemma using mathematical induction.

When $t = 1$, we have $\mathbf{U}^1 = \mathbf{X}^1, \mathbf{V}^1 = \mathbf{X}^1, \mathbf{X}^2 = T_{\sqrt{2\lambda s}}(\mathbf{X}^1 - s\nabla g(\mathbf{X}^1))$, using the argument in the proof of Lemma A.1, we have

$$\text{supp}(\sigma(\mathbf{X}^2)) \subseteq \text{supp}(\sigma(\mathbf{X}^1)). \tag{52}$$

Suppose $\text{supp}(\sigma(\mathbf{X}^{t+1})) \subseteq \text{supp}(\sigma(\mathbf{X}^t))$ holds for all $t \leq t'$ with $t' \geq 1$, now consider the case that $t = t' + 1$. Based on the update rule for $\mathbf{V}^t$, we have

$$\text{supp}(\sigma(\mathbf{V}^{t'+1})) \subseteq \text{supp}(\sigma(\mathbf{X}^{t'+1})). \tag{53}$$

Let $\bar{\mathbf{X}}^{t'+2} = \mathbf{V}^{t'+1} - s\nabla g(\mathbf{V}^{t'+1})$, then $\sigma_i(\mathbf{X}^{t'+2}) = 0$ for any $i \notin \text{supp}(\sigma(\mathbf{V}^{t'+1}))$ since $\sigma_i(\bar{\mathbf{X}}^{t'+2}) \leq \sqrt{2\lambda s}$ for such i. So the zero elements in $\sigma(\mathbf{V}^{t'+1})$ remain unchanged in $\sigma(\mathbf{X}^{t'+2})$, and it follows that

$$\text{supp}(\sigma(\mathbf{X}^{t'+2})) \subseteq \text{supp}(\sigma(\mathbf{V}^{t'+1})) \subseteq \text{supp}(\sigma(\mathbf{X}^{t'+1})), \tag{54}$$

therefore $\text{supp}(\sigma(\mathbf{X}^{t+1})) \subseteq \text{supp}(\sigma(\mathbf{X}^t))$ holds for $t = t' + 1$. Based on mathematical induction it holds for all $t \geq 1$. □

Theorem B.2. *Suppose $s \leq \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$. and $\mathbf{X}^*$ is a limit point of $\{\mathbf{X}^t\}_{t=0}^{\infty}$ generated by Algorithm 2, then there exists $t_0 \geq 1$ such that for all $m \geq t_0$, we have*

$$F(\mathbf{X}^{m+1}) - F(\mathbf{X}^*) \leq \frac{4}{(m+1)^2} V^{t_0}, \tag{55}$$

where V^{t_0} is a value defined as

$$V^{t_0} = \frac{1}{2s}\|(\alpha^{t_0-1} - 1)\mathbf{X}^{t_0-1} - \alpha^{t_0-1}\mathbf{X}^{t_0} + \mathbf{X}^*\|_F^2 + (\alpha^{t_0-1})^2(F(\mathbf{X}^{t_0}) - F(\mathbf{X}^*)). \tag{56}$$

Proof. According to Lemma B.1, there exists $t_0 \geq 0$ such that $\{\mathbf{X}^t\}_{t=t_0}^{\infty} \subseteq \mathfrak{X}^K$. It follows that $\text{supp}(\sigma(\mathbf{X}^*)) = S^*$.

When $\text{supp}(\sigma(\mathbf{W})) = \text{supp}(\sigma(\mathbf{X}^{t+1}))$ for $t \geq t_0$, with the similar process in Eq. (40), we get

$$F(\mathbf{X}^{t+1}) \leq F(\mathbf{W}) + \frac{1}{s}\langle \mathbf{X}^{t+1} - \mathbf{V}^t, \mathbf{W} - \mathbf{V}^t\rangle - (\frac{1}{s} - \frac{L}{2})\|\mathbf{X}^{t+1} - \mathbf{V}^t\|_F^2, \tag{57}$$

