Journal of Machine Learning Research 26 (2025) 1-55 Submitted 7/23; Revised 1/24; Published 7/25

Sparse SVM with Hard-Margin Loss: a Newton-Augmented Lagrangian Method in Reduced Dimensions

Penghe Zhang PENGHE.ZHANG@POLYU.EDU.HK
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hung Hom, Hong Kong

Naihua Xiu NHXIU@BJTU.EDU.CN
School of Mathematics and Statistics
Beijing Jiaotong University
Beijing, China

Hou-Duo Qi* HOUDUO.QI@POLYU.EDU.HK
Department of Applied Mathematics
Department of Data Science and Artificial Intelligence
The Hong Kong Polytechnic University
Hung Hom, Hong Kong

Editor: John Shawe-Taylor

Abstract

The hard-margin loss function has been at the core of the support vector machine research from the very beginning due to its generalization capability. On the other hand, the cardinality constraint has been widely used for feature selection, leading to sparse solutions. This paper studies the sparse SVM with the hard-margin loss that integrates the virtues of both worlds, resulting in one of the most challenging models to solve. We cast the problem as a composite optimization with the cardinality constraint. We characterize its local minimizers in terms of pseudo KKT point that well captures the combinatorial structure of the problem, and investigate a sharper P-stationary point with a concise representation for algorithm design. We further develop an inexact proximal augmented Lagrangian method (iPAL). The different parts of the inexactness measurements from the P-stationarity are controlled at different scales in a way that the generated sequence converges both globally and at a linear rate. To make iPAL practically efficient, we propose a gradient-Newton method in a subspace for the iPAL subproblem. This is accomplished by detecting active samples and features with the help of the proximal operator of the hard margin loss and the projection of the cardinality constraint. Extensive numerical results on both simulated and real data sets demonstrate that the proposed method is fast, produces sparse solution of high accuracy, and can lead to effective reduction on active samples and features when compared with several leading solvers.

Keywords: support vector machine, hard-margin loss, sparse feature selection, P-stationary point, inexact proximal augmented Lagrangian method, Newton's method.

1. Introduction

This paper is concerned with one of the most challenging formulations in the study of support vector machines (SVM):

$$\min_{\mathbf{w}\in\mathbb{R}^n, b\in\mathbb{R}} \ \frac{1}{2}\|\mathbf{w}\|^2+\frac{1}{2}b^2+\lambda\sum_{i=1}^{m} h\Big(1-y_i(\langle\mathbf{w},\mathbf{x}_i\rangle+b)\Big), \ \text{ s.t. } \mathbb{S}:=\{\mathbf{w}\in\mathbb{R}^n \mid \|\mathbf{w}\|_0\le s\}, \quad (1)$$

where $\{(\mathbf{x}_i, y_i)\}_{i=1}^m$ are the sample data with $\mathbf{x}_i\in\mathbb{R}^n$ and $y_i\in\{1,-1\}$ being its label. The separating hyperplane is $\langle\mathbf{w},\mathbf{x}\rangle+b=0$ and the loss function is the hard-margin loss:

$$h(t)=\begin{cases} 1, & \text{if } t>0,\\ 0, & \text{if } t\le 0.\end{cases}$$

Furthermore, the model aims to seek a hyperplane of sparse features selected by the ℓ_0-norm $\|\cdot\|_0$ with a user-specified sparsity level $s\ge 1$ and $\mathbb{S}$ is known as the s-sparse set. Vapnik (1998) discussed the hard-margin loss (also known as the 0/1-loss), which is to construct the hyperplane that makes the smallest number of separating errors. However, the optimization of it is NP-complete. The use of ℓ_0-norm is getting popular in selecting sparse features. The first two terms in the objective is to maximize the separation gap in the $(\mathbf{w}, b)$ space rather than in the feature space of $\mathbf{w}$. This objective has been promoted by Mangasarian and his collaborators (see, Mangasarian and Musicant (2001); Fung and Mangasarian (2001); Lee and Mangasarian (2001)). Due to its strong convexity in both $\mathbf{w}$ and b, Newton's method has been the core of those studies for the ridge/hinge-loss function. The purpose of this paper is to extend Newton's method to the sparse SVM with hard-margin loss under the framework of augmented Lagrangian method with proved convergence. This section is organized as follows. We will first conduct a literature review, followed by an explanation of our numerical approach.

1.1 Related work

There exists extensive research on SVMs. We refer to Vapnik (1998); Cristianini et al. (2000); Smola and Schölkopf (2004); Steinwart and Christmann (2008); Chang and Lin (2011) for many of the models and the solvers. We restrict our review to the sparse SVM with the hard-margin loss and the related numerical methods. We split the papers into three groups. The first is the reformulation and relaxation approach. The second group is to treat (1) as a composite optimization and the augmented Lagrangian method is a natural choice. The last group is on Newton's method for such composite optimization.

(A) Reformulation and convex relaxation. The advantage of simultaneously addressing the 0/1-loss and the ℓ_0-norm for feature selection was thoroughly justified by Ustun and Rudin (2016) for a medical scoring problem. In this application, both the solution accuracy (controlled by the 0/1-loss) and solution sparsity (controlled by the ℓ_0-norm) are crucial to yield a reliable medical score. The solution method is to reformulate the problem as a mixed integer programming (MIP) by using the old trick: Big-M constraint on both the 0/1-loss and the ℓ_0-norm. We refer to Liittschwager and Wang (1978); Bajgier and Hill (1982); Brooks (2011) for earlier works along this line. Another trick for MIP reformulation

is based on the following fact:

$$\text{(complementarity reformulation)} \quad \|\mathbf{w}\|_0 = \min_{\mathbf{v}\in\mathbb{R}^n} \sum_{i=1}^{n} v_i, \;\; s.t.\; w_i(1-v_i) = 0,\; v_i \in [0,1], \tag{2}$$

see Feng et al. (2018); Kanzow et al. (2022). One potential drawback for the smooth approach is the drastic increase in the dimensionality, especially when Newton's method is applied, see Section 7.4 of Kanzow et al. (2022) for a numerical example. One can imagine that this drawback would get worse when the 0/1-loss is also represented by the complementarity reformulation. It is worth pointing out that exciting progress has been made in a recent MIP approach (e.g., via Big-M constraint) by Dedieu et al. (2021), who cleverly combines a continuous approach and MIP to develop a fast algorithm for an ℓ_0-norm minimization problem. It remains to be seen how the approach would be adapted to problem (1), which involves both ℓ_0-norm and the 0/1-loss.

More recently, Cui et al. (2023); Han et al. (2024) systematically studied a class of composite optimization involving 0/1-loss and developed several lifted reformulations, which are tractable with computationally verifiable stationary points. Among many is an important result that local minimizers of such nonconvex problems can be completely characterized in terms of epi-stationary points under a convex-like property. We will show that the concept of the epi-stationary point is intrinsically related to our proposed pseudo KKT points for (1). The development leads to one of the strong claims of this paper that our algorithm is capable of computing local minimizers of (1).

Extensive work has been done in relaxing the ℓ_0-norm by its convex surrogate ℓ_1-norm see, e.g., Zhu et al. (2003); Fung and Mangasarian (2004); Shao et al. (2019); Yuan et al. (2010); Dedieu et al. (2022). Although the approximation models are easier to tackle, they may not exactly recover the solution to the original ℓ_0-based model. For example, comparison studies on linear regression and convex quantile regression show that ℓ_0-norm has better performance than ℓ_1-norm on feature selection, see Johnson et al. (2015); Dai (2023). Therefore, for applications that require higher solution accuracy, solving problem (1) directly seems necessary as done in Ustun and Rudin (2016). However, MIP approach has drawbacks on scalability and computational speed for problem (1).

(B) Augmented Lagrangian methods for nonconvex problems. From the perspective of constrained optimization, it is natural to consider the augmented Lagrangian method (ALM) of Hestenes (1969); Powell (1969) for problem (1). ALMs have become standard textbook material (see, e.g., Bertsekas (1996); Nocedal and Wright (2006); Birgin and Martínez (2014)). However, direct application is not possible due to the problem being a type of nonsmooth, nonconvex, and composite optimization. Despite this, significant progress has been recently made for this type of problems by Bolte et al. (2018):

$$\min \;\; f(\mathbf{x}) + \theta(F(\mathbf{x})), \tag{3}$$

where $f : \mathbb{R}^n \mapsto \mathbb{R}$ is continuously differentiable (C^1 class), $F : \mathbb{R}^n \mapsto \mathbb{R}^m$ ($m \le n$) is also C^1, and $\theta : \mathbb{R}^m \mapsto (-\infty, +\infty]$ is a proper and lower-semicontinuous (lsc) function. A key message delivered in Bolte et al. (2018) was that adaptive Lagrangian-based multiplier methods can be developed with guaranteed convergence properties. An essential requirement is that the primal iterates are kept close to the so-called information zone, where certain regularity

conditions are assumed. This requirement is often met when the subproblems are solved exactly. Other developments also appear in Li and Pong (2015); Wang et al. (2018); Boţ and Nguyen (2020) for unconstrained composite optimization.

Another possible solution method for (1) is to follow the framework of the augmented Lagrangian method of Kanzow et al. (2021); De Marchi et al. (2023); Jia et al. (2023) for composite optimization covering the cardinality constraint (i.e., ℓ_0-norm constraint). One of the techniques used is to represent the cardinality constraint as a smooth complementarity system in the spirit of (2). Similarly, the hard-margin loss can also be represented by a system of complementarity. This would drastically increase the dimensions of the resulting formulation.

Our problem (1) can be put in the framework of (3) by making use of the indicator function on the sparse constraint. The number of smooth functions in $F(\cdot)$ would be $(n+m)$, violating the requirement of $m \leq n$ in Bolte et al. (2018). It is also not clear how the primal iterates would be kept close to the problem information zone as we are simultaneously dealing with both the hard-margin loss and the ℓ_0-norm constraint. Furthermore, Mangasarian's original proposal for introducing the quadratic objective in the $(\mathbf{w}, b)$ space is for Newton's method to be used as its Hessian matrix is diagonal (i.e., sparse). Therefore, our proposal in this paper is to develop an augmented Lagrangian method sharing similar convergence properties as in Bolte et al. (2018) while allowing Newton's method to be used. Additionally, we allow the subproblem to be solved inexactly, enhancing the practicality.

(C) Newton's method for composite optimization. We briefly discuss our own work on this aspect. For the application of compressed sensing with cardinality constraint, we developed a Newton-based hard-thresholding method in Zhou et al. (2021b), which is also proved to be globally convergent. For the hard-margin loss, we were only able to prove its local quadratic convergence in Zhou et al. (2021a). Our recent attempt of Zhang et al. (2023) studies an ALM for a hard-margin loss composite optimization without any constraints. The current paper can be seen as an extension to the constrained case with the cardinality constraint. Extension of optimization methods from unconstrained optimization to constrained counterpart is sometime very challenging. The difficulty lies with the challenge of simultaneously handling both the sparse set and the hard-margin loss, both of which are of combinatorial nature. This paper resolved this difficulty in the venue of SVMs.

1.2 Main contributions

The review above establishes that (1) is a very useful yet challenging model to solve. There lacks efficient numerical methods for it especially for large data sets. Since we are not following the MIP approach, we are contented with being capable of computing a local minimizer. Our first contribution is on the characterization of local minimizers of (1). This is explained below with other innovative contributions.

(i) On the concept of stationarity. Since problem (1) is essentially a nonconvex composite optimization with a cardinality constraint, it has various formulations (e.g., via the complementarity systems as we review above). One good example to follow is the recent paper of Cui et al. (2022). In a series of papers by Gómez et al. (2023); Cui et al. (2023); Han et al. (2024), various stationary points including pseudo-minimizer, B-stationarity and epi-stationarity have been proposed for a class of composite optimization with 0/1-loss. A

very encouraging message is that local minimizers of such non-convex, nonsmooth problems can be completely characterized under a convex-like property. We prove a similar result in terms of pseudo-KKT point for (1). Furthermore, among the pseudo-KKT points, we investigate a group of points, termed as P-stationary points, that have more concise structural representation and can be computed efficiently. The relationship between those stationary points has been thoroughly studied and summarized in Figure 2. Moreover, the P-stationarity extends the previous stationarity concepts of Beck and Eldar (2013); Pan et al. (2015); Zhou et al. (2021b) on sparse optimization to the hard-margin case.

(ii) Inexact framework of proximal augmented Lagrangian method. To make the proposed ALM implementable, we solve its subproblem inexactly in a way that the generated iterates should enjoy the best known convergence properties, namely global convergence to a stationary point with a linear rate. It turns out that the accuracy of different parts of the stationarity measurement of the iterates should satisfy certain relationship between them. In other words, a new set of computable stopping criteria for solving each subproblem of ALM is developed. Unlike the case where each subproblem is solved exactly in terms of satisfying its optimality condition, the inexactness of the approximate solution creates some unavoidable obstacles in applying the traditional convergence analysis tools. A new Lyapunov function is constructed by adding a proximal term to the standard augmented Lagrangian to prove the global convergence as well as the linear rate of convergence under certain regularity conditions often met by data with $n \gg m$.

(iii) Optimization methods in reduced dimensions. Since problem (1) is highly combinatorial defined by the both hard-margin loss and the sparse set, a (local) solution should stay in a subspace when the iterates are close to it. This raises the question whether we can develop a subspace-based optimization method for each of the subproblems in the ALM framework. Intuitively, it is possible. However, for thus generated sequences to have good convergence properties as stated in (ii) above requires delicate tracking of the true underlying space. We achieved this tracking by making use of a sharp observation that the optimal solution should satisfy some complementarity conditions. Those conditions naturally define a subspace at each iteration. We then apply a gradient descent method in this subspace to get a sufficient decrease in the Lyapunov function. To speed up the convergence, we further update the iterate by Newton's method in the same subspace. The generated iterate is guaranteed to meet the stopping criteria discussed in (ii). The Newton method enjoys the quadratic convergence under the assumption of strict complementarity condition.

The resulting algorithm is highly efficient and is benchmarked against several leading SVM solvers on both simulated and real data sets. The proposed method is capable of computing a sparse solution with high classification accuracy and a smaller number of support vectors. And it is fast due to the fact that subproblems were often solved in a much smaller subspace rather than the full space.

1.3 Organization

In next section, we explain the notations used in the paper and present the basic properties of the projection operator to the s-sparse set and the positive hard-thresholding operator for the hard-margin loss function. Section 3 introduces the P-stationary and pseudo KKT point and presents their relationship with other stationarity and local minimizer of problem

(1). Section 4 develops the inexact framework of the proximal augmented Lagrangian method (iPAL) and conducts its convergence analysis. In Section 5, we propose an efficient numerical strategy to solve the subproblem in iPAL in a subspace. The strategy consists of two parts: first apply a gradient descent to guarantee a sufficient decrease, followed by a Newton step. Both are computed in a well defined subspace. We also conduct convergence analysis of this numerical strategy. We report extensive numerical experiments in Section 6.

The new algorithmic framework does not rely on any external optimization solvers for its subproblems. The design of the algorithm is constructive and is active-set based. It requires a new set of convergence analysis. We provide all the detailed proofs in Appendix.

2. Preliminaries and Positive Hard-Thresholding Operator

2.1 Notation and Definitions

We use boldfaced lowercase letters to denote vectors. For example, $\mathbf{w} \in \mathbb{R}^n$ is a column vector of size n and $\mathbf{w}^\top$ is its transpose. Let w_i or $[\mathbf{w}]_i$ denote the ith element of $\mathbf{w}$. The norm $\|\mathbf{w}\|$ denotes the Euclidean norm of $\mathbf{w}$ and for a matrix A, and $\|A\|$ is the induced norm by the Euclidean norm so that we always have $\|A\mathbf{w}\| \leq \|A\|\|\mathbf{w}\|$. For two column vectors $\mathbf{w}$ and $\boldsymbol{\xi}$, we will commonly use the shorthand symbol $\mathbf{u} := (\mathbf{w}, \boldsymbol{\xi})$ to denote the new vector pair concatenating $\mathbf{w}$ and $\boldsymbol{\xi}$ (similarly, $\mathbf{u}^* := (\mathbf{w}^*, \boldsymbol{\xi}^*)$). The neighborhood of $\mathbf{w}^* \in \mathbb{R}^n$ with radius $\delta > 0$ is denoted by $\mathcal{N}(\mathbf{w}^*, \delta) := \{\mathbf{w} \in \mathbb{R}^n \mid \|\mathbf{w} - \mathbf{w}^*\| \leq \delta\}$, where ":=" means "define". We let I denote the identity matrix of appropriate dimension. $\mathbb{N}$ (resp. $\mathbb{N}^+$) denotes the set of all natural (resp. positive natural) numbers.

Let $[n]$ denote the set of indices $\{1, \ldots, n\}$. For a subset $T \subset [n]$, $|T|$ denotes the number of elements in T (cardinality of T) and $\mathbf{w}_T$ denotes the subvector of $\mathbf{w}$ indexed by T. We also let $\overline{T}$ denote the set of indices not in T (i.e., $\overline{T} = [n] \setminus T$). Given $\Gamma \in [m]$ and $A \in \mathbb{R}^{m \times n}$, $A_{\Gamma,T}$ denotes a submatrix of A with row and column indexed by Γ and T respectively. Particularly, $A_{\Gamma:}$ (resp. $A_{:T}$) is the submatrix with full column (resp. row) index.

We recall from (Rockafellar, 1976, Definition 1.22) that the Moreau envelop for a proper and lower semi-continuous function $f : \mathbb{R}^n \to \mathbb{R}$ with $\varrho > 0$ is defined as

$$\Phi^{\varrho}_{f(\cdot)}(\boldsymbol{\xi}) := \min_{\mathbf{q} \in \mathbb{R}^n} \; f(\mathbf{q}) + \frac{1}{2\varrho}\|\mathbf{q} - \boldsymbol{\xi}\|^2.$$

The set of the solutions achieving the value $\Phi^{\varrho}_{f(\cdot)}(\boldsymbol{\xi})$ is denoted by $\mathrm{Prox}_{\varrho f(\cdot)}(\boldsymbol{\xi})$ (the proximal operator of f). Throughout the paper, we only deal with functions whose Moreau envelop is always achieved.

2.2 Projection onto the s-sparse set

The orthogonal projection onto the s-sparse set $\mathbb{S}$ is known, see (Beck, 2017, Sect. 6.8.3). We use a different (but equivalent) description below. For a given $\mathbf{w} \in \mathbb{R}^n$, let $|\mathbf{w}|$ be the vector whose element is the absolute value of the corresponding element in $\mathbf{w}$. Let $|\mathbf{w}|_{(i)}$ denote the ith largest value in $|\mathbf{w}|$. Define $\mathcal{T}_s(\mathbf{w})$ to be the collection of all sets, each

consisting the s indices which give rise to the largest s elements in $|\mathbf{w}|$:

$$\mathcal{T}_s(\mathbf{w}) := \left\{\{i_1, \ldots, i_s\} \mid |w_{i_1}| = |\mathbf{w}|_{(1)}, \ldots, |w_{i_s}| = |\mathbf{w}|_{(s)}\right\} \tag{4}$$

For example, for $\mathbf{w} = [10; 20; 10]$, we have $\mathcal{T}_2(\mathbf{w}) = \{\{2, 1\},\ \{2, 3\}\}$. The orthogonal projection onto $\mathbb{S}$ is given by

$$\mathrm{Proj}_{\mathbb{S}}(\mathbf{w}) = \left\{\mathbf{q} \in \mathbb{R}^n \mid \mathbf{q} = \sum_{i\in T} w_i \mathbf{e}_i,\ \ T \in \mathcal{T}_s(\mathbf{w})\right\}, \tag{5}$$

where $\mathbf{e}_i$ is the ith standard unit vector in $\mathbb{R}^n$. An easy consequence of this description is the following result, see also (Pan et al., 2015, Table 1).

Lemma 1 *(Fixed-point characterization of the s-sparse set) Given vectors $\mathbf{w}, \mathbf{q} \in \mathbb{R}^n$ and $\alpha > 0$, we have*

$$\mathbf{w} \in \mathrm{Proj}_{\mathbb{S}}(\mathbf{w} - \alpha\mathbf{q}) \Longleftrightarrow \begin{cases} q_i = 0, & \text{if } w_i \neq 0, \\ |q_i| \leq |\mathbf{w}|_{(s)}/\alpha, & \text{if } w_i = 0, \end{cases} \tag{6}$$

Moreover, for such pair $(\mathbf{w}, \mathbf{q})$, the complementarity condition holds:

$$w_i \times q_i = 0, \quad \forall\ i \in [n].$$

2.3 Positive hard-thresholding operator

For the ease of description, we define the following function $J : \mathbb{R}^m \mapsto \mathbb{R}$:

$$J(\boldsymbol{\xi}) := \sum_{i=1}^{n} h(\xi_i),$$

where $h(t)$ is the hard-margin loss. The proximal operator of the $h(t)$ has a simple characterization (it can be computed directly through its definition) for $\beta > 0$:

$$\mathrm{Prox}_{\beta h(\cdot)}(t) = \mathcal{H}_{\sqrt{2\beta}}(t), \quad \text{where} \quad \mathcal{H}_\nu(t) := \begin{cases} \min\{0,\ t\}, & \text{if } t < \nu, \\ t, & \text{if } t > \nu, \\ \{0,\ t\}, & \text{if } t = \nu. \end{cases} \tag{7}$$

The operator $\mathcal{H}_\nu(t)$ with $\nu > 0$ treats small positive values t as zero and is very similar to the well-known hard-thresholding operator that treats small absolute values of t as zero, see (Beck, 2017, Example 6.10). We call $\mathcal{H}_\nu(t)$ the *positive hard-thresholding operator*. Consequently, the proximal operator of $J(\cdot)$ is given by

$$\mathrm{Prox}_{\beta J(\cdot)}(\boldsymbol{\xi}) = \mathcal{H}_{\sqrt{2\beta}}(\xi_1) \times \cdots \times \mathcal{H}_{\sqrt{2\beta}}(\xi_n). \tag{8}$$

Noting that $\mathcal{H}_\nu$ can be further simplified as

$$\mathcal{H}_\nu(t) := \begin{cases} t, & \text{if } t \in (-\infty, 0] \cup (\nu, +\infty), \\ \{0,\ t\}, & \text{if } t = \nu, \\ 0, & \text{if } t \in (0, \nu), \end{cases} \tag{9}$$

we can derive $t \notin \mathcal{H}_\nu(t)$ whenever $t \in (0, \nu)$. Consequently, we have

$\boldsymbol{\xi} \notin \mathrm{Prox}_{\beta J(\cdot)}(\boldsymbol{\xi})$ if and only if there exists an index $i \in [m]$ such that $\xi_i \in (0, \sqrt{2\beta})$.

Equivalently, we have

$$\boldsymbol{\xi} \in \mathrm{Prox}_{\beta J(\cdot)}(\boldsymbol{\xi}) \quad \text{if and only if } \ \xi_i \in (-\infty, 0] \cup [\sqrt{2\beta},\ \infty) \text{ for all } i \in [m]. \tag{10}$$

We extend this result to a more general situation and it will be used in characterizing the stationary point of our problem (1) .

Lemma 2 *(Fix-point characterization of the hard-margin loss) Suppose β, λ are two positive constants. Let $\boldsymbol{\xi}, \mathbf{v} \in \mathbb{R}^m$ be given. It holds that*

$$\boldsymbol{\xi} \in \mathrm{Prox}_{\beta\lambda J(\cdot)}(\boldsymbol{\xi} + \beta\mathbf{v})$$

if and only if

$$\boldsymbol{\xi} \in \mathrm{Prox}_{\beta\lambda J(\cdot)}(\boldsymbol{\xi}) \quad \textit{and} \quad \begin{cases} v_i = 0, & \textit{if } \xi_i \in (-\infty, 0) \cup [\sqrt{2\beta\lambda},\ \infty) \\ v_i \in [0,\ \sqrt{2\lambda/\beta}], & \textit{if } \xi_i = 0. \end{cases}$$

Consequently, the complementarity condition holds for such pair $(\boldsymbol{\xi}, \mathbf{v})$:

$$\xi_i \times v_i = 0,\ \forall\ i \in [m].$$

3. Stationarity Characterization of Local Minimizers

For the sake of simplicity, it is without loss of generality that we merge the variable b into $\mathbf{w}$ in (1): $\mathbf{w} := [\mathbf{w}; b]$ (Matlab notation). Define the corresponding matrix A with its ith row being $A_{i:} = -y_i[\mathbf{x}_i^\top, 1]$, $i = 1, \ldots, m$. We still treat thus defined vector $\mathbf{w}$ as n-dimensional vector (to save us from using $(n+1)$) and A is $m \times n$ data matrix. $\mathbf{1}$ is a vector with appropriate dimension and all entries being one. problem (1) then becomes

$$\min_{\mathbf{w}} f(\mathbf{w}) := \frac{1}{2}\|\mathbf{w}\|^2 + \lambda J(A\mathbf{w} + \mathbf{1}), \quad \text{s.t. } \ \|\mathbf{w}\|_0 \le s. \tag{11}$$

By introducing the auxiliary variable $\boldsymbol{\xi} \in \mathbb{R}^m$, we consider the following reformulation:

$$\min_{\mathbf{w}, \boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 + \lambda J(\boldsymbol{\xi}) + \delta_{\mathbb{S}}(\mathbf{w}), \quad \text{s.t. } \ A\mathbf{w} + \mathbf{1} = \boldsymbol{\xi}, \tag{12}$$

where $\delta_{\mathbb{S}}(\cdot)$ is the indicator function of the set $\mathbb{S}$. The augmented Lagrangian function of (12) is

$$\mathcal{L}_\rho(\mathbf{w}, \boldsymbol{\xi}, \mathbf{z}) := \frac{1}{2}\|\mathbf{w}\|^2 + \langle \mathbf{z}, A\mathbf{w} + \mathbf{1} - \boldsymbol{\xi}\rangle + \frac{\rho}{2}\|A\mathbf{w} + \mathbf{1} - \boldsymbol{\xi}\|^2 + \lambda J(\boldsymbol{\xi}) + \delta_{\mathbb{S}}(\mathbf{w}),$$

where $\mathbf{z} \in \mathbb{R}^m$ is the Lagrange multiplier and $\rho > 0$ is a penalty parameter. We will interchangeably refer to (11) and (12) depending on the situation whether $\boldsymbol{\xi}$ is needed or not. Next we will define a pseudo KKT point of (12) for optimality analysis. This definition is based on a class of nonlinear programming.

Given a reference point $\mathbf{u}^* := (\mathbf{w}^*, \boldsymbol{\xi}^*)$ belonging to the feasible region of (12), let us define the following index sets

$$\mathcal{S}^* := \{i \in [n] : w_i^* \neq 0\}, \quad \mathcal{I}_-^* := \{i \in [m] : \xi_i^* \leq 0\}, \quad \mathbb{T}^* := \{T \subseteq [n] : T \supseteq \mathcal{S}^*, |T| = s\}.$$

Taking $T^* \in \mathbb{T}^*$, we consider the following nonlinear programming associated with T^* (abbreviated as NLP-T^*)

$$\min_{\mathbf{w},\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2, \quad \text{s.t.} \quad \mathbf{w}_{\overline{T}^*} = 0, \quad \boldsymbol{\xi}_{\mathcal{I}_-^*} \leq 0, \quad A\mathbf{w} + \mathbf{1} = \boldsymbol{\xi}. \tag{NLP-T^*}$$

We notice that when $\|\mathbf{w}^*\|_0 = s$, $\mathbb{T}^*$ is singleton, whereas it has more than one elements if $\|\mathbf{w}^*\|_0 < s$. For each $T^* \in \mathbb{T}^*$, the Lagrange function of (NLP-T^*) is denoted by

$$\mathcal{L}_{T^*}(\mathbf{u}, \mathbf{q}_w, \mathbf{q}_\xi, \mathbf{z}) := \frac{1}{2}\|\mathbf{w}\|^2 + \langle \mathbf{q}_w, \mathbf{w}_{\overline{T}^*} \rangle + \langle \mathbf{q}_\xi, \boldsymbol{\xi}_{\mathcal{I}_-^*} \rangle + \langle \mathbf{z}, A\mathbf{w} + \mathbf{1} - \boldsymbol{\xi} \rangle,$$

where $(\mathbf{q}_w, \mathbf{q}_\xi, \mathbf{z}) \in \mathbb{R}^{|\overline{T}^*|} \times \mathbb{R}^{|\mathcal{I}_-^*|} \times \mathbb{R}^m$ are multipliers associated with the corresponding three constraints in (NLP-T^*). Thereby, the KKT system of (NLP-T^*) can be represented as

$$\begin{cases} (\mathbf{w} + A^\top \mathbf{z})_{T^*} = 0, \ \mathbf{w}_{\overline{T}^*} = 0, \\ \mathbf{z}_{\mathcal{I}_-^*} \geq 0, \ \boldsymbol{\xi}_{\mathcal{I}_-^*} \leq 0, \ \langle \mathbf{z}_{\mathcal{I}_-^*}, \boldsymbol{\xi}_{\mathcal{I}_-^*} \rangle = 0, \ \mathbf{z}_{\overline{\mathcal{I}}_-^*} = 0, \\ A\mathbf{w} + \mathbf{1} - \boldsymbol{\xi} = 0, \ \mathbf{q}_w = -(\mathbf{w} + A^\top \mathbf{z})_{\overline{T}^*}, \ \mathbf{q}_\xi = \mathbf{z}_{\mathcal{I}_-^*}. \end{cases} \tag{13}$$

We say $(\mathbf{w}, \boldsymbol{\xi})$ satisfying (13) is a KKT point of (NLP-T^*) with Lagrange multipliers $(\mathbf{q}_w, \mathbf{q}_\xi, \mathbf{z})$. For convenience, we give the following definition.

Definition 3 *A point $\mathbf{u}^* = (\mathbf{w}^*, \boldsymbol{\xi}^*)$ is called a pseudo KKT point of* (12) *if it is a KKT point of (NLP-T^*) for all $T^* \in \mathbb{T}^*$.*

Remark 1 *The definition of pseudo KKT point is similar to that of the pseudo B-stationary point proposed by Cui et al. (2023) for a type of Heaviside composite optimization. In the Proposition 14 of Appendix B, we show that for problem* (12)*, these two stationary points are equivalent when $\|\mathbf{w}^*\|_0 = s$, whereas the pseudo KKT point is stronger when $\|\mathbf{w}^*\|_0 < s$.*

Han et al. (2024) recently defined an epi-stationary point for a class of convex-like Heaviside optimization, and further showed that it is equivalent to the local minimizer (see Han et al. 2024, Corollary 3). Actually, their model includes (12) as a special case. The next theorem establishes the equivalence of epi-stationary point and pseudo KKT point for problem (12).

Theorem 4 *For problem* (12)*, $\mathbf{u}^* = (\mathbf{w}^*, \boldsymbol{\xi}^*)$ is an epi-stationary point (see Han et al. 2024, Definition 1) if and only if it is a pseudo KKT point.*

We note that it is not a trivial task to establish the equivalence. However, its implication is very important because together with (Han et al., 2024, Corollary 3), the pseudo KKT points are the local minimizers for (12). Furthermore, the characterization leads to an important property, which is illustrated in the following example.

Example 1 *(i) Consider problem* (12) *with* $m = 2, n = 3, s = 2$, $\lambda = 1$, *and*

$$A = A_1 = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & -1 \end{bmatrix}.$$

The index set T^* *for the associated (NLP-*T^**) can only be taken from* $\{\{1,2\},\{1,3\},\{2,3\}\}$, *while* $\mathcal{I}_-^*$ *can only be taken from* $\{\{1,2\},\{1\},\{2\}\}$. *Then we can use* (13) *to verify that there are only four pseudo KKT points. The corresponding* $\mathbf{w}$*-components are included in the following set*

$$\Omega_{\mathbf{w}}^* = \{(0,0,0)^\top,\ (1,0,0)^\top,\ (0.5,0,-0.5)^\top,\ (0.5,0,0.5)^\top\}.$$

Next let us consider the local minimizers of (12). *We notice that* $w_2 = 0$ *must hold, and thus we just need to compute the local minimizer of following simplified problem*

$$\min_{w_1,w_3} (w_1^2 + w_3^2)/2 + h(-w_1 + w_3 + 1) + h(-w_1 - w_3 + 1). \tag{14}$$

This problem is graphed in Figure 1, from which we can check that the local minimizers of (14) *are* $(0,0),(1,0),(0.5,-0.5),(0.5,0.5)$. *Therefore, we can conclude that in this case, the local minimizer and pseudo KKT point sets are identical.*

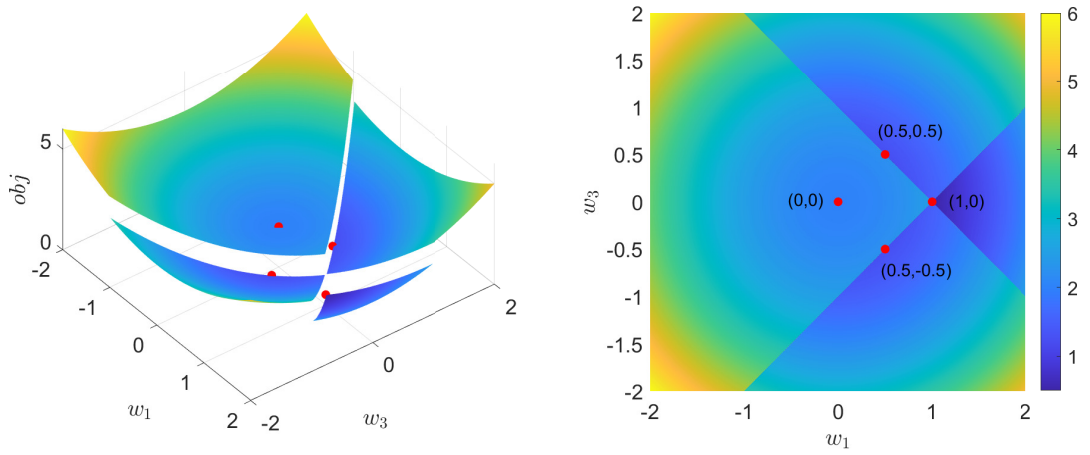

Figure 1: The objective function of problem (14) and overhead view

(ii) Now let us pick up one pseudo-KKT point, say $\mathbf{w}^* = (1,0,0)^\top$, *for closer examination. Taking* $\mathbf{u} = (\mathbf{w}, \boldsymbol{\xi} = A_1\mathbf{w} + \mathbf{1}) \in \mathcal{N}(\mathbf{u}^*, 1/2)$ *with* $\|\mathbf{w}\|_0 \le 2$, *we must have* $w_1 \neq 0$ *while either* $w_2 = 0$ *or* $w_3 = 0$. *We compare the objective values* $f(\mathbf{w}^*)$ *and* $f(\mathbf{w})$. *We consider two cases:* $A_1\mathbf{w} + \mathbf{1} = \boldsymbol{\xi} \le 0$ *and* $A_1\mathbf{w} + \mathbf{1} = \boldsymbol{\xi} \not\le 0$. *For both cases, we can verify the following bound:*

$$\|\mathbf{w}\|^2/2 + \lambda J(\boldsymbol{\xi}) \ge \|\mathbf{w}^*\|^2/2 + \lambda J(\boldsymbol{\xi}^*) + \|\mathbf{w} - \mathbf{w}^*\|^2/2.$$

Such bound is known as the quadratic growth condition. It turns out that the growth condition holds at any pseudo-KKT point and this important result is stated below.

Theorem 5 *(Equivalent characterization of local minimizers) Consider problem (12). The point $\mathbf{u}^* = (\mathbf{w}^*, \boldsymbol{\xi}^*)$ is a local minimizer if and only if it is a pseudo KKT point. Moreover, each local minimizer $\mathbf{u}^*$ is a strict local minimizer satisfying the following quadratic growth condition*

$$f(\mathbf{w}) \geq f(\mathbf{w}^*) + c_*\Big(\|\mathbf{w}-\mathbf{w}^*\|^2 + \|A(\mathbf{w}-\mathbf{w}^*)\|^2\Big), \quad \forall\ \mathbf{w} \in \mathcal{N}(\mathbf{w}^*, \epsilon_*) \cap \mathbb{S}, \tag{15}$$

where c_* and ϵ_* are some positive constants.

Although pseudo KKT point is an equivalent characterization of local minimizer for problem (12), its definition involves unknown index set $T^* \in \mathbb{T}^*$. When $\mathbb{T}^*$ is not a singleton, for each $T^* \in \mathbb{T}^*$, the Lagrange multplier of system (13) might be different. Taking these factors into account, we next define a proximal-type (P-)stationary point with more concise structure for our algorithm design.

Definition 6 *A point $\mathbf{u}^* := (\mathbf{w}^*, \boldsymbol{\xi}^*)$ is called a* P*-stationary point of problem (12) if there exists a Lagrange multiplier $\mathbf{z}^*$ and two positive constants $\alpha > 0$ and $\beta > 0$ such that*

$$\begin{cases} \mathbf{w}^* \in \mathrm{Proj}_{\mathbb{S}}(\mathbf{w}^* - \alpha(\mathbf{w}^* + A^\top \mathbf{z}^*)), \\ \boldsymbol{\xi}^* \in \mathrm{Prox}_{\beta\lambda J(\cdot)}(\boldsymbol{\xi}^* + \beta\mathbf{z}^*), \\ A\mathbf{w}^* + \mathbf{1} - \boldsymbol{\xi}^* = 0, \end{cases} \tag{16}$$

For convenience, we also call $(\mathbf{u}^*, \mathbf{z}^*)$ a P-stationary pair of problem (12).

Remark 2 *(i) The notation of* P*-stationarity has its reference to the projection and proximal operators used in its definition. The first inclusion relationship in (16) characterizes the stationarity with regarding to the s-sparse set $\mathbb{S}$. The projection operator is actually the proximal operator of the indicator function $\delta_{\mathbb{S}}(\cdot)$. The second inclusion relationship is about the hard-margin loss function. Proximal operators have been used to characterize stationary points in sparse optimization, see Beck and Eldar (2013); Zhou et al. (2021b). We also note that if the* P*-stationary condition (16) is satisfied for some $\alpha = \alpha_0$ and $\beta = \beta_0$, then it is also satisfied with any $\alpha \leq \alpha_0$ and $\beta \leq \beta_0$. This follows from the fixed-point characterizations in Lemmas 1 and 2. Therefore, the proper α and β can be taken on whole intervals whose upper bounds are determined by $\mathbf{w}^*$, $\boldsymbol{\xi}^*$ and $\mathbf{z}^*$.*

(ii) If we denote $\mathcal{S}^* := \{i \in [n] : w_i^* \neq 0\}$ and $\Gamma^* := \{i \in [m] : \xi_i^* \neq 0\}$, then by using Lemmas 1 and 2, we can derive $\mathbf{z}^*_{\Gamma^*} = 0$ and $(\mathbf{w}^* + A^\top \mathbf{z}^*)_{\mathcal{S}^*} = 0$ from (16), which further leads to

$$\mathbf{w}^*_{\mathcal{S}^*} = -A^\top_{\overline{\Gamma}^*,\mathcal{S}^*}\mathbf{z}^*_{\overline{\Gamma}^*} \quad \textit{and} \quad \mathbf{w}^*_{\overline{\mathcal{S}}^*} = 0.$$

This means that $\overline{\Gamma}^*$ actually includes all the support vectors of $\mathbf{z}^*$.

The subsequent theorem shows that a P-stationary point is stronger than a pseudo KKT point (or local minimizer) of problem (12), unless the KKT systems (13) share a common set of Lagrange multipliers for all $T^* \in \mathbb{T}^*$.

Theorem 7 *Let* $(\mathbf{u}^*, \mathbf{z}^*) \in \mathbb{R}^{n+m} \times \mathbb{R}^m$ *be a P-stationary pair of* (12)*, then* $\mathbf{u}^*$ *is a pseudo KKT point. Moreover, all the (NLP-*T^**) with* $T^* \in \mathbb{T}^*$ *share common multipliers* $(\mathbf{q}_w^*, \mathbf{q}_\xi^*, \mathbf{z}^*)$*, where*

$$\mathbf{q}_w^* = \begin{cases} [\mathbf{w}^* + A^\top \mathbf{z}^*]_{\overline{\mathcal{S}}^*}, & \text{if } \|\mathbf{w}^*\|_0 = s, \\ 0, & \text{if } \|\mathbf{w}^*\|_0 < s \end{cases} \quad \text{and } \mathbf{q}_\xi^* = \mathbf{z}^*_{\mathcal{I}_-^*}. \tag{17}$$

Now let us give an example to intuitively demonstrate the relationship of pseudo KKT point and P-stationary point.

Example 2 *Let us first consider problem* (12) *with* $m = 2$*,* $n = 3$*,* $s = 2$ *and*

$$A = A_2 = \begin{bmatrix} -1 & -2\sqrt{2} & 1 \\ -1 & 0 & -1 \end{bmatrix}$$

Given $\mathbf{w}^* = (1, 0, 0)^\top$*, we can identify* $\mathbb{T}^* = \{\{1, 2\}, \{1, 3\}\}$ *and* $\mathcal{I}_-^* = \{1, 2\}$*. The associated (NLP-*$T^*$*) problems are as follows:*

$$\min_{\mathbf{w}, \boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 \quad s.t. \quad w_3 = 0, \quad A_2\mathbf{w} + \mathbf{1} = \boldsymbol{\xi} \leq 0, \tag{18}$$

$$\min_{\mathbf{w}, \boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 \quad s.t. \quad w_2 = 0, \quad A_2\mathbf{w} + \mathbf{1} = \boldsymbol{\xi} \leq 0. \tag{19}$$

We can check that $\mathbf{w}^*$ *is KKT point of the above two problems, and thus it is a pseudo KKT point. However, it is not a* P*-stationary point because the optimal multiplier of* (18) *and* (19) *are* $(1, 0)^\top$ *and* $(1/2, 1/2)^\top$ *respectively.*

To end this section, let us give the relationships of all the aforementioned stationary points for problem (12)

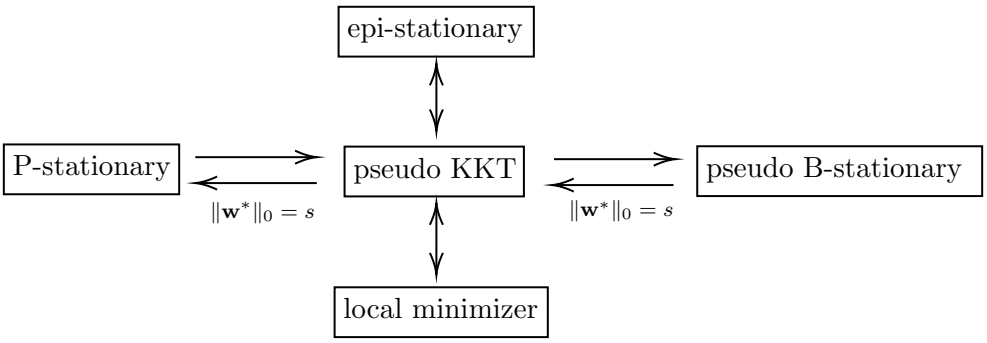

Figure 2: Relationships of five types of stationary points for problem (12).

4. Inexact Proximal Augmented Lagrangian Method

As mentioned in Introduction, problem (1) can be put in the framework of composite optimization. Therefore, general principle for developing augmented Lagrangian methods (ALM) set in Bolte et al. (2018) serves a guidance for us. In this part, we develop an implementable ALM, which is based on the following important innovations.

(i) The subproblems of our ALM are solved inexactly. Computable stopping criteria are designed and are sufficient for the generated sequence to have both global and local linear convergence rate. This is the most challenging part of our method.

(ii) In general, ALM generates infeasible iterates. Our problem has two constraints:

$$\mathbf{w} \in \mathbb{S} \quad \text{and} \quad A\mathbf{w} + \mathbf{1} = \boldsymbol{\xi}.$$

We treat the first constraint as "hard" constraint, which mus be met. In other words, we will generate feasible iterates $\mathbf{w}^k \in \mathbb{S}$. However, we allow the second constraint be only approximately satisfied. This gives us much freedom to control the quality of the iterates that satisfy some decrease condition.

(iii) We take the advantage of the combinatorial nature of the hard-margin loss function to define a subspace sufficiently big enough to contain a local minimizer of problem (12). This subspace is potentially much smaller than the full space at each iteration. The benefit is that the ALM subproblems can be efficiently solved by Newton's method.

The consideration above results in a new ALM. We first describe the framework of the ALM and then state its convergence properties.

4.1 Framework of iPAL.

Throughout, we denote $\mathbf{u} := (\mathbf{w}, \boldsymbol{\xi}) \in \mathbb{R}^{n+m}$ and $\mathbf{u}^k := (\mathbf{w}^k, \boldsymbol{\xi}^k)$ for each iterate. We further define the Lyapunov function $\mathcal{M}_{\rho,\mu} : \mathbb{R}^{n+m} \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ by

$$\begin{aligned}
&\mathcal{M}_{\rho,\mu}(\mathbf{u}, \mathbf{z}, \mathbf{v}) := \mathcal{L}_\rho(\mathbf{u}, \mathbf{z}) + \frac{\mu}{2}\|\mathbf{w} - \mathbf{v}\|^2 \\
&= \underbrace{\frac{1}{2}\|\mathbf{w}\|^2 + \langle \mathbf{z}, A\mathbf{w} + \mathbf{1} - \boldsymbol{\xi}\rangle + \frac{\rho}{2}\|A\mathbf{w} + \mathbf{1} - \boldsymbol{\xi}\|^2 + \frac{\mu}{2}\|\mathbf{w} - \mathbf{v}\|^2}_{=:g(\mathbf{u},\mathbf{z},\mathbf{v})} + \delta_{\mathbb{S}}(\mathbf{w}) + \lambda J(\boldsymbol{\xi}),
\end{aligned}$$

where $\mathbf{z}$ represents the Lagrangian multiplier and $\mathbf{v}$ is a point that acts as a proximal term to $\mathbf{w}$. The function $g(\mathbf{u}, \mathbf{z}, \mathbf{v})$ is the smooth part of the Lyapunov function.

Suppose the current iterate is $(\mathbf{u}^k, \mathbf{z}^k)$. We obtain $\mathbf{u}^{k+1}$ by

$$\mathbf{u}^{k+1} \approx \arg\min_{\mathbf{u}} \mathcal{M}_{\rho,\mu}(\mathbf{u}, \mathbf{z}^k, \mathbf{w}^k) = \arg\min_{\mathbf{u}} \underbrace{g(\mathbf{u}, \mathbf{z}^k, \mathbf{w}^k)}_{:=g_k(\mathbf{u})} + \delta_{\mathbb{S}}(\mathbf{w}) + \lambda J(\boldsymbol{\xi}), \tag{20}$$

and the Lagrange multiplier is updated according to the usual rule. The question now is how accurate $\mathbf{u}^{k+1}$ should be calculated. We must come up with a reasonable and computable criterion for it. Suppose problem (20) were to be solved exactly and let $\widehat{\mathbf{u}}^{k+1}$ denote its solution. Then it must satisfy the following first-order optimality condition for some $\alpha > 0$ and $\beta > 0$:

$$\begin{cases} \widehat{\mathbf{w}}^{k+1} \in \mathrm{Proj}_{\mathbb{S}}(\widehat{\mathbf{w}}^{k+1} - \alpha\nabla_{\mathbf{w}} g_k(\widehat{\mathbf{u}}^{k+1}))), \\ \widehat{\boldsymbol{\xi}}^{k+1} \in \mathrm{Prox}_{\beta\lambda J(\cdot)}(\widehat{\boldsymbol{\xi}}^{k+1} - \beta\nabla_{\boldsymbol{\xi}} g_k(\widehat{\mathbf{u}}^{k+1})). \end{cases} \tag{21}$$

Both the projection and the proximal operators in (21) have been well studied in Lemmas 1 and 2, where the complementarity relationships show the different magnitudes of the

quantities involved. Let us expand those quantities in order to derive a good approximation to (21).

Given a point $\mathbf{u}$, let us define its gradient step by

$$\widetilde{\mathbf{w}}^k(\mathbf{u}) := \mathbf{w} - \alpha\nabla_{\mathbf{w}} g_k(\mathbf{u}) \quad \text{and} \quad \widetilde{\boldsymbol{\xi}}^k(\mathbf{u}) := \boldsymbol{\xi} - \beta\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}).$$

Pick the index sets $T_{\mathbf{u}}$ and $\Gamma_{\mathbf{u}}$ respectively by

$$T_{\mathbf{u}} \in \mathcal{T}_s(\widetilde{\mathbf{w}}^k(\mathbf{u})) \quad \text{and} \quad \Gamma_{\mathbf{u}} = \{i \in [m] \mid [\widetilde{\boldsymbol{\xi}}^k(\mathbf{u})]_i \in (-\infty, 0) \cup (\sqrt{2\beta\lambda}, \infty)\},$$

where $\mathcal{T}_s$ is defined in (4). We simply use T and Γ instead of $T_{\mathbf{u}}$ and $\Gamma_{\mathbf{u}}$ when no confusion is caused. Using the representation of projection on $\mathbb{S}$ (see (5)) and proximal operator of $J(\cdot)$ (see (8) and (9)), we know that a sufficient condition of (21) is

$$\mathcal{R}_1(\widehat{\mathbf{u}}^{k+1}) = 0, \quad \mathcal{R}_2(\widehat{\mathbf{u}}^{k+1}) = 0, \quad \text{and} \quad \mathcal{R}_3(\widehat{\mathbf{u}}^{k+1}) = 0,$$

where

$$\begin{cases} \mathcal{R}_1(\mathbf{u}) := \|[\nabla_T g_k(\mathbf{u}); \mathbf{w}_{\overline{T}}]\|, \\ \mathcal{R}_2(\mathbf{u}) := \|[\nabla_\Gamma g_k(\mathbf{u}); \boldsymbol{\xi}_{\overline{\Gamma}}]\|, \\ \mathcal{R}_3(\mathbf{u}) := (\beta/2)\|\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u})\|^2 + \lambda J(\boldsymbol{\xi}) - \Phi^{\beta}_{\lambda J(\cdot)}(\boldsymbol{\xi} - \beta\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u})), \end{cases}$$

and we denote $\nabla_T g_k(\mathbf{u}) := [\nabla_{\mathbf{w}} g_k(\mathbf{u})]_T$ and $\nabla_\Gamma g_k(\mathbf{u}) := [\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u})]_\Gamma$. We note that the residual $\mathcal{R}_3$ involves the Moreau envelop of the hard-margin loss $\lambda J(\boldsymbol{\xi})$ and plays an important role in our analysis. We now present our inexact ALM in Alg. 1.

Algorithm 1 (iPAL: inexact Proximal Augmented Lagrangian Method)

Initialization: Given positive constants c_1, c_2 and initial point $(\mathbf{u}^0,\ \mathbf{z}^0)$. Select a positive sequence $\{\vartheta_k\}_{k\in\mathbb{N}}$ converging to zero.

for $k = 0, 1, \cdots$ **do**

1. Primal step: Starting with $(\mathbf{u}^k, \mathbf{z}^k)$, solve the subproblem (20) for $\mathbf{u}^{k+1}$, which satisfies the following criteria:

$$\begin{cases} \mathcal{M}_{\rho,\mu}(\mathbf{u}^{k+1}, \mathbf{z}^k, \mathbf{w}^k) \le \mathcal{M}_{\rho,\mu}(\mathbf{u}^k, \mathbf{z}^k, \mathbf{w}^k) \text{ and } \|\mathbf{w}^{k+1}\|_0 \le s \\ \mathcal{R}_1(\mathbf{u}^{k+1}) \le c_1\|\mathbf{w}^{k+1} - \mathbf{w}^k\|, \\ \mathcal{R}_2(\mathbf{u}^{k+1}) \le c_2\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2, \\ \mathcal{R}_3(\mathbf{u}^{k+1}) \le \vartheta_k. \end{cases} \tag{22}$$

2. Multiplier step:

$$\mathbf{z}^{k+1} = \mathbf{z}^k + \rho(A\mathbf{w}^{k+1} + \mathbf{1} - \boldsymbol{\xi}^{k+1}). \tag{23}$$

end for

Remark 3 *The algorithm iPAL follows the standard framework of ALM having both the primal and the multiplier steps. The only difference is that the subproblem was solved inexactly, but increasingly accurate. In particular, the residual $\mathcal{R}_2$ is one order more accurate*

than $\mathcal{R}_1$. This requirement is crucial in ensuring the generated sequence to converge linearly. We will design Newton's method for the subproblem in the next section to meet those criteria. For now, we present the convergence results.

4.2 Convergence of iPAL

As rightly emphasized in Bolte et al. (2018), certain regularity is needed on the constraints in composite optimization for global convergence of ALMs. We need the following regularity assumption. Let $r := \lfloor s/2 \rfloor$ and define $\Theta := \{T \subseteq [n] : |T| = r\}$, where $\lfloor \cdot \rfloor$ is the floor function.

Assumption 1 *For any $T \in \Theta$, $A_{:T}$ has full row rank. Consequently, there exists $\gamma > 0$ satisfying $\gamma^2 = \min_{T\in\Theta} \lambda_{\min}(A_{:,T}A_{:,T}^\top)$.*

The assumption is particularly useful when the sample data is small (i.e., $m \ll n$). This has been confirmed in our numerical experiments for such data. The assumption can be weakened to certain rows of A associated with $\overline{\Gamma}_k$ in Alg. 1. A further result (see (70)) indicates that $\overline{\Gamma}_k$ can be seen as an approximation of the support vector index set $\overline{\Gamma}^*$ (defined in Remark 2), which is usually much smaller than m. This increases the chance for the assumption to hold. The general assumption significantly simplifies our analysis.

Parameter Setup: Let c_1 and c_2 be two constants used in Alg. 1. Given $\mu > 0$, set ρ and η as follows:

$$\rho \geq \left\{ \frac{2}{\gamma^2}, \frac{8(c_3^2 + c_4^2)}{\mu} \right\}, \quad \eta = \frac{4c_4^2}{\rho}, \tag{24}$$

where $c_3 := (2c_1 + \mu + 2)/\gamma$ and $c_4 := (2c_1 + \mu)/\gamma$.

Our first result states that Alg. 1 leads to a sufficient decrease in the function value of the Lyapunov function $\mathcal{M}_{\rho,\eta}(\cdot)$. Let

$$\mathcal{M}_{k+1} := \mathcal{M}_{\rho,\eta}(\mathbf{u}^{k+1}, \mathbf{z}^{k+1}, \mathbf{w}^k), \quad \text{for } k = 0, 1, \ldots, .$$

Proposition 8 *Suppose that Assumption 1 holds and parameters are chosen as in (24). If $\{(\mathbf{u}^k; \mathbf{z}^k)\}_{k\in\mathbb{N}}$ is a sequence generated by iPAL. The following hold.*

(i) (Sufficient decrease) The sequence $\{\mathcal{M}_k\}_{k\in\mathbb{N}^+}$ is nonincreasing and

$$\mathcal{M}_k - \mathcal{M}_{k+1} \geq \frac{\mu}{4}\|\mathbf{u}^{k+1} - \mathbf{u}^k\|^2. \tag{25}$$

(ii) (Sequence boundedness) The sequence $\{(\mathbf{u}^k; \mathbf{z}^k)\}_{k\in\mathbb{N}}$ is bounded. Moreover

$$\lim_{k\to\infty} \|\mathbf{u}^{k+1} - \mathbf{u}^k\| = 0 \quad \textit{and} \quad \lim_{k\to\infty} \|\mathbf{z}^{k+1} - \mathbf{z}^k\| = 0. \tag{26}$$

Remark 4 *If the Lyapunov function $\mathcal{M}_{\rho,\eta}(\mathbf{u}, \mathbf{z}, \mathbf{v})$ is bounded from below by a constant M_∞, then (25) would imply*

$$\frac{\mu}{4}\sum_k \|\mathbf{u}^{k+1} - \mathbf{u}^k\|^2 \leq \sum_k (\mathcal{M}_k - \mathcal{M}_{k+1}) \leq \mathcal{M}_1 - M_\infty \leq \infty.$$

Then (26) would be a direct consequence.

Those results ensures the global convergence as well as linear convergence rate of iPAL.

Theorem 9 *(Global Convergence) Suppose that Assumption 1 holds and parameters are chosen as* (24)*. Let* $\{(\mathbf{u}^k;\mathbf{z}^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by iPAL. Then the whole sequence converges to a* P*-stationary pair* $(\mathbf{u}^*,\mathbf{z}^*)$ *of (12). Furthermore,* $\mathbf{u}^*$ *is a strict minimizer of (12).*

Since the whole sequence $\{(\mathbf{u}^k;\mathbf{z}^k)\}_{k\in\mathbb{N}}$ converges and the Lyapunov sequence $\{\mathcal{M}_k\}_{k\in\mathbb{N}^+}$ is nonincreasing, there must exist a limit, denoted by $\mathcal{M}_*$. Actually, we can prove $\mathcal{M}_* = \mathcal{M}_{\rho,\eta}(\mathbf{u}^*,\mathbf{z}^*,\mathbf{w}^*)$. For more details, please refer to Corollary 17 in Appendix.

Theorem 10 *(Linear rate of convergence) Under the premise in Theorem 9, the following estimations hold with a constant* $q\in(0,1)$*.*

(i) (Linear convergence in Lyapunov function) There exists a positive constant c_m *and a sufficiently large index* k^* *such that*

$$\mathcal{M}_k - \mathcal{M}_* \le c_m q^k, \quad \forall k \ge k^*. \tag{27}$$

(ii) (Linear convergence in iterative sequence) There exist a sufficiently large index k^* *and positive constants* c_w*,* c_ξ *and* c_z *such that for any* $k \ge k^*$*, it holds*

$$\|\mathbf{w}^k - \mathbf{w}^*\| \le c_w\sqrt{q}^k, \quad \|\boldsymbol{\xi}^k - \boldsymbol{\xi}^*\| \le c_\xi\sqrt{q}^k, \quad \textit{and} \quad \|\mathbf{z}^k - \mathbf{z}^*\| \le c_z\sqrt{q}^k. \tag{28}$$

5. Projected Gradient-Newton Method for Subproblems

The algorithmic framework of iPAL looks promising in terms of its global and linear convergence. To make it practically effective, we need to address how the subproblem (20) can be efficiently solved so as to meet the stopping criteria (22). As mentioned earlier, our ultimate purpose is to apply Newton's method in reduced dimensions. However, it is widely known that Newton's method is a local method. This motivates us to use a gradient descent method to initialize the computation. We put those considerations in precise formulation.

First, the subproblem (20) takes the following form:

$$\min_{\mathbf{u}:=(\mathbf{w},\boldsymbol{\xi})} G(\mathbf{u}) := g(\mathbf{u}) + \delta_{\mathbb{S}}(\mathbf{w}) + \lambda J(\boldsymbol{\xi}), \tag{29}$$

where we dropped the dependence of g on the iterate k. The main purpose is to solve (29). It is very important to note that (i) the gradient $\nabla g(\mathbf{u})$ is Lipschitzian continuous with constant ℓ_g:

$$\|\nabla g(\mathbf{u}) - \nabla g(\mathbf{v})\| \le \ell_g\|\mathbf{u}-\mathbf{v}\| \quad \forall\ \mathbf{u},\mathbf{v}\in\mathbb{R}^n$$

and (ii) $g(\mathbf{u})$ is strongly convex with constant σ_g:

$$g(\mathbf{u}) \ge g(\mathbf{v}) + \langle \nabla g(\mathbf{v}),\ \mathbf{u}-\mathbf{v}\rangle + \frac{\sigma_g}{2}\|\mathbf{u}-\mathbf{v}\|^2 \quad \forall\ \mathbf{u},\mathbf{v}\in\mathbb{R}^n.$$

Now suppose $\mathbf{u}^j = (\mathbf{w}^j, \boldsymbol{\xi}^j)$ be the current iterate. For given two constants $\alpha > 0$ and $\beta > 0$ (they serve as stepsizes respectively for $\mathbf{w}$ and $\boldsymbol{\xi}$), the new iterate by the gradient step is given by

$$\widehat{\mathbf{w}}^j := \mathbf{w}^j - \alpha \nabla_{\mathbf{w}} g(\mathbf{u}^j) \quad \text{and} \quad \widehat{\boldsymbol{\xi}}^j := \boldsymbol{\xi}^j - \beta \nabla_{\boldsymbol{\xi}} g(\mathbf{u}^j). \tag{30}$$

We then project $\widehat{\mathbf{w}}^j$ to the s-sparse set $\mathbb{S}$ and compute the hard-margin proximal of $\widehat{\boldsymbol{\xi}}^j$ and denote them by $\mathbf{u}^{j+1/2} = (\mathbf{w}^{j+1/2}, \boldsymbol{\xi}^{j+1/2})$

$$\mathbf{w}^{j+1/2} = \mathrm{Proj}_{\mathbb{S}}(\widehat{\mathbf{w}}^j) \quad \text{and} \quad \boldsymbol{\xi}^{j+1/2} \in \mathrm{Prox}_{\lambda\beta J(\cdot)}(\widehat{\boldsymbol{\xi}}^j). \tag{31}$$

We only consider those indices where $\mathbf{w}^{j+1/2}$ and $\boldsymbol{\xi}^{j+1/2}$ are not zero:

$$T_j \in \mathcal{T}_s(\widehat{\mathbf{w}}^j) \quad \text{and} \quad \Gamma_j = \left\{ i \in [m] \mid [\widehat{\boldsymbol{\xi}}^j]_i \in (-\infty, 0) \cup (\sqrt{2\lambda\beta}, \infty) \right\}. \tag{32}$$

Consequently, when restricting to the subspace:

$$\left\{ \mathbf{u} = (\mathbf{w}, \boldsymbol{\xi}) \in \mathbb{R}^n \times \mathbb{R}^m \mid \mathbf{w}_{\overline{T}_j} = 0, \ \ \boldsymbol{\xi}_{\overline{\Gamma}_j} = 0 \right\},$$

the objective function $G(\mathbf{u})$ is locally twice continuously differentiable at $\mathbf{u}^{j+1/2} = (\mathbf{w}^{j+1/2}, \boldsymbol{\xi}^{j+1/2})$. Newton's method is well defined at $\mathbf{u}^{j+1/2}$ on this subspace. The resulting algorithm is called the projected gradient-Newton method, which is detailed in Alg. 2

Algorithm 2 (PGN: Projected Gradient-Newton Method)

Initialization: Set $\alpha, \beta \in (0, 1/\ell_g)$, take initial point $\mathbf{u}^0 := (\mathbf{w}^0, \boldsymbol{\xi}^0) \in \mathbb{R}^{n+m}$ with $\|\mathbf{w}^0\|_0 \leq s$.

for $j = 0, 1, \cdots$ **do**

1. Identification step: Compute $\widehat{\mathbf{w}}^j$ and $\widehat{\boldsymbol{\xi}}^j$ by (30). Select T_j and Γ_j by (32)

2. Gradient step: Compute $\mathbf{u}^{j+1/2} = (\mathbf{w}^{j+1/2}, \boldsymbol{\xi}^{j+1/2})$ by (31).

3. Newton step: Denote $\Upsilon_j := T_j \cup \Gamma_j$ and compute $\widetilde{\mathbf{u}}^{j+1} := (\widetilde{\mathbf{w}}^{j+1}, \widetilde{\boldsymbol{\xi}}^{j+1})$ by solving the following reduced Newton equation in $\mathbf{u} := (\mathbf{w}, \boldsymbol{\xi})$

$$\begin{cases} H^{j+1/2}(\mathbf{u} - \mathbf{u}^{j+1/2})_{\Upsilon_j} = -\nabla_{\Upsilon_j} g(\mathbf{u}^{j+1/2}) \\ \mathbf{w}_{\overline{T}_j} = 0, \ \boldsymbol{\xi}_{\overline{\Gamma}_j} = 0, \end{cases} \tag{33}$$

where $H^{j+1/2} := [\nabla^2 g(\mathbf{u}^{j+1/2})]_{\Upsilon_j, \Upsilon_j}$.

4. Update step: Update $\mathbf{u}^j$ either by the Newton step or the gradient step as follows:

$$\mathbf{u}^{j+1} = \begin{cases} \widetilde{\mathbf{u}}^{j+1}, & \text{if } G(\mathbf{u}^{j+1/2}) - G(\widetilde{\mathbf{u}}^{j+1}) \geq (\sigma_g/4)\|\mathbf{u}^{j+1/2} - \widetilde{\mathbf{u}}^{j+1}\|^2 \\ \mathbf{u}^{j+1/2}, & \text{otherwise} \end{cases} \tag{34}$$

end for

Remark 5 *(i) Computational complexity of the gradient step. Assuming the gradient of $g(\mathbf{u})$ is available, the complexity of selecting T_j and Γ_j is $O(ns)$. According to Lemmas 1 and 2, the gradient update is computed by*

$$\begin{aligned} \mathbf{w}_i^{j+1/2} &= \begin{cases} [\mathbf{w}^j - \alpha \nabla_{\mathbf{w}} g(\mathbf{u}^j)]_i, & \text{if } i \in T_j, \\ 0, & \text{otherwise.} \end{cases} \\ \boldsymbol{\xi}_i^{j+1/2} &= \begin{cases} [\boldsymbol{\xi}^j - \beta \nabla_{\boldsymbol{\xi}} g(\mathbf{u}^j)]_i, & \text{if } i \in \Gamma_j, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{35}$$

Therefore, the overall complexity for computing $\mathbf{u}^{j+1/2}$ is $O(ns)$.

(ii) Computational complexity of the Newton step. We expand the Newton equation (33) *as follows:*

$$\begin{bmatrix} [(\mu+1)I + \rho A^\top A]_{T_j,T_j} & -\rho(A_{\Gamma_j,T_j})^\top \\ -\rho A_{\Gamma_j,T_j} & \rho I \end{bmatrix} \begin{bmatrix} \mathbf{d}_w \\ \mathbf{d}_\xi \end{bmatrix} = -\nabla_{\Upsilon_j} g_k(\mathbf{u}^{j+1/2}) = \begin{bmatrix} \mathbf{b}_w \\ \mathbf{b}_\xi \end{bmatrix}$$

with variable $\mathbf{d} = [\mathbf{d}_w; \mathbf{d}_\xi] \in \mathbb{R}^{|\Upsilon_j|}$ to be computed. By using Schur complement theorem, it is equivalent to

$$\begin{cases} \left((\mu+1)I + \rho A_{\overline{\Gamma}_j,T_j}^\top A_{\overline{\Gamma}_j,T_j}\right) \mathbf{d}_w = \mathbf{b}_w + (A_{\Gamma_j,T_j})^\top \mathbf{b}_\xi, \\ \mathbf{d}_\xi = \frac{1}{\rho}\mathbf{b}_\xi + A_{\Gamma_j,T_j}\mathbf{d}_w. \end{cases} \tag{36}$$

The computational complexity for solving this linear system is $O(|\overline{\Gamma}_j||T_j|^2)$. We can also apply Sherman-Morrison-Woodbury formula to this linear equation when $|\overline{\Gamma}_j| \ll |T_j|$ and the corresponding computational complexity will be $O(|\overline{\Gamma}_j|^2|T_j|)$.

Theorem 11 (Global Convergence of PGN) *Let $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$ be the sequence produced by PGN. The following statements hold.*

(i) (Sufficient decrease) We have

$$G(\mathbf{u}^j) - G(\mathbf{u}^{j+1}) \geq \zeta \|\mathbf{u}^{j+1/2} - \mathbf{u}^j\|^2 + (\sigma_g/4)\|\mathbf{u}^{j+1} - \mathbf{u}^{j+1/2}\|^2, \tag{37}$$

where $\zeta := \min\{(1/\alpha - \ell_g)/2, (1/\beta - \ell_g)/2\}$. This further leads to

$$\lim_{j\to\infty} \|\mathbf{u}^{j+1} - \mathbf{u}^j\| = 0 \quad \text{and} \quad \lim_{j\to\infty} \|\mathbf{u}^{j+1/2} - \mathbf{u}^j\| = 0 \tag{38}$$

(ii) (Convergence to stationary point) The sequence $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$ converges to a P*-stationary point $\widehat{\mathbf{u}} := (\widehat{\mathbf{w}}, \widehat{\boldsymbol{\xi}})$ of problem* (29) *satisfying*

$$\begin{cases} \widehat{\mathbf{w}} \in \mathrm{Proj}_{\mathbb{S}}(\widehat{\mathbf{w}} - \alpha \nabla_{\mathbf{w}} g(\widehat{\mathbf{u}}))) \\ \widehat{\boldsymbol{\xi}} \in \mathrm{Prox}_{\beta\lambda J(\cdot)}(\widehat{\boldsymbol{\xi}} - \beta \nabla_{\boldsymbol{\xi}} g(\widehat{\mathbf{u}})) \end{cases} \tag{39}$$

where $\alpha, \beta \in (0, \ell_g)$ are consist with the parameter setting in Alg. 2.

The global convergence theorem indicates that when PGN is applied to solving the $(k+1)$-th subproblem produced by iPAL, the generated sequence $\{\mathbf{u}^{k,j}\}_{j\in\mathbb{N}}$ with initial point $\mathbf{u}^{k,0} = \mathbf{u}^k$ will converge to a P-stationary point $\widehat{\mathbf{u}}^{k+1}$ of (21). In practice, for each iterate $\mathbf{u}^{k,j}$ with $j \geq 1$, the first line of stopping criteria (22) must hold by the sufficient decent property (37), and thus we just need to check whether $\mathbf{u}^{j,k}$ satisfies

$$\begin{cases} \mathcal{R}_1(\mathbf{u}) \leq c_1\|\mathbf{w} - \mathbf{w}^{k,0}\|, \\ \mathcal{R}_2(\mathbf{u}) \leq c_2\|\mathbf{w} - \mathbf{w}^{k,0}\|^2, \\ \mathcal{R}_3(\mathbf{u}) \leq \vartheta_k. \end{cases}$$

For a fixed $k \in \mathbb{N}$, we can prove that $\lim_{j\to\infty} \mathcal{R}_i(\mathbf{u}^{k,j}) = 0$ for $i = 1, 2, 3$ (see proof of Corollary 12 in Appendix I). If $\mathbf{w}^{k,0} \neq \widehat{\mathbf{w}}^{k+1}$ (initial point does not equal to the limit point), then $\lim_{j\to\infty} \|\mathbf{w}^{k,j} - \mathbf{w}^{k,0}\| > 0$ holds. This means the stopping criteria must be met in finite iterates when k is fixed.

Corollary 12 *(iPAL is well defined) When PGN is applied to solving the $(k+1)$-th subproblem generated by iPAL, if $\mathbf{w}^{k,0} \neq \widehat{\mathbf{w}}^{k+1}$, then there exists a sufficiently large index j_k such that $\mathbf{u}^{j_k}$ satisfies the stopping criteria (22).*

Remark 6 *The above corollary ensures that each element of $\{j_k\}_{k\in\mathbb{N}}$ is finite. However, it is still unclear whether this sequence is bounded, i.e. the existence of a uniform maximal iterates for all the inner subproblems. Generally, this kind of result relies on the global convergence rate of subroutine and the upper bound of residuals. The interested reader can refer to the iterate complexity analysis of ALM for convex or differentiable constrained optimization problems (see e.g. (Xu, 2021, Lemma 6), (Xie and Wright, 2021, Theorem 3)). However, for the nonconvex and nonsmooth constrained programming, the global convergence rate analysis of subroutine is a challenging task, and thus most related works only discuss the finiteness of j_k when k is fixed (e.g. (Song et al., 2020, Theorem 3.3), (Chen et al., 2017, Theorem 3.4)). Due to problem* (29) *involving nonconvex and discontinuous term $\delta_{\mathbb{S}}(\cdot)$ and $J(\cdot)$, the associated global convergence rate analysis of PGN requires further investigation. We will provide a local quadratic convergence of PGN in Theorem 13.*

We consider the situation near the stationary point $\widehat{\mathbf{u}}$ in (39). It follows from Lemma 2 that $\widehat{\boldsymbol{\xi}}$ and $\nabla_{\boldsymbol{\xi}} g(\widehat{\mathbf{u}})$ must satisfy the complementarity condition. We assume further that they satisfy the strict complementarity condition:

$$\widehat{\boldsymbol{\xi}}_i + [\nabla_{\boldsymbol{\xi}} g(\widehat{\mathbf{u}})]_i \neq 0, \quad \forall i \in [m]. \tag{40}$$

Under this assumption, we can prove that Newton's step is always accepted when $j \geq j_k$ and hence PGN is quadratically convergent.

Theorem 13 (Local Quadratic Convergence of PGN) *Let $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$ be a sequence converging to a* P*-stationary point $\widehat{\mathbf{u}}$ of* (20)*. Suppose that $\widehat{\boldsymbol{\xi}}$ and $\nabla_{\boldsymbol{\xi}} g(\widehat{\mathbf{u}})$ satisfy strictly complementary condition* (40)*, then there exists sufficiently large integer j_k such that Newton's step will always be accepted for all iterations $j \geq j_k$. Moreover, we have*

$$\|\mathbf{u}^{j+1} - \widehat{\mathbf{u}}\| \leq O(\|\mathbf{u}^j - \widehat{\mathbf{u}}\|^2) \quad \textit{for} \quad j \geq j_k.$$

This may be the best result one may hope for when Newton's method is used. The question now is whether the Newton equation can be efficiently solved. Our numerical results demonstrate that it is the case for many types of data.

6. Numerical Experiments

In this section, extensive numerical experiments will be conducted by using Matlab 2022a on a laptop with 32GB memory and Intel CORE i7 2.6 GHz CPU.

6.1 Benchmark Methods and Experimental Setting

To implement iPAL, we need to set up two types of parameters. One type called model parameters of (12) contains λ, ρ, μ and s. To simplify the parameter tuning, we will set $\lambda = \rho$. The best choices are often dependent on data, and thus we will give more details about the selection in the subsequent experiments. Another type of parameters appearing in Alg. 1 and Alg. 2 is called algorithmic parameters. We set

$$c_1 = c_2 = 0.1, \quad \gamma = 0.1 \min\{\|\mathbf{a}_i\| | i \in [m]\}, \quad \epsilon_k = \lambda/k \tag{41}$$

and η is taken as (24). We adopt $(\mathbf{w}^0, \boldsymbol{\xi}^0, \mathbf{z}^0) = \mathbf{0}$ as initial point and iPAL will stop if the following criterion holds

$$\frac{\|\mathbf{w}^k - \mathbf{w}^{k-1}\| + \|\boldsymbol{\xi}^k - \boldsymbol{\xi}^{k-1}\| + \|\mathbf{z}^k - \mathbf{z}^{k-1}\|}{\|\mathbf{w}^k\| + \|\boldsymbol{\xi}^k\| + \|\mathbf{z}^k\|} < 10^{-3}$$

We also select six efficient algorithms for numerical comparison. Together with iPAL, the associated SVM models to be solved are summarized in Table 1

Table 1: Benchmark Algorithms and Their Models

Algorithm	Reference	Loss Function	Regularizer	Constraint
iPAL	This work	Hard margin	ℓ_2	ℓ_0
ADMM0/1	Wang et al. (2021)	Hard margin	ℓ_2	–
LISVM	Yuan et al. (2010)	Squared hinge	ℓ_1	–
NLPSVM	Fung and Mangasarian (2004)	Hinge	ℓ_1	–
PDLSVM	Shao et al. (2019)	Least square	ℓ_1	Linear
ZFPR	Themelis et al. (2018)	Squared hinge	ℓ_0	–
NMAPG	Li and Lin (2015)	Squared hinge	ℓ_0 and ℓ_2	–

Four metrics are used for evaluating performance of the algorithms. They are classification accuracy: `Acc` $:= 1 - J(A\mathbf{w})/m$, CPU time (`Time`), the number of support vectors (`nSV`), and the number of nonzero elements `nnz` $:= \|\mathbf{w}\|_0$. As LISVM, ZFPR and NMAPG solve the primal optimization without introducing dual variables, these three solvers does not provide a dual solution and thus we do not record the `nSV` for them.

6.2 Experiments on Simulated Data

In this subsection, we will test all the solvers on data sets generated by the following example.

Example 3 *Samples with positive (resp. negative) labels are drawn from the normal distribution $N(\mu_1, \Sigma_1)$ (resp. $N(\mu_2, \Sigma_2)$), where the parameters $\mu_1 \in \mathbb{R}^n$ (resp. μ_2) are mean vectors, and $\Sigma_1 \in \mathbb{R}^{n\times n}$ (resp. Σ_2) are diagonal covariance matrices. We then flip r percentage (noise ratio) of those samples, making them be marked with reverse labels.*

6.2.1 Convergence Test

In this part, we will observe how the model parameters (λ, ρ, μ and s) influence the convergence of iPAL. We will use the following metric to judge the violation of first-order optimality condition of (12) for an iterate

$$\texttt{VFC} := \max\{r_1, \quad r_2, \quad r_3\},$$

where

$$\begin{array}{rcl} r_1 & := & \mathrm{dist}(\mathbf{w}^k, \mathrm{Proj}_{\mathbb{S}}(\mathbf{w}^k - \alpha(\mathbf{w}^k + A^\top \mathbf{z}^k))) \\ r_2 & := & \mathrm{dist}(\boldsymbol{\xi}^k, \mathrm{Prox}_{\beta\lambda \mathrm{J}(\cdot)}(\boldsymbol{\xi}^k + \beta \mathbf{z}^k)), \\ r_3 & := & \|A\mathbf{w}^k + \mathbf{1} - \boldsymbol{\xi}^k\|, \end{array}$$

where dist is the distance from a point to a nonempty set. A simulated data set with $m = 1000$ and $n = 2000$ is generated as the way described in Ex. 3. As mentioned at the beginning of Subsection 6.1, we will set $\lambda = \rho$ with the model parameters selected from the following sets:

$$\Omega_\rho = \{10^{-3}, 10^{-2}, \cdots, 10^3\},\ \Omega_\mu := 10^{-2} \times \{2^0, 2^1, \cdots, 2^{10}\},\ \Omega_s = \{20, 40, \cdots, 200\}$$

We have the following comments.

(i) From Fig. 3, we can observe that `VFC` decreases faster when ρ grows. However, the `Time` v.s. Iteration graph in Fig 3 shows that a large ρ does not always leads to a smaller `Time`. In fact, when ρ increases, the conditional number of linear equation (36) becomes bigger and thus it takes more time to solve.

(ii) We illustrate how the change of μ influence the convergence of iPAL. As shown in Fig 4, iPAL with larger μ tends to converge slower. But it might spend less `Time` because the linear system (36) admits smaller conditional number. For example, a medium value $\mu = 0.32$ leads to the least `Time` in this simulation.

(iii) We can see from Fig. 5 that the convergence rate shows a faster decreasing trend when the s grows. This is because the matrix dimension in linear system (36) is $s \times s$. A smaller s will lead to a significant reduction in dimension and computation. That is why iPAL with $s = 20$ (the smallest value of s) runs much faster than other cases (see `Time` v.s. Iteration in Fig. 5).

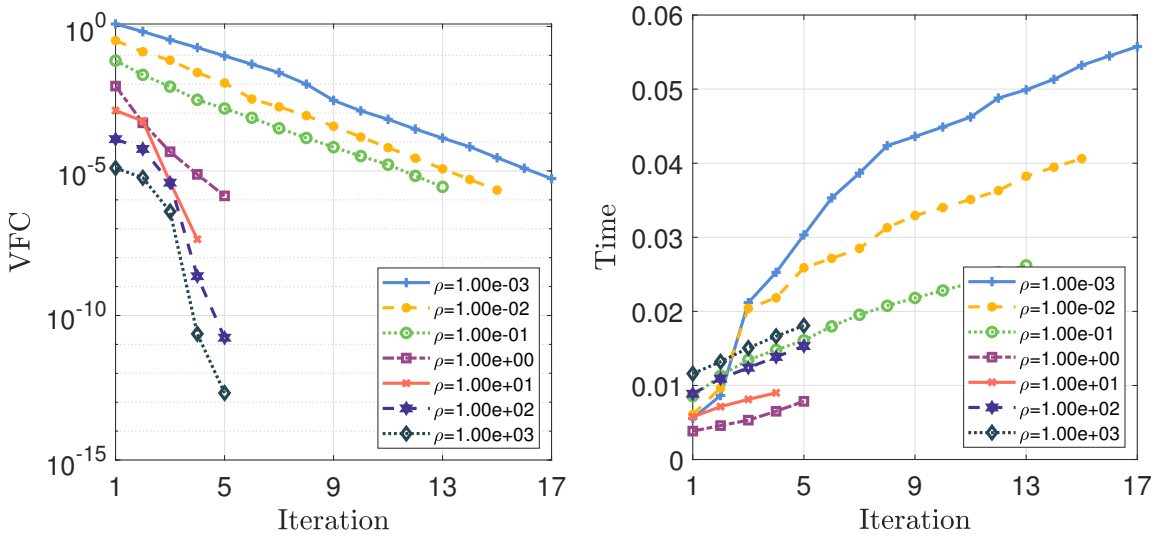

Figure 3: `VFC` and `Time` of iPAL along with iteration when $\mu = 10^{-2}$, $s = 20$ and $\rho = \lambda \in \Omega_\rho$.

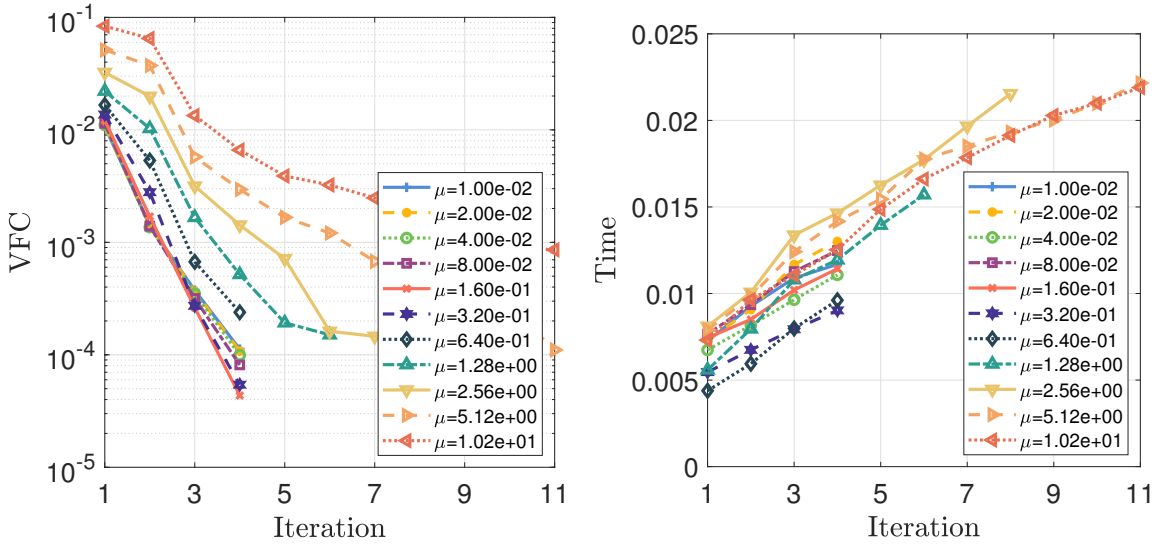

Figure 4: `VFC` and `Time` of iPAL along with iteration when $\rho = \lambda = 1$, $s = 20$ and $\mu \in \Omega_\mu$.

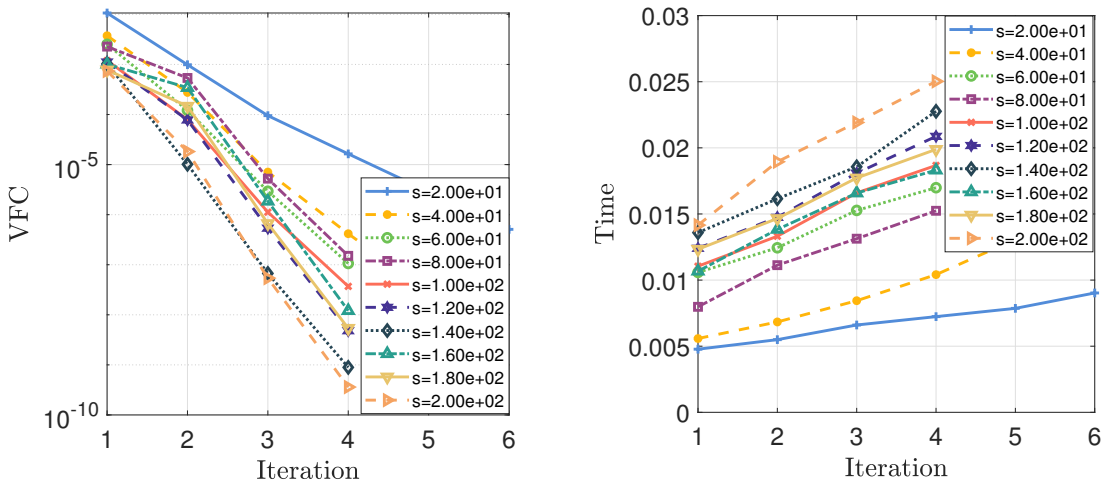

Figure 5: `VFC` and `Time` of iPAL along with iteration when $\mu = 10^{-2}$, $\rho = \lambda = 1$ and $s \in \Omega_s$.

6.2.2 Numerical Comparison

In this part, we will generate data sets with various m, n and r (noise rate) by the method in Ex. 3. The performance of all the seven algorithms will be compared. Half of the samples will be chosen as training set, and the rest of the samples are used for testing. In the following three tests, for iPAL, we set $\lambda = 1$, $\rho = 1$, $\mu = 10^{-2}$. As s will influence the `Time` and `nnz` of iPAL, we will set $s = 10, 20, 30, 40$ and then record the corresponding performance of iPAL in this subsection. Other algorithms used their default parameter settings.

Test I. We fix $m = 1000$, $r = 0.1$ and vary $n \in \{5000, 10000, \cdots, 30000\}$. In this test, we can see from Fig. 6 that `Acc` of NLPSVM and NMAPG are lower than other algorithms. iPAL spends the least amount of `Time` with the fewest `nSV`. Those algorithms adopting ℓ_0 norm for feature selection, including iPAL, ZFPR and NMAPG, compute solutions with smaller `nnz`. ADMM0/1 is the second fastest solver in this test, but its `nnz` is much larger and increases as n grows. This is because this algorithm is designed for a SVM problem without a sparsity constraint on its solutions. PDLSVM also shows a significant increase on `nnz` when n rises, whereas `nnz` of the other three solvers remain stable. As the number of samples m is fixed, the numbers for `nSV` of all the algorithms are steady.

Test II. We fix $n = 1000$, $r = 0.1$ and alter $m \in \{5000, 10000, \cdots, 30000\}$. Please refer to Fig. 7 for the discussion below. Our iPAL spends the shortest `Time` computing the solution with highest `Acc`. ZFPR, iPAL and NMAPG all have smaller `nnz`, but the `Time` of iPAL is significantly shorter than the other two algorithms in all values of s. A possible reason is that, in addition to the sparse cardinality constraint contributing to a low `nnz`, the proximal operator of hard margin loss simultaneously helps iPAL reduce `nSV`, consequently leading to a diminished computational cost. In particular, `Time` of iPAL is almost one order

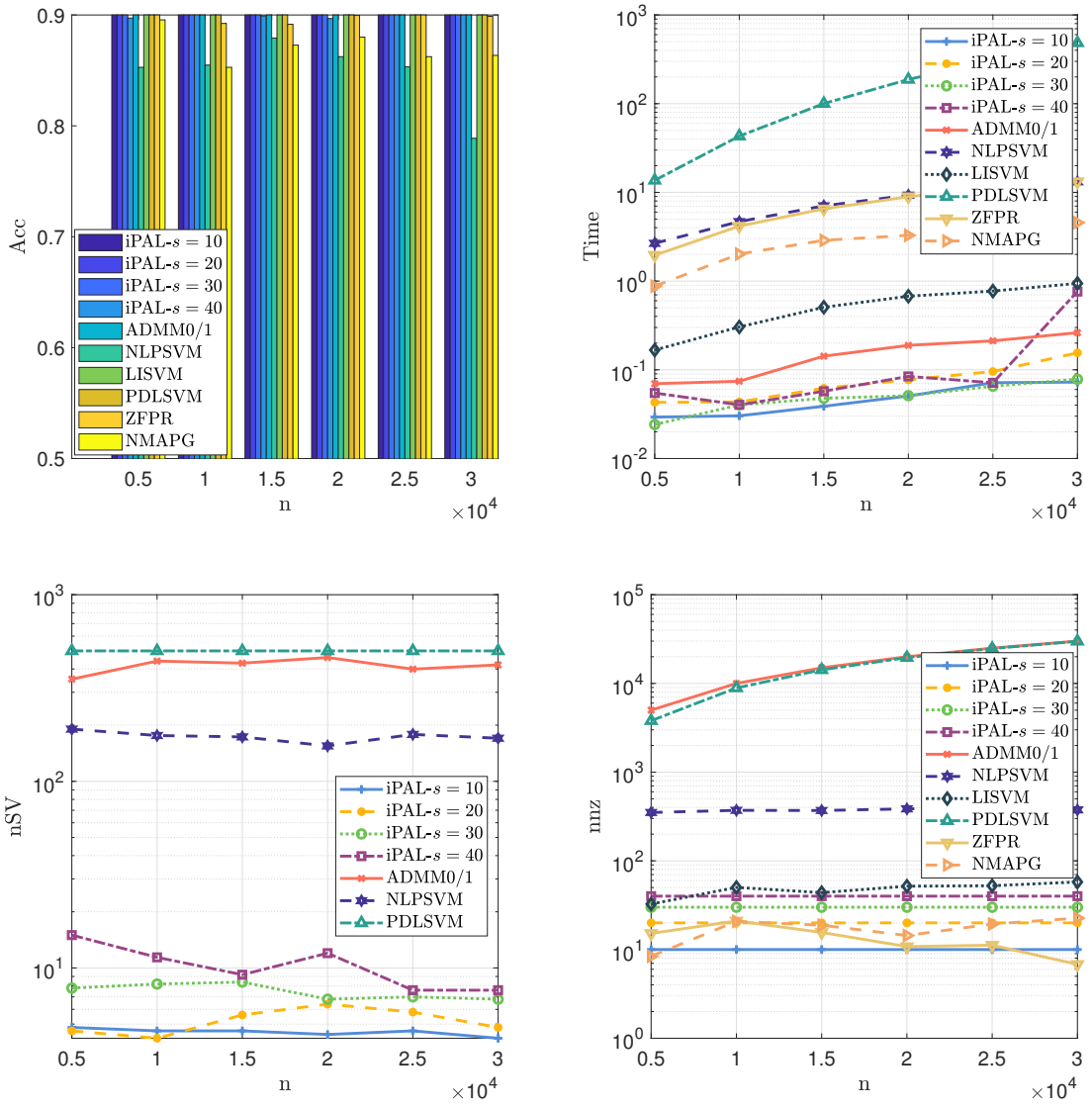

Figure 6: Comparison results on simulated data set with $m = 1000$, $r = 0.1$ and $n \in \{5000, 10000, \cdots, 30000\}$.

faster than that of ADMM0/1 and LISVM. When m becomes larger, there are significant increases on `nSV` of NLPSVM and PDLSVM, as well as on `nnz` of LISVM and NLPSVM.

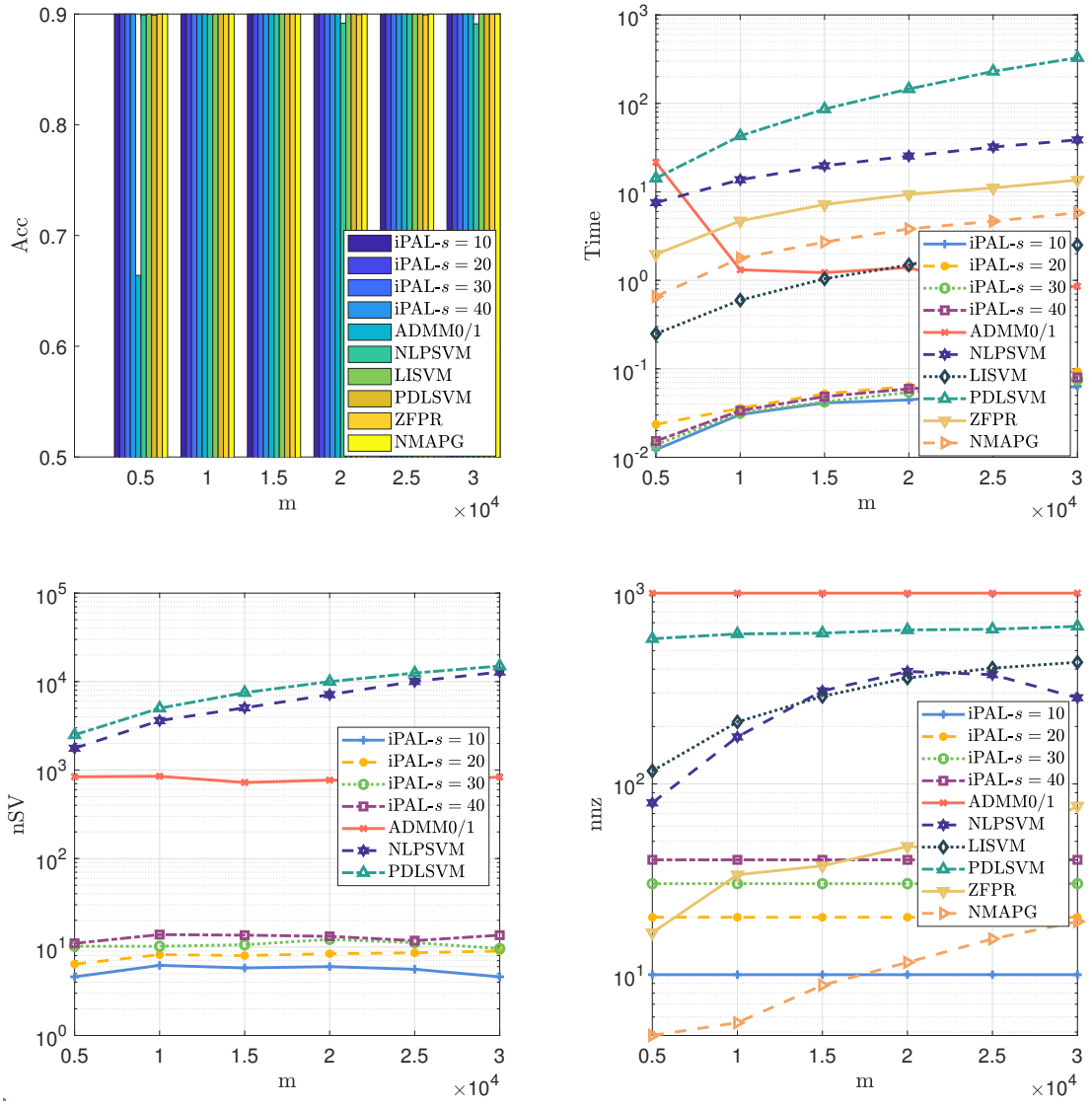

Figure 7: Comparison results on simulated data set with $n = 1000$, $r = 0.1$ and $m \in \{5000, 10000, \cdots, 30000\}$.

Test III. We fix $m = 1000$, $n = 10000$ and vary $r \in \{0.11, 0.12, \cdots, 0.16\}$. The numerical results are illustrated in Fig. 8. It can be observed that with the increase of noise rate, the `Acc` of all the algorithms drops. Particularly, the `Acc` of NLPSVM, ZFPR and NMAPG are more sensitive to noise rate than other solvers. The `nSV`, `nnz` and `Time` of all the algorithms are relatively stable with the change of r. We can see that when we increase

s from 10 to 40, the `nSV`, `Time` and `nnz` of iPAL will slightly rise, but the `Acc` remains stable. The similar phenomenons can be also observed in Test I and II.

The numerical experiments on the simulated data seem to suggest that iPAL is very competitive in terms of the four evaluating metrics. Similar behaviour of iPAL has also been consistently observed with the real data as we report below.

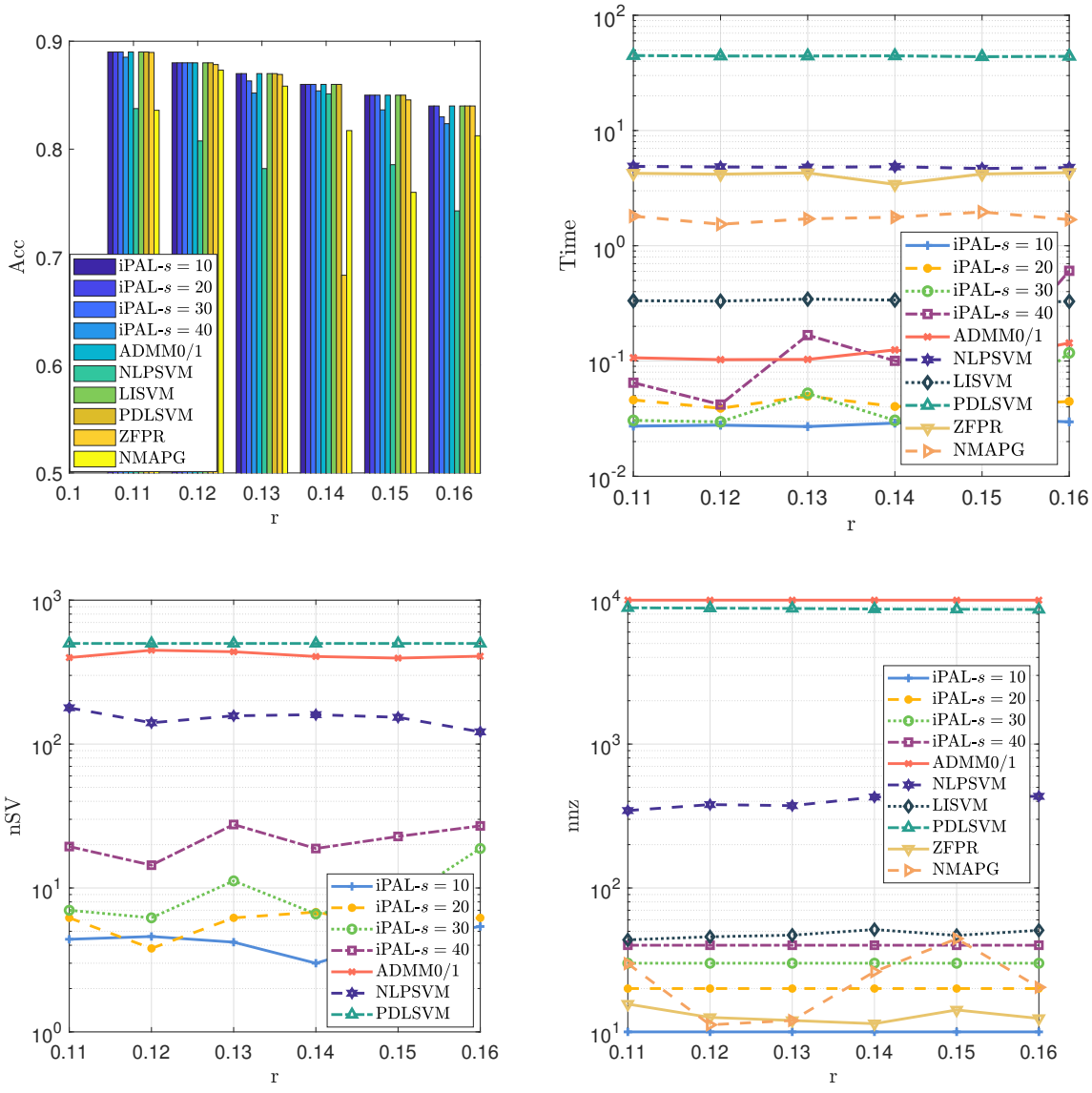

Figure 8: Comparison results on simulated data set with $m = 1000$, $n = 10000$ and $r \in \{0.11, 0.12, \cdots, 0.16\}$.

6.3 Experiments on Real Data

In this section, we will conduct numerical comparison on the real data sets listed in Table 2.

Example 4 *We select the data sets in Tables 2 and 3 with large number of features. Apart from* `gli` *and* `dex`*, all other data sets are preprocessed by feature-wise scaling to [-1,1].*

Table 2: Real binary classification data sets with $n > m$

ID	data set	Source	number of features	number of instances
`all`	ALLAML	feature selection database[1]	7129	72
`col`	Colon	feature selection database	2000	62
`gli`	GLI85	feature selection database	22283	85
`pro`	Prostate	feature selection database	5966	102
`smk`	SMK187	feature selection database	19993	187
`dor`	Dorothea	uci[2]	100000	1950
`dex`	Dexter	uci	20000	2600
`dbb`	Dbworld_bodies	uci	3721	64
`dbs`	Dbworld_subjects	uci	229	64
`abu`	AP_Breast_Uterus	openML[3]	10936	468
`alk`	AP_Lung_Kidney	openML	10936	368
`aou`	AP_Ovary_Uterus	openML	10936	322
`ove`	OVA_Endometrium	openML	10936	1545
`ovk`	OVA_Kidney	openML	10936	1545
`ovo`	OVA_Ovary	openML	10936	1545
`bre`	Breast	openML	24482	97
`ova`	Ovarian	openML	15155	253
`dbc`	Duke_breast_cancer	openML	7129	44
`ge1`	Gse2280	refine.bio [4]	11868	27
`ge2`	Gse7670	refine.bio	11868	54
`ge3`	Gse25099	refine.bio	16738	79
`ge4`	Gse27612	refine.bio	11868	195

In the following experiments, for iPAL, we set $\lambda = 1$, $\rho = 1$, $\mu = 10^{-2}$ and s was chosen from $\{\lceil 0.001n\rceil, \lceil 0.002n\rceil, \cdots, \lceil 0.01n\rceil, \lceil 0.02n\rceil, \cdots, \lceil 0.1n\rceil, \lceil 0.2n\rceil \cdots, n\}$, where $\lceil\cdot\rceil$ is the ceil function. For ZFPR and NMAPG, the regularization parameter of ℓ_0 norm will be selected from $\{2^{-10}, 2^{-9}, \cdots, 2^{10}\}$. For all other algorithms, the regularization parameters used for trade-off between regularizer and loss function were chosen from $\{10^{-5}, 10^{-4}, \cdots, 10^{5}\}$. We conduct five-fold cross-validation on all the data sets in Tables 2 and 3, and the average results are summarized into Table 4, 5 and 6. Some comments about these results are given as follows.

[1] https://jundongl.github.io/scikit-feature/
[2] https://archive-beta.ics.uci.edu/datasets
[3] https://www.openml.org/
[4] https://www.refine.bio/

Table 3: Real binary classification data sets with $n < m$

ID	data set	Source	number of features	number of instances
chr	Christine	openML	1636	5418
jas	Jasmine	openML	144	2984
mad	Madeline	openML	259	3140
phi	Philippine	openML	308	5382
hiv	Hiva_Agnostic	openML	1617	4229
gui	Guillermo	openML	4296	20000
evi	Evita	openML	3000	20000
bio	Bioresponse	openML	1776	3751
sw1	Swarm_Aligned	uci	2400	24016
sw2	Swarm_Flocking	uci	2400	24016
sw3	Swarm_Grouped	uci	2400	24016
int	Internet-Advertisements	uci	1559	3279
qar	QSAR_aquatic_receptor	uci	1024	8992
qot	QSAR_oral_toxity	uci	1024	1687

(i) **On** `Acc`. iPAL has the best `Acc` on most of data sets. When compared with ADMM0/1, iPAL achieves higher `Acc` with much smaller `nnz`. This means that the cardinality constraint is beneficial to improving performance of SVM when the number of features is large. ZFPR or NMAPG also achieve best `Acc` on some data sets with $n > m$, for example `chr` and `qar`.

(ii) **On** `Time`. iPAL shows competitive `Time` in this comparison. For example, on the `all`, `gli` and `ge4`, `Time` of iPAL is less than 1/3 that of LISVM and even less than 1/5 that of ZFPR and NMAPG with ℓ_0 regularizer. The high speed of iPAL mainly benefits from the reduction on `nSV` and `nnz`, which are accomplished by the proximal operator of hard margin loss function and projection of cardinality constraint.

(iii) **On** `nSV`. Both iPAL and NLPSVM have small `nSV`. However, NLPSVM tends to be aggressive on reducing `nSV` and causes the low `Acc`, see, for instance, `all`, `dbb` and `bre`. We can also observe that iPAL has smaller `nSV` than that of ADMM0/1 on almost all the data sets. A possible explanation is that when the redundant features are eliminated, it is easier for a classifier to identify support vectors.

(iv) **On** `nnz`. We can see that iPAL, LISVM, NLPSVM, ZFPR and NMAPG show significant reduction on `nnz` of solution. Particularly, iPAL has better performance on feature selection in the case of $n \gg m$, because it finds solution with smaller `nnz` while higher `Acc`. NLPSVM or LISVM has smaller `nnz` than that of iPAL in some cases such as `chr`, `gui` and `qar`, but iPAL has better `Acc` in those cases. ZFPR and NMAPG also show competitive performance on feature selection, but they do not find a sparse solution on some data sets such as `bre`, `chr` and `gui`.

Table 4: Experiment results in terms of `Acc` and `Time` on real data sets with $n > m$

	Acc (%)							Time (sec)						
	iPAL	ADMM0/1	NLPSVM	LISVM	PDLSVM	ZFPR	NMAPG	iPAL	ADMM0/1	NLPSVM	LISVM	PDLSVM	ZFPR	NMAPG
all	98.57	96.07	91.79	91.61	98.57	97.14	97.14	9.788e-3	2.168e-1	1.826e-1	3.498e-2	1.287e+1	5.299e-2	2.298e-1
col	90.00	87.38	80.71	80.95	85.71	79.52	87.14	3.817e-3	2.407e+0	6.580e-2	3.344e-3	8.048e-1	3.525e-2	9.844e-2
gli	88.24	88.24	82.35	88.24	88.24	83.53	88.24	2.027e-1	9.550e-1	3.502e-1	7.980e-1	2.147e+2	1.132e+0	1.457e+0
pro	95.00	93.00	93.00	94.00	90.18	94.00	94.00	6.852e-2	2.171e+0	1.176e-1	5.537e-2	9.259e+0	6.675e-1	2.018e-1
smk	77.46	74.79	72.11	73.68	75.93	78.06	75.87	8.418e-1	1.146e+1	3.974e-1	8.095e+0	1.621e+2	3.153e+0	1.648e+0
dor	93.39	92.52	80.00	93.30	–	91.22	93.91	1.113e+0	1.918e+0	3.405e+0	2.305e-1	–	1.537e+0	1.927e+0
dex	95.00	94.67	70.17	91.33	94.33	93.17	92.17	1.759e-2	4.538e-1	8.497e-1	1.375e-1	1.808e+2	1.031e-1	8.649e-1
dbb	90.42	89.17	78.75	86.25	90.83	87.92	89.17	2.326e-2	1.163e+0	1.867e-1	2.325e-2	5.512e+0	3.822e-2	2.088e-1
dbs	88.75	88.75	84.17	87.08	77.92	88.33	88.33	1.458e-3	2.811e-3	4.442e-3	1.951e-3	1.714e-2	1.392e-2	4.630e-2
abu	96.38	95.94	86.77	95.93	94.88	93.37	94.01	5.454e-1	1.064e+1	3.232e+0	6.317e-1	4.638e+1	3.932e+0	3.681e+0
alk	97.92	97.4	90.66	97.92	96.88	96.11	95.59	4.640e-1	1.707e+0	2.578e+0	8.298e-1	4.403e+1	2.812e+0	2.041e+0
aou	90.06	88.48	84.21	89.73	85.10	85.40	83.24	8.096e-1	3.056e+0	2.006e+0	1.629e+0	4.216e+1	2.748e+0	2.605e+0
ove	96.63	96.50	96.05	96.38	96.05	96.50	94.69	2.057e+0	1.218e+1	1.997e+1	1.247e+0	7.398e+1	1.244e+1	7.675e+0
ovk	98.71	98.71	88.22	98.58	97.86	98.19	94.89	2.738e+0	1.028e+1	1.927e+1	2.750e+0	7.328e+1	1.252e+1	1.042e+1
ovo	92.49	91.78	87.18	92.36	89.26	91.13	91.20	3.259e+0	1.596e+1	1.916e+1	1.985e+1	7.285e+1	1.251e+1	5.096e+0
bre	79.35	70.93	75.04	76.19	75.34	63.11	64.16	3.984e-2	4.084e-2	2.025e-1	9.198e-2	2.634e+2	2.172e+0	2.775e+0
ova	100.0	100.0	98.80	100.0	99.20	100.0	100.0	5.618e-2	1.148e+0	5.429e-1	1.978e-1	8.554e+1	2.476e+0	1.238e+0
dbc	90.83	90.83	81.67	88.33	86.67	91.67	90.83	2.517e-2	3.153e-1	1.502e-1	2.063e-2	1.340e+1	2.451e-1	5.899e-1
ge1	85.14	81.14	76.00	82.29	81.14	81.14	81.14	7.778e-3	1.978e+0	1.116e-1	8.129e-2	4.198e+1	1.320e-1	2.132e-1
ge2	98.00	98.00	90.57	96.00	96.00	94.00	96.00	9.653e-2	1.506e+0	3.762e-1	8.858e-2	4.275e+1	3.547e-1	4.397e-1
ge3	100.0	100.0	96.56	100.0	100.0	100.0	100.0	1.489e-2	8.836e-1	7.921e-1	9.266e-2	1.005e+2	4.980e-1	5.996e-1
ge4	100.0	100.0	90.26	100.0	100.0	100.0	100.0	2.440e-2	1.184e+0	1.318e+0	9.131e-2	4.588e+1	4.564e-1	6.675e-1

Note: PDLSVM fails to give a solution within 2 hours when solving `dor`. The `nSV` of LISVM with ℓ_1 regularizer, ZFPR and NMAPG are not recorded because they solves primal SVM without introducing dual variables.

Table 5: Experiment results in terms of `nSV` and `nnz` on real data sets with $n > m$

	nSV							nnz						
	iPAL	ADMM0/1	NLPSVM	LISVM	PDLSVM	ZFPR	NMAPG	iPAL	ADMM0/1	NLPSVM	LISVM	PDLSVM	ZFPR	NMAPG
all	22	49	11	–	20	–	–	43	7130	35	63	4206	307	24
col	11	36	16	–	43	–	–	7	2001	42	8	1485	186	23
gli	18	37	20	–	13	–	–	45	22284	56	1012	10846	672	292
pro	14	59	41	–	28	–	–	36	5967	52	159	3503	840	255
smk	57	127	38	–	145	–	–	800	19994	175	14732	13804	3155	791
dor	121	810	275	–	–	–	–	301	85488	2525	343	–	858	34
dex	362	409	189	–	480	–	–	1200	9244	634	1291	19999	2128	1389
dbb	33	49	12	–	51	–	–	189	3971	747	135	2340	256	439
dbs	39	45	19	–	26	–	–	73	193	44	44	138	60	56
abu	65	124	167	–	97	–	–	219	10937	209	355	4874	309	9050
alk	61	87	130	–	72	–	–	329	10937	154	1010	5208	431	5941
aou	122	157	120	–	72	–	–	547	10937	201	1759	5728	436	3492
ove	111	162	749	–	353	–	–	438	10937	571	129	3601	487	9497
ovk	107	159	543	–	381	–	–	766	10937	439	371	4331	789	8924
ovo	227	272	479	–	200	–	–	657	10937	500	3244	4539	672	7962
bre	43	63	9	–	13	–	–	74	24482	10	25	18099	1614	5790
ova	27	50	12	–	46	–	–	46	15155	12	10	7908	622	11
dbc	17	30	8	–	8	–	–	58	7130	34	32	4159	51	91
ge1	10	20	13	–	21	–	–	36	11869	22	412	10982	542	40
ge2	13	26	20	–	10	–	–	60	11869	251	416	525	116	103
ge3	12	26	17	–	15	–	–	51	16739	33	40	7053	637	54
ge4	25	38	8	–	54	–	–	24	11869	157	10	3501	191	8

Note: PDLSVM fails to give a solution within 2 hours when solving `dor`. The `nSV` of LISVM with ℓ_1 regularizer, ZFPR and NMAPG are not recorded because they solves primal SVM without introducing dual variables.

Table 6: Experiment results on real data sets with $n < m$

	Acc (%)							Time (sec)						
	iPAL	ADMMO/1	NLPSVM	LISVM	PDLSVM	ZFPR	NMAPG	iPAL	ADMMO/1	NLPSVM	LISVM	PDLSVM	ZFPR	NMAPG
chr	73.29	54.32	68.83	72.90	70.27	74.73	68.97	4.227e+0	5.572e+1	2.936e+1	3.388e+0	4.324e+1	6.595e+0	2.165e+0
jas	79.79	77.72	77.01	77.88	77.75	79.19	78.82	1.367e-1	4.826e-1	2.242e-1	2.231e-1	7.480e+0	8.402e-1	8.453e-2
mad	61.97	61.88	59.20	61.88	56.56	62.20	58.38	2.287e-1	2.415e-1	1.282e+0	9.529e-2	9.029e+0	9.130e-1	1.090e-1
phi	71.16	70.47	72.31	72.46	70.35	69.55	61.97	5.671e-1	4.977e-1	2.921e+0	2.795e+0	2.844e+1	7.212e-1	3.387e-1
hiv	96.48	93.26	96.48	96.69	96.50	96.41	96.48	8.283e-2	1.545e+2	8.204e+0	5.345e-1	2.903e+1	6.163e+0	2.183e+0
gui	72.91	70.08	60.16	72.00	70.02	64.97	64.07	2.483e+0	1.892e+2	4.979e+2	2.388e+0	5.304e+2	6.469e+1	2.551e+1
evi	96.59	96.59	96.70	96.59	96.80	96.43	96.59	3.457e-1	1.129e+0	9.467e+0	1.287e+0	4.608e+2	1.321e+1	6.250e+0
bio	73.69	61.72	52.07	76.73	74.94	73.39	73.39	5.817e-2	1.683e+1	2.196e+1	4.263e-1	2.547e+1	4.583e+0	1.753e+0
sw1	100.00	99.99	72.56	100.00	100.00	99.45	98.20	1.261e+0	3.073e+2	1.867e+2	3.132e+0	6.132e+2	4.322e+1	4.272e+1
sw2	99.97	99.98	72.20	99.99	99.94	97.38	97.07	1.232e+0	4.101e+2	1.963e+2	3.884e+0	6.085e+2	9.050e+1	8.002e+1
sw3	100.00	99.93	72.12	100.00	99.94	98.97	99.32	1.347e+0	3.767e+2	1.873e+2	2.871e+0	6.232e+2	4.380e+1	2.260e+1
int	97.10	97.07	86.00	95.88	90.30	93.05	91.86	6.209e-1	1.090e+2	2.427e+0	2.151e-2	1.983e+1	4.105e+0	8.568e-1
qar	89.03	77.77	88.21	86.72	89.75	89.33	88.62	1.527e-1	1.660e+1	6.141e-1	4.209e-1	5.874e+0	1.467e+0	3.619e-1
qot	92.39	92.19	92.38	91.16	92.26	91.86	92.87	8.777e-1	7.737e+0	9.038e+0	3.165e+0	8.099e+1	7.890e+0	2.516e+0
	nSV							nnz						
	iPAL	ADMMO/1	NLPSVM	LISVM	PDLSVM	ZFPR	NMAPG	iPAL	ADMMO/1	NLPSVM	LISVM	PDLSVM	ZFPR	NMAPG
chr	189	3272	2275	–	1645	–	–	492	1611	1221	364	905	646	1554
jas	251	128	1077	–	1919	–	–	44	137	33	130	90	94	56
mad	17	184	411	–	2512	–	–	24	260	119	11	253	10	259
phi	4	213	1781	–	4316	–	–	13	309	214	229	194	19	308
hiv	16	3234	2932	–	1210	–	–	12	1618	6	189	651	16	31
gui	1638	1161	6811	–	15850	–	–	258	4281	1165	108	2796	3923	633
evi	12	8	13686	–	15635	–	–	10	495	106	160	2027	33	26
bio	546	489	1406	–	605	–	–	8	1748	990	336	979	14	5
sw1	235	2750	17201	–	7612	–	–	481	2401	337	112	1298	842	623
sw2	346	6766	9830	–	7182	–	–	481	2401	1278	347	1133	1921	1103
sw3	290	5849	10174	–	7156	–	–	481	2401	1269	801	1155	788	1256
int	154	2623	2256	–	201	–	–	312	1559	131	430	440	130	385
qar	34	1350	1190	–	655	–	–	52	1025	50	342	572	46	60
qot	82	258	5860	–	2509	–	–	93	740	232	777	539	95	247

Note: The nSV of LISVM with ℓ_1 regularizer, ZFPR and NMAPG are not recorded because they solves primal SVM without introducing dual variables.

7. Conclusion

This paper aims to solve a nonsmooth and nonconvex problem (1). We define a pseudo KKT point to equivalently characterize its local minimizer. A sharper P-stationary point is further defined for algorithm design. To find such a stationary point, we develop an inexact proximal augmented Lagrangian method (iPAL), which comprises a primal and multiplier step. Based on the P-stationarity of the primal step, the inexactness measurement is carefully designed to ensure iPAL converges both globally and at a linear rate. To make the iPAL practically efficient, we design a projected gradient-Newton method (PGN) for computing the primal step with global and local quadratic rate. By the virtue of proximal operator of hard margin loss function and the projection of cardinality constraint, active samples and features can be identified to reduce the dimension of data matrix in PGN. In the extensive numerical comparison, iPAL shows effective reduction on active samples and features while ensuring high classification accuracy and fast computational speed.

This research brings new insights on nonconvex composite optimization with cardinality constraint. An interesting question is how to extend the convergence result to a more general model in which the quadratic term of (1) is replaced by a smooth function. In such an extension, the nice features of the strong convexity as well as the separable property of the quadratic function would be lost. Therefore, some proof techniques developed in this paper would not be applicable anymore. We leave the extension to future research.

Acknowledgements

We are grateful to the Action Editor and the anonymous reviewers for their valuable comments, which have significantly improved the quality of our paper. This paper was supported by Hong Kong RGC General Research Fund (PolyU/15309223), PolyU AMA Projects (P0044200, P0045347), the National Key R&D Program of China (2023YFA1011100), 111 Project of China (B16002), and the National Natural Science Foundation of China (12131004).

Appendix A. Proof of Lemma 2

Proof It follows from (8) and (9) that $\boldsymbol{\xi} \in \mathrm{Prox}_{\beta\lambda J(\cdot)}(\boldsymbol{\xi} + \beta\mathbf{v})$ if and only if one of the following cases occurs for each $i = 1, \ldots, n$:

$$\begin{cases} (i) & \xi_i = 0, & 0 \le v_i < \sqrt{2\lambda/\beta} \\ (ii) & \xi_i < 0, & v_i = 0 \\ (iii) & \xi_i > \sqrt{2\beta\lambda}, & v_i = 0 \\ (iv) & \xi_i = 0, & v_i = \sqrt{2\lambda/\beta} \\ (v) & \xi_i = \sqrt{2\beta\lambda}, & v_i = 0. \end{cases}$$

Combing those cases leads to

$$\begin{cases} v_i = 0, & \text{if } \xi_i \in (-\infty, 0) \cup [\sqrt{2\beta\lambda},\ \infty) \\ v_i \in [0,\ \sqrt{2\lambda/\beta}], & \text{if } \xi_i = 0. \end{cases}$$

This means that ξ_i must satisfy

$$\xi_i \in (-\infty, 0] \cup [\sqrt{2\beta\lambda},\ \infty)$$

The characterization (10) implies that $\boldsymbol{\xi} \in \mathrm{Prox}_{\beta\lambda J(\cdot)}(\boldsymbol{\xi})$. This proves the necessity part of the lemma. The sufficiency part is by direct verification. ■

Appendix B. Relationship of pseudo KKT and pseudo B-stationary points

Proposition 14 *For problem* (12)*, let us consider a reference point* $\mathbf{u}^* = (\mathbf{w}^*, \boldsymbol{\xi}^*)$.

(i) When $\|\mathbf{w}^*\|_0 = s$*, the point* $\mathbf{u}^*$ *is a pseudo KKT point if and only if it is a pseudo B-stationary point.*

(ii) When $\|\mathbf{w}^*\|_0 < s$*, if* $\mathbf{u}^*$ *is a pseudo KKT point, then it is a pseudo B-stationary point, while the converse is not necessarily true.*

Proof From (Cui and Pang, 2021, Definition 6.1.1), $\mathbf{u}^* = (\mathbf{w}^*, \boldsymbol{\xi}^*)$ is a pseudo B-stationary point of (12) if it is a B-stationary point of the following problem

$$\min_{\mathbf{w},\boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2, \text{s.t. } \ \mathbf{w}_{\mathcal{S}^*_>} \geq 0, \ \ \mathbf{w}_{\mathcal{S}^*_<} \leq 0, \ \ \mathbf{w}_{\overline{\mathcal{S}}^*} = 0, \ \ \boldsymbol{\xi}_{\mathcal{I}^*_-} \leq 0, \ \ \boldsymbol{\xi}_{\overline{\mathcal{I}}^*_-} \geq 0, \ A\mathbf{w} + \mathbf{1} = \boldsymbol{\xi}, \quad (42)$$

where $\mathcal{S}^*_> := \{i \in [n] : w^*_i > 0\}$ and $\mathcal{S}^*_< := \{i \in [n] : w^*_i < 0\}$. A direct observation is that (42) has stricter constraints than (NLP-T^*) for each $T^* \in \mathbb{T}^*$. Since (42) is convex programming, $\mathbf{u}^*$ is also a KKT point of this problem. Let us denote the corresponding optimal multipliers as $\mathbf{q}^{*(i)}, i = 1, \cdots, 5$ and $\mathbf{z}^*$ in sequential order. Considering $\mathbf{w}^*_{\mathcal{S}^*_>} > 0$, $\mathbf{w}^*_{\mathcal{S}^*_<} < 0$ and $\boldsymbol{\xi}^*_{\overline{\mathcal{I}}^*_-} > 0$, the KKT system of (42) at $\mathbf{u}^*$ can be simplified as

$$\begin{cases} (\mathbf{w}^* + A^\top \mathbf{z}^*)_{\mathcal{S}^*} = 0, \ \mathbf{w}^*_{\overline{\mathcal{S}}^*} = 0, \\ \mathbf{z}^*_{\mathcal{I}^*_-} \geq 0, \ \boldsymbol{\xi}^*_{\mathcal{I}^*_-} \leq 0, \ \langle \mathbf{z}^*_{\mathcal{I}^*_-}, \boldsymbol{\xi}^*_{\mathcal{I}^*_-} \rangle = 0, \ \mathbf{z}^*_{\overline{\mathcal{I}}^*_-} = 0, \\ A\mathbf{w}^* + \mathbf{1} - \boldsymbol{\xi}^* = 0, \ \mathbf{q}^{*(1)} = 0, \ \mathbf{q}^{*(2)} = 0, \\ \mathbf{q}^{*(3)} = -(\mathbf{w}^* + A^\top \mathbf{z}^*)_{\overline{\mathcal{S}}^*}, \ \mathbf{q}^{*(4)} = \mathbf{z}^*_{\mathcal{I}^*_-}, \ \mathbf{q}^{*(5)} = 0. \end{cases} \quad (43)$$

We can derive the relationships in (i) and (ii) by comparing (43) with (13). ■

Appendix C. Proof of Theorem 4

Let C be a nonempty closed set and $\mathbf{u} \in C$, then we denote the tangent cone of C at $\mathbf{u}$ as $\Xi(\mathbf{u}, C)$.

Proof Define $\varphi(\mathbf{u}) := \|\mathbf{w}\|^2/2 + \lambda J(\boldsymbol{\xi})$, then according to (Han et al., 2024, Definition 1), $\mathbf{u}^*$ is an epi-stationary point of (12) if $(\mathbf{u}^*, \varphi(\mathbf{u}^*))$ is a B-stationary of the following problem

$$\min_{(\mathbf{u},t)} \ t \quad s.t. \ \ \|\mathbf{w}\|^2/2 + \lambda J(\boldsymbol{\xi}) \leq t, \ \ \|\mathbf{w}\|_0 \leq s, \ \ A\mathbf{w} + \mathbf{1} = \boldsymbol{\xi}. \quad (44)$$

Let us denote the feasible region of the above optimization problem as Ω. Taking $T^* \in \mathbb{T}^*$, let us define

$$\Omega_{T^*} := \left\{ (\mathbf{u}, t) \ \middle| \ \frac{1}{2}\|\mathbf{w}\|^2 + \lambda|\overline{\mathcal{I}}^*_-| \leq t, \ A\mathbf{w} + \mathbf{1} = \boldsymbol{\xi}, \ \boldsymbol{\xi}_{\mathcal{I}^*_-} \leq 0, \ \mathbf{w}_{\overline{T}^*} = 0 \right\}.$$

Taking a sufficiently small $\delta^* > 0$, the following relationships hold for any $(\mathbf{u}, t) \in \mathcal{N}((\mathbf{u}^*, \varphi(\mathbf{u}^*)), \delta^*)$

$$\mathcal{S}^* \subseteq \{i \in [m] : w_i \neq 0\} \tag{45}$$

$$\{i \in [m] : \xi_i^* > 0\} \subseteq \{i \in [m] : \xi_i > 0\} \tag{46}$$

$$|t - \|\mathbf{w}\|^2/2 - (\varphi(\mathbf{u}^*) - \|\mathbf{w}^*\|^2/2)| \leq \lambda/2 \tag{47}$$

Now let us prove that there exists a neighborhood $\mathcal{N}((\mathbf{u}^*, \varphi(\mathbf{u}^*)), \delta^*)$ such that

$$\mathcal{N}((\mathbf{u}^*, \varphi(\mathbf{u}^*)), \delta^*) \cap \Omega = \mathcal{N}((\mathbf{u}^*, \varphi(\mathbf{u}^*)), \delta^*) \cap (\cup_{T^* \in \mathbb{T}^*} \ \Omega_{T^*}) \tag{48}$$

The " $\supseteq$ " conclusion can be verified by $T \in \mathbb{T}^*$ and (46). To prove " $\subseteq$ ", for any $(\mathbf{u}, t)$ taken from the left-hand side of the above relationship, $\|\mathbf{w}\|_0 \leq s$ and (45) imply that there exists $T^* \in \mathbb{T}^*$ such that $\mathbf{w}_{\overline{T}^*} = 0$. Moreover, $\lambda J(\boldsymbol{\xi}) \leq t - \|\mathbf{w}\|^2/2$ and (47) leads to $\lambda J(\boldsymbol{\xi}) \leq \lambda J(\boldsymbol{\xi}^*) + \lambda/2$. This together with (46) implies $J(\boldsymbol{\xi}) = J(\boldsymbol{\xi}^*)$ and $\boldsymbol{\xi}_{\mathcal{I}_-^*} \leq 0$. Therefore, the relationship (48) holds. Then following from (Ban et al., 2011, Proposition 3.1), the tangent cone of Ω at $(\mathbf{u}^*, \varphi(\mathbf{u}^*))$ can be represented as

$$\Xi((\mathbf{u}^*, \varphi(\mathbf{u}^*)), \Omega) = \bigcup_{T^* \in \mathbb{T}^*} \Xi((\mathbf{u}^*, \varphi(\mathbf{u}^*), \Omega_{T^*}),$$

Then according to (Han et al., 2024, Proposition 5), $\mathbf{u}^*$ is an epi-stationary point of (12) if and only if it is a B-stationary point for all (NLP-T^*) with $T^* \in \mathbb{T}^*$. Since (NLP-T^*) is convex programming, we can conclude that the epi-stationary point is equivalent to the pseudo KKT stationary point for (12). ■

Appendix D. Proof of Theorem 5

Before the proof, let us first show that (NLP-T^*) naturally satisfies the second-order necessary condition (SOSC, see e.g. Nocedal and Wright 2006, Theorem 12.5), which is well defined for smooth optimization.

Lemma 15 *Given a KKT pair $(\mathbf{u}^*, \mathbf{q}_w^*, \mathbf{q}_\xi^*, \mathbf{z}^*)$ of (NLP-T^*) with $T^* \in \mathbb{T}^*$, the following SOSC naturally holds*

$$[\mathbf{d}^w; \mathbf{d}^\xi]^\top \nabla^2_{\mathbf{u},\mathbf{u}} \mathcal{L}_{T^*}(\mathbf{u}^*, \mathbf{q}_w^*, \mathbf{q}_\xi^*, \mathbf{z}^*)[\mathbf{d}^w; \mathbf{d}^\xi] > 0, \quad \forall \ [\mathbf{d}^w; \mathbf{d}^\xi] \in \mathcal{C}^* \backslash \{0\}, \tag{49}$$

where $\mathcal{C}^* := \{(\mathbf{d}^w, \mathbf{d}^\xi) \in \mathbb{R}^{m+n} : A\mathbf{d}^w = \mathbf{d}^\xi, \ \mathbf{d}^w_{\overline{T}^*} = 0, \mathbf{d}^\xi_{\mathcal{I}_0^*} \leq 0, \ \mathbf{d}^\xi_{\mathcal{I}_+^*} = 0\}$ is the critical cone of (NLP-T^*), $\mathcal{I}_0^* := \{i \in [m] : \xi_i^* = 0, z_i^* = 0\}$ and $\mathcal{I}_+^* := \{i \in [m] : \xi_i^* = 0, z_i^* > 0\}$.

Proof The Hessian of the Lagrangian of (NLP-T^*) with respect to $\mathbf{u}$ can be written as

$$\nabla^2_{\mathbf{u},\mathbf{u}} \mathcal{L}_{T^*}(\mathbf{u}^*, \mathbf{z}_w^*, \mathbf{z}_\xi^*, \mathbf{z}^*) = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

Thus, (49) actually means

$$\|\mathbf{d}^w\|^2 > 0 \quad \forall \ [\mathbf{d}^w; \mathbf{d}^\xi] \in \mathcal{C}^* \backslash \{0\}.$$

Notice that $\mathbf{d}^w \neq 0$ must hold. Otherwise $A\mathbf{d}^w = \mathbf{d}^\xi$ would imply $0 = (\mathbf{d}^w, \mathbf{d}^\xi)$, contradicting with the assumption that it is not zero. Therefore $\|\mathbf{d}^w\|^2 > 0$ and the SOSC naturally holds. ■

Proof of Theorem 5. We define the feasible regions of (12) and (NLP-T^*) as

$$\mathcal{F} := \{\mathbf{u} = (\mathbf{w}, \boldsymbol{\xi}) : \|\mathbf{w}\|_0 \leq s, A\mathbf{w} + \mathbf{1} - \boldsymbol{\xi} = 0\},$$
$$\mathcal{F}_{T^*} := \{\mathbf{u} = (\mathbf{w}, \boldsymbol{\xi}) : \mathbf{w}_{\overline{T}^*} = 0, \boldsymbol{\xi}_{\mathcal{I}_-^*} \leq 0, A\mathbf{w} + \mathbf{1} - \boldsymbol{\xi} = 0\}.$$

(a) local minimizer $\Longrightarrow$ pseudo KKT point. Let $\mathbf{u}^* := (\mathbf{w}^*, \boldsymbol{\xi}^*)$ with $\boldsymbol{\xi}^* = A\mathbf{w}^* + \mathbf{1}$ be a local minimizer of (12), there exists $\epsilon^* > 0$ such that

$$\frac{1}{2}\|\mathbf{w}\|^2 + \lambda J(\boldsymbol{\xi}) \geq \frac{1}{2}\|\mathbf{w}^*\|^2 + \lambda J(\boldsymbol{\xi}^*), \text{ for all } \mathbf{u} \in \mathcal{N}(\mathbf{u}^*, \epsilon^*) \cap \mathcal{F}. \tag{50}$$

Given $T^* \in \mathbb{T}^*$, we have $\mathcal{F}_{T^*} \subseteq \mathcal{F}$ and let us consider $\mathbf{u} \in \mathcal{N}(\mathbf{u}^*, \epsilon^*) \cap \mathcal{F}_{T^*}$. Since $\mathcal{F}_{T^*} \subseteq \mathcal{F}$ and $J(\boldsymbol{\xi}^*) \geq J(\boldsymbol{\xi})$, we have the following inequality from (50)

$$\frac{1}{2}\|\mathbf{w}\|^2 \geq \frac{1}{2}\|\mathbf{w}^*\|^2, \text{ for all } \mathbf{u} \in \mathcal{N}(\mathbf{u}^*, \epsilon^*) \cap \mathcal{F}_{T^*},$$

which means that $\mathbf{u}^*$ is also a local minimizer of (NLP-T^*). Noticing that for each $T^* \in \mathbb{T}$, (NLP-T^*) is a smooth nonlinear optimization problem with linear constraints, we can further deduce that for any $T^* \in \mathbb{T}^*$, $\mathbf{u}^*$ is a KKT point of (NLP-T^*). Thus, $\mathbf{u}^*$ is a pseudo KKT point of (12).

(b) pseudo KKT point $\Longrightarrow$ local minimizer. Let $\mathbf{u}^* = (\mathbf{w}^*, \boldsymbol{\xi}^*)$ with $\boldsymbol{\xi}^* := A\mathbf{w}^* + \mathbf{1}$ be a KKT point of (NLP-T^*) for each $T^* \in \mathbb{T}^*$. Meanwhile, noticing that the SOSC (49) holds, it follows from (Nocedal and Wright, 2006, Theorem 12.6) that there exists $\epsilon_{T^*} > 0$ and $c_{T^*} > 0$ such that

$$\frac{1}{2}\|\mathbf{w}\|^2 \geq \frac{1}{2}\|\mathbf{w}^*\|^2 + c_{T^*}\|\mathbf{u} - \mathbf{u}^*\|^2, \ \forall \mathbf{u} \in \mathcal{N}(\mathbf{u}^*, \epsilon_{T^*}) \cap \mathcal{F}_{T^*}. \tag{51}$$

Denote $c^* := \min_{T^* \in \mathbb{T}^*} c_{T^*}$. Now we take a radius ϵ^* satisfying

$$\epsilon^* < \min_{T^* \in \mathbb{T}^*} \epsilon_{T^*} \quad \text{and} \quad c^* {\epsilon^*}^2 < \lambda/2. \tag{52}$$

We also assume that ϵ^* is small enough such that for any $\mathbf{u} \in \mathcal{N}(\mathbf{u}^*, \epsilon^*)$, the following relationships hold

$$\mathcal{S}^* \subseteq \{i \in [m] : w_i \neq 0\} \text{ and } \{i \in [m] : \xi_i^* > 0\} \subseteq \{i \in [m] : \xi_i > 0\}, \tag{53}$$
$$|\|\mathbf{w}\|^2 - \|\mathbf{w}^*\|^2| < \lambda, \tag{54}$$

where the inequality follows from the continuity of $\|\cdot\|^2$. Particularly, (53) further leads to

$$J(\boldsymbol{\xi}) \geq J(\boldsymbol{\xi}^*). \tag{55}$$

Denoting $\mathcal{F}^* := \bigcup_{T^* \in \mathbb{T}^*} \mathcal{F}_{T^*} \subseteq \mathcal{F}$, then from (51) and (55), we can obtain

$$\frac{1}{2}\|\mathbf{w}\|^2 + \lambda J(\boldsymbol{\xi}) \geq \frac{1}{2}\|\mathbf{w}^*\|^2 + \lambda J(\boldsymbol{\xi}^*) + c^*\|\mathbf{u} - \mathbf{u}^*\|^2, \ \forall \mathbf{u} \in \mathcal{N}(\mathbf{u}^*, \epsilon^*) \cap \mathcal{F}^*.$$

If we take $\mathbf{u} \in \mathcal{N}(\mathbf{u}^*, \epsilon^*) \cap (\mathcal{F} \backslash \mathcal{F}^*)$, considering $\mathcal{S}^* \subseteq \{i \in [n] : w_i \neq 0\}$ in (54) and $\|\mathbf{w}\|_0 \leq s$, there must exists $T^* \in \mathbb{T}^*$ such that $\mathbf{w}_{\overline{T}^*} = 0$. This together with $\mathbf{u} \notin \mathcal{F}^*$ lead to $\boldsymbol{\xi}_{\mathcal{I}_-^*} \not\leq 0$. There exists an index $i_0 \in \mathcal{I}_-^*$ such that $\xi_{i_0} > 0$. Combining this with (53) leads to $J(\boldsymbol{\xi}) \geq J(\boldsymbol{\xi}^*) + 1$. Then taking (54) and (55) into consideration, we have

$$\frac{1}{2}\|\mathbf{w}\|^2 + \lambda J(\boldsymbol{\xi}) \geq \frac{1}{2}\|\mathbf{w}^*\|^2 + \lambda J(\boldsymbol{\xi}^*) + \lambda/2 \overset{(52)}{\geq} \frac{1}{2}\|\mathbf{w}^*\|^2 + \lambda J(\boldsymbol{\xi}^*) + c^*\|\mathbf{u} - \mathbf{u}^*\|^2.$$

Overall, we have obtained

$$\frac{1}{2}\|\mathbf{w}\|^2 + \lambda J(\boldsymbol{\xi}) \geq \frac{1}{2}\|\mathbf{w}^*\|^2 + \lambda J(\boldsymbol{\xi}^*) + c^*\|\mathbf{u} - \mathbf{u}^*\|^2, \ \forall \mathbf{u} \in \mathcal{N}(\mathbf{u}^*, \epsilon^*) \cap \mathcal{F}.$$

Finally, (15) follows from the definition of $\mathcal{F}$. ■

Appendix E. Proof of Theorem 7

Proof If $\mathbf{u}^*$ is a P-stationary point of (12), then there exists a P-stationary multiplier $\mathbf{z}^*$ such that $(\mathbf{u}^*, \mathbf{z}^*)$ satisfies (16). Let us first prove $(\mathbf{w}^* + A^\top \mathbf{z}^*)_{T^*} = 0$ and $\mathbf{w}^*_{\overline{T}^*} = 0$.

The claim $\mathbf{w}^*_{\overline{T}^*} = 0$ directly follows from $T^* \supseteq \mathcal{S}^*$. If $\|\mathbf{w}^*\|_0 = s$, then $T^* \in \mathbb{T}^* = \mathcal{S}^*$. By the definition of $\mathcal{S}^*$, (6) implies $(\mathbf{w}^* + A^\top \mathbf{z}^*)_{T^*} = 0$. We can take the multipliers $\mathbf{q}^*_w = [\mathbf{w}^* + A^\top \mathbf{z}^*]_{\overline{\mathcal{S}}^*}$ and $\mathbf{q}^*_\xi = \mathbf{z}^*_{\mathcal{I}_-^*}$. If $\|\mathbf{w}^*\|_0 < s$, then $|\mathbf{w}^*|_{(s)} = 0$ and $\mathbf{w}^* + A^\top \mathbf{z}^* = 0$ holds from (6). The second line of (13) can be obtained from Lemma 2. We can take the multipliers $\mathbf{q}^*_w = 0$ and $\mathbf{q}^*_\xi = \mathbf{z}^*_{\mathcal{I}_-^*}$. ■

Appendix F. Proofs on Global Convergence of iPAL

In this part, our ultimate goal is to prove Theorem 9. It is beneficial to briefly explain the main ideas behind our proofs.

- First, we will prove Proposition 8, including the sufficient decrease of Lyapunov function in (25), boundedness of the sequence $\{(\mathbf{u}^k, \mathbf{z}^k)\}_{k \in \mathbb{N}}$, and the convergence of difference of successive iterates (26).

- The boundedness of sequence ensures that there must exist an accumulated point. The inexact criteria (22) actually means that each iterate approximately satisfies a P-stationary system and the degree of approximation can be measured by $\|\mathbf{w}^{k+1} - \mathbf{w}^k\|$. For such a sequence, each accumulated point is a P-stationary point of (12) by using (26) and the proximal behavior (Rockafellar, 1976, Theorem 1.25). This result is referred to as a subsequence convergence property (see Lemma 16).

- We will mainly use (Kanzow and Qi, 1999, Proposition 7) to prove that the whole sequence generated by iPAL is convergent. The requirements for using this proposition are (26) and the isolatedness of accumulation points. The isolatedness property follows from Theorem 5.

Proof of Proposition 8. By the definition of g_k and (23), we have

$$\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}) = \mathbf{w}^{k+1} + \mu(\mathbf{w}^{k+1} - \mathbf{w}^k) + A^\top \mathbf{z}^{k+1} \tag{56}$$
$$\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^{k+1}) = -\mathbf{z}^{k+1}. \tag{57}$$

These facts will be frequently used in the following proofs.

(i) First, we need to estimate an upper bound for $\|\mathbf{z}^{k+1} - \mathbf{z}^k\|$. If $|T_{k+1} \cap T_k| \geq r$, from (56), we have

$$\begin{aligned} A^\top_{:,T_{k+1}\cap T_k}(\mathbf{z}^{k+1} - \mathbf{z}^{k+1}) =& [\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}) - \nabla_{\mathbf{w}} g_{k-1}(\mathbf{u}^k) - (\mathbf{w}^{k+1} - \mathbf{w}^k) \\ & - \mu(\mathbf{w}^{k+1} - \mathbf{w}^k) + \mu(\mathbf{w}^k - \mathbf{w}^{k-1})]_{T_{k+1}\cap T_k} \end{aligned}$$

Using Assumption 1, we can further estimate

$$\begin{aligned} \gamma\|\mathbf{z}^{k+1} - \mathbf{z}^k\| \leq& \|A^\top_{:,T_{k+1}\cap T_k}(\mathbf{z}^{k+1} - \mathbf{z}^k)\| \leq \|\nabla_{T_{k+1}} g_k(\mathbf{u}^{k+1})\| + \|\nabla_{T_k} g_{k-1}(\mathbf{u}^k)\| \\ & + \|\mathbf{w}^{k+1} - \mathbf{w}^k\| + \mu\|\mathbf{w}^{k+1} - \mathbf{w}^k\| + \mu\|\mathbf{w}^k - \mathbf{w}^{k-1}\| \\ \overset{(22)}{\leq}& (c_1 + \mu + 1)\|\mathbf{w}^{k+1} - \mathbf{w}^k\| + (c_1 + \mu)\|\mathbf{w}^k - \mathbf{w}^{k-1}\|. \end{aligned} \tag{58}$$

If $|T_{k+1} \cap T_k| < r$, then taking $|T_{k+1}| = |T_k| = s$ into account, $|T_{k+1} \cap \overline{T}_k| = |\overline{T}_{k+1} \cap T_k| \geq r$ holds. By (56) and Assumption 1, we can obtain

$$\begin{aligned} \gamma\|\mathbf{z}^{k+1}\| \leq& \|A^\top_{:,T_{k+1}\cap \overline{T}_k}\mathbf{z}^{k+1}\| \\ \overset{(56)}{\leq}& \|\nabla_{T_{k+1}\cap\overline{T}_k} g_k(\mathbf{u}^{k+1})\| + \|[\mathbf{w}^{k+1} + \mu(\mathbf{w}^{k+1} - \mathbf{w}^k)]_{T_{k+1}\cap\overline{T}_k}\| \\ \overset{(22)}{\leq}& c_1\|\mathbf{w}^{k+1} - \mathbf{w}^k\| + \|\mathbf{w}^{k+1}_{T_{k+1}\cap\overline{T}_k}\| + \mu\|\mathbf{w}^{k+1} - \mathbf{w}^k\| \\ \leq& (c_1 + \mu)\|\mathbf{w}^{k+1} - \mathbf{w}^k\| + \|[\mathbf{w}^{k+1} - \mathbf{w}^k]_{T_{k+1}\cap\overline{T}_k}\| + \|\mathbf{w}^k_{T_{k+1}\cap\overline{T}_k}\| \\ \overset{(22)}{\leq}& (c_1 + \mu + 1)\|\mathbf{w}^{k+1} - \mathbf{w}^k\| + c_1\|\mathbf{w}^k - \mathbf{w}^{k-1}\|. \end{aligned} \tag{59}$$

$$\begin{aligned} \gamma\|\mathbf{z}^k\| \leq& \|A^\top_{:,\overline{T}_{k+1}\cap T_k}\mathbf{z}^k\| \overset{(56)}{\leq} \|\nabla_{\overline{T}_{k+1}\cap T_k} g_{k-1}(\mathbf{u}^k)\| + \|[\mathbf{w}^k + \mu(\mathbf{w}^k - \mathbf{w}^{k-1})]_{\overline{T}_{k+1}\cap T_k}\| \\ \overset{(22)}{\leq}& c_1\|\mathbf{w}^k - \mathbf{w}^{k-1}\| + \|\mathbf{w}^k_{\overline{T}_{k+1}\cap T_k}\| + \mu\|\mathbf{w}^k - \mathbf{w}^{k-1}\| \\ \leq& (c_1 + \mu)\|\mathbf{w}^k - \mathbf{w}^{k-1}\| + \|[\mathbf{w}^{k+1} - \mathbf{w}^k]_{\overline{T}_{k+1}\cap T_k}\| + \|\mathbf{w}^{k+1}_{\overline{T}_{k+1}\cap T_k}\| \\ \overset{(22)}{\leq}& (c_1 + \mu)\|\mathbf{w}^k - \mathbf{w}^{k-1}\| + (c_1 + 1)\|\mathbf{w}^{k+1} - \mathbf{w}^k\|. \end{aligned}$$

Adding the two inequalities above yields

$$\begin{aligned}\gamma\|\mathbf{z}^{k+1}-\mathbf{z}^k\| \leq & \gamma\|\mathbf{z}^{k+1}\|+\gamma\|\mathbf{z}^k\| \\ \leq & (2c_1+\mu+2)\|\mathbf{w}^{k+1}-\mathbf{w}^k\|+(2c_1+\mu)\|\mathbf{w}^k-\mathbf{w}^{k-1}\|.\end{aligned}$$

Combining this inequality and (58) leads to

$$\|\mathbf{z}^{k+1}-\mathbf{z}^k\| \leq c_3\|\mathbf{w}^{k+1}-\mathbf{w}^k\|+c_4\|\mathbf{w}^k-\mathbf{w}^{k-1}\|. \tag{60}$$

By using arithmetic mean and quadratic mean inequality, we can obtain

$$\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2 \leq 2c_3^2\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2+2c_4^2\|\mathbf{w}^k-\mathbf{w}^{k-1}\|^2. \tag{61}$$

From the definition of Lyapunov function and the first line of (22), we have the following chain of inequalities

$$\begin{aligned}\mathcal{L}_\rho(\mathbf{u}^k,\mathbf{z}^k)-\mathcal{L}_\rho(\mathbf{u}^{k+1},\mathbf{z}^{k+1}) = & \mathcal{L}_\rho(\mathbf{u}^k,\mathbf{z}^k)-\mathcal{L}_\rho(\mathbf{u}^{k+1},\mathbf{z}^k)+\mathcal{L}_\rho(\mathbf{u}^{k+1},\mathbf{z}^k)-\mathcal{L}_\rho(\mathbf{u}^{k+1},\mathbf{z}^{k+1}) \\ \overset{(22,23)}{\geq} & \frac{\mu}{2}\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2-\frac{1}{\rho}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|^2 \\ \overset{(61)}{\geq} & (\frac{\mu}{2}-\frac{2c_3^2}{\rho})\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2-\frac{2c_4^2}{\rho}\|\mathbf{w}^k-\mathbf{w}^{k-1}\|^2.\end{aligned}$$

Then we can further estimate

$$\begin{aligned}\mathcal{M}_k-\mathcal{M}_{k+1} = & \mathcal{L}_\rho(\mathbf{u}^k,\mathbf{z}^k)+\frac{\eta}{2}\|\mathbf{w}^k-\mathbf{w}^{k-1}\|-\mathcal{L}_\rho(\mathbf{u}^{k+1},\mathbf{z}^{k+1})-\frac{\eta}{2}\|\mathbf{w}^{k+1}-\mathbf{w}^k\| \\ \geq & (\frac{\mu}{2}-\frac{2c_3^2}{\rho}-\frac{\eta}{2})\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2+(\frac{\eta}{2}-\frac{2c_4^2}{\rho})\|\mathbf{w}^k-\mathbf{w}^{k-1}\|^2 \\ \overset{(24)}{\geq} & \frac{\mu}{4}\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2\end{aligned}$$

(ii) From (56), we can obtain

$$A_{:T_{k+1}}^\top\mathbf{z}^{k+1}=[\nabla_{\mathbf{w}}g_k(\mathbf{u}^{k+1})-\mathbf{w}^{k+1}-\mu(\mathbf{w}^{k+1}-\mathbf{w}^k)]_{T_{k+1}}$$

Using Assumption 1 and (22), we derive

$$\gamma\|\mathbf{z}^{k+1}\| \leq c_1\|\mathbf{w}^{k+1}-\mathbf{w}^k\|+\|\mathbf{w}^{k+1}\|+\mu\|\mathbf{w}^{k+1}-\mathbf{w}^k\| \leq \|\mathbf{w}^{k+1}\|+(c_1+\mu)\|\mathbf{w}^{k+1}-\mathbf{w}^k\|.$$

By using arithmetic and quadratic mean inequality, we have

$$\|\mathbf{z}^{k+1}\|^2 \leq \frac{2}{\gamma^2}\|\mathbf{w}^{k+1}\|^2+\frac{2(c_1+\mu)^2}{\gamma^2}\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2 \tag{62}$$

The following chain of inequalities holds by (25) and the definition of the Lyapunov function

$$\begin{aligned}\mathcal{M}_1 \geq & \mathcal{M}_{k+1} = \frac{1}{2}\|\mathbf{w}^{k+1}\|^2 + \langle \mathbf{z}^{k+1}, A\mathbf{w}^{k+1} + \mathbf{1} - \boldsymbol{\xi}^{k+1}\rangle + \frac{\rho}{2}\|A\mathbf{w}^{k+1} + \mathbf{1} - \boldsymbol{\xi}^{k+1}\|^2 \\ & + \frac{\eta}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 + \delta_{\mathbb{S}}(\mathbf{w}^{k+1}) + \lambda J(\boldsymbol{\xi}^{k+1}) \\ \geq & \frac{1}{2}\|\mathbf{w}^{k+1}\|^2 + \frac{\rho}{2}\|A\mathbf{w}^{k+1} + \mathbf{1} - \boldsymbol{\xi}^{k+1} + \mathbf{z}^{k+1}/\rho\|^2 + \frac{\eta}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 - \frac{1}{2\rho}\|\mathbf{z}^{k+1}\|^2 \\ \overset{(62)}{\geq} & \left(\frac{1}{2} - \frac{1}{\rho\gamma^2}\right)\|\mathbf{w}^{k+1}\|^2 + \left(\frac{\eta}{2} - \frac{(c_1+\mu)^2}{\rho\gamma^2}\right)\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 \\ & + \frac{\rho}{2}\|A\mathbf{w}^{k+1} + \mathbf{1} - \boldsymbol{\xi}^{k+1} + \frac{1}{\rho}\mathbf{z}^{k+1}\|^2. \end{aligned} \tag{63}$$

Taking (24) into account, both quantities $(1/2 - 1/(\rho\gamma^2))$ and $(\eta/2 - (c_1+\mu)^2/(\rho\gamma^2))$ are positive. Thus the sequences $\{\mathbf{w}^{k+1}\}_{k\in\mathbb{N}}$, $\{\mathbf{w}^{k+1} - \mathbf{w}^k\}_{k\in\mathbb{N}}$ and $\{A\mathbf{w}^{k+1} + \mathbf{1} - \boldsymbol{\xi}^{k+1} + \mathbf{z}^{k+1}/\rho\}_{k\in\mathbb{N}}$ are bounded. Then (62) leads to the boundedness of $\{\mathbf{z}^{k+1}\}_{k\in\mathbb{N}}$. The bound

$$\|\boldsymbol{\xi}^{k+1}\| \leq \|A\mathbf{w}^{k+1} + \mathbf{1} - \boldsymbol{\xi}^{k+1} + \mathbf{z}^{k+1}/\rho\| + \|A\|\|\mathbf{w}^{k+1}\| + \|\mathbf{1}\| + \|\mathbf{z}^{k+1}\|/\rho$$

implies the boundedness of $\{\boldsymbol{\xi}^{k+1}\}_{k\in\mathbb{N}}$. Overall, the generated sequence $\{(\mathbf{u}^k, \mathbf{z}^k)\}_{k\in\mathbb{N}}$ is bounded.

Finally, let us prove the successive changes of the sequence converge to zero. Actually, (63) implies $\mathcal{M}_{k+1} \geq 0$ for all $k \in \mathbb{N}$. Combining this and the nonincreasing property (25), it follows from the monotone convergence theorem that sequence $\{\mathcal{M}_{k+1}\}_{k\in\mathbb{N}}$ must be convergent. Therefore, $\lim_{k\to\infty}\|\mathbf{w}^{k+1} - \mathbf{w}^k\| = 0$. Considering that (60) holds, we have $\lim_{k\to\infty}\|\mathbf{z}^{k+1} - \mathbf{z}^k\| = 0$. Finally, by using (23), we can obtain

$$\|\boldsymbol{\xi}^{k+1} - \boldsymbol{\xi}^k\| \leq (\|\mathbf{z}^{k+1} - \mathbf{z}^k\| + \|\mathbf{z}^k - \mathbf{z}^{k-1}\|)/\rho + \|A\|\|\mathbf{w}^{k+1} - \mathbf{w}^k\|.$$

which implies $\lim_{k\to\infty}\|\boldsymbol{\xi}^{k+1} - \boldsymbol{\xi}^k\| = 0$ by $\lim_{k\to\infty}\|\mathbf{w}^{k+1} - \mathbf{w}^k\| = 0$ and $\lim_{k\to\infty}\|\mathbf{z}^{k+1} - \mathbf{z}^k\| = 0$. ■

Lemma 16 *(Subsequence Convergence) Suppose that Assumption 1 holds and parameters are chosen as* (24)*. Let* $\{(\mathbf{u}^k; \mathbf{z}^k)\}_{k\in\mathbb{N}}$ *be a sequence generated by iPAL, then each of its accumulations points is a* P*-stationary pair of* (12)*. Furthermore,* $\mathbf{u}^*$ *is a strict local minimizer of* (12)*.*

Proof Suppose that $(\mathbf{u}^*, \mathbf{z}^*)$ is an accumulation point of $\{(\mathbf{u}^k; \mathbf{z}^k)\}_{k\in\mathbb{N}}$. Then there exists a subsequence $\{(\mathbf{u}^k; \mathbf{z}^k)\}_{k\in\mathcal{K}}$ with $\lim_{k\in\mathcal{K},\, k\to\infty}(\mathbf{u}^k, \mathbf{z}^k) = (\mathbf{u}^*, \mathbf{z}^*)$. It follows from (26) that $\{(\mathbf{u}^{k+1}, \mathbf{z}^{k+1})\}_{k\in K}$ also converges to $(\mathbf{u}^*, \mathbf{z}^*)$. Let us take

$$\overline{\mathbf{w}}^{k+1} := \begin{bmatrix} [\mathbf{w}^{k+1} - \alpha\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1})]_{T_{k+1}} \\ \mathbf{0} \end{bmatrix} \text{ and } \overline{\boldsymbol{\xi}}^{k+1} := \begin{bmatrix} [\boldsymbol{\xi}^{k+1} - \beta\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^{k+1})]_{\Gamma_{k+1}} \\ \mathbf{0} \end{bmatrix}.$$

By the definition of T_{k+1} and Γ_{k+1}, $\overline{\mathbf{w}}^{k+1}$ and $\overline{\boldsymbol{\xi}}^{k+1}$ actually satisfy

$$\overline{\mathbf{w}}^{k+1} \in \mathrm{Proj}_{\mathbb{S}}(\mathbf{w}^{k+1} - \alpha\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1})) \text{ and } \overline{\boldsