
A Globally Optimal Portfolio for m -Sparse Sharpe Ratio Maximization

Yizun Lin¹ Zhao-Rong Lai^{1*} Cheng Li¹

¹Department of Mathematics
College of Information Science and Technology
Jinan University, Guangzhou, China
{linyizun, laizhr}@jnu.edu.cn
licheng@stu2020.jnu.edu.cn

Abstract

The Sharpe ratio is an important and widely-used risk-adjusted return in financial engineering. In modern portfolio management, one may require an m -sparse (no more than m active assets) portfolio to save managerial and financial costs. However, few existing methods can optimize the Sharpe ratio with the m -sparse constraint, due to the nonconvexity and the complexity of this constraint. We propose to convert the m -sparse fractional optimization problem into an equivalent m -sparse quadratic programming problem. The semi-algebraic property of the resulting objective function allows us to exploit the Kurdyka-Łojasiewicz property to develop an efficient Proximal Gradient Algorithm (PGA) that leads to a portfolio which achieves the globally optimal m -sparse Sharpe ratio under certain conditions. The convergence rates of PGA are also provided. To the best of our knowledge, this is the first proposal that achieves a globally optimal m -sparse Sharpe ratio with a theoretically-sound guarantee.

1 Introduction

The Sharpe ratio (SR) [33] is an important and widely-used performance metric in finance. Suppose an investing strategy is represented by a portfolio $w \in \mathbb{R}^N$ of N assets from a financial market. $\mu \in \mathbb{R}^N$ and $\Sigma \in \mathbb{R}^{N \times N}$ denote the expected return vector (in excess of the risk-free rate) and its covariance matrix for the N assets, respectively. It can be seen that $w^\top \mu$ and $\sqrt{w^\top \Sigma w}$ represent the expected return and its standard deviation (i.e., risk) for the portfolio w . The original definition of SR is given as the follow quotient between return and risk:

$$S_0(w) = \frac{w^\top \mu}{\sqrt{w^\top \Sigma w}}. \quad (1.1)$$

Ever since the proposal of SR, how to maximize it becomes an attractive research topic. Ordinary portfolio optimization methods based on either the mean-variance approach [5, 10] or the exponential growth rate approach [22, 24] can reduce the portfolio risk and increase the portfolio return to some extent [23], and hence improve the SR. On the other hand, direct SR optimization methods are also proposed. Hung et al. [18], Yu and Xu [35] consider the SR as a differentiable function of the portfolio, which can be solved via the augmented Lagrangian method. Pang [29] converts the SR maximization under the self-financing and long-only constraints into a linear complementarity problem, which can be solved via the Parametric Linear Complementarity Technique (PLCT) and the principle pivoting algorithm [12]. Note that PLCT requires $\mu_i > 0$ for at least some asset i in order to be feasible.

*Correspondence to: Zhao-Rong Lai

In modern portfolio management, it is widely-recognized that the number of selected assets should be restricted to a manageable size, in order to keep simplicity and save time and financial costs. Managerial strategies provide an approach to achieve this objective, such as the revenue driven resource allocation [11], the endowment model [15], and selling stocks after market crashes [27]. However, the managerial approaches still require intensive administration and abundant experience in management and finance. Hence researchers turn to sparsity models for solutions via the computational approaches. In [10], a Sparse and Stable Markowitz Portfolio (SSMP) is proposed by imposing ℓ_1 -regularization on the portfolio. Ao et al. [2] further propose a mean-variance model with an ℓ_1 constraint based on a maximum-Sharpe-ratio estimation. In [24], the exponential growth rate (EGR) criterion [1, 13, 19, 23] is exploited to develop a Short-term Sparse Portfolio Optimization (SSPO). Furthermore, a Short-term Sparse Portfolio Optimization with ℓ_0 -regularization (SSPO- ℓ_0) is developed in [28]. In [25], a nonlinear shrinkage of the covariance matrix is proposed to obtain an appropriate size of free parameters. Motivated by this strategy, Lai et al. [22] characterize a sparse structure for covariance estimation to construct a portfolio via the machine learning approach.

The ℓ_1 -regularization and the ℓ_1 constraint cannot control the exact number of selected assets. One has to tune the sparsity parameter to roughly adjust this number. On the other hand, suppose we want to select no more than m active assets out of N assets to construct a portfolio, then this can be exactly represented by the m -sparse (or ℓ_0) constraint $\|\mathbf{w}\|_0 \leq m$, where the ℓ_0 norm $\|\cdot\|_0$ denotes the number of nonzero components of a vector. Although many sparsity models are established for the Markowitz portfolio, few existing methods can optimize the SR (1.1) with the ℓ_0 constraint, due to the nonconvexity and the complexity of this constraint. In addition to the ℓ_0 constraint, other realistic constraints should also be imposed to ensure feasibility. For example, the self-financing constraint represents full re-investment and no external loans; the long-only constraint represents no short position. If all these constraints are imposed, the whole model becomes even much more difficult to solve.

To overcome these difficulties, we observe that this optimization is essentially an m -sparse fractional optimization that can be transformed into an equivalent m -sparse quadratic programming. Then the resulting objective function is semi-algebraic, so that the Kurdyka-Łojasiewicz (KL) property [3] can be exploited to develop an efficient Proximal Gradient Algorithm (PGA) [30]. It converges to a portfolio which achieves the globally optimal m -sparse SR under certain conditions. To the best of our knowledge, this is the first proposal that achieves a globally optimal m -sparse SR with a theoretically-sound guarantee. Our main contributions can be summarized as follows.

- 1) We propose to directly maximize the SR with the ℓ_0 constraint, the self-financing constraint and the long-only constraint on the portfolio. This model aims to obtain a feasible and realistic portfolio that optimizes the SR with exact sparsity.
- 2) SR maximization is essentially a fractional optimization. We convert this m -sparse fractional optimization problem into an equivalent m -sparse quadratic programming problem, which reduces the difficulty of solving it.
- 3) We observe that the resulting objective function is semi-algebraic, thus exploit the KL property to develop an efficient PGA that leads to a globally or at least a locally optimal solution of the m -sparse SR maximization model. The convergence rates of PGA are also provided.

Besides the above contributions, our approach also has several advantages: (i) It can be extended to a wide range of optimization problems with semi-algebraic objective functions and constraints. (ii) The actual sparsity is robust to the choice of m . (iii) It needs very little parameter tuning. (iv) It does not require any external algorithms or commercial optimizers.

2 Related Works and Existing Problems

There are some existing works that indirectly or directly optimize the SR to some extent via the computational approach. We introduce some examples and then analyze some unsolved problems.

2.1 Ordinary Portfolio Optimization

An intuitive approach is to directly optimize the portfolio, so that the expected return is maximized and/or the risk is minimized. These methods can be categorized into the mean-variance approach and the exponential growth rate approach [23]. Let $\mathbf{R} \in \mathbb{R}^{T \times N}$ be the sample asset return matrix with T trading times and N assets, and $\mathbf{1}_n$ denotes the vector of n ones. Brodie et al. [10] propose

to impose ℓ_1 -regularization on the mean-variance model, forming a Sparse and Stable Markowitz Portfolio (SSMP):

$$\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^N} \left\{ \frac{1}{T} \|\mathbf{R}\mathbf{w} - \rho \mathbf{1}_T\|_2^2 + \tau \|\mathbf{w}\|_1 \right\}, \quad \text{s.t.} \quad \mathbf{w}^\top \hat{\boldsymbol{\mu}} = \rho, \quad \mathbf{w}^\top \mathbf{1}_N = 1, \quad (2.1)$$

where $\hat{\boldsymbol{\mu}} := \frac{1}{T} \mathbf{R}^\top \mathbf{1}_T$ denotes the column vector of sample mean returns, and $\tau \geq 0$ is the regularization parameter. $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the ℓ_2 -norm and the ℓ_1 -norm, respectively. The quadratic form $\frac{1}{T} \|\mathbf{R}\mathbf{w} - \rho \mathbf{1}_T\|_2^2$ actually computes the mean squared error between the sample portfolio return $\mathbf{r}^{(t)} \mathbf{w}$ ($\mathbf{r}^{(t)}$ is the t -th row of \mathbf{R}) and the given return level ρ . Equations $\mathbf{w}^\top \hat{\boldsymbol{\mu}} = \rho$ and $\mathbf{w}^\top \mathbf{1}_N = 1$ are the expected return constraint and the self-financing constraint, respectively. Model (2.1) can be approximately solved by a surrogate model [10] and the Least Absolute Shrinkage and Selection Operator (Lasso) [34]. Goto and Xu [17] also exploit the Lasso to solve a mean-variance model through sparse hedging restrictions.

Ao et al. [2] propose a maximum-Sharpe-ratio estimated and sparse regression (MAXER) to approach mean-variance efficiency. Assume there are sufficient observations $T > (N + 2)$. MAXER first computes the maximum-Sharpe-ratio estimated regression response \hat{r}_c as follows:

$$\hat{\theta}_s = \hat{\boldsymbol{\mu}}^\top \hat{\boldsymbol{\Sigma}}^{-1} \hat{\boldsymbol{\mu}}, \quad \hat{\theta} := \frac{(T - N - 2)\hat{\theta}_s - N}{T}, \quad \hat{r}_c := \sigma \frac{1 + \hat{\theta}}{\sqrt{\hat{\theta}}}, \quad (2.2)$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ denote the sample mean and the sample covariance, respectively, σ is the risk constraint parameter. Then it adopts the Lasso to obtain the portfolio:

$$\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^N} \frac{1}{T} \|\mathbf{R}\mathbf{w} - \hat{r}_c \mathbf{1}_T\|_2^2, \quad \text{s.t.} \quad \|\mathbf{w}\|_1 \leq \tau. \quad (2.3)$$

Instead of the mean-variance approach, our method takes an essentially different objective that directly maximizes the SR in (1.1). Besides, our method does not require $T > (N + 2)$.

Based on the exponential growth rate criterion [1, 13, 19, 23], Lai et al. [24] propose to minimize a kind of negative potential return $\mathbf{w}^\top \boldsymbol{\varphi}$ with ℓ_1 -regularization but without any risk term, forming a Short-term Sparse Portfolio Optimization (SSPO) model

$$\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \mathbb{R}^N} \{ \mathbf{w}^\top \boldsymbol{\varphi} + \tau \|\mathbf{w}\|_1 \}, \quad \text{s.t.} \quad \mathbf{w}^\top \mathbf{1}_N = 1. \quad (2.4)$$

It develops an unconstrained augmented Lagrangian with the existence of a saddle point that can be solved by the alternating direction method of multipliers (ADMM). Luo et al. [28] further propose the SSPO- ℓ_0 model

$$\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w} \in \Delta} \{ \mathbf{w}^\top \boldsymbol{\varphi} + \tau \|\mathbf{w}\|_0 \}, \quad \Delta := \{ \mathbf{w} \in \mathbb{R}^N \mid \mathbf{w} \geq \mathbf{0}_N \text{ and } \mathbf{w}^\top \mathbf{1}_N = 1 \}, \quad (2.5)$$

where Δ is the N -dimensional simplex. This simplex constraint $\mathbf{w} \in \Delta$ is the combination of the long-only and the self-financing constraints. Under this constraint, the ℓ_0 -regularization problem (2.5) has a closed-form solution $\tilde{\mathbb{I}}_{\boldsymbol{\varphi}}^{\min} := \{ i \in \mathbb{N}_N \mid \varphi_i \leq \min_{j \in \mathbb{N}_N} \varphi_j + \epsilon \}$ with a tolerance $\epsilon \geq 0$, where $\mathbb{N}_N := \{1, 2, \dots, N\}$.

On the other hand, Lai et al. [22] propose a rank-one covariance estimator based on the operator space decomposition, in order to capture the rapidly-changing risk structure in the financial market:

$$\mathbf{X} = \mathbf{V}_J \Xi \mathbf{U}_J^\top, \quad \mathbf{D} = \Xi^2 - \frac{1}{T} \Xi \mathbf{V}_J^\top \mathbf{1}_T \mathbf{1}_T^\top \mathbf{V}_J \Xi, \quad \zeta_1^* = \left(\frac{\operatorname{tr}(\mathbf{D})}{N(T-1)} \right)^{-\frac{1}{2}} \theta_1, \quad \hat{\boldsymbol{\Sigma}}_{RO} := \mathbf{u}_1 \zeta_1^* \mathbf{u}_1^\top,$$

where $\mathbf{X} = \mathbf{R} + \mathbf{1}_{T \times N}$ denotes the price relative matrix and $\mathbf{V}_J \Xi \mathbf{U}_J^\top$ is its singular value decomposition (SVD), θ_1 and \mathbf{u}_1 are the largest eigenvalue and its eigenvector, respectively.

Although the above portfolio optimization methods may partly improve the SR, they may not be competitive to direct SR optimization. Hence direct SR optimization methods should still be developed and investigated.

2.2 Sharpe Ratio Optimization

Pang [29] proposes to optimize the following SR model:

$$\max_{\mathbf{w} \in \mathbb{R}^N} S_0(\mathbf{w}), \quad \text{s.t. } \mathbf{w} \in \Delta, \mathbf{C}\mathbf{w} \leq \mathbf{d}, \quad (2.6)$$

where $S_0(\mathbf{w})$ is defined by (1.1), $\mathbf{C} \in \mathbb{R}^{l \times N}$ and $\mathbf{d} \in \mathbb{R}^l$ form a linear constraint for \mathbf{w} . It can be transformed into the following equivalent parametric linear complementarity problem:

$$\begin{cases} \mathbf{u} = -\boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{w} + (\mathbf{C}^\top - \mathbf{1}_N \mathbf{d}^\top) \mathbf{y} \geq \mathbf{0}_N, & \mathbf{w} \geq \mathbf{0}_N, \\ \mathbf{v} = -(\mathbf{C} - \mathbf{d} \mathbf{1}_N^\top) \mathbf{w} \geq \mathbf{0}_l, & \mathbf{y} \geq \mathbf{0}_l, \\ \mathbf{u}^\top \mathbf{w} = \mathbf{v}^\top \mathbf{y} = 0. \end{cases} \quad (2.7)$$

Problem (2.7) can be efficiently solved by the principle pivoting algorithm [12], but it requires $\mu_i > 0$ for at least some asset i in order to be feasible. Moreover, if we aim to construct an m -sparse \mathbf{w} , this approach becomes invalid. In Section 3, we will convert the m -sparse SR optimization into an equivalent m -sparse quadratic programming, and the latter is still a nonconvex optimization. We further elaborate a proximal gradient algorithm to obtain a globally or locally optimal SR.

Another viable approach is to consider the SR as a function of the portfolio and directly optimize it under some realistic constraints. Hung et al. [18] propose the following IPSRM-D model to optimize the SR:

$$\max_{\mathbf{w} \in \Delta} \left\{ \mathcal{S}(\mathbf{w}) := \frac{\mathbf{w}^\top \boldsymbol{\mu} + \kappa_1 \mathbf{w}^\top \mathbf{U} \mathbf{w}}{\mathbf{w}^\top \mathbf{D} \mathbf{w}} + \kappa_2 \mathbf{w}^\top (\mathbf{1}_N - \mathbf{w}) \right\}, \quad (2.8)$$

where $\mathbf{U} \in \mathbb{R}^{N \times N}$ and $\mathbf{D} \in \mathbb{R}^{N \times N}$ are upside and downside risk matrices, respectively, $\mathbf{w}^\top (\mathbf{1}_N - \mathbf{w})$ is a diversification term. κ_1 and κ_2 are hyperparameters that control the strength of upside risk and diversification, respectively. Interested readers can further refer to [35] for some practical estimators for $\boldsymbol{\mu}$, \mathbf{U} and \mathbf{D} .

However, $\mathcal{S}(\mathbf{w})$ in model (2.8) is different from the original SR $S_0(\mathbf{w})$ (1.1) in several significant parts. First, $\mathcal{S}(\mathbf{w})$ uses second-order moments $\mathbf{w}^\top \mathbf{U} \mathbf{w}$ and $\mathbf{w}^\top \mathbf{D} \mathbf{w}$ as risk metrics, but $S_0(\mathbf{w})$ uses the first-order moment $\sqrt{\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}}$ instead. In general, a first-order moment is more appropriate because the expected return $\mathbf{w}^\top \boldsymbol{\mu}$ should remain in the same order of magnitude as $\sqrt{\mathbf{w}^\top \boldsymbol{\Sigma} \mathbf{w}}$. Second, the numerator of $S_0(\mathbf{w})$ does not contain any risk term, while the numerator of $\mathcal{S}(\mathbf{w})$ contains $\mathbf{w}^\top \mathbf{U} \mathbf{w}$. This may change the meaning of SR as an equilibrium point in the efficient frontier based on the CAPM theory [32]. These facts may affect the performance of SR optimization.

Another problem is the lack of effective solving algorithms that could really maximize the SR under constraints. A conventional way is to adopt gradient methods, since $\mathcal{S}(\mathbf{w})$ is a differentiable function when $\mathbf{w} \neq \mathbf{0}_N$. Hung et al. [18], Yu and Xu [35] propose to adopt the augmented Lagrangian method to optimize (2.8). Though they do not specify which form of Lagrangian models is used, we give the following one without loss of generality:

$$\mathcal{L}(\mathbf{w}, \boldsymbol{\lambda}) := \mathcal{S}(\mathbf{w}) + \frac{\varrho}{2} (\mathbf{w}^\top \mathbf{1}_N - 1)^2 + \boldsymbol{\lambda}^\top \mathbf{w}, \quad (2.9)$$

where $\frac{\varrho}{2} (\mathbf{w}^\top \mathbf{1}_N - 1)^2$ with hyperparameter $\varrho \leq 0$ is a regularization term for the self-financing constraint, and $\boldsymbol{\lambda} \in \mathbb{R}_+^N$ is the dual variable with respect to (w.r.t.) \mathbf{w} for the long-only constraint $\mathbf{w} \geq \mathbf{0}_N$. \mathbb{R}_+^N denotes the set of all N -dimensional nonnegative vectors. The update scheme is

$$\begin{cases} \mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \eta_1 \nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}^{(k)}, \boldsymbol{\lambda}^{(k)}), \\ \boldsymbol{\lambda}^{(k+1)} = \boldsymbol{\lambda}^{(k)} - \eta_2 \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\mathbf{w}^{(k+1)}, \boldsymbol{\lambda}^{(k)}), \end{cases} \quad (2.10)$$

where $\eta_1, \eta_2 \geq 0$ are update step sizes. Note that \mathcal{S} is a nonconvex function w.r.t. \mathbf{w} , and the augmented Lagrangian method is a surrogate method that approximates model (2.8). Hence (2.10) does not necessarily lead to the maximum SR. Worse still, due to the augmented term $\frac{\varrho}{2} (\mathbf{w}^\top \mathbf{1}_N - 1)^2$, (2.10) may not even decrease the objective function \mathcal{S} . Moreover, the Lagrangian $\mathcal{L}(\mathbf{w}^{(k)}, \boldsymbol{\lambda}^{(k)})$ is increased by the $\mathbf{w}^{(k)}$ updates but decreased by the $\boldsymbol{\lambda}^{(k)}$ updates, hence (2.10) cannot guarantee convergence to a point $(\mathbf{w}^*, \boldsymbol{\lambda}^*)$ without a thorough investigation of the update scheme. To summarize, the augmented Lagrangian method and most existing gradient methods cannot guarantee global or local optimality for model (2.8).