Let $\mathbf{W} = \mathbf{X}^t$ and $\mathbf{W} = \mathbf{X}^*$, we get

$$F(\mathbf{X}^{t+1}) \leq F(\mathbf{X}^t) + \frac{1}{s}\langle \mathbf{X}^{t+1} - \mathbf{V}^t, \mathbf{X}^t - \mathbf{V}^t\rangle - (\frac{1}{s} - \frac{L}{2})\|\mathbf{X}^{t+1} - \mathbf{V}^t\|_F^2, \tag{58}$$

and

$$F(\mathbf{X}^{t+1}) \leq F(\mathbf{X}^*) + \frac{1}{s}\langle \mathbf{X}^{t+1} - \mathbf{V}^t, \mathbf{X}^* - \mathbf{V}^t\rangle - (\frac{1}{s} - \frac{L}{2})\|\mathbf{X}^{t+1} - \mathbf{V}^t\|_F^2, \tag{59}$$

$(\alpha^t - 1)\times$Eq. (58) + Eq. (59), we obtain

$$\begin{aligned}&\alpha^t F(\mathbf{X}^{t+1}) - (\alpha^t - 1)F(\mathbf{X}^t) - F(\mathbf{X}^*)\\ \leq&\frac{1}{s}\langle \mathbf{X}^{t+1} - \mathbf{V}^t, (\alpha^t - 1)(\mathbf{X}^t - \mathbf{V}^t) + \mathbf{X}^* - \mathbf{V}^t\rangle - \alpha^t(\frac{1}{s} - \frac{L}{2})\|\mathbf{X}^{t+1} - \mathbf{V}^t\|_F^2.\end{aligned} \tag{60}$$

Multiply both sides of Eq. (60) by α^t, and use the fact that $(\alpha^t)^2 - \alpha^t = (\alpha^{t-1})^2$, we have

$$\begin{aligned}&(\alpha^t)^2(F(\mathbf{X}^{t+1}) - F(\mathbf{X}^*)) - (\alpha^{t-1})^2(F(\mathbf{X}^t) - F(\mathbf{X}^*))\\ \leq&\frac{1}{2s}(\|(\alpha^t - 1)\mathbf{X}^t - \alpha^t\mathbf{V}^t + \mathbf{X}^*\|_F^2 - \|(\alpha^t - 1)\mathbf{X}^t - \alpha^t\mathbf{X}^{t+1} + \mathbf{X}^*\|_F^2).\end{aligned} \tag{61}$$

Since for any matrix $\mathbf{A}, \mathbf{B}, \mathbf{C}$, when $\mathrm{supp}(\sigma(\mathbf{A})) \subseteq \mathrm{supp}(\sigma(\mathbf{C}))$, we have $\|\mathbf{A} - P_{\mathrm{supp}(\sigma(\mathbf{C}))}(\mathbf{B})\|_F \leq \|\mathbf{A} - \mathbf{B}\|_F$, and $\mathbf{V}^t = P_{\mathrm{supp}(\sigma(\mathbf{X}^t))}(\mathbf{U}^t)$, it follows that

$$\begin{aligned}&(\alpha^t)^2(F(\mathbf{X}^{t+1}) - F(\mathbf{X}^*)) - (\alpha^{t-1})^2(F(\mathbf{X}^t) - F(\mathbf{X}^*))\\ \leq&\frac{1}{2s}(\|(\alpha^t - 1)\mathbf{X}^t - \alpha^t\mathbf{U}^t + \mathbf{X}^*\|_F^2 - \|(\alpha^t - 1)\mathbf{X}^t - \alpha^t\mathbf{X}^{t+1} + \mathbf{X}^*\|_F^2).\end{aligned} \tag{62}$$

For simplicity, we define $\mathbf{A}^{t+1} = (\alpha^t - 1)\mathbf{X}^t - \alpha^t\mathbf{X}^{t+1} + \mathbf{X}^*$, $\mathbf{A}^t = (\alpha^{t-1} - 1)\mathbf{X}^{t-1} - \alpha^{t-1}\mathbf{X}^t + \mathbf{X}^*$, according to the update rule for $\mathbf{U}^t$, we can get $\mathbf{A}^t = (\alpha^t - 1)\mathbf{X}^t - \alpha^t\mathbf{U}^t + \mathbf{X}^*$, then based on Eq. (62), we obtain

$$(\alpha^t)^2(F(\mathbf{X}^{t+1}) - F(\mathbf{X}^*)) - (\alpha^{t-1})^2(F(\mathbf{X}^t) - F(\mathbf{X}^*)) \leq \frac{1}{2s}(\|\mathbf{A}^t\|_F^2 - \|\mathbf{A}^{t+1}\|_F^2). \tag{63}$$

Sum Eq. (63) over $t = t_0, \ldots, m$ for $m \geq t_0$, we have

$$(\alpha^m)^2(F(\mathbf{X}^{m+1}) - F(\mathbf{X}^*)) - (\alpha^{t_0-1})^2(F(\mathbf{X}^{t_0}) - F(\mathbf{X}^*)) \leq \frac{1}{2s}(\|\mathbf{A}^{t_0}\|_F^2 - \|\mathbf{A}^{m+1}\|_F^2) \leq \frac{1}{2s}\|\mathbf{A}^{t_0}\|_F^2, \tag{64}$$

therefore, with $\alpha^t \geq \frac{t+1}{2}$, we get

$$\begin{aligned}F(\mathbf{X}^{m+1}) - F(\mathbf{X}^*) \leq&\frac{1}{2s(\alpha^m)^2}\|\mathbf{A}^{t_0}\|_F^2 + \frac{(\alpha^{t_0-1})^2}{(\alpha^m)^2}(F(\mathbf{X}^{t_0}) - F(\mathbf{X}^*))\\ \leq&\frac{4}{(m+1)^2}(\frac{1}{2s}\|\mathbf{A}^{t_0}\|_F^2 + (\alpha^{t_0-1})^2(F(\mathbf{X}^{t_0}) - F(\mathbf{X}^*))),\end{aligned} \tag{65}$$

where $\mathbf{A}^{t_0} = (\alpha^{t_0-1} - 1)\mathbf{X}^{t_0-1} - \alpha^{t_0-1}\mathbf{X}^{t_0} + \mathbf{X}^*$. □

C Proofs for Subsection 2.3

Lemma C.1. *The sequence $\{\mathbf{Z}^t\}_{t=1}^\infty$ and $\{\mathbf{X}^t\}_{t=1}^\infty$ generated by Algorithm 3 satisfies*

$$\mathrm{supp}(\sigma(\mathbf{Z}^{t+1})) \subseteq \mathrm{supp}(\sigma(\mathbf{Z}^t)), \mathrm{supp}(\sigma(\mathbf{X}^{t+1})) \subseteq \mathrm{supp}(\sigma(\mathbf{X}^t)), t \geq 1. \tag{66}$$

Proof. We will prove the above lemma using mathematical induction.

It can be easily verified that $\mathrm{supp}(\sigma(\mathbf{Z}^2)) \subseteq \mathrm{supp}(\sigma(\mathbf{Z}^1))$.

Suppose $\mathrm{supp}(\sigma(\mathbf{Z}^{t+1})) \subseteq \mathrm{supp}(\sigma(\mathbf{Z}^t))$ holds for all $t \leq t'$ with $t' \geq 1$, now consider the case that $t = t' + 1$. With the similar thought process in the proof for Lemma B.1, based on the update rule for $\mathbf{W}^t$, the zero elements in $\sigma(\mathbf{V}^{t'+1})$ remain unchanged in $\sigma(\mathbf{Z}^{t'+2})$, thus we have $\mathrm{supp}(\sigma(\mathbf{Z}^{t'+2})) \subseteq \mathrm{supp}(\sigma(\mathbf{V}^{t'+1})) \subseteq \mathrm{supp}(\sigma(\mathbf{Z}^{t'+1}))$.

Therefore, $\text{supp}(\sigma(\mathbf{Z}^{t+1})) \subseteq \text{supp}(\sigma(\mathbf{Z}^t))$ holds for all $t \geq 1$.

We already show that for all $t \geq 1$, $\text{supp}(\sigma(\mathbf{X}^t)) = \text{supp}(\sigma(\mathbf{Z}^{\bar{t}}))$ for some $\bar{t} \leq t$. And based on the update rule for $\mathbf{X}$, we have $\mathbf{X}^{t+1} = \mathbf{Z}^{t+1}$ or $\mathbf{X}^{t+1} = \mathbf{X}^t$. If $\mathbf{X}^{t+1} = \mathbf{Z}^{t+1}$, $\text{supp}(\sigma(\mathbf{X}^{t+1})) = \text{supp}(\sigma(\mathbf{Z}^{t+1})) \subseteq \text{supp}(\sigma(\mathbf{Z}^{\bar{t}})) = \text{supp}(\sigma(\mathbf{X}^t))$ since $\bar{t} \leq t < t+1$. If $\mathbf{X}^{t+1} = \mathbf{X}^t$, it's easy to see $\text{supp}(\sigma(\mathbf{X}^{t+1})) = \text{supp}(\sigma(\mathbf{X}^t))$. Therefore, $\text{supp}(\sigma(\mathbf{X}^{t+1})) \subseteq \text{supp}(\sigma(\mathbf{X}^t))$ holds for all $t \geq 1$.

□

Theorem C.2. *Suppose $s \leq \min\{\frac{2\lambda}{G^2}, \frac{1}{L}\}$. and $\mathbf{X}^*$ is a limit point of $\{\mathbf{X}^t\}_{t=0}^{\infty}$ generated by Algorithm 3, then there exists $t_0 \geq 1$ such that for all $m \geq t_0$, we have*

$$F(\mathbf{X}^{m+1}) - F(\mathbf{X}^*) \leq \frac{4}{(m+1)^2} W^{t_0}, \tag{67}$$

where W^{t_0} is a value defined as

$$W^{t_0} = \frac{1}{2s}\|(\alpha^{t_0-1}-1)\mathbf{X}^{t_0-1} - \alpha^{t_0-1}\mathbf{Z}^{t_0} + \mathbf{X}^*\|_F^2 + (\alpha^{t_0-1})^2(F(\mathbf{X}^{t_0}) - F(\mathbf{X}^*)). \tag{68}$$

Proof. Based on Lemma C.1, $\{\mathbf{X}^t\}_{t=0}^{\infty}$ forms at most $K_1 \leq |S|+1$ subsequences with shrinking support $\{\mathcal{X}^k\}_{k=1}^{K_1}$, and $\{\mathbf{Z}^t\}_{t=0}^{\infty}$ forms at most $K_2 \leq |S|+1$ subsequences with shrinking support $\{\mathcal{Z}^k\}_{k=1}^{K_2}$. Based on Lemma A.6, there exists $t_1 \geq 0$ such that $\{\mathbf{X}^t\}_{t=t_1}^{\infty} \subseteq \mathcal{X}^{K_1}$, and there exists $t_2 \geq 0$ such that $\{\mathbf{Z}^t\}_{t=t_2}^{\infty} \subseteq \mathcal{Z}^{K_2}$. Let all the elements of $\sigma(\mathcal{X}^{K_1})$ have support S_1, let all the elements of $\sigma(\mathcal{Z}^{K_2})$ have support S_2, we show that $S_1 = S_2$: let $t_0 = \max\{t_1, t_2\}$, then there exists $t' \geq t_0$ such that $\mathbf{X}^{t'} = \mathbf{Z}^{t'}$, and due to the fact that $\{\mathbf{X}^t\}_{t=t_1}^{\infty} \subseteq \mathcal{X}^{K_1}$ and $\{\mathbf{Z}^t\}_{t=t_2}^{\infty} \subseteq \mathcal{Z}^{K_2}$, we have $S_1 = \text{supp}(\sigma(\mathbf{X}^{t'})) = \text{supp}(\sigma(\mathbf{Z}^{t'})) = S_2$.

Let $S_1 = S_2 = S^*$, then the singular value vectors of all the elements of $\{\mathbf{X}^t\}_{t=t_0}^{\infty}$ and $\{\mathbf{Z}^t\}_{t=t_0}^{\infty}$ have the same support S^*, and $\text{supp}(\sigma(\mathbf{X}^*)) = S^*$.

Following the same process in the proof for Theorem B.2, we get

$$F(\mathbf{Z}^{t+1}) \leq F(\mathbf{X}^t) + \frac{1}{s}\langle \mathbf{Z}^{t+1} - \mathbf{V}^t, \mathbf{X}^t - \mathbf{V}^t\rangle - (\frac{1}{s} - \frac{L}{2})\|\mathbf{Z}^{t+1} - \mathbf{V}^t\|_F^2, \tag{69}$$

and

$$F(\mathbf{X}^{t+1}) \leq F(\mathbf{X}^*) + \frac{1}{s}\langle \mathbf{Z}^{t+1} - \mathbf{V}^t, \mathbf{X}^* - \mathbf{V}^t\rangle - (\frac{1}{s} - \frac{L}{2})\|\mathbf{Z}^{t+1} - \mathbf{V}^t\|_F^2, \tag{70}$$

and $(\alpha^t - 1)\times$Eq. (69) + Eq. (70), multiply both sides by α^t, and use the fact that $(\alpha^t)^2 - \alpha^t = (\alpha^{t-1})^2$, we have

$$\begin{aligned}&(\alpha^t)^2(F(\mathbf{X}^{t+1}) - F(\mathbf{X}^*)) - (\alpha^{t-1})^2(F(\mathbf{X}^t) - F(\mathbf{X}^*))\\ \leq&\frac{1}{2s}(\|(\alpha^t-1)\mathbf{X}^t - \alpha^t\mathbf{V}^t + \mathbf{X}^*\|_F^2 - \|(\alpha^t-1)\mathbf{X}^t - \alpha^t\mathbf{Z}^{t+1} + \mathbf{X}^*\|_F^2).\end{aligned} \tag{71}$$

Based on the update rule for $\mathbf{V}^t$, we have

$$\begin{aligned}&(\alpha^t)^2(F(\mathbf{Z}^{t+1}) - F(\mathbf{X}^*)) - (\alpha^{t-1})^2(F(\mathbf{X}^t) - F(\mathbf{X}^*))\\ \leq&\frac{1}{2s}(\|(\alpha^t-1)\mathbf{X}^t - \alpha^t\mathbf{U}^t + \mathbf{X}^*\|_F^2 - \|(\alpha^t-1)\mathbf{X}^t - \alpha^t\mathbf{Z}^{t+1} + \mathbf{X}^*\|_F^2).\end{aligned} \tag{72}$$

Define $\mathbf{A}^{t+1} = (\alpha^t - 1)\mathbf{X}^t - \alpha^t\mathbf{Z}^{t+1} + \mathbf{X}^*$, $\mathbf{A}^t = (\alpha^{t-1}-1)\mathbf{X}^{t-1} - \alpha^{t-1}\mathbf{Z}^t + \mathbf{X}^*$, we can get $\mathbf{A}^t = (\alpha^t-1)\mathbf{X}^t - \alpha^t\mathbf{U}^t + \mathbf{X}^*$, therefore,

$$(\alpha^t)^2(F(\mathbf{Z}^{t+1}) - F(\mathbf{X}^*)) - (\alpha^{t-1})^2(F(\mathbf{X}^t) - F(\mathbf{X}^*)) \leq \frac{1}{2s}(\|\mathbf{A}^t\|_F^2 - \|\mathbf{A}^{t+1}\|_F^2). \tag{73}$$

Sum Eq. (73) over $t = t_0, \ldots, m$ for $m \geq t_0$, we have

$$(\alpha^m)^2(F(\mathbf{Z}^{m+1}) - F(\mathbf{X}^*)) - (\alpha^{t_0-1})^2(F(\mathbf{X}^{t_0}) - F(\mathbf{X}^*)) \leq \frac{1}{2s}(\|\mathbf{A}^{t_0}\|_F^2 - \|\mathbf{A}^{m+1}\|_F^2) \leq \frac{1}{2s}\|\mathbf{A}^{t_0}\|_F^2, \tag{74}$$

therefore, with $\alpha^t \geq \frac{t+1}{2}$, we get

$$F(\mathbf{Z}^{m+1}) - F(\mathbf{X}^*) \leq \frac{4}{(m+1)^2}(\frac{1}{2s}\|\mathbf{A}^{t_0}\|_F^2 + (\alpha^{t_0-1})^2(F(\mathbf{X}^{t_0}) - F(\mathbf{X}^*))), \tag{75}$$

where $\mathbf{A}^{t_0} = (\alpha^{t_0-1}-1)\mathbf{X}^{t_0-1} - \alpha^{t_0-1}\mathbf{Z}^{t_0} + \mathbf{X}^*$. □