3 PGA for m -Sparse Sharpe Ratio Maximization

In this section, we formulate the maximization problem of SR as a nonconvex fractional optimization under constraints. Instead of directly solving the proposed model, we develop an efficient proximal gradient algorithm to solve a simpler surrogate model (subtraction form) that is equivalent to the original constrained fractional optimization model.

3.1 m -Sparse Sharpe Ratio Maximization Model

In order to retain the risk premium meaning of SR in finance, we directly maximize the original SR in (1.1) instead of a variant like (2.8). In the perspective of statistical estimation, suppose we have a sample asset return (in excess of the risk-free rate) matrix $\mathbf{R} \in \mathbb{R}^{T \times N}$ with T trading times and N assets. Then the original SR (1.1) can be represented by

$$S(\mathbf{w}) := \frac{\frac{1}{T} \mathbf{1}_T^\top \mathbf{R} \mathbf{w}}{\sqrt{\frac{1}{T-1} \|\mathbf{R} \mathbf{w} - (\frac{1}{T} \mathbf{1}_T^\top \mathbf{R} \mathbf{w}) \mathbf{1}_T\|_2^2 + \epsilon \|\mathbf{w}\|_2^2}}, \quad (3.1)$$

where $\epsilon \|\mathbf{w}\|_2^2$ is a regularization term for a positive definite \mathbf{Q}_ϵ defined in (3.2). The parameter ϵ can be an arbitrarily-small positive parameter, whose effect on the risk term can be negligible. To simplify the notation, we define

$$\mathbf{p} := \frac{1}{T} \mathbf{R}^\top \mathbf{1}_T, \quad \mathbf{Q} := \frac{1}{\sqrt{T-1}} \left(\mathbf{R} - \frac{1}{T} \mathbf{1}_T \mathbf{1}_T^\top \mathbf{R} \right) \quad \text{and} \quad \mathbf{Q}_\epsilon := \mathbf{Q}^\top \mathbf{Q} + \epsilon \mathbf{I}. \quad (3.2)$$

Then the maximization of SR under the m -sparse, long-only and self-financing constraints is given by

$$\max_{\substack{\mathbf{w} \in \Delta \\ \|\mathbf{w}\|_0 \leq m}} S(\mathbf{w}) := \frac{\mathbf{p}^\top \mathbf{w}}{\sqrt{\mathbf{w}^\top \mathbf{Q}_\epsilon \mathbf{w}}}, \quad (3.3)$$

where the simplex Δ is defined in (2.5). Note that minimizing $-S(\mathbf{w})$ under the constraint $\|\mathbf{w}\|_0 \leq m$ is essentially quite different from minimizing the ℓ_0 -regularization version $-S(\mathbf{w}) + \tau \|\mathbf{w}\|_0$ with some positive τ . In general, the latter is easier because it incorporates the ℓ_0 norm into the objective function and enlarges the feasible set by dropping the constraint $\|\mathbf{w}\|_0 \leq m$. We simply call (3.3) the m -Sparse Sharpe Ratio Maximization (mSSRM) model. In fact, to solve the mSSRM model, it suffices to solve the following simpler constrained minimization model

$$\min_{\substack{\mathbf{v} \geq \mathbf{0}_N \\ \|\mathbf{v}\|_0 \leq m}} \left\{ \frac{1}{2} \mathbf{v}^\top \mathbf{Q}_\epsilon \mathbf{v} - \mathbf{p}^\top \mathbf{v} \right\}. \quad (3.4)$$

To see this, we establish the relation between the solutions of these two models in the following theorem, whose proof is provided in Appendix A.1. We define the constraint sets in model (3.3) and (3.4) by

$$\Omega_1 := \{\mathbf{w} \in \Delta \mid \|\mathbf{w}\|_0 \leq m\} \quad \text{and} \quad \Omega := \{\mathbf{v} \in \mathbb{R}^N \mid \mathbf{v} \geq \mathbf{0}_N \text{ and } \|\mathbf{v}\|_0 \leq m\}, \quad (3.5)$$

respectively. It is obvious that $\Omega_1 \subsetneq \Omega$.

Theorem 1 *Suppose that there exists some $\tilde{\mathbf{w}} \in \Omega_1$ such that $\mathbf{p}^\top \tilde{\mathbf{w}} > 0$. If $\hat{\mathbf{v}}$ is an optimal solution of model (3.3), then $\frac{\mathbf{p}^\top \hat{\mathbf{v}}}{\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}} \hat{\mathbf{v}}$ is an optimal solution of model (3.4). Conversely, if $\hat{\mathbf{v}}$ is an optimal solution of model (3.4), then $\frac{\hat{\mathbf{v}}}{\hat{\mathbf{v}}^\top \mathbf{1}_N}$ is an optimal solution of model (3.3).*

Defining the indicator function ι_Ω by

$$\iota_\Omega(\mathbf{v}) := \begin{cases} 0, & \text{if } \mathbf{v} \in \Omega; \\ +\infty, & \text{otherwise,} \end{cases} \quad (3.6)$$

we can rewrite model (3.4) as the following two-term unconstrained minimization model:

$$\min_{\mathbf{v} \in \mathbb{R}^N} \{f(\mathbf{v}) + \iota_\Omega(\mathbf{v})\}, \quad \text{where } f(\mathbf{v}) := \frac{1}{2} \mathbf{v}^\top \mathbf{Q}_\epsilon \mathbf{v} - \mathbf{p}^\top \mathbf{v}. \quad (3.7)$$

We then turn to solving model (3.7) instead of the mSSRM model in (3.3).

3.2 Proximal Gradient Algorithm

To develop a proximal gradient algorithm for solving model (3.7), we recall the notion of proximity operator, and then establish the relation between the solution of model (3.7) and the proximity characterization (3.8) in Theorem 2. For a proper function $\psi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$, its proximity operator at $\mathbf{x} \in \mathbb{R}^n$ is defined by $\text{prox}_\psi(\mathbf{x}) := \text{argmin}_{\mathbf{u} \in \mathbb{R}^n} \left\{ \frac{1}{2} \|\mathbf{u} - \mathbf{x}\|_2^2 + \psi(\mathbf{u}) \right\}$. We remark that for function ψ that is nonconvex, $\text{prox}_\psi(\mathbf{x})$ may not be unique. Throughout this paper, the formula $\mathbf{h} = \text{prox}_\psi(\mathbf{x})$ represents $\mathbf{h} \in \text{prox}_\psi(\mathbf{x})$. For $\mathbf{v}^* \in \mathbb{R}^N$ and $\delta > 0$, we denote by $B(\mathbf{v}^*; \delta)$ the neighborhood of \mathbf{v}^* with radius δ . If there exists some $\delta > 0$ such that $f(\mathbf{v}^*) \leq f(\mathbf{v})$ holds for all $\mathbf{v} \in \Omega \cap B(\mathbf{v}^*; \delta)$, then we say that \mathbf{v}^* is a locally optimal solution of model (3.7).

Theorem 2 *Let function ι_Ω and f be defined by (3.6) and (3.7), respectively. If \mathbf{v}^* is a globally optimal solution of model (3.7), then for any $\alpha \in \left(0, \frac{1}{\|\mathbf{Q}_\varepsilon\|_2}\right]$,*

$$\mathbf{v}^* = \text{prox}_{\iota_\Omega}(\mathbf{v}^* - \alpha \nabla f(\mathbf{v}^*)). \quad (3.8)$$

Conversely, we have the following two statements:

- (i) *If $\alpha \geq \frac{1}{\varepsilon}$ and (3.8) holds, then \mathbf{v}^* is a globally optimal solution of model (3.7).*
- (ii) *For any $\alpha > 0$, if (3.8) holds, then \mathbf{v}^* is a locally optimal solution of model (3.7).*

The proof of Theorem 2 is provided in Appendix A.2. Based on this theorem, the Proximal Gradient Algorithm (PGA) for solving model (3.7) can be given by the following iterative scheme:

$$\mathbf{v}^{(k+1)} = \text{prox}_{\iota_\Omega}(\mathbf{v}^{(k)} - \alpha \nabla f(\mathbf{v}^{(k)})), \quad \text{where } k \in \mathbb{N}, \alpha > 0, \mathbf{v}^{(0)} \in \Omega. \quad (3.9)$$

We then compute the closed form of $\text{prox}_{\iota_\Omega}$. For a vector $\mathbf{v} \in \mathbb{R}^N$, we denote by $m_{\mathbf{v}}$ and $J_{\text{pos}}^{\mathbf{v}}$ the number of positive components and the index set of positive components of \mathbf{v} . If $m_{\mathbf{v}} \geq m$, then we denote by $J_{m\text{-pos}}^{\mathbf{v}}$ an index set of the m -largest positive components of \mathbf{v} . Specifically, by letting $\{v_{j_i}\}_{i \in \mathbb{N}_N}$ be an rearrangement of $\{v_j\}_{j \in \mathbb{N}_N}$ such that $v_{j_1} \geq v_{j_2} \geq \dots \geq v_{j_N}$, then $J_{m\text{-pos}}^{\mathbf{v}} := \{j_1, j_2, \dots, j_m\}$. Throughout this paper, for a given vector $\mathbf{v} \in \mathbb{R}^N$, we shall always compute $\text{prox}_{\iota_\Omega}(\mathbf{v})$ according to the following proposition. Its proof is given in Appendix A.3.

Proposition 3 *Let ι_Ω be defined by (3.6), $\mathbf{v} \in \mathbb{R}^N$, and define the index set $J^{\mathbf{v}}$ by*

$$J^{\mathbf{v}} = \begin{cases} J_{m\text{-pos}}^{\mathbf{v}}, & \text{if } m_{\mathbf{v}} > m; \\ J_{\text{pos}}^{\mathbf{v}}, & \text{if } m_{\mathbf{v}} \leq m. \end{cases}$$

Then the vector \mathbf{h} given by $h_j = \begin{cases} v_j, & \text{if } j \in J^{\mathbf{v}}; \\ 0, & \text{if } j \in \mathbb{N}_N \setminus J^{\mathbf{v}} \end{cases}$ satisfies that $\mathbf{h} \in \text{prox}_{\iota_\Omega}(\mathbf{v})$.

3.3 Convergence Analysis of PGA

In this subsection, we delve into the convergence analysis of PGA. We aim to demonstrate that PGA possesses the capability to converge to a globally optimal solution of model (3.4). The limit point obtained by PGA can also yield a globally optimal solution of the original model (3.3), under certain conditions. We also demonstrate the convergence rates of PGA.

Firstly, we introduce a proposition that illustrates the convergence and monotonic decreasing behavior of the objective function values for the iterative sequence, as well as the vanishing gap between consecutive iterates. The proof of this proposition is provided in Appendix A.4.

Proposition 4 *Let function ι_Ω and f be defined by (3.6) and (3.7), respectively, and let $F := f + \iota_\Omega$. If $\alpha \in \left(0, \frac{1}{\|\mathbf{Q}_\varepsilon\|_2}\right)$, then for arbitrary initial vector $\mathbf{v}^{(0)} \in \mathbb{R}^N$, the sequence $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ generated by PGA satisfies the following properties:*

- (i) $\mathbf{v}^{(k)} \in \Omega$, for all $k \in \mathbb{N}$;
- (ii) $F(\mathbf{v}^{(k+1)}) + a \|\mathbf{v}^{(k+1)} - \mathbf{v}^{(k)}\|_2^2 \leq F(\mathbf{v}^{(k)})$ for all $k \in \mathbb{N}$, where $a := \frac{1}{2} \left(\frac{1}{\alpha} - \|\mathbf{Q}_\varepsilon\|_2 \right) > 0$;
- (iii) $\lim_{k \rightarrow \infty} F(\mathbf{v}^{(k)})$ exists;
- (iv) $\lim_{k \rightarrow \infty} \|\mathbf{v}^{(k+1)} - \mathbf{v}^{(k)}\|_2 = 0$.

Though we have established the convergence of $\{F(\mathbf{v}^{(k)})\}_{k \in \mathbb{N}}$ and the vanishing gap between consecutive iterates, further efforts are necessary to rigorously confirm the convergence of the iterative sequence $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$. We demonstrate the convergence of $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ to a local minimizer of function F and the corresponding convergence rates in the following theorem. In order to maintain consistency with the original SR maximization model (as outlined in Theorem 1), we further define sequence $\{\mathbf{w}^{(k)}\}_{k \in \mathbb{N}}$ based on $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ and conduct an analysis of the convergence rate of $\{\mathbf{w}^{(k)}\}_{k \in \mathbb{N}}$. To prove the theorem, we need to introduce the notions of subdifferential, semi-algebraic function and Kurdyka-Łojasiewicz property, along with several technical lemmas. Detailed proofs and relevant content can be found in Appendix A.5.

Theorem 5 *Suppose that there exists some $\tilde{\mathbf{w}} \in \Omega$ such that $\mathbf{p}^\top \tilde{\mathbf{w}} > 0$. For arbitrary initial vector $\mathbf{v}^{(0)} \in \mathbb{R}^N$, let $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ be generated by PGA, and let $\{\mathbf{w}^{(k)}\}_{k \in \mathbb{N}}$ be defined by*

$$\mathbf{w}^{(k)} := \begin{cases} \frac{\mathbf{v}^{(k)}}{(\mathbf{v}^{(k)})^\top \mathbf{1}_N}, & \text{if } (\mathbf{v}^{(k)})^\top \mathbf{1}_N \neq 0; \\ \mathbf{0}_N, & \text{otherwise.} \end{cases}$$

If $\alpha \in \left(0, \frac{1}{\|\mathbf{Q}_\epsilon\|_2}\right)$, then the following statements hold:

- (i) $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ converge to a locally optimal solution \mathbf{v}^* of model (3.4) with convergence rates $\|\mathbf{v}^{(k)} - \mathbf{v}^*\|_2 = O(1/\sqrt{k})$ and $|f(\mathbf{v}^{(k)}) - f(\mathbf{v}^*)| = O(1/k)$.
- (ii) The limit point \mathbf{v}^* of $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ satisfies that $\mathbf{v}^* \geq \mathbf{0}_N$ and $\mathbf{v}^* \neq \mathbf{0}_N$.
- (iii) $\{\mathbf{w}^{(k)}\}_{k \in \mathbb{N}}$ converge to $\mathbf{w}^* := \frac{\mathbf{v}^*}{(\mathbf{v}^*)^\top \mathbf{1}_N}$ with convergence rates $\|\mathbf{w}^{(k)} - \mathbf{w}^*\|_2 = O(1/\sqrt{k})$ and $|S(\mathbf{w}^{(k)}) - S(\mathbf{w}^*)| = O(1/\sqrt{k})$, where S is defined in (3.3).

In the remainder of this section, we always let $\mathbf{v}^* \in \Omega$ be the locally optimal solution of model (3.7) that sequence $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ converges to. We recall that $m_{\mathbf{v}^*}$ and $J_{\text{pos}}^{\mathbf{v}^*}$ denote the number of positive components and the index set of positive components of \mathbf{v}^* , respectively. Suppose that there exists some $\tilde{\mathbf{w}} \in \Omega$ such that $\mathbf{p}^\top \tilde{\mathbf{w}} > 0$. Then item (ii) in Theorem 5 together with $\mathbf{v}^* \in \Omega$ yields that $1 \leq m_{\mathbf{v}^*} \leq m$. In fact, \mathbf{v}^* is also the globally optimal solution of the convex model

$$\min_{\mathbf{v} \in \hat{\Omega}} \left\{ \frac{1}{2} \mathbf{v}^\top \mathbf{Q}_\epsilon \mathbf{v} - \mathbf{p}^\top \mathbf{v} \right\}, \text{ where } \hat{\Omega} := \{ \mathbf{v} \in \mathbb{R}^N \mid \mathbf{v} \geq \mathbf{0}_N \text{ and } v_j = 0 \text{ for all } j \in \mathbb{N}_N \setminus J_{\text{pos}}^{\mathbf{v}^*} \}. \quad (3.10)$$

Certainly, $\hat{\Omega} \neq \emptyset$ due to the condition $m_{\mathbf{v}^*} \geq 1$. According to the definition of $\hat{\Omega}$, it is straightforward to observe that $\mathbf{v}^* \in \hat{\Omega}$. Furthermore, $\hat{\Omega}$ is a closed convex set and $\hat{\Omega} \subset \Omega$. To analyze the relation between \mathbf{v}^* and the original m -sparse Sharpe ratio maximization model (3.3), we define

$$\hat{\Omega}_1 := \{ \mathbf{v} \in \Delta \mid v_j = 0 \text{ for all } j \in \mathbb{N}_N \setminus J_{\text{pos}}^{\mathbf{v}^*} \}, \quad (3.11)$$

where Δ is given by (2.5). It is easy to see that $\hat{\Omega}_1 \subset \Omega_1$, where Ω_1 defined in (3.5) is the constraint set of model (3.3). We then have the following theorem, whose proof is provided in Appendix A.6.

Theorem 6 *Suppose that there exists some $\tilde{\mathbf{w}} \in \hat{\Omega}$ such that $\mathbf{p}^\top \tilde{\mathbf{w}} > 0$, where $\hat{\Omega}$ is defined in (3.10), and let $\mathbf{w}^* := \frac{\mathbf{v}^*}{(\mathbf{v}^*)^\top \mathbf{1}_N}$. Then the following statements hold:*

- (i) \mathbf{v}^* is the unique globally optimal solution of model (3.10).
- (ii) \mathbf{w}^* is a globally optimal solution of model $\max_{\mathbf{w} \in \hat{\Omega}_1} S(\mathbf{w})$.
- (iii) If $m_{\mathbf{v}^*} = m$, then \mathbf{w}^* is a locally optimal solution of model (3.3).

Item (iii) in Theorem 6 demonstrates that the limit point of the sequence obtained by PGA can yield a locally optimal solution of model (3.3). In fact, according to item (i) in Theorem 2, we have the following theorem that provides sufficient conditions for obtaining a globally optimal solution of model (3.3), whose proof is provided in Appendix A.7.

Theorem 7 *Suppose that there exists some $\tilde{\mathbf{w}} \in \Omega_1$ such that $\mathbf{p}^\top \tilde{\mathbf{w}} > 0$, and let $\mathbf{w}^* := \frac{\mathbf{v}^*}{(\mathbf{v}^*)^\top \mathbf{1}_N}$. If one of the following two conditions holds:*

- (i) $m_{\mathbf{v}^*} < m$;

(ii) $m_{\mathbf{v}^*} = m$ and $\nabla_i f(\mathbf{v}^*) > -\epsilon \cdot \min\{v_i^* | i \in \text{supp}(\mathbf{v}^*)\}$ for all $i \in \mathbb{N}_N \setminus \text{supp}(\mathbf{v}^*)$,

then \mathbf{w}^* is a globally optimal solution of model (3.3).

Combining Theorem 7 and item (iii) in Theorem 6, we see that the proposed method can obtain a globally optimal solution of model (3.3) when $m_{\mathbf{w}^*} < m$. Even if this condition does not hold, we can obtain at least a locally optimal solution. To test validation of PGA's global optimality, we conduct a set of simulation experiments, whose details are presented in Appendix A.8. The codes for the simulation experiments are accessible via the link: https://github.com/linyizun2024/mSSRM/tree/main/Codes_for_Simulation.

We call the existence of $\tilde{\mathbf{w}} \in \hat{\Omega}$ such that $\mathbf{p}^\top \tilde{\mathbf{w}} > 0$ in Theorem 6 the Existence of Positive Expected Return (EPER) condition. Although the EPER condition is required to guarantee the convergence of \mathbf{w}^* to a locally optimal solution of the original model (3.3), the proposed method is still of high practical significance in the case that the EPER condition does not hold. From the proofs of item (i) in Theorem 5 and item (i) in Theorem 6, we see that even if the EPER condition does not hold, the sequence generated by PGA still converges to a locally optimal solution \mathbf{v}^* of model (3.4), which is also the globally optimal solution of model (3.10). In these two models, the objective function $\frac{1}{2}\mathbf{v}^\top \mathbf{Q}_\epsilon \mathbf{v} - \mathbf{p}^\top \mathbf{v}$ (subtraction form) w.r.t. \mathbf{v} represents risk minus expected return, whose minimization gives smaller risk and less loss in revenue, even if the expected return is not positive. For the case that the expected return is not positive, compared with the failure of Sharpe ratio of fractional form, the globally or locally optimal solution of the subtraction form seems to have more realistic significance. We recall from item (ii) in Theorem 5 that \mathbf{v}^* may be equal to $\mathbf{0}_N$ if the EPER condition does not hold. In this case, we shall set $\mathbf{w}^* = \mathbf{0}_N$ and keep all the wealth in the risk-free asset to avoid loss in revenue. To close this section, we summarize the whole m -sparse Sharpe ratio maximization method, which we abbreviate to mSSRM-PGA, in Appendix A.9.

4 Experimental results

Extensive experiments with real-world financial data sets are conducted to evaluate the performance of the proposed mSSRM-PGA. Moreover, we also consider one baseline method: 1/N [14], as well as 9 state-of-the-art methods: IPSRM-D [18], PLCT [29], SSMP [10], MAXER [2], SSPO [24], SPOLC [22], S1, S2 and S3 [28], as competitors in the experiments. We use 6 real-world monthly benchmark data sets: FF25, FF25EU, FF32, FF49, FF100 and FF100MEINV to compare different methods. These data sets are collected from the baseline and commonly-used Kenneth R. French's Real-world Data Library². Details regarding these competitors and data sets are given in Appendix A.10. As for mSSRM-PGA, we examine three levels of sparsity $m = 10$, $m = 15$, $m = 20$ and set $\epsilon = 10^{-3}$. The setting of other parameters are presented in Appendix A.9. The codes of mSSRM-PGA are accessible via the link: https://github.com/linyizun2024/mSSRM/tree/main/Codes_for_Experiments_in_Paper.

4.1 Results for Sharpe ratios

We adopt the moving-window trading framework in [23] to imitate real-world portfolio management. For each method, we use the asset returns $\{\mathbf{r}_{(t)}\}_{t=1}^T$ or the price relatives $\{\mathbf{x}_{(t)} := \mathbf{r}_{(t)} + \mathbf{1}_N\}_{t=1}^T$ in the time window $t = [1 : T]$ to update the portfolio $\hat{\mathbf{w}}_{(T+1)}$ for the next trading time. On the $(T + 1)$ -th time, we compute the portfolio return by $\hat{r}_{(T+1), \hat{\mathbf{w}}} = \mathbf{x}_{(T+1)}^\top \hat{\mathbf{w}}_{(T+1)} - 1$ and then turn to the next round where the time window moves to $t = [2 : (T + 1)]$ and a new portfolio $\hat{\mathbf{w}}_{(T+2)}$ is computed. This procedure is repeated till the last trading time \mathcal{T} , which yields a return sequence $\{\hat{r}_{(t), \hat{\mathbf{w}}}\}_{t=1}^{\mathcal{T}}$. This sequence can be used to compute the test SR:

$$\widehat{SR} = \frac{(\sum_{t=T+1}^{\mathcal{T}} \hat{r}_{(t), \hat{\mathbf{w}}}) / (\mathcal{T} - T)}{\sqrt{(\sum_{s=T+1}^{\mathcal{T}} (\hat{r}_{(s), \hat{\mathbf{w}}} - (\sum_{t=T+1}^{\mathcal{T}} \hat{r}_{(t), \hat{\mathbf{w}}}) / (\mathcal{T} - T))^2) / (\mathcal{T} - T - 1)}}.$$

The 1/N strategy does not involve the time window size T . For all other methods, we examine two conventional settings for the time window size in the finance industry [2, 17]: $T = 60$ and $T = 120$.

²http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Table 1 shows the (monthly) SRs of the 11 compared methods. Because MAXER requires $T > (N + 2)$, it is unavailable on FF100 and FF100MEINV when $T = 60$. It is worth noting that the trivial strategy 1/N outperforms most competitors in most situations. The reason is that 1/N diversifies the risk over all the assets, which is also an effective risk control approach [14]. However, mSSRM-PGA outperforms all the competitors including 1/N on all the 6 data sets when $T = 60$ and on 5 data sets when $T = 120$. For example, its SR is more than 70% higher than that of 1/N on FF25EU whether $T = 60$ or $T = 120$. Hence mSSRM-PGA achieves competitive SRs with sparse portfolios, which saves much managerial cost.

Table 1: Sharpe ratios of different portfolio optimization methods on 6 benchmark data sets.

Strategy	T = 60						T = 120					
	FF25	FF25EU	FF32	FF49	FF100	FF100MEINV	FF25	FF25EU	FF32	FF49	FF100	FF100MEINV
1/N	0.2276	0.1574	0.2234	0.2057	0.2087	0.2151	0.2276	0.1574	0.2234	0.2057	0.2087	0.2151
SPOLC	0.1452	0.0315	0.1734	0.0752	0.0562	0.1009	0.1545	0.0350	0.1830	0.1291	0.0988	0.1218
SSPO	0.1544	0.0411	0.1181	0.0588	0.0425	0.0872	0.1789	0.0719	0.1557	0.0601	0.0529	0.1109
S1	0.1497	0.0369	0.1169	0.0559	0.0327	0.0879	0.1789	0.0736	0.1525	0.0648	0.0467	0.0999
S2	0.1382	0.0633	0.1225	0.0573	0.0456	0.1034	0.1578	0.0725	0.1438	0.0605	0.0602	0.1203
S3	0.1428	0.0607	0.1238	0.0570	0.0469	0.1100	0.1609	0.0709	0.1463	0.0617	0.0603	0.1215
SSMP	0.1934	0.1596	0.1535	0.1658	0.0883	0.1448	0.1920	0.0849	0.1512	0.1581	0.0573	0.1495
MAXER	0.1825	0.2229	0.1625	0.1581	N/A	N/A	0.1921	0.2379	0.1465	0.1433	0.1351	0.1479
IPSRM-D	0.2239	0.1994	0.1952	0.1436	0.1766	0.1662	0.2439	0.2358	0.2240	0.1410	0.2012	0.1712
PLCT	0.2475	0.2708	0.2600	0.2119	0.2270	0.2220	0.2468	0.2796	0.2577	0.2025	0.2369	0.2279
mSSRM-PGA(m=10)	0.2481	0.2712	0.2612	0.2151	0.2290	0.2217	0.2472	0.2796	0.2592	0.2041	0.2391	0.2271
mSSRM-PGA(m=15)	0.2481	0.2708	0.2615	0.2135	0.2289	0.2232	0.2474	0.2796	0.2592	0.2040	0.2381	0.2293
mSSRM-PGA(m=20)	0.2481	0.2708	0.2615	0.2134	0.2285	0.2234	0.2474	0.2796	0.2592	0.2041	0.2384	0.2292

4.2 Results for Cumulative Wealths

Ordinary investors are also concerned about how much they gain when using an investing strategy. Without loss of generality, we can set the initial wealth for an investing strategy as $S_{(0)} = 1$, then the final cumulative wealth can be conveniently computed by $S_{(\mathcal{T})} = \prod_{t=1}^{\mathcal{T}} (\hat{r}_{(t), \tilde{w}} + 1)$. The results of final cumulative wealths are shown in Table 2. The two competitors 1/N and PLCT perform well in general. Nevertheless, mSSRM-PGA achieves the best final cumulative wealths on 4 out of the 6 data sets. Besides, it outperforms each competitor on at least 5 out of the 6 data sets. For example, mSSRM-PGA is about 20% higher than the second best competitor PLCT on FF49 when $T = 60$ and $m = 10$. On the data sets where mSSRM-PGA is not the best method, it is still the second best method. These results indicate that mSSRM-PGA is an effective strategy for pursuing return gain in a practical perspective.

Table 2: Cumulative wealths of different portfolio optimization methods on 6 benchmark data sets.

Strategy	T = 60						T = 120					
	FF25	FF25EU	FF32	FF49	FF100	FF100MEINV	FF25	FF25EU	FF32	FF49	FF100	FF100MEINV
1/N	355.98	13.05	424.42	235.48	364.87	428.70	355.98	13.05	424.42	235.48	364.87	428.70
SPOLC	57.53	0.96	169.58	5.44	2.39	14.05	70.46	1.03	259.74	100.49	16.03	36.20
SSPO	129.35	1.22	30.20	1.33	0.89	8.98	286.51	2.67	130.21	1.61	1.74	25.62
S1	100.76	1.08	29.47	1.09	0.54	9.25	265.82	2.78	121.47	2.23	1.27	15.57
S2	66.24	2.17	39.27	1.39	1.15	20.45	130.31	2.73	93.13	1.89	2.67	43.66
S3	70.88	2.01	36.88	1.28	1.20	23.73	129.90	2.61	89.51	1.92	2.62	38.70
SSMP	248.67	13.47	158.98	186.79	10.09	154.27	237.45	3.25	149.65	143.18	2.26	222.35
MAXER	173.39	47.56	200.03	142.31	N/A	N/A	216.94	55.71	117.42	98.85	79.82	188.54
IPSRM-D	398.55	37.25	243.83	69.57	240.12	146.40	567.76	77.79	507.47	50.04	457.86	188.34
PLCT	581.41	126.04	918.62	238.27	471.44	354.70	608.65	148.19	854.83	157.50	552.41	399.48
mSSRM-PGA (m=10)	615.34	126.02	991.89	285.02	527.09	375.75	640.89	147.17	928.19	188.38	635.65	421.97
mSSRM-PGA (m=15)	614.71	125.19	996.32	262.54	522.28	383.44	643.44	147.17	927.21	172.95	597.67	435.01
mSSRM-PGA (m=20)	614.70	125.19	996.23	262.06	515.50	384.65	643.44	147.17	927.16	173.27	603.05	433.15

4.3 Results for Transaction Costs

Cumulative wealth with transaction cost can also be tested to see how the transaction cost influences the performance of different methods. We adopt the proportional transaction cost model [8, 26, 21]

$$S^\nu = S_{(0)} \prod_{t=1}^{\mathcal{T}} \left[(\hat{\mathbf{w}}_{(t)}^\top \mathbf{x}_{(t)}) \cdot \left(1 - \frac{\nu}{2} \sum_{i=1}^N |\hat{w}_{(t),i} - \tilde{w}_{(t-1),i}| \right) \right], \quad \tilde{w}_{(t-1),i} = \frac{\hat{w}_{(t-1),i} \mathbf{x}_{(t-1),i}}{\hat{\mathbf{w}}_{(t-1)}^\top \mathbf{x}_{(t-1)}},$$

where $\tilde{w}_{(t-1),i}$ is the evolved portfolio weight of the i -th asset at the end of the $(t-1)$ -th period, and ν is the bidirectional transaction cost rate. When the cost rate of buying is the same as that

of selling, updating the evolved portfolio $\tilde{w}_{(t-1)}$ as the next portfolio $\hat{w}_{(t)}$ yields a proportional transaction cost of $\frac{\nu}{2} \sum_{i=1}^N |\hat{w}_{(t),i} - \tilde{w}_{(t-1),i}|$. Figure 2 in Appendix A.10 shows the final cumulative wealths of different methods as ν varies from 0 to 0.5% with $T = 60$. mSSRM-PGA outperforms all other competitors on FF25, FF25EU and FF32 for all $\nu \in [0, 0.5\%]$, and on FF100 for $\nu \leq 0.45\%$. mSSRM-PGA is the second best method on FF100MEINV, following 1/N. This is because 1/N naturally keeps a small trading volume. Note that a manager for a mutual fund with sufficient trades and capital is able to negotiate for a small enough ν . Thus mSSRM-PGA is applicable to scenarios with a certain level of transaction cost.

4.4 Sparsity for mSSRM-PGA

In this subsection, we examine the sparsity of the portfolios $\{\hat{w}_{(t)}\}$ generated by mSSRM-PGA. The sparsity can be measured by the cardinality of the support set of $\hat{w}_{(t)}$: $|\text{supp}(\hat{w}_{(t)})|$. For each data set and each setting of m , the mean and the standard deviation (STD) of $\{|\text{supp}(\hat{w}_{(t)})|\}$ are computed to provide a general description, shown in Table 3. It indicates that mSSRM-PGA further increases sparsity compared with the presetted sparsity level m . Moreover, mSSRM-PGA keeps stable sparsity w.r.t. the change of m . For example, the average sparsity for mSSRM-PGA is about 4.9 when $T = 60$ (or 4.4 when $T = 120$) on FF25EU, for all the settings $m = 10, 15, 20$. As the total number of assets N increases, mSSRM-PGA gets more advantageous in sparsity. For example, the average sparsity for mSSRM-PGA is about 8 ~ 11 on FF100 and FF100MEINV, compared with $N = 100$. It indicates that mSSRM-PGA selects only 8% ~ 11% of the assets in the whole asset pool, while the widely-used 1/N strategy has to maintain the whole asset pool. Therefore, mSSRM-PGA can save much managerial cost by reducing the proportion of selected assets, while keeping a competitive performance in SR optimization.

Table 3: Sparsity of the portfolios generated by mSSRM-PGA: $|\text{supp}(\hat{w}_{(t)})|$.

m	FF25 FF25EU FF32 FF49 FF100 FF100MEINV						FF25 FF25EU FF32 FF49 FF100 FF100MEINV						
	$T = 60$						$T = 120$						
10	Mean	6.3511	4.8342	6.4159	8.1214	8.0097	7.9175	7.1359	4.4560	7.0825	8.4790	8.3706	8.8722
	STD	2.4164	2.1763	2.3654	1.9918	2.2473	2.3915	2.2221	1.4645	2.4464	1.9080	2.3343	2.0117
15	Mean	6.4746	4.9352	7.4286	9.0000	8.9709	9.0825	7.1637	4.4430	7.4919	9.6003	9.8754	10.5906
	STD	3.0451	2.3573	3.0590	3.0713	3.3964	3.5906	2.2995	1.4462	2.2567	3.1731	3.6169	3.4845
20	Mean	6.4763	4.9352	7.4692	9.0421	9.1974	9.2994	7.1637	4.4430	7.4919	9.6828	10.0437	10.7621
	STD	3.0462	2.3573	3.1734	3.1620	3.8949	4.0125	2.2995	1.4462	2.2567	3.3349	3.9160	3.7460

5 Concluding Remarks

The Sharpe ratio (SR) is a very important measurement for the performance of returns attributable to risk in finance. On the other hand, modern portfolio management usually restricts the number of selected assets to a relatively small size, in order to save managerial and financial costs. The m -sparse (ℓ_0) constraint is an exact constraint for a sparse portfolio, but it is nonconvex and complex. Thus few existing methods can optimize the SR with the m -sparse constraint. In this study, we convert the m -sparse fractional optimization problem into an equivalent m -sparse quadratic programming problem. Then we develop an efficient, easy-to-implement and mathematically sound proximal gradient algorithm to solve this nonconvex problem. We theoretically prove that this algorithm yields a portfolio that achieves the globally optimal m -sparse Sharpe ratio under certain conditions.

We conduct extensive experiments on 6 real-world monthly benchmark data sets built on the Kenneth R. French’s widely-used public data library. The numerical results demonstrate that the proposed mSSRM-PGA improves the SR, compared with 9 state-of-the-art portfolio optimization methods including SPOLC, SSPO, S1, S2, S3, SSMP, MAXER, IPSRM-D, PLCT and one baseline method 1/N. For another evaluating metric cumulative wealth, mSSRM-PGA outperforms each competitor on at least 5 out of the 6 data sets. Besides, mSSRM-PGA can withstand a considerable level of transaction cost rate. Sparsity experiments indicate that mSSRM-PGA successfully generates portfolios with stable sparsity, and its advantage increases as the size of the whole asset pool increases. In summary, the proposed mSSRM-PGA is a promising approach in managing portfolios or other financial issues, which is worth further investigations. A limitation of this research lies in its inability to directly apply to fractional optimization models featuring nondifferentiable numerator and denominator. Future work will strive to broaden the theoretical and methodological foundations, ultimately enabling its application to a broader spectrum of fractional optimization models in machine learning.

Acknowledgements

The authors thank the anonymous reviewers for their constructive comments and valuable suggestions in improving this paper. This work was supported in part by National Natural Science Foundation of China under Grants 12401120 and 62176103, in part by Guangdong Basic and Applied Basic Research Foundation under Grants 2021A1515110541 and 2023B1515120064, in part by the Science and Technology Planning Project of Guangdong under Grant 2023A0505030013, and in part by the Science and Technology Planning Project of Guangzhou under Grants 2024A04J3940, 2024A04J9896, 202206030007, Nansha District: 2023ZD001 and Development District: 2023GH01.

References

- [1] Paul H. Algoet and Thomas M. Cover. Asymptotic optimality and asymptotic equipartition properties of log-optimum investment. *The Annals of Probability*, 16(2):876–898, 1988.
- [2] Mengmeng Ao, Li Yingying, and Xinghua Zheng. Approaching mean-variance efficiency for large portfolios. *The Review of Financial Studies*, 32(7):2890–2919, 2019.
- [3] Hedy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.
- [4] Hedy Attouch, Jérôme Bolte, and Benar Fux Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.
- [5] Gah-Yi Ban, Noureddine El Karoui, and Andrew E. B. Lim. Machine learning and portfolio optimization. *Management Science*, 64(3):1136–1154, 2018.
- [6] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Space*. Springer, New York, 2nd edition, 2017.
- [7] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 2nd edition, 1999.
- [8] Avrim Blum and Adam Kalai. Universal portfolios with and without transaction costs. *Machine Learning*, 35(3):193–205, 1999.
- [9] Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014.
- [10] Joshua Brodie, Ingrid Daubechies, Christine De Mol, Domenico Giannone, and Ignace Loris. Sparse and stable Markowitz portfolios. *Proceedings of the National Academy of Sciences of the United States of America*, 106(30):12267–12272, 2009.
- [11] Raul O. Chao, Stylianos Kavadias, and Cheryl Gaimon. Revenue driven resource allocation: Funding authority, incentives, and new product development portfolio management. *Management Science*, 55(9):1556–1569, 2009.
- [12] Richard W. Cottle. Monotone solutions of the parametric linear complementarity problem. *Mathematical Programming*, 3(1):210–224, 1972.
- [13] Thomas M. Cover. Universal portfolios. *Mathematical Finance*, 1(1):1–29, 1991.
- [14] Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *The Review of Financial Studies*, 22(5):1915–1953, 2009.
- [15] Stephen G. Dimmock, Neng Wang, and Jinqiang Yang. The endowment model and modern portfolio theory. *Management Science*, 70(3):1554–1579, 2024.
- [16] John Duchi, Shai Shalev-Shwartz, Yoram Singer, and Tushar Chandra. Efficient projections onto the ℓ^1 -ball for learning in high dimensions. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 272–279, 2008.

- [17] Shingo Goto and Yan Xu. Improving mean variance optimization through sparse hedging restrictions. *The Journal of Financial and Quantitative Analysis*, 50(6):1415–1441, 2015.
- [18] Kei Keung Hung, Chi Chiu Cheung, and Lei Xu. New Sharpe-ratio-related methods for portfolio selection. In *Proceedings of the IEEE/IAFE/INFORMS 2000 Conference on Computational Intelligence for Financial Engineering (CIFEr)*, pages 34–37, 2000.
- [19] John L. Kelly. A new interpretation of information rate. *The Bell System Technical Journal*, 35(4):917–926, 1956.
- [20] Min Jeong Kim, Yongjae Lee, Jang Ho Kim, and Woo Chang Kim. Sparse tangent portfolio selection via semi-definite relaxation. *Operations Research Letters*, 44(4):540–543, 2016.
- [21] Zhao-Rong Lai, Dao-Qing Dai, Chuan-Xian Ren, and Ke-Kun Huang. Radial basis functions with adaptive input and composite trend representation for portfolio selection. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):6214–6226, 2018.
- [22] Zhao-Rong Lai, Liming Tan, Xiaotian Wu, and Liangda Fang. Loss control with rank-one covariance estimate for short-term portfolio optimization. *Journal of Machine Learning Research*, 21(97):1–37, 2020.
- [23] Zhao-Rong Lai and Haisheng Yang. A survey on gaps between mean-variance approach and exponential growth rate approach for portfolio optimization. *ACM Computing Surveys*, 55(2):1–36, 2023. Article No. 25.
- [24] Zhao-Rong Lai, Pei-Yi Yang, Liangda Fang, and Xiaotian Wu. Short-term sparse portfolio optimization based on alternating direction method of multipliers. *Journal of Machine Learning Research*, 19(63):1–28, 2018.
- [25] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets Goldilocks. *The Review of Financial Studies*, 30(12):4349–4388, 2017.
- [26] Bin Li, Steven C.H. Hoi, Doyen Sahoo, and Zhi-Yong Liu. Moving average reversion strategy for on-line portfolio selection. *Artificial Intelligence*, 222:104–123, 2015.
- [27] Hong Liu and Mark Loewenstein. Market crashes, correlated illiquidity, and portfolio choice. *Management Science*, 59(3):715–732, 2013.
- [28] Ziyang Luo, Xiaotong Yu, Naihua Xiu, and Xingyuan Wang. Closed-form solutions for short-term sparse portfolio optimization. *Optimization*, 71(7):1937–1953, 2022.
- [29] Jong-Shi Pang. A parametric linear complementarity technique for optimal portfolio selection with a risk-free asset. *Operations Research*, 28(4):927–941, 1980.
- [30] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.
- [31] R. Tyrrell Rockafellar and Roger J-B. Wets. *Variational Analysis*, volume 317. Springer Science & Business Media, 2009.
- [32] William F. Sharpe. Capital asset prices: A theory of market equilibrium under conditions of risk. *Journal of Finance*, 19(3):425–442, 1964.
- [33] William F. Sharpe. Mutual fund performance. *Journal of Business*, 39(1):119–138, 1966.
- [34] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.
- [35] Xiaohui Yu and Lei Xu. Adaptive improved portfolio Sharpe ratio maximization with diversification. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN)*, pages 472–476, 2000.

A Appendix

A.1 Proof of Theorem 1

To prove Theorem 1, we need the following lemma.

Lemma 8 *Suppose that there exists some $\tilde{\mathbf{w}} \in \Omega_1$ such that $\mathbf{p}^\top \tilde{\mathbf{w}} > 0$. If $\hat{\mathbf{v}}$ is an optimal solution of model (3.4), then $\hat{\mathbf{v}} \neq \mathbf{0}_N$ and $\mathbf{p}^\top \hat{\mathbf{v}} = \hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}} > 0$.*

Proof. Since $\hat{\mathbf{v}}$ is an optimal solution of model (3.4), we know that $\hat{\mathbf{v}} \in \Omega$. Let $\mathbf{w} := \frac{\mathbf{p}^\top \tilde{\mathbf{w}}}{\tilde{\mathbf{w}}^\top \mathbf{Q}_\epsilon \tilde{\mathbf{w}}} \tilde{\mathbf{w}}$. The facts $\tilde{\mathbf{w}} \in \Omega_1$ and $\mathbf{p}^\top \tilde{\mathbf{w}} > 0$ imply that $\mathbf{w} \in \Omega$. Then it follows that

$$\frac{1}{2} \hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}} - \mathbf{p}^\top \hat{\mathbf{v}} \leq \frac{1}{2} \mathbf{w}^\top \mathbf{Q}_\epsilon \mathbf{w} - \mathbf{p}^\top \mathbf{w} = -\frac{1}{2} \frac{(\mathbf{p}^\top \tilde{\mathbf{w}})^2}{\tilde{\mathbf{w}}^\top \mathbf{Q}_\epsilon \tilde{\mathbf{w}}} < 0.$$

Hence $\mathbf{p}^\top \hat{\mathbf{v}} > \frac{1}{2} \hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}} \geq 0$ and $\hat{\mathbf{v}} \neq \mathbf{0}_N$. Now by letting $\mathbf{v} := \frac{\mathbf{p}^\top \hat{\mathbf{v}}}{\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}} \hat{\mathbf{v}}$, then $\mathbf{v} \in \Omega$ and

$$\frac{1}{2} \hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}} - \mathbf{p}^\top \hat{\mathbf{v}} \leq \frac{1}{2} \mathbf{v}^\top \mathbf{Q}_\epsilon \mathbf{v} - \mathbf{p}^\top \mathbf{v} = -\frac{1}{2} \frac{(\mathbf{p}^\top \hat{\mathbf{v}})^2}{\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}}. \quad (\text{A.1})$$

Multiplying both sides of (A.1) by $2\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}$ yields

$$(\mathbf{p}^\top \hat{\mathbf{v}} - \hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}})^2 \leq 0,$$

which implies that $\mathbf{p}^\top \hat{\mathbf{v}} = \hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}} > 0$. \square

Proof of Theorem 1. Let $\hat{\mathbf{v}}$ be an optimal solution of model (3.3). Then $\hat{\mathbf{v}} \in \Omega_1$ and

$$\frac{\mathbf{p}^\top \hat{\mathbf{v}}}{\sqrt{\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}}} \geq \frac{\mathbf{p}^\top \tilde{\mathbf{w}}}{\sqrt{\tilde{\mathbf{w}}^\top \mathbf{Q}_\epsilon \tilde{\mathbf{w}}}} > 0.$$

Defining $\tilde{\mathbf{v}} := \frac{\mathbf{p}^\top \hat{\mathbf{v}}}{\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}} \hat{\mathbf{v}}$, we see that $\tilde{\mathbf{v}} \in \Omega$ and

$$\frac{1}{2} \tilde{\mathbf{v}}^\top \mathbf{Q}_\epsilon \tilde{\mathbf{v}} - \mathbf{p}^\top \tilde{\mathbf{v}} = \frac{1}{2} \frac{(\mathbf{p}^\top \hat{\mathbf{v}})^2}{(\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}})^2} \hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}} - \frac{(\mathbf{p}^\top \hat{\mathbf{v}})^2}{\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}} = -\frac{1}{2} \frac{(\mathbf{p}^\top \hat{\mathbf{v}})^2}{\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}} < 0. \quad (\text{A.2})$$

For any $\mathbf{u} \in \Omega$ such that $\frac{1}{2} \mathbf{u}^\top \mathbf{Q}_\epsilon \mathbf{u} - \mathbf{p}^\top \mathbf{u} < 0$, we have $\mathbf{p}^\top \mathbf{u} > 0$, $\mathbf{u} \neq \mathbf{0}_N$ and $\tilde{\mathbf{u}} := \frac{\mathbf{u}}{\mathbf{u}^\top \mathbf{1}_N} \in \Omega_1$. Then the fact $\hat{\mathbf{v}}$ is an optimal solution of model (3.3) implies that

$$\frac{(\mathbf{p}^\top \hat{\mathbf{v}})^2}{\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}} \geq \frac{(\mathbf{p}^\top \tilde{\mathbf{u}})^2}{\tilde{\mathbf{u}}^\top \mathbf{Q}_\epsilon \tilde{\mathbf{u}}} = \frac{(\mathbf{p}^\top \mathbf{u})^2}{\mathbf{u}^\top \mathbf{Q}_\epsilon \mathbf{u}}. \quad (\text{A.3})$$

Note that

$$\frac{1}{2} \mathbf{u}^\top \mathbf{Q}_\epsilon \mathbf{u} - \mathbf{p}^\top \mathbf{u} + \frac{1}{2} \frac{(\mathbf{p}^\top \mathbf{u})^2}{\mathbf{u}^\top \mathbf{Q}_\epsilon \mathbf{u}} = \frac{1}{2\mathbf{u}^\top \mathbf{Q}_\epsilon \mathbf{u}} (\mathbf{u}^\top \mathbf{Q}_\epsilon \mathbf{u} - \mathbf{p}^\top \mathbf{u})^2 \geq 0,$$

which combined with (A.3) and (A.2) yields

$$\frac{1}{2} \mathbf{u}^\top \mathbf{Q}_\epsilon \mathbf{u} - \mathbf{p}^\top \mathbf{u} \geq -\frac{1}{2} \frac{(\mathbf{p}^\top \mathbf{u})^2}{\mathbf{u}^\top \mathbf{Q}_\epsilon \mathbf{u}} \geq -\frac{1}{2} \frac{(\mathbf{p}^\top \hat{\mathbf{v}})^2}{\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}} = \frac{1}{2} \tilde{\mathbf{v}}^\top \mathbf{Q}_\epsilon \tilde{\mathbf{v}} - \mathbf{p}^\top \tilde{\mathbf{v}}.$$

Therefore, $\tilde{\mathbf{v}}$ is an optimal solution of model (3.4).

Conversely, let $\hat{\mathbf{v}}$ be an optimal solution of model (3.4). It follows from Lemma 8 that $\hat{\mathbf{v}} \neq \mathbf{0}_N$ and $\mathbf{p}^\top \hat{\mathbf{v}} = \hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}} > 0$. Thus $\hat{\mathbf{v}} = \frac{\mathbf{p}^\top \hat{\mathbf{v}}}{\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}} \hat{\mathbf{v}}$. For any $\mathbf{v} \in \Omega$ such that $\mathbf{p}^\top \mathbf{v} > 0$, we let $\mathbf{u} := \frac{\mathbf{p}^\top \mathbf{v}}{\mathbf{v}^\top \mathbf{Q}_\epsilon \mathbf{v}} \mathbf{v}$. Then $\mathbf{u} \in \Omega$ and

$$-\frac{1}{2} \frac{(\mathbf{p}^\top \hat{\mathbf{v}})^2}{\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}} = \frac{1}{2} \hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}} - \mathbf{p}^\top \hat{\mathbf{v}} \leq \frac{1}{2} \mathbf{u}^\top \mathbf{Q}_\epsilon \mathbf{u} - \mathbf{p}^\top \mathbf{u} = -\frac{1}{2} \frac{(\mathbf{p}^\top \mathbf{v})^2}{\mathbf{v}^\top \mathbf{Q}_\epsilon \mathbf{v}}. \quad (\text{A.4})$$

Now we let $\bar{\mathbf{v}} := \frac{\hat{\mathbf{v}}}{\hat{\mathbf{v}}^\top \mathbf{1}_N}$. Then $\bar{\mathbf{v}} \in \Omega_1$. Inequality (A.4) yields that

$$\frac{\mathbf{p}^\top \bar{\mathbf{v}}}{\sqrt{\bar{\mathbf{v}}^\top \mathbf{Q}_\epsilon \bar{\mathbf{v}}}} = \frac{\mathbf{p}^\top \hat{\mathbf{v}}}{\sqrt{\hat{\mathbf{v}}^\top \mathbf{Q}_\epsilon \hat{\mathbf{v}}}} \geq \frac{\mathbf{p}^\top \mathbf{v}}{\sqrt{\mathbf{v}^\top \mathbf{Q}_\epsilon \mathbf{v}}}.$$

Note that $\Omega_1 \subset \Omega$. Therefore, $\bar{\mathbf{v}}$ is an optimal solution of model (3.3). \square

A.2 Proof of Theorem 2

To prove Theorem 2, we first investigate the properties of function f in Proposition 9, and then recall two well-known results as Lemmas 10 and 11. Let ψ be a function from \mathbb{R}^n to $[-\infty, +\infty]$. Then ψ is proper if $-\infty \notin \psi(\mathbb{R}^n)$ and $\{\mathbf{x} \in \mathbb{R}^n \mid \psi(\mathbf{x}) < +\infty\} \neq \emptyset$. Let $\psi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be a proper function. We say that ψ is convex if for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and any $\lambda \in (0, 1)$, $\psi(\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda\psi(\mathbf{x}) + (1 - \lambda)\psi(\mathbf{y})$. If there exists $\beta > 0$ such that $\psi - \frac{\beta}{2}\|\cdot\|_2^2$ is convex, then ψ is said to be β -strongly convex.

Proposition 9 *Let $f : \mathbb{R}^N \rightarrow \mathbb{R}$ be defined in (3.7). Then the following hold:*

- (i) f is ϵ -strongly convex on \mathbb{R}^N ;
- (ii) ∇f is $\|\mathbf{Q}_\epsilon\|_2$ -Lipschitz continuous on \mathbb{R}^N .

Proof. Let $\tilde{f}(\mathbf{v}) := f(\mathbf{v}) - \frac{\epsilon}{2}\|\mathbf{v}\|_2^2 = \frac{1}{2}\mathbf{v}^\top \mathbf{Q}^\top \mathbf{Q} \mathbf{v} - \mathbf{p}^\top \mathbf{v}$, $\mathbf{v} \in \mathbb{R}^N$. Since the Hessian matrix $\mathbf{Q}^\top \mathbf{Q}$ of \tilde{f} is positive semidefinite, we know that \tilde{f} is convex on \mathbb{R}^N (see Proposition B.4 of [7]). Thus item (i) holds. The gradient of f is given by $\nabla f(\mathbf{v}) = \mathbf{Q}_\epsilon \mathbf{v} - \mathbf{p}$. For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2 \leq \|\mathbf{Q}_\epsilon\|_2 \|\mathbf{x} - \mathbf{y}\|_2$, which implies item (ii). \square

Lemma 10 (Proposition A.24 of [7]) *Let function $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable with an L -Lipschitz continuous gradient, where $L > 0$. Then*

$$\psi(\mathbf{y}) - \psi(\mathbf{x}) \leq \langle \nabla \psi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$$

holds for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Lemma 11 (Exercise 17.5 of [6]) *Let $\psi : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable and $\beta > 0$. Then ψ is β -strongly convex if and only if*

$$\psi(\mathbf{y}) - \psi(\mathbf{x}) \geq \langle \nabla \psi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\beta}{2} \|\mathbf{y} - \mathbf{x}\|_2^2$$

holds for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Proof of Theorem 2. We first show that (3.8) holds when \mathbf{v}^* is a globally optimal solution of model (3.7). By the definition of proximity operator, (3.8) is equivalent to

$$\mathbf{v}^* = \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^N} \iota_\Omega(\mathbf{u}) + \frac{1}{2} \|\mathbf{u} - \mathbf{v}^* + \alpha \nabla f(\mathbf{v}^*)\|_2^2,$$

that is,

$$\iota_\Omega(\mathbf{u}) + \frac{1}{2} \|\mathbf{u} - \mathbf{v}^* + \alpha \nabla f(\mathbf{v}^*)\|_2^2 \geq \iota_\Omega(\mathbf{v}^*) + \frac{1}{2} \|\alpha \nabla f(\mathbf{v}^*)\|_2^2, \text{ for all } \mathbf{u} \in \mathbb{R}^N.$$

According to the definition of ι_Ω in (3.6) and the fact $\mathbf{v}^* \in \Omega$, the above inequality can be simply rewritten as

$$\langle \nabla f(\mathbf{v}^*), \mathbf{u} - \mathbf{v}^* \rangle + \frac{1}{2\alpha} \|\mathbf{u} - \mathbf{v}^*\|_2^2 \geq 0, \text{ for all } \mathbf{u} \in \Omega. \quad (\text{A.5})$$

To prove (3.8), it suffices to show that (A.5) holds. From Proposition 9, we know that f is ϵ -strongly convex and ∇f is $\|\mathbf{Q}_\epsilon\|_2$ -Lipschitz continuous on \mathbb{R}^N . Then Lemma 10 yields that

$$f(\mathbf{u}) - f(\mathbf{v}^*) \leq \langle \nabla f(\mathbf{v}^*), \mathbf{u} - \mathbf{v}^* \rangle + \frac{\|\mathbf{Q}_\epsilon\|_2^2}{2} \|\mathbf{u} - \mathbf{v}^*\|_2^2, \text{ for all } \mathbf{u} \in \Omega. \quad (\text{A.6})$$

Since \mathbf{v}^* is a globally optimal solution of model (3.7), $f(\mathbf{u}) - f(\mathbf{v}^*) \geq 0$ for all $\mathbf{u} \in \Omega$, which together with (A.6) and the fact $\alpha \in \left(0, \frac{1}{\|\mathbf{Q}_\epsilon\|_2}\right]$ implies (A.5). This proves that (3.8) holds.

Conversely, if $\alpha \geq \frac{1}{\epsilon}$ and (3.8) holds, then we have (A.5). Recall that f is ϵ -strongly convex. It follows from Lemma 11 that

$$f(\mathbf{u}) - f(\mathbf{v}^*) \geq \langle \nabla f(\mathbf{v}^*), \mathbf{u} - \mathbf{v}^* \rangle + \frac{\epsilon}{2} \|\mathbf{u} - \mathbf{v}^*\|_2^2,$$

which together with the fact $\alpha \geq \frac{1}{\epsilon}$ and (A.5) implies that $f(\mathbf{u}) - f(\mathbf{v}^*) \geq 0$ for all $\mathbf{u} \in \Omega$. Thus the assertion in item (i) holds.

We then prove item (ii). The fact (3.8) holds implies (A.5). For $\delta > 0$, we define

$$\tilde{\Omega}_\delta := \{\mathbf{u} \in B(\mathbf{v}^*; \delta) \mid \langle \nabla f(\mathbf{v}^*), \mathbf{u} - \mathbf{v}^* \rangle = 0\}.$$

Note that when \mathbf{u} tends to \mathbf{v}^* , the quadratic term $\frac{1}{2\alpha} \|\mathbf{u} - \mathbf{v}^*\|_2^2$ is of higher order infinitesimal than the linear term $|\langle \nabla f(\mathbf{v}^*), \mathbf{u} - \mathbf{v}^* \rangle|$. There must be some $\delta > 0$ such that

$$|\langle \nabla f(\mathbf{v}^*), \mathbf{u} - \mathbf{v}^* \rangle| > \frac{1}{2\alpha} \|\mathbf{u} - \mathbf{v}^*\|_2^2, \quad \text{for all } \mathbf{u} \in B(\mathbf{v}^*; \delta) \setminus \tilde{\Omega}_\delta. \quad (\text{A.7})$$

We then show that

$$\langle \nabla f(\mathbf{v}^*), \mathbf{u} - \mathbf{v}^* \rangle \geq 0, \quad \text{for all } \mathbf{u} \in (B(\mathbf{v}^*; \delta) \setminus \tilde{\Omega}_\delta) \cap \Omega. \quad (\text{A.8})$$

Otherwise, there exists some $\tilde{\mathbf{u}} \in (B(\mathbf{v}^*; \delta) \setminus \tilde{\Omega}_\delta) \cap \Omega$ such that $\langle \nabla f(\mathbf{v}^*), \tilde{\mathbf{u}} - \mathbf{v}^* \rangle < 0$. It follows from (A.7) that

$$\langle \nabla f(\mathbf{v}^*), \tilde{\mathbf{u}} - \mathbf{v}^* \rangle + \frac{1}{2\alpha} \|\tilde{\mathbf{u}} - \mathbf{v}^*\|_2^2 < 0,$$

which contradicts (A.5). Hence (A.8) holds. This together with the definition of $\tilde{\Omega}_\delta$ yields that

$$\langle \nabla f(\mathbf{v}^*), \mathbf{u} - \mathbf{v}^* \rangle \geq 0, \quad \text{for all } \mathbf{u} \in B(\mathbf{v}^*; \delta) \cap \Omega. \quad (\text{A.9})$$

Recall that f is convex and differentiable on \mathbb{R}^N . According to (A.9) and the first order condition for convexity (Proposition B.3 of [7]),

$$f(\mathbf{u}) - f(\mathbf{v}^*) \geq \langle \nabla f(\mathbf{v}^*), \mathbf{u} - \mathbf{v}^* \rangle \geq 0, \quad \text{for all } \mathbf{u} \in B(\mathbf{v}^*; \delta) \cap \Omega,$$

which implies that \mathbf{v}^* is a locally optimal solution of model (3.7). \square

A.3 Proof of Proposition 3

Proof. By the definitions of ι_Ω and its proximity operator, we have

$$\text{prox}_{\iota_\Omega}(\mathbf{v}) = \underset{\mathbf{u} \in \Omega}{\text{argmin}} \|\mathbf{u} - \mathbf{v}\|_2.$$

To prove that $\mathbf{h} \in \text{prox}_{\iota_\Omega}(\mathbf{v})$, it is equivalent to show that

$$\|\mathbf{h} - \mathbf{v}\|_2^2 \leq \|\mathbf{u} - \mathbf{v}\|_2^2, \quad \text{for all } \mathbf{u} \in \Omega. \quad (\text{A.10})$$

For any $\mathbf{u} \in \Omega$, there exists an index set $J_{\mathbf{u}} \in \mathbb{N}_N$ with m elements such that $u_j = 0$ for all $j \in \mathbb{N}_N \setminus J_{\mathbf{u}}$. Let $J_{\text{neg}}^{\mathbf{v}}$ be the index set of negative components in \mathbf{v} and $J_{\mathbf{u}}' := (\mathbb{N}_N \setminus J_{\mathbf{u}}) \cup J_{\text{neg}}^{\mathbf{v}}$. Since $\mathbf{u} \geq \mathbf{0}_N$, $\|\mathbf{u} - \mathbf{v}\|_2^2 \geq \sum_{j \in J_{\mathbf{u}}'} v_j^2$. Let $J_{\mathbf{h}}' = \mathbb{N}_N \setminus J^{\mathbf{v}}$. Then $J_{\text{neg}}^{\mathbf{v}} \subset J_{\mathbf{h}}'$ and $\|\mathbf{h} - \mathbf{v}\|_2^2 = \sum_{j \in J_{\mathbf{h}}'} v_j^2$. If $m_{\mathbf{v}} > m$, then $J^{\mathbf{v}} = J_{m\text{-pos}}^{\mathbf{v}}$. We are easy to see from the definition of $J_{m\text{-pos}}^{\mathbf{v}}$ that

$$\sum_{j \in J_{\mathbf{u}}'} v_j^2 - \sum_{j \in J_{\mathbf{h}}'} v_j^2 = \sum_{j \in \mathbb{N}_N \setminus (J_{\mathbf{u}} \cup J_{\text{neg}}^{\mathbf{v}})} v_j^2 - \sum_{j \in \mathbb{N}_N \setminus (J_{m\text{-pos}}^{\mathbf{v}} \cup J_{\text{neg}}^{\mathbf{v}})} v_j^2 \geq 0.$$

If $m_{\mathbf{v}} \leq m$, then $J^{\mathbf{v}} = J_{\text{pos}}^{\mathbf{v}}$ and $\sum_{j \in J_{\mathbf{u}}'} v_j^2 - \sum_{j \in J_{\mathbf{h}}'} v_j^2 = \sum_{j \in (\mathbb{N}_N \setminus J_{\mathbf{u}}) \cup J_{\text{neg}}^{\mathbf{v}}} v_j^2 - \sum_{j \in J_{\text{neg}}^{\mathbf{v}}} v_j^2 \geq 0$. Now we conclude from the above two cases that

$$\|\mathbf{u} - \mathbf{v}\|_2^2 - \|\mathbf{h} - \mathbf{v}\|_2^2 \geq \sum_{j \in J_{\mathbf{u}}'} v_j^2 - \sum_{j \in J_{\mathbf{h}}'} v_j^2 \geq 0,$$

that is, (A.10) holds. This completes the proof. \square

A.4 Proof of Proposition 4

Proof. Item (i) follows from (3.9) and the definition of $\text{prox}_{\iota_\Omega}$ directly. Then we have that $\iota_\Omega(\mathbf{v}^{(k)}) = 0$ for all $k \in \mathbb{N}$. To prove item (ii), it suffices to show that

$$f(\mathbf{v}^{(k+1)}) + a\|\mathbf{v}^{(k+1)} - \mathbf{v}^{(k)}\|_2^2 \leq f(\mathbf{v}^{(k)}), \quad \text{for all } k \in \mathbb{N}. \quad (\text{A.11})$$

Note that $a = \frac{1}{\alpha} - \|\mathbf{Q}_\epsilon\|_2 > 0$, since $\alpha \in \left(0, \frac{1}{\|\mathbf{Q}_\epsilon\|_2}\right)$. Let

$$\varphi(\mathbf{u}) := \frac{1}{2} \left\| \mathbf{u} - \mathbf{v}^{(k)} + \alpha \nabla f(\mathbf{v}^{(k)}) \right\|_2^2 + \iota_\Omega(\mathbf{u}), \quad \mathbf{u} \in \mathbb{R}^N. \quad (\text{A.12})$$

Then (3.9) implies that $\varphi(\mathbf{v}^{(k+1)}) \leq \varphi(\mathbf{v}^{(k)})$, that is,

$$\langle \nabla f(\mathbf{v}^{(k)}), \mathbf{v}^{(k+1)} - \mathbf{v}^{(k)} \rangle \leq -\frac{1}{2\alpha} \|\mathbf{v}^{(k+1)} - \mathbf{v}^{(k)}\|_2^2, \quad \text{for all } k \in \mathbb{N}, \quad (\text{A.13})$$

It follows from Lemma 10 that

$$f(\mathbf{v}^{(k+1)}) - f(\mathbf{v}^{(k)}) \leq \langle \nabla f(\mathbf{v}^{(k)}), \mathbf{v}^{(k+1)} - \mathbf{v}^{(k)} \rangle + \frac{\|\mathbf{Q}_\epsilon\|_2}{2} \|\mathbf{v}^{(k+1)} - \mathbf{v}^{(k)}\|_2^2. \quad (\text{A.14})$$

Combining (A.13) and (A.14) yields (A.11). Thus item (ii) holds. Now that F is monotonically decreasing, according to the monotone convergence theorem, to prove item (iii), it suffices to show that function F is bounded below on Ω . Solving $\nabla f(\mathbf{v}^*) = 0$ gives $\mathbf{v}^* = \mathbf{Q}_\epsilon^{-1}\mathbf{p}$. Since f is convex and differentiable on \mathbb{R}^N , f attains the minimum value at $\mathbf{Q}_\epsilon^{-1}\mathbf{p}$ on \mathbb{R}^N . Hence $f(\mathbf{v}) \geq f(\mathbf{Q}_\epsilon^{-1}\mathbf{p}) = -\frac{1}{2}\mathbf{p}^\top \mathbf{Q}_\epsilon^{-1}\mathbf{p}$ for all $\mathbf{v} \in \mathbb{R}^N$, which implies that $F(\mathbf{v}) \geq -\frac{1}{2}\mathbf{p}^\top \mathbf{Q}_\epsilon^{-1}\mathbf{p}$ for all $\mathbf{v} \in \Omega$. Therefore, item (iii) holds. Now taking the limit on both sides of the inequality in item (ii) yields item (iv) immediately. This completes the proof. \square

A.5 Proof of Theorem 5

In order to prove Theorem 5, it is necessary to review several definitions and establish several preliminary results. First, We recall the notions of subdifferentials and critical point. The lower limit of function ψ at \mathbf{x} and the domain of ψ are defined by

$$\liminf_{\mathbf{y} \rightarrow \mathbf{x}} \psi(\mathbf{y}) := \lim_{\delta \rightarrow 0^+} \left(\inf_{\mathbf{y} \in B(\mathbf{x}; \delta)} \psi(\mathbf{y}) \right) \quad (\text{A.15})$$

and

$$\text{dom } \psi := \{\mathbf{x} \in \mathbb{R}^n \mid \psi(\mathbf{x}) < +\infty\},$$

respectively. We say that ψ is lower semicontinuous at $\mathbf{x} \in \mathbb{R}^n$ if $\psi(\mathbf{x}) \leq \liminf_{\mathbf{u} \rightarrow \mathbf{x}} \psi(\mathbf{u})$. If ψ is lower semicontinuous at every $\mathbf{x} \in \mathbb{R}^n$, then ψ is lower semicontinuous on \mathbb{R}^n [31].

Definition 1 (Subdifferentials and critical point) Let $\psi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be a proper lower semicontinuous function.

- (i) For each $\mathbf{x} \in \text{dom } \psi$, the Fréchet subdifferential of ψ at \mathbf{x} , written by $\hat{\partial}\psi(\mathbf{x})$, is the set of all vectors $\mathbf{u} \in \mathbb{R}^n$ which satisfy

$$\liminf_{\substack{\mathbf{y} \rightarrow \mathbf{x} \\ \mathbf{y} \neq \mathbf{x}}} \frac{\psi(\mathbf{y}) - \psi(\mathbf{x}) - \langle \mathbf{u}, \mathbf{y} - \mathbf{x} \rangle}{\|\mathbf{y} - \mathbf{x}\|_2} \geq 0.$$

When $\mathbf{x} \notin \text{dom } \psi$, we set $\hat{\partial}\psi(\mathbf{x}) = \emptyset$.

- (ii) The limiting-subdifferential, or simply the subdifferential of ψ at $\mathbf{x} \in \text{dom } \psi$, written by $\partial\psi(\mathbf{x})$, is defined through the following closure process

$$\partial\psi(\mathbf{x}) := \{\mathbf{u} \in \mathbb{R}^n \mid \exists \mathbf{x}^k \rightarrow \mathbf{x}, \psi(\mathbf{x}^k) \rightarrow \psi(\mathbf{x}) \text{ and } \mathbf{u}^k \in \hat{\partial}\psi(\mathbf{x}^k) \rightarrow \mathbf{u} \text{ as } k \rightarrow +\infty\}.$$

We call an element in $\partial\psi(\mathbf{x})$ subgradient of ψ at \mathbf{x} . We say that \mathbf{x} is a critical point of ψ if $\mathbf{0}_n \in \partial\psi(\mathbf{x})$.

We also recall the following known results about subdifferential from Theorem 8.6, Exercise 8.8 (c) and Theorem 10.1 of [31], respectively.

Fact 12 For $\mathbf{x} \in \text{dom } \psi$, $\hat{\partial}\psi(\mathbf{x}) \subset \partial\psi(\mathbf{x})$.

Fact 13 Let $\psi_1 : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ and $\psi_2 : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be two proper lower semicontinuous functions and $\mathbf{x} \in \mathbb{R}^n$. If ψ_1 is differentiable on a neighborhood of \mathbf{x} and ψ_2 is finite at \mathbf{x} , then

$$\partial(\psi_1 + \psi_2)(\mathbf{x}) = \nabla\psi_1(\mathbf{x}) + \partial\psi_2(\mathbf{x}).$$

Fact 14 (Fermat's rule) If $\mathbf{x} \in \mathbb{R}^n$ is a local minimizer of ψ , then $\mathbf{0}_n \in \partial\psi(\mathbf{x})$.

We shall use Theorem 2.9 of [4], which is recalled as Proposition 15, to prove the convergence of the PGA. For this purpose, we recall the notions of Kurdyka-Łojasiewicz (KL) property and KL function.

Definition 2 (KL property) Let $\psi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be a proper semicontinuous function. We say that ψ satisfies the KL property at $\hat{\mathbf{x}} \in \text{dom } \partial\psi$ if there exist $\eta \in (0, +\infty]$, a neighborhood U of $\hat{\mathbf{x}}$ and a continuous concave function $\varphi : [0, \eta] \rightarrow [0, +\infty]$ such that

- (i) $\varphi(0) = 0$;
- (ii) φ is continuously differentiable on $(0, \eta)$ with $\varphi' > 0$;
- (iii) $\varphi'(\psi(\mathbf{x}) - \psi(\hat{\mathbf{x}})) \cdot \text{dist}(\mathbf{0}, \partial\psi(\mathbf{x})) \geq 1$ for any $\mathbf{x} \in U \cap \{\mathbf{x} \in \mathbb{R}^n : \psi(\hat{\mathbf{x}}) < \psi(\mathbf{x}) < \psi(\hat{\mathbf{x}}) + \eta\}$.

Definition 3 (KL function) We call a proper lower semicontinuous function $\psi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ KL function if ψ satisfies the KL property at all points in $\text{dom } \partial\psi$.

Proposition 15 Let $\psi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be a proper lower semicontinuous function. Consider a sequence $\{\mathbf{x}^{(k)}\}_{k \in \mathbb{N}} \subset \mathbb{R}^n$ satisfying the following conditions:

- (i) There exists $a > 0$ such that

$$\psi(\mathbf{x}^{(k+1)}) + a\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2^2 \leq \psi(\mathbf{x}^{(k)}), \text{ for all } k \in \mathbb{N}.$$

- (ii) There exist $b > 0$ and $\mathbf{y}^{(k+1)} \in \partial\psi(\mathbf{x}^{(k+1)})$ such that

$$\|\mathbf{y}^{(k+1)}\|_2 \leq b\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_2, \text{ for all } k \in \mathbb{N}.$$

- (iii) There exist a subsequence $\{\mathbf{x}^{(k_j)}\}_{j \in \mathbb{N}_+}$ and $\mathbf{x}^* \in \mathbb{R}^n$ such that

$$\lim_{j \rightarrow \infty} \mathbf{x}^{(k_j)} = \mathbf{x}^* \text{ and } \lim_{j \rightarrow \infty} \psi(\mathbf{x}^{(k_j)}) = \psi(\mathbf{x}^*).$$

If ψ satisfies the KL property at \mathbf{x}^* , then

$$\lim_{k \rightarrow \infty} \mathbf{x}^{(k)} = \mathbf{x}^* \text{ and } \mathbf{0}_n \in \partial\psi(\mathbf{x}^*).$$

We then focus on verifying that the sequence $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ generated by PGA satisfies all the conditions in Proposition 15. The satisfaction of item (i) has been shown in Proposition 4. We next consider the satisfaction of item (ii) in Proposition 15.

Proposition 16 Let $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ be generated by PGA. Then there exist $\mathbf{q}^{(k+1)} \in \partial F(\mathbf{v}^{(k+1)})$ and $b > 0$ such that

$$\|\mathbf{q}^{(k+1)}\|_2 \leq b\|\mathbf{v}^{(k+1)} - \mathbf{v}^{(k)}\|_2, \text{ for } k \in \mathbb{N}. \quad (\text{A.16})$$

Proof. Let

$$\mathbf{q}^{(k+1)} := \frac{1}{\alpha}(\mathbf{v}^{(k)} - \mathbf{v}^{(k+1)}) + \nabla f(\mathbf{v}^{(k+1)}) - \nabla f(\mathbf{v}^{(k)}), \quad k \in \mathbb{N},$$

and function φ be defined by (A.12). We first prove that $\mathbf{q}^{(k+1)} \in \partial F(\mathbf{v}^{(k+1)})$. It follows from (3.9) and Fact 14 that $\mathbf{0}_N \in \partial\varphi(\mathbf{v}^{(k+1)})$, which together with Fact 13 yields that

$$\mathbf{v}^{(k)} - \mathbf{v}^{(k+1)} - \alpha \nabla f(\mathbf{v}^{(k)}) \in \partial\iota_\Omega(\mathbf{v}^{(k+1)}), \text{ for all } k \in \mathbb{N}.$$

Note that $\iota_\Omega = \alpha\iota_\Omega$. The above inclusion relation can be rewritten as

$$\frac{1}{\alpha}(\mathbf{v}^{(k)} - \mathbf{v}^{(k+1)}) - \nabla f(\mathbf{v}^{(k)}) \in \partial\iota_\Omega(\mathbf{v}^{(k+1)}), \text{ for all } k \in \mathbb{N}. \quad (\text{A.17})$$

Now combining (A.17) and the fact $\partial F(\mathbf{v}^{(k+1)}) = \nabla f(\mathbf{v}^{(k+1)}) + \partial\iota_\Omega(\mathbf{v}^{(k+1)})$ yields that $\mathbf{q}^{(k+1)} \in \partial F(\mathbf{v}^{(k+1)})$, $k \in \mathbb{N}$.

We next prove that (A.16) holds. Since ∇f is \mathbf{Q}_ϵ -Lipschitz continuous,

$$\|\mathbf{q}^{(k+1)}\|_2 \leq \frac{1}{\alpha} \|\mathbf{v}^{(k+1)} - \mathbf{v}^{(k)}\|_2 + \|\nabla f(\mathbf{v}^{(k+1)}) - \nabla f(\mathbf{v}^{(k)})\|_2 \leq b \|\mathbf{v}^{(k+1)} - \mathbf{v}^{(k)}\|_2,$$

where $b := (\frac{1}{\alpha} + \|\mathbf{Q}_\epsilon\|_2)$. This completes the proof. \square

We then consider the satisfaction of item (iii) in Proposition 15. To this end, we need the following two lemmas.

Lemma 17 *Let $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ be generated by PGA. If $\alpha \in (0, \frac{1}{\|\mathbf{Q}_\epsilon\|_2})$, then $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ is bounded.*

Proof. We let $\gamma := \|\mathbf{I} - \alpha\mathbf{Q}_\epsilon\|_2$, and denote by $\lambda_{\max}(\mathbf{Q}_\epsilon)$, $\lambda_{\min}(\mathbf{Q}_\epsilon)$ the maximum and the minimum eigenvalues of \mathbf{Q}_ϵ , respectively. Since \mathbf{Q}_ϵ is symmetric positive definite and $\alpha < \frac{1}{\|\mathbf{Q}_\epsilon\|_2} = \frac{1}{\lambda_{\max}(\mathbf{Q}_\epsilon)}$, we have $\gamma = 1 - \alpha \cdot \lambda_{\min}(\mathbf{Q}_\epsilon) \in (0, 1)$. From Proposition 3, we are easy to see that $\|\text{prox}_{\iota_\Omega}(\mathbf{v})\|_2 \leq \|\mathbf{v}\|_2$ for all $\mathbf{v} \in \mathbb{R}^N$, which together with (3.9) yields that

$$\|\mathbf{v}^{(k+1)}\|_2 \leq \|\mathbf{v}^{(k)} - \alpha \nabla f(\mathbf{v}^{(k)})\|_2 = \|(\mathbf{I} - \alpha\mathbf{Q}_\epsilon)\mathbf{v}^{(k)} + \alpha\mathbf{p}\|_2 \leq \gamma \|\mathbf{v}^{(k)}\|_2 + \alpha \|\mathbf{p}\|_2,$$

for all $k \in \mathbb{N}$. The above inequality implies that

$$\|\mathbf{v}^{(k+1)}\|_2 \leq \gamma^{k+1} \|\mathbf{v}^{(0)}\|_2 + \alpha \|\mathbf{p}\|_2 \sum_{j=0}^k \gamma^j = \gamma^{k+1} \|\mathbf{v}^{(0)}\|_2 + \alpha \|\mathbf{p}\|_2 \cdot \frac{1 - \gamma^{k+1}}{1 - \gamma},$$

for all $k \in \mathbb{N}$. Therefore, $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ is bounded, since $\gamma \in (0, 1)$. \square

Lemma 18 *Let $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ be generated by PGA. If \mathbf{v}^* is an accumulation point of $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$, then $\mathbf{v}^* \in \Omega$.*

Proof. Since \mathbf{v}^* is an accumulation point of $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$, there exists a subsequence $\{\mathbf{v}^{(k_j)}\}_{j \in \mathbb{N}_+}$ of $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ such that $\lim_{j \rightarrow \infty} \mathbf{v}^{(k_j)} = \mathbf{v}^*$. Note that the set Ω is closed and $\mathbf{v}^{(k_j)} \in \Omega$ for all $j \in \mathbb{N}_+$.

Hence $\mathbf{v}^* \in \Omega$, which completes the proof. \square

Proposition 19 *Let $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ be generated by PGA and $F := f + \iota_\Omega$. If $\alpha \in (0, \frac{1}{\|\mathbf{Q}_\epsilon\|_2})$, then there exist a subsequence $\{\mathbf{v}^{(k_j)}\}_{j \in \mathbb{N}_+}$ of $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ and $\mathbf{v}^* \in \Omega$ such that*

$$\lim_{j \rightarrow \infty} \mathbf{v}^{(k_j)} = \mathbf{v}^* \text{ and } \lim_{j \rightarrow \infty} F(\mathbf{v}^{(k_j)}) = F(\mathbf{v}^*).$$

Proof. It follows from Lemma 17 that $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ is bounded. So there exists a subsequence $\{\mathbf{v}^{(k_j)}\}_{j \in \mathbb{N}_+}$ of $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ converges to some $\mathbf{v}^* \in \mathbb{R}^N$. It follows from Lemma 18 that $\mathbf{v}^* \in \Omega$. By the continuity of f on \mathbb{R}^N , we have $\lim_{j \rightarrow \infty} f(\mathbf{v}^{(k_j)}) = f(\mathbf{v}^*)$. We also know that $\iota_\Omega(\mathbf{v}^*) = \iota_\Omega(\mathbf{v}^{(k)}) = 0$ for all $k \in \mathbb{N}$. Therefore, $\lim_{j \rightarrow \infty} F(\mathbf{v}^{(k_j)}) = F(\mathbf{v}^*)$, which completes the proof. \square

To employ Proposition 15 for the convergence of PGA. We still need to show that F satisfies the KL property at \mathbf{v}^* . To this end, we recall the notions of semi-algebraic sets and functions, and recall a known result in [4, 9] that establishes the relation between semi-algebraic property and KL property as Lemma 20.

Definition 4 (Semi-algebraic sets and functions) A subset $\mathcal{S} \subset \mathbb{R}^n$ is called real semi-algebraic if it can be represented by

$$\mathcal{S} = \bigcup_{j=1}^s \bigcap_{i=1}^t \{ \mathbf{x} \in \mathbb{R}^n \mid p_{ij}(\mathbf{x}) = 0, q_{ij}(\mathbf{x}) < 0 \}, \quad (\text{A.18})$$

where p_{ij} and q_{ij} are real polynomial functions for $i \in \mathbb{N}_t$, $j \in \mathbb{N}_s$, for some $s, t \in \mathbb{N}_+$. A function $\psi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is called semi-algebraic if its graph $\{(\mathbf{x}, \psi(\mathbf{x})) : \mathbf{x} \in \text{dom } \psi\}$ is a semi-algebraic subset of \mathbb{R}^{n+1} .

Lemma 20 Let $\psi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be a proper lower semicontinuous function. If ψ is semi-algebraic, then it satisfies the KL property at any point of $\text{dom } \partial\psi := \{\mathbf{u} \in \mathbb{R}^n \mid \partial\psi(\mathbf{u}) \neq \emptyset\}$.

Proposition 21 Let $F := f + \iota_\Omega$, where function ι_Ω and f are defined by (3.6) and (3.7), respectively. Then $\text{dom } \partial F = \Omega$ and F is semi-algebraic.

Proof. We first prove that $\text{dom } \partial F = \Omega$. Note that $\text{dom } F = \Omega$ is closed, which together with the definition of limiting-subdifferential implies that $\text{dom } \partial F \subset \Omega$. For any $\mathbf{v} \in \Omega$, it is easy to verify that $\mathbf{0}_N \in \hat{\partial}\iota_\Omega(\mathbf{v})$. Then we see from Fact 12 that $\mathbf{0}_N \in \partial\iota_\Omega(\mathbf{v})$. Now by Fact 13, we have $\nabla f(\mathbf{v}) \in \partial F(\mathbf{v})$, which means that $\partial F(\mathbf{v}) \neq \emptyset$, that is, $\mathbf{v} \in \text{dom } \partial F$. Hence $\Omega \subset \text{dom } \partial F$. This proves $\text{dom } \partial F = \Omega$.

We then prove that F is semi-algebraic. From the definition of semi-algebraic function, we are easy to see that the sum of semi-algebraic functions is still semi-algebraic. It is obvious that function f is semi-algebraic. To prove that F is semi-algebraic, it suffices to show that ι_Ω is semi-algebraic. The graph of ι_Ω is given by

$$\text{gra } \iota_\Omega = \{ \mathbf{x} \in \mathbb{R}^{N+1} \mid \mathbf{x}_{1:N} \geq \mathbf{0}_N, \|\mathbf{x}_{1:N}\|_0 \leq m \text{ and } x_{N+1} = 0 \}, \quad (\text{A.19})$$

where $\mathbf{x}_{1:N} := (x_1, x_2, \dots, x_N)^\top$. Note that there are $K := \binom{N}{N-m}$ combinations to choose an index set with $(N-m)$ elements out of the set $\{1, 2, \dots, N\}$. We denote these index sets with size $(N-m)$ by J_1, J_2, \dots, J_K , and let $\tilde{J}_i := J_i \cup \{N+1\}$, $i \in \mathbb{N}_K$. Then the graph of ι_Ω in (A.19) can be represented by

$$\text{gra } \iota_\Omega = \bigcup_{j=1}^K \left[\left(\bigcap_{i \in \tilde{J}_j} \{ \mathbf{x} \in \mathbb{R}^{N+1} \mid x_i = 0 \} \right) \cap \left(\bigcap_{i \notin \tilde{J}_j} \{ \mathbf{x} \in \mathbb{R}^{N+1} \mid -x_i \leq 0 \} \right) \right],$$

which implies that ι_Ω is a semi-algebraic function. This completes the proof. \square

We show in the following proposition that the objective function $F := f + \iota_\Omega$ satisfies the KL property at any accumulation point of sequence $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$.

Proposition 22 Let $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ be generated by PGA. If \mathbf{v}^* is an accumulation point of $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$, then F satisfies the KL property at \mathbf{v}^* .

Proof. Since \mathbf{v}^* is an accumulation point of $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$, it follows from Lemma 18 that $\mathbf{v}^* \in \Omega$. By Proposition 21, we know that F is semi-algebraic and $\mathbf{v}^* \in \text{dom } \partial F$. Thus the desired result follows from Lemma 20 immediately. \square

We then show the continuity of the proximity operator $\text{prox}_{\iota_\Omega}$ in the following proposition, which is also required to prove the convergence of PGA.

Proposition 23 Let $\{\mathbf{x}^{(k)}\}_{k \in \mathbb{N}} \subset \mathbb{R}^N$ be a sequence converges to some $\mathbf{x}^* \in \Omega$, and let $\mathbf{h}^{(k)} = \text{prox}_{\iota_\Omega}(\mathbf{x}^{(k)})$ for $k \in \mathbb{N}$ and $\mathbf{h}^* = \text{prox}_{\iota_\Omega}(\mathbf{x}^*)$ be given according to Proposition 3. Then $\lim_{k \rightarrow \infty} \mathbf{h}^{(k)} = \mathbf{h}^*$.

Proof. Since $\mathbf{x}^* \in \Omega$, $m_{\mathbf{x}^*} \leq m$. We first consider the case $m_{\mathbf{x}^*} = 0$, that is, $x_j^* \leq 0$ for all $j \in \mathbb{N}_N$. Then $\mathbf{h}^* = \mathbf{0}_N$. For all $\varepsilon > 0$, we let $\delta_1 := \frac{\varepsilon}{\sqrt{N}}$. Then there exists $K_1 \in \mathbb{N}$ such that $\mathbf{x}_j^{(k)} \leq \delta_1$ for

all $j \in \mathbb{N}_N$ and $k \geq K_1$. By the definition of $\mathbf{h}^{(k)}$, we know that $0 \leq h_j^{(k)} \leq \delta_1$ for all $j \in \mathbb{N}_N$ and $k \geq K_1$. Hence

$$\|\mathbf{h}^{(k)} - \mathbf{h}^*\|_2 = \|\mathbf{h}^{(k)}\|_2 \leq \sqrt{N}\delta_1 = \varepsilon, \text{ for all } k \geq K_1,$$

which implies $\lim_{k \rightarrow \infty} \mathbf{h}^{(k)} = \mathbf{h}^*$.

We then consider the case $0 < m_{\mathbf{x}^*} \leq m$. In this case, the set $\{j \in \mathbb{N}_N \mid x_j^* > 0\}$ is nonempty. For all $\varepsilon > 0$, we let $\delta_2 := \min \left\{ \frac{1}{3}x_{\min\text{-pos}}^*, \frac{\varepsilon}{\sqrt{N}} \right\}$, where

$$x_{\min\text{-pos}}^* := \min_{j \in \mathbb{N}_N} \{x_j^* \mid x_j^* > 0\}.$$

There exists $K_2 \in \mathbb{N}$ such that for all $k \geq K_2$, $\|\mathbf{x}^{(k)} - \mathbf{x}^*\|_2 \leq \delta_2$, which indicates that

$$x_j^{(k)} \geq x_j^* - \delta_2 \geq \frac{2}{3}x_{\min\text{-pos}}^* > 0, \text{ for } j \in J_{\text{pos}}^{\mathbf{x}^*} \quad (\text{A.20})$$

and

$$x_j^{(k)} \leq x_j^* + \delta_2 \leq \delta_2 \leq \frac{1}{3}x_{\min\text{-pos}}^*, \text{ for } j \in \mathbb{N}_N \setminus J_{\text{pos}}^{\mathbf{x}^*}. \quad (\text{A.21})$$

By the fact $m_{\mathbf{x}^*} \leq m$ and the definitions of $\mathbf{h}^{(k)}$ and \mathbf{h}^* , we can conclude from (A.20) and (A.21) that for all $k \geq K_2$,

$$h_j^{(k)} = x_j^{(k)}, h_j^* = x_j^*, \text{ for } j \in J_{\text{pos}}^{\mathbf{x}^*}$$

and

$$0 \leq h_j^{(k)} \leq \delta_2, h_j^* = 0, \text{ for } j \in \mathbb{N}_N \setminus J_{\text{pos}}^{\mathbf{x}^*}.$$

Thus $\|\mathbf{h}^{(k)} - \mathbf{h}^*\|_2 \leq \sqrt{N}\delta_2 \leq \varepsilon$, which yields $\lim_{k \rightarrow \infty} \mathbf{h}^{(k)} = \mathbf{h}^*$. This completes the proof. \square

We are now in a position to utilize Proposition 15 and Theorem 2 to prove Theorem 5.

Proof of Theorem 5. We first prove item (i). According to Propositions 4, 16, 19 and 22, the convergence of $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ to a critical point $\mathbf{v}^* \in \Omega$ of $F := f + \iota_\Omega$ follows from Proposition 15 immediately. By item (ii) in Theorem 2, to prove that \mathbf{v}^* is a locally optimal solution of model (3.7) (or model (3.4)), it suffices to show that (3.8) holds. Since $\lim_{k \rightarrow \infty} \mathbf{v}^{(k)} = \mathbf{v}^*$, the Lipschitz continuity of ∇f yields that

$$\lim_{k \rightarrow \infty} \left(\mathbf{v}^{(k)} - \alpha \nabla f(\mathbf{v}^{(k)}) \right) = \mathbf{v}^* - \alpha \nabla f(\mathbf{v}^*).$$

Now by letting $k \rightarrow \infty$ on both side of (3.9) and employing Proposition 23, we obtain (3.8), which proves the convergence of $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ to a locally optimal solution \mathbf{v}^* of model (3.4).

We then prove the convergence rates of $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$. Let $\Phi_k(\mathbf{v}) := \|\mathbf{v} - \mathbf{v}^{(k)} + \alpha \nabla f(\mathbf{v}^{(k)})\|_2^2 + \iota_\Omega(\mathbf{v})$, $\mathbf{v} \in \mathbb{R}^N$. It is obvious that Φ_k is 2-strongly convex, since the Hessian matrix of function $\Phi_k - \|\cdot\|_2^2$ is positive semidefinite. To prove the convergence rate, we first show that there exists $K \in \mathbb{N}$ such that

$$\langle \nabla f(\mathbf{v}^*), \mathbf{v}^{(k)} - \mathbf{v}^* \rangle \geq 0 \text{ and } \langle \nabla \Phi_k(\mathbf{v}^{(k+1)}), \mathbf{v}^* - \mathbf{v}^{(k+1)} \rangle \geq 0, \quad (\text{A.22})$$

for all $k \geq K$. From the proof of Theorem 2 in Appendix A.2, we see that (A.9) holds. It has been shown that $\lim_{k \rightarrow \infty} \mathbf{v}^{(k)} = \mathbf{v}^*$ and $\mathbf{v}^{(k)} \in \Omega$ for all $k \in \mathbb{N}$ (see item (i) in Proposition 4). Then there exists $K_1 \in \mathbb{N}$ such that $\mathbf{v}^{(k)} \in B(\mathbf{v}^*; \delta) \cap \Omega$ for all $k \geq K_1$, which together with (A.9) implies that the first inequality in (A.22) holds for all $k \geq K_1$. According to the definition of Φ_k and (3.9), we see that

$$\mathbf{v}^{(k+1)} = \operatorname{argmin}_{\mathbf{v} \in \mathbb{R}^N} \frac{1}{2} \|\mathbf{v} - \mathbf{v}^{(k)} + \alpha \nabla f(\mathbf{v}^{(k)})\|_2^2 + \iota_\Omega(\mathbf{v}) = \operatorname{argmin}_{\mathbf{v} \in \Omega} \Phi_k(\mathbf{v}).$$

By a procedure similar to the first paragraph of the proof of Theorem 2, we can establish that

$$\mathbf{v}^{(k+1)} = \operatorname{prox}_{\iota_\Omega} \left(\mathbf{v}^{(k+1)} - \alpha \nabla \Phi_k(\mathbf{v}^{(k+1)}) \right).$$

Note that Φ_k is also strongly convex. Using a proof analogous to that of the first inequality in (A.22), we can establish the existence of $K_2 \in \mathbb{N}$ such that the second inequality in (A.22) holds for all $k \geq K_2$. By setting $K = \max\{K_1, K_2\}$, we conclude that (A.22) holds for all $k \geq K$.

It follows from Lemma 11 that

$$\Phi_k(\mathbf{v}^*) \geq \Phi_k(\mathbf{v}^{(k+1)}) + \langle \nabla \Phi_k(\mathbf{v}^{(k+1)}), \mathbf{v}^* - \mathbf{v}^{(k+1)} \rangle + \|\mathbf{v}^* - \mathbf{v}^{(k+1)}\|_2^2, \quad (\text{A.23})$$

which together with the second inequality in (A.22) implies that

$$\Phi_k(\mathbf{v}^*) \geq \Phi_k(\mathbf{v}^{(k+1)}) + \|\mathbf{v}^* - \mathbf{v}^{(k+1)}\|_2^2,$$

that is,

$$\|\mathbf{v}^* - \mathbf{v}^{(k)} + \alpha \nabla f(\mathbf{v}^{(k)})\|_2^2 \geq \|\mathbf{v}^{(k+1)} - \mathbf{v}^{(k)} + \alpha \nabla f(\mathbf{v}^{(k)})\|_2^2 + \|\mathbf{v}^* - \mathbf{v}^{(k+1)}\|_2^2, \quad (\text{A.24})$$

for all $k \geq K$. For simplicity of notation, we define $\mathbf{z}^{(k)} := \mathbf{v}^{(k+1)} - \mathbf{v}^{(k)}$, $k \in \mathbb{N}$. Expanding (A.24) and dividing the resulting inequality by 2α yields

$$\langle \nabla f(\mathbf{v}^{(k)}), \mathbf{z}^{(k)} \rangle + \frac{1}{2\alpha} \|\mathbf{z}^{(k)}\|_2^2 + \frac{1}{2\alpha} \|\mathbf{v}^* - \mathbf{v}^{(k+1)}\|_2^2 \leq \frac{1}{2\alpha} \|\mathbf{v}^* - \mathbf{v}^{(k)}\|_2^2 + \langle \nabla f(\mathbf{v}^{(k)}), \mathbf{v}^* - \mathbf{v}^{(k)} \rangle, \quad (\text{A.25})$$

for all $k \geq K$. Recall that (A.14) in the proof of Proposition 4 holds. Since $\alpha \in \left(0, \frac{1}{\|\mathbf{Q}_\varepsilon\|_2}\right)$, (A.14) implies that

$$f(\mathbf{v}^{(k+1)}) - f(\mathbf{v}^{(k)}) \leq \langle \nabla f(\mathbf{v}^{(k)}), \mathbf{z}^{(k)} \rangle + \frac{1}{2\alpha} \|\mathbf{z}^{(k)}\|_2^2, \quad \text{for all } k \in \mathbb{N}. \quad (\text{A.26})$$

Combining (A.26) and (A.25), we obtain that

$$f(\mathbf{v}^{(k+1)}) - f(\mathbf{v}^{(k)}) + \frac{1}{2\alpha} \|\mathbf{v}^* - \mathbf{v}^{(k+1)}\|_2^2 \leq \frac{1}{2\alpha} \|\mathbf{v}^* - \mathbf{v}^{(k)}\|_2^2 + \langle \nabla f(\mathbf{v}^{(k)}), \mathbf{v}^* - \mathbf{v}^{(k)} \rangle, \quad (\text{A.27})$$

for all $k \geq K$. It follows from the first order condition for convexity (Proposition B.3 of [7]) that

$$f(\mathbf{v}^*) \geq f(\mathbf{v}^{(k)}) + \langle \nabla f(\mathbf{v}^{(k)}), \mathbf{v}^* - \mathbf{v}^{(k)} \rangle,$$

which together with (A.27) yields

$$f(\mathbf{v}^{(k+1)}) + \frac{1}{2\alpha} \|\mathbf{v}^* - \mathbf{v}^{(k+1)}\|_2^2 \leq f(\mathbf{v}^*) + \frac{1}{2\alpha} \|\mathbf{v}^* - \mathbf{v}^{(k)}\|_2^2,$$

that is,

$$f(\mathbf{v}^{(k+1)}) - f(\mathbf{v}^*) \leq \frac{1}{2\alpha} \left(\|\mathbf{v}^{(k)} - \mathbf{v}^*\|_2^2 - \|\mathbf{v}^{(k+1)} - \mathbf{v}^*\|_2^2 \right), \quad \text{for all } k \geq K.$$

We see from (A.11) that $\{f(\mathbf{v}^{(k)})\}_{k \in \mathbb{N}}$ is monotonically decreasing. Then for all $j \in \mathbb{N}_+$,

$$\begin{aligned} j \left(f(\mathbf{v}^{(K+j)}) - f(\mathbf{v}^*) \right) &\leq \sum_{i=K}^{K+j-1} \left(f(\mathbf{v}^{(i+1)}) - f(\mathbf{v}^*) \right) \\ &\leq \frac{1}{2\alpha} \sum_{i=K}^{K+j-1} \left(\|\mathbf{v}^{(i)} - \mathbf{v}^*\|_2^2 - \|\mathbf{v}^{(i+1)} - \mathbf{v}^*\|_2^2 \right) \\ &= \frac{1}{2\alpha} \left(\|\mathbf{v}^{(K)} - \mathbf{v}^*\|_2^2 - \|\mathbf{v}^{(K+j)} - \mathbf{v}^*\|_2^2 \right). \end{aligned}$$

Hence

$$f(\mathbf{v}^{(K+j)}) - f(\mathbf{v}^*) \leq \frac{1}{2\alpha j} \|\mathbf{v}^{(K)} - \mathbf{v}^*\|_2^2. \quad (\text{A.28})$$

Note that $\|\mathbf{v}^{(K)} - \mathbf{v}^*\|_2^2$ is a constant. Let $k = K + j$ and $C := \frac{1}{\alpha} \|\mathbf{v}^{(K)} - \mathbf{v}^*\|_2^2$. Then (A.28) implies that

$$f(\mathbf{v}^{(k)}) - f(\mathbf{v}^*) \leq C \cdot \frac{1}{2j} \leq C \cdot \frac{1}{k}, \quad \text{for } j \geq K.$$

Thus

$$|f(\mathbf{v}^{(k)}) - f(\mathbf{v}^*)| = O\left(\frac{1}{k}\right). \quad (\text{A.29})$$

Combining Lemma 11 and the first inequality in (A.22), we obtain that

$$f(\mathbf{v}^{(k)}) - f(\mathbf{v}^*) \geq \frac{\epsilon}{2} \|\mathbf{v}^{(k)} - \mathbf{v}^*\|_2^2, \text{ for all } k \geq K.$$

This together with (A.29) yields that $\|\mathbf{v}^{(k)} - \mathbf{v}^*\|_2 = O\left(\frac{1}{\sqrt{k}}\right)$.

We next prove item (ii). The fact $\mathbf{v}^* \geq \mathbf{0}_N$ follows from (3.8) and Proposition 3 directly. Assume, to reach a contradiction, that $\mathbf{v}^* = \mathbf{0}_N$. Item (i) in this theorem shows that \mathbf{v}^* is a locally optimal solution of model (3.7). Then there exists $\delta > 0$ such that

$$f(\mathbf{v}) \geq f(\mathbf{v}^*) = f(\mathbf{0}_N) = 0, \text{ for all } \mathbf{v} \in \Omega \cap B(\mathbf{v}^*; \delta). \quad (\text{A.30})$$

We recall from the assumption of this theorem that there exists $\tilde{\mathbf{w}} \in \Omega$ such that $\mathbf{p}^\top \tilde{\mathbf{w}} > 0$. Let $\tilde{\mathbf{w}}_\alpha := \alpha \tilde{\mathbf{w}}$, where $\alpha > 0$. Then $\tilde{\mathbf{w}}_\alpha \in \Omega$ and $\mathbf{p}^\top \tilde{\mathbf{w}}_\alpha > 0$. Note that when α tends to 0, the quadratic term $\frac{1}{2} \tilde{\mathbf{w}}_\alpha^\top \mathbf{Q}_\epsilon \tilde{\mathbf{w}}_\alpha$ is of higher order infinitesimal than the linear term $\mathbf{p}^\top \tilde{\mathbf{w}}_\alpha$. There exists some sufficient small $\alpha > 0$ such that $\tilde{\mathbf{w}}_\alpha \in B(\mathbf{v}^*; \delta)$ and

$$f(\tilde{\mathbf{w}}_\alpha) = \frac{1}{2} \tilde{\mathbf{w}}_\alpha^\top \mathbf{Q}_\epsilon \tilde{\mathbf{w}}_\alpha - \mathbf{p}^\top \tilde{\mathbf{w}}_\alpha < 0,$$

which contradicts (A.30). Therefore, $\mathbf{v}^* \neq \mathbf{0}_N$.

Lastly, we prove item (iii). Since $\mathbf{v}^* \geq \mathbf{0}_N$ and $\mathbf{v}^* \neq \mathbf{0}_N$. There exist $\varepsilon_0 > 0$ and $K' \in \mathbb{N}$ such that $(\mathbf{v}^*)^\top \mathbf{1}_N > \varepsilon_0$ and $(\mathbf{v}^{(k)})^\top \mathbf{1}_N > \varepsilon_0$ for all $k \geq K'$, and hence $\mathbf{w}^{(k)} = \frac{\mathbf{v}^{(k)}}{(\mathbf{v}^{(k)})^\top \mathbf{1}_N}$ for all $k \geq K'$.

Note that \mathbf{v}^* and $\{\mathbf{v}^{(k)}\}_{k \in \mathbb{N}}$ are both bounded. There exist $C_1 > 0$ and $C_2 > 0$ such that

$$\frac{\|\mathbf{v}^{(k)}\|_2}{|(\mathbf{v}^{(k)})^\top \mathbf{1}_N| \cdot |(\mathbf{v}^*)^\top \mathbf{1}_N|} \leq C_1 \text{ and } C_1 \sqrt{N} + \left| \frac{1}{(\mathbf{v}^*)^\top \mathbf{1}_N} \right| \leq C_2.$$

Then for all $k \geq K'$,

$$\begin{aligned} \left\| \mathbf{w}^{(k)} - \frac{\mathbf{v}^{(k)}}{(\mathbf{v}^*)^\top \mathbf{1}_N} \right\|_2 &\leq \left| \frac{1}{(\mathbf{v}^{(k)})^\top \mathbf{1}_N} - \frac{1}{(\mathbf{v}^*)^\top \mathbf{1}_N} \right| \cdot \|\mathbf{v}^{(k)}\|_2 \\ &= \frac{\|\mathbf{v}^{(k)}\|_2}{|(\mathbf{v}^{(k)})^\top \mathbf{1}_N| \cdot |(\mathbf{v}^*)^\top \mathbf{1}_N|} \cdot |(\mathbf{v}^{(k)} - \mathbf{v}^*)^\top \mathbf{1}_N| \\ &\leq C_1 \sqrt{N} \|\mathbf{v}^{(k)} - \mathbf{v}^*\|_2, \end{aligned}$$

and hence

$$\begin{aligned} \|\mathbf{w}^{(k)} - \mathbf{w}^*\|_2 &= \left\| \mathbf{w}^{(k)} - \frac{\mathbf{v}^{(k)}}{(\mathbf{v}^*)^\top \mathbf{1}_N} + \frac{\mathbf{v}^{(k)}}{(\mathbf{v}^*)^\top \mathbf{1}_N} - \mathbf{w}^* \right\|_2 \\ &\leq C_1 \sqrt{N} \|\mathbf{v}^{(k)} - \mathbf{v}^*\|_2 + \left| \frac{1}{(\mathbf{v}^*)^\top \mathbf{1}_N} \right| \|\mathbf{v}^{(k)} - \mathbf{v}^*\|_2 \\ &\leq C_2 \|\mathbf{v}^{(k)} - \mathbf{v}^*\|_2. \end{aligned}$$

This implies that $\|\mathbf{w}^{(k)} - \mathbf{w}^*\|_2 = O\left(\frac{1}{\sqrt{k}}\right)$. Similarly, we can prove that there exists $C_3 > 0$ such that

$$|S(\mathbf{w}^{(k)}) - S(\mathbf{w}^*)| \leq C_3 \|\mathbf{w}^{(k)} - \mathbf{w}^*\|_2, \text{ for all } k \geq K',$$

which implies that $|S(\mathbf{w}^{(k)}) - S(\mathbf{w}^*)| = O\left(\frac{1}{\sqrt{k}}\right)$. This completes the proof. \square

A.6 Proof of Theorem 6

To prove Theorem 6, we recall Proposition 11.4 of [6] as the following lemma.

Lemma 24 *Let $\psi : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ be proper and convex. Then every local minimizer of ψ is a global minimizer.*

Proof of Theorem 6. We first prove item (i). Let $\iota_{\hat{\Omega}}$ be defined by

$$\iota_{\hat{\Omega}}(\mathbf{v}) := \begin{cases} 0, & \text{if } \mathbf{v} \in \hat{\Omega}; \\ +\infty, & \text{otherwise.} \end{cases}$$

Then model (3.10) is equivalent to $\min_{\mathbf{v} \in \mathbb{R}^N} \hat{F}(\mathbf{v})$, where $\hat{F} := f + \iota_{\hat{\Omega}}$. Of course, $\iota_{\hat{\Omega}}$ is proper. The convexity of $\hat{\Omega}$ implies that $\iota_{\hat{\Omega}}$ is convex (see Example 8.3 of [6]). Recall that f is strictly convex, and hence \hat{F} is proper and strictly convex. Since \mathbf{v}^* is a locally optimal solution of model (3.7) and $\hat{\Omega} \subset \Omega$, there exists $\delta > 0$ such that

$$f(\mathbf{u}) \geq f(\mathbf{v}^*), \quad \text{for all } \mathbf{u} \in B(\mathbf{v}^*; \delta) \cap \hat{\Omega}. \quad (\text{A.31})$$

The fact $\mathbf{v}^* \in \hat{\Omega}$ gives $\iota_{\hat{\Omega}}(\mathbf{v}^*) = 0$. Then (A.31) implies that $\hat{F}(\mathbf{u}) \geq \hat{F}(\mathbf{v}^*)$ for all $\mathbf{u} \in B(\mathbf{v}^*; \delta)$, that is, \mathbf{v}^* is a local minimizer of \hat{F} . Now it follows from Lemma 24 that \mathbf{v}^* is also a global minimizer of \hat{F} . The strict convexity of \hat{F} implies the uniqueness of its global minimizer.

We next prove item (ii). From item (i) in this theorem and item (ii) in Theorem 5, we see that $\mathbf{v}^* \in \hat{\Omega}$ is a globally optimal solution of model (3.10) and $\mathbf{v}^* \neq \mathbf{0}_N$. Then we are able to prove that $\mathbf{p}^\top \mathbf{v}^* = (\mathbf{v}^*)^\top \mathbf{Q}_\epsilon \mathbf{v}^* > 0$. We omit this proof here since it is very similar to the proof of Lemma 8. Now we have $\mathbf{v}^* = \frac{\mathbf{p}^\top \mathbf{v}^*}{(\mathbf{v}^*)^\top \mathbf{Q}_\epsilon \mathbf{v}^*} \mathbf{v}^*$. For any $\mathbf{v} \in \hat{\Omega}$ such that $\mathbf{p}^\top \mathbf{v} > 0$, we let $\mathbf{u} := \frac{\mathbf{p}^\top \mathbf{v}}{\mathbf{v}^\top \mathbf{Q}_\epsilon \mathbf{v}} \mathbf{v}$. Then $\mathbf{u} \in \hat{\Omega}$ and $\mathbf{p}^\top \mathbf{u} > 0$. Hence

$$-\frac{1}{2} \frac{(\mathbf{p}^\top \mathbf{v}^*)^2}{(\mathbf{v}^*)^\top \mathbf{Q}_\epsilon \mathbf{v}^*} = \frac{1}{2} (\mathbf{v}^*)^\top \mathbf{Q}_\epsilon \mathbf{v}^* - \mathbf{p}^\top \mathbf{v}^* \leq \frac{1}{2} \mathbf{u}^\top \mathbf{Q}_\epsilon \mathbf{u} - \mathbf{p}^\top \mathbf{u} = -\frac{1}{2} \frac{(\mathbf{p}^\top \mathbf{v})^2}{\mathbf{v}^\top \mathbf{Q}_\epsilon \mathbf{v}},$$

which implies that

$$\frac{\mathbf{p}^\top \mathbf{v}^*}{\sqrt{(\mathbf{v}^*)^\top \mathbf{Q}_\epsilon \mathbf{v}^*}} \geq \frac{\mathbf{p}^\top \mathbf{v}}{\sqrt{\mathbf{v}^\top \mathbf{Q}_\epsilon \mathbf{v}}}, \quad \text{for all } \mathbf{v} \in \hat{\Omega}.$$

Since $\mathbf{v}^* \in \hat{\Omega}$ and $\mathbf{v}^* \neq \mathbf{0}_N$, by the definition of $\hat{\Omega}_1$, we see that $\mathbf{w}^* \in \hat{\Omega}_1$. Note that $\hat{\Omega}_1 \subset \hat{\Omega}$. For all $\mathbf{w} \in \hat{\Omega}_1$,

$$\frac{\mathbf{p}^\top \mathbf{w}}{\sqrt{\mathbf{w}^\top \mathbf{Q}_\epsilon \mathbf{w}}} \leq \frac{\mathbf{p}^\top \mathbf{v}^*}{\sqrt{(\mathbf{v}^*)^\top \mathbf{Q}_\epsilon \mathbf{v}^*}} = \frac{\mathbf{p}^\top \mathbf{w}^*}{\sqrt{(\mathbf{w}^*)^\top \mathbf{Q}_\epsilon \mathbf{w}^*}},$$

which implies that \mathbf{w}^* is a globally optimal solution of model $\max_{\mathbf{w} \in \hat{\Omega}_1} S(\mathbf{w})$.

Lastly, we prove item (iii). Note that $w_j^* > 0$ for all $j \in J_{\text{pos}}^{v^*}$. Let $w_{\text{min-pos}}^* := \min_{j \in J_{\text{pos}}^{v^*}} \{w_j^*\}$, $\delta := \frac{1}{3} w_{\text{min-pos}}^*$, and let \mathbf{w} be any vector in $B(\mathbf{w}^*; \delta) \cap \Omega_1$. Then $w_j \geq 2\delta > 0$ for $j \in J_{\text{pos}}^{v^*}$ and $|w_j| \leq \delta$ for $j \in \mathbb{N}_N \setminus J_{\text{pos}}^{v^*}$. Since $\mathbf{w} \in \Omega_1$ and $m_{v^*} = m$, we conclude that $w_j = 0$ for $j \in \mathbb{N}_N \setminus J_{\text{pos}}^{v^*}$, which implies that $\mathbf{w} \in \hat{\Omega}_1$. It has been shown that \mathbf{w}^* is an optimal solution of model $\max_{\mathbf{w} \in \hat{\Omega}_1} S(\mathbf{w})$.

Therefore, $S(\mathbf{w}) \leq S(\mathbf{w}^*)$ for all $\mathbf{w} \in B(\mathbf{w}^*; \delta) \cap \Omega_1$, that is, \mathbf{w}^* is a locally optimal solution of model (3.3). This completes the proof. \square

A.7 Proof of Theorem 7

Proof. According to Theorem 1 and item (i) in Theorem 2, to prove the desired result, it suffices to show that

$$\mathbf{v}^* = \text{prox}_{\iota_{\Omega}} \left(\mathbf{v}^* - \frac{1}{\epsilon} \nabla f(\mathbf{v}^*) \right). \quad (\text{A.32})$$

From the proof of Theorem 5, we know that (3.8) holds. According to the computation of $\text{prox}_{\iota_{\Omega}}$ in Proposition 3, to guarantee the validity of (3.8), we have $\nabla_i f(\mathbf{v}^*) = 0$ for all $i \in \text{supp}(\mathbf{v}^*)$. Otherwise, there exists some $i_0 \in \text{supp}(\mathbf{v}^*)$ such that $v_{i_0}^* \neq v_{i_0}^* - \alpha \nabla_{i_0} f(\mathbf{v}^*)$, which together with Proposition 3 implies that $\mathbf{v}^* \neq \text{prox}_{\iota_{\Omega}}(\mathbf{v}^* - \alpha \nabla f(\mathbf{v}^*))$, a contradiction to (3.8).

Suppose that $m_{v^*} < m$. Then we have $\nabla_i f(\mathbf{v}^*) \geq 0$ for all $i \in \mathbb{N}_N \setminus \text{supp}(\mathbf{v}^*)$. Otherwise, there exists some $i_1 \in \mathbb{N}_N \setminus \text{supp}(\mathbf{v}^*)$ such that $v_{i_1}^* - \alpha \nabla_{i_1} f(\mathbf{v}^*) > 0$. Note that $m_{v^*} < m$. The operation

of $\text{prox}_{\iota_\Omega}$ will preserve the positive value $v_{i_1}^* - \alpha \nabla_{i_1} f(\mathbf{v}^*)$ instead of truncating it as 0, which violates (3.8). In this case, we now have $v_i^* - \frac{1}{\epsilon} \nabla_i f(\mathbf{v}^*) = v_i^*$ for $i \in \text{supp}(\mathbf{v}^*)$ and $v_i^* - \frac{1}{\epsilon} \nabla_i f(\mathbf{v}^*) \leq 0$ for $i \in \mathbb{N}_N \setminus \text{supp}(\mathbf{v}^*)$, which imply that (A.32) holds.

Suppose that item (ii) holds. Let $\delta := \min\{v_i^* | i \in \text{supp}(\mathbf{v}^*)\} > 0$. For $i \in \mathbb{N}_N \setminus \text{supp}(\mathbf{v}^*)$, since $\frac{1}{\epsilon} \nabla_i f(\mathbf{v}^*) > -\delta$ and $v_i = 0$, we have $v_i^* - \frac{1}{\epsilon} \nabla_i f(\mathbf{v}^*) < \delta$. Note that $m_{\mathbf{v}^*} = m$. The operation of $\text{prox}_{\iota_\Omega}$ makes $v_i^* - \frac{1}{\epsilon} \nabla_i f(\mathbf{v}^*) = v_i^*$ for $i \in \text{supp}(\mathbf{v}^*)$ and $v_i^* - \frac{1}{\epsilon} \nabla_i f(\mathbf{v}^*) = 0$ for $i \in \mathbb{N}_N \setminus \text{supp}(\mathbf{v}^*)$, that is, (A.32) holds. This completes the proof. \square

A.8 Validation of PGA's Global Optimality Through Simulation Experiments

To test the validation of PGA's global optimality, we conduct a set of simulation experiments by considering model (3.7), where $\mathbf{Q}_\epsilon := \mathbf{Q}^\top \mathbf{Q} + \epsilon \mathbf{I}$. The iterative scheme of PGA for solving this model is given by (3.9) with $\alpha = \frac{0.99}{\|\mathbf{Q}\|_2}$.

In the simulation experiments, we set $\Sigma \in \mathbb{R}^{10 \times 10}$ by $\Sigma_{ij} := 0.5^{|i-j|}$, and randomly generate a matrix $\mathbf{Q} \in \mathbb{R}^{50 \times 10}$ from the multivariate normal distribution, with mean vector $\mathbf{0}_{10}$ and covariance matrix Σ . We set \mathbf{p} as a random vector with components that are randomly generated numbers in the range $[-10, 10]$, and casually set $\epsilon = 0.001$ and the sparsity $m = 3$.

The direct exhaustive approach enumerates all possible support set configurations, totaling $C_{10}^3 = 120$ cases. In each case, we solve a 3-dimension quadratic programming problem. By comparing the optimal solutions corresponding to these 120 cases, we obtain the exact globally optimal solution of model (3.7). After that, we can evaluate the optimality of PGA's convergence.

For each experiment, we performed 500 iterations of PGA. To ensure the robustness of our findings, we used three different initializations: $\mathbf{0}_N$, $\mathbf{1}_N/N$ and $\mathbf{1}_N$. We repeated the experiments 10^4 times for each initialization, with different \mathbf{Q} and \mathbf{p} in each run. We found that in over 7,200 of the 10^4 trials, for any of the three initializations, both the normalized error of the iterative sequence $\|\mathbf{v}^{(k)} - \mathbf{v}^*\|_2 / \|\mathbf{v}^*\|_2$ and the normalized error of the function value $|f(\mathbf{v}^{(k)}) - f(\mathbf{v}^*)| / |f(\mathbf{v}^*)|$ were smaller than 10^{-10} . Here \mathbf{v}^* denotes the globally optimal solution, and $\mathbf{v}^{(k)}$ represents the iterative sequence at the k -th iteration of PGA. We show the plots of $\|\mathbf{v}^{(k)} - \mathbf{v}^*\|_2 / \|\mathbf{v}^*\|_2$ and $|f(\mathbf{v}^{(k)}) - f(\mathbf{v}^*)| / |f(\mathbf{v}^*)|$ in the following Figure 1, and show in Table 4 these two normalized errors obtained at 500 iterations of PGA in ten simulation experiments.

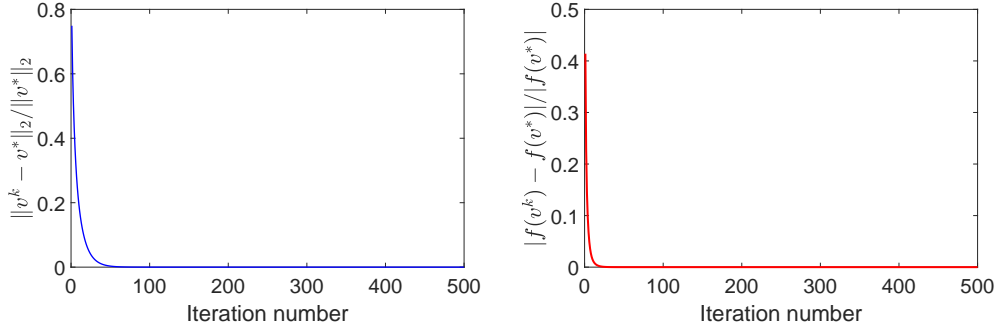


Figure 1: Simulation results of PGA for model (3.7). Left: normalized error of the iterative sequence versus number of iterations. Right: normalized error of function value versus number of iterations.

Table 4: The normalized errors obtained at 500 iterations of PGA in 10 simulation experiments.

k	1	2	3	4	5	6	7	8	9	10
$\frac{\ \mathbf{v}^{(k)} - \mathbf{v}^*\ _2}{\ \mathbf{v}^*\ _2}$	2.45E-16	1.62E-16	3.03E-16	6.23E-01	1.14E-16	7.68E-16	7.41E-16	9.55E-17	1.50E-16	2.52E-16
$\frac{ f(\mathbf{v}^{(k)}) - f(\mathbf{v}^*) }{ f(\mathbf{v}^*) }$	0.00	1.48E-16	1.68E-16	1.17E-01	0.00	1.40E-16	4.73E-16	1.36E-16	0.00	0.00

From the simulation experiments, we conclude that the proposed PGA has a high probability (over 72%) of directly converging to a globally optimal solution of model (3.7). This finding is consistent with the sufficient conditions for global optimality in Theorem 7.

A.9 Solving Algorithm: mSSRM-PGA

Algorithm A1 mSSRM-PGA

Input: Given the sample asset return matrix $\mathbf{R} \in \mathbb{R}^{T \times N}$ and the positive parameter ϵ .

Preparation: Let $\mathbf{p} = \frac{1}{T} \mathbf{R}^\top \mathbf{1}_T$, $\mathbf{Q} = \frac{1}{\sqrt{T-1}} (\mathbf{R} - \frac{1}{T} \mathbf{1}_T \mathbf{1}_T^\top \mathbf{R})$ and $\mathbf{Q}_\epsilon = \mathbf{Q}^\top \mathbf{Q} + \epsilon \mathbf{I}$. Compute the largest eigenvalue λ_1 of \mathbf{Q}_ϵ , and set $\alpha = \frac{0.999}{\lambda_1}$.

Initialization: Set $\mathbf{v}^{(0)} = \mathbf{p}$, $tol = 10^{-5}$, $MaxIter = 10^4$ and $k = 0$.

repeat

1. $\mathbf{v}^{(k+1)} = \text{prox}_{\ell_\Omega} (\mathbf{v}^{(k)} - \alpha (\mathbf{Q}_\epsilon \mathbf{v}^{(k)} - \mathbf{p}))$

2. $k = k + 1$

until $\frac{\|\mathbf{v}^{(k)} - \mathbf{v}^{(k-1)}\|_2}{\|\mathbf{v}^{(k-1)}\|_2} \leq tol$ or $k > MaxIter$.

if $\mathbf{v}^{(k)} \neq \mathbf{0}_N$

3. $\mathbf{w}^* = \frac{\mathbf{v}^{(k)}}{(\mathbf{v}^{(k)})^\top \mathbf{1}_N}$

else

4. $\mathbf{w}^* = \mathbf{0}_N$

Output: The portfolio \mathbf{w}^* .

A.10 Additional Experimental Results

The 1/N strategy rebalances to the equally-weighted portfolio on each trading time. S1, S2 and S3 are different versions of SSPO- ℓ_0 in (2.5), among which S1 is deterministic but S2 and S3 are randomized. The hyperparameters of these competitors are set according to the original papers.

FF25 contains 25 portfolios developed by BE/ME (book equity to market equity) and investment in the US market. FF25EU contains 25 portfolios developed by ME and prior return in the European market. FF32 contains 32 portfolios developed by BE/ME and investment in the US market. FF49 contains 49 industry portfolios in the US market. FF100 contains 100 portfolios developed by ME and BE/ME, while FF100MEINV contains 100 portfolios developed by ME and investment, both in the US market. The information of these data sets are given in Table 5.

Table 5: Information of 6 real-world monthly benchmark data sets.

Data Set	Region	Time	Months	Assets
FF25	US	<i>Jul/1971 ~ May/2023</i>	623	25
FF25EU	EU	<i>Nov/1990 ~ May/2023</i>	391	25
FF32	US	<i>Jul/1971 ~ May/2023</i>	623	32
FF49	US	<i>Jul/1971 ~ May/2023</i>	623	49
FF100	US	<i>Jul/1971 ~ May/2023</i>	623	100
FF100MEINV	US	<i>Jul/1971 ~ May/2023</i>	623	100

There is a relaxation approach based on the semi-definite programming (SDP Relaxation, [20]) that intends to address nearly the same mSSRM model (3.3) of this paper, except for relaxing the cardinality constraint and the long-only constraint. Therefore, this method fails to control cardinality exactly and a simplex projection [16] should be implemented to ensure feasibility. Its experimental results are also provided in Table 6, which are not so good as those of mSSRM-PGA.

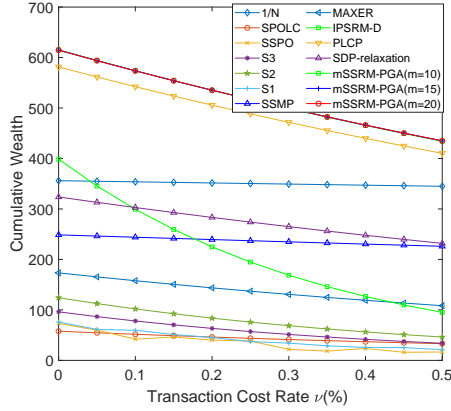
Table 6: Final cumulative wealths (CW) and Sharpe Ratios (SR) of SDP Relaxation and mSSRM-PGA on 6 data sets ($T = 60$).

Data Set	FF25		FF25EU		FRENCH32		FF49		FF100		FF100MEINV	
	CW	SR	CW	SR	CW	SR	CW	SR	CW	SR	CW	SR
SDP Relaxation	323.76	0.2340	14.25	0.1674	290.24	0.2224	280.46	0.2151	0.51	0.0218	194.09	0.1528
mSSRM-PGA (m=10)	615.34	0.2481	126.02	0.2712	991.89	0.2612	285.02	0.2151	527.09	0.2290	375.75	0.2217
mSSRM-PGA (m=15)	614.71	0.2481	125.19	0.2708	996.32	0.2615	262.54	0.2135	522.28	0.2289	383.44	0.2232
mSSRM-PGA (m=20)	614.70	0.2481	125.19	0.2708	996.23	0.2615	262.06	0.2134	515.50	0.2285	384.65	0.2234

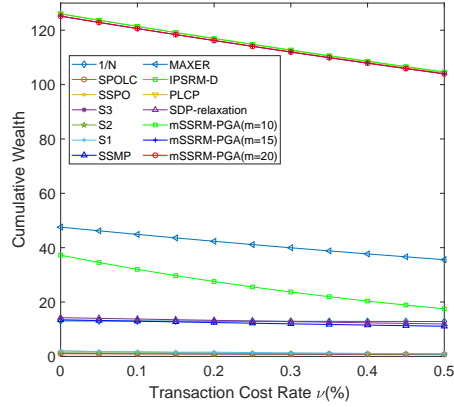
As for practical issues, Table 7 shows the running times for different methods with $T = 60$, where mSSRM-PGA achieves competitive computational efficiency. Figure 2 shows the final cumulative wealths of different methods as the transaction cost rate ν varies from 0 to 0.5% with $T = 60$, which indicates that mSSRM-PGA can withstand considerable levels of transaction cost rates.

Table 7: The average running times (in seconds) of different portfolio optimization models for one period on 6 data sets.

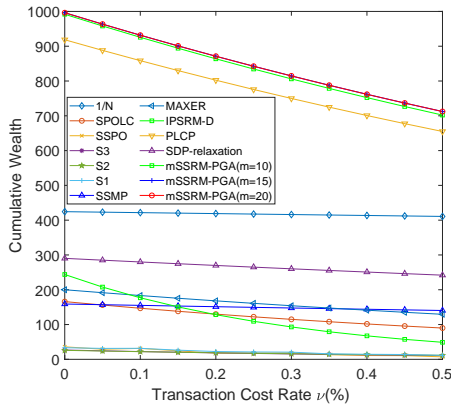
Data Set	SPOLC	SSPO	S1	S2	S3	SSMP	MAXER	IPSRM-D	PLCT	SDP Relaxation	mSSRM-PGA
FF25	0.0263	0.0234	5.72E-05	5.63E-05	8.46E-05	0.0122	0.0525	0.0009	0.0020	1.0115	0.0052
FF25EU	0.0222	0.0239	1.70E-05	2.89E-05	2.59E-05	0.0316	0.0588	0.0009	0.0015	0.8178	0.0059
FRENCH32	0.0239	0.0250	2.93E-05	2.81E-05	3.08E-05	0.0075	0.0392	0.0012	0.0021	1.2862	0.0075
FF49	0.0252	0.0458	2.94E-05	3.51E-05	2.82E-05	0.0083	0.0270	0.0029	0.0034	12.3780	0.0114
FF100	0.0306	0.0854	5.38E-05	4.48E-05	4.27E-05	0.0451	0.0458	0.0132	0.0052	24.5852	0.0713
FF100MEINV	0.0296	0.0864	5.08E-05	4.81E-05	4.53E-05	0.0145	0.0152	0.0144	0.0059	23.9911	0.0696



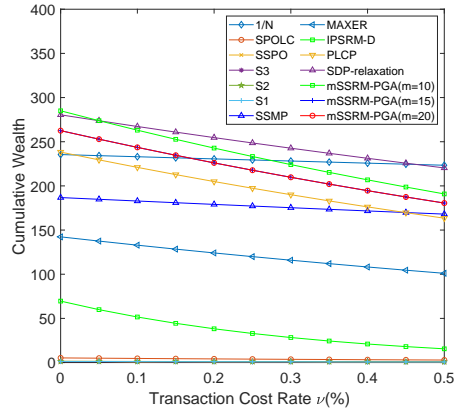
(a) FF25



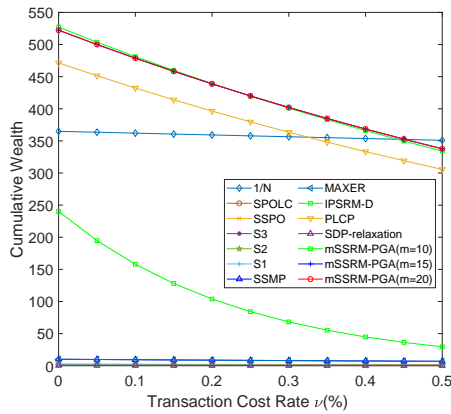
(b) FF25EU



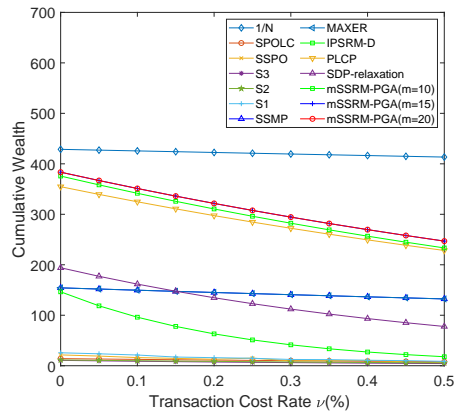
(c) FF32



(d) FF49



(e) FF100



(f) FF100MEINV

Figure 2: Final cumulative wealths of portfolio optimization methods w.r.t. transaction cost rate ν on 6 benchmark data sets.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The whole project with all the codes for reproductions would be made public if this paper were to be accepted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This work has no social risk of misuse.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

14. Crowdsourcing and Research with Human Subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

15. Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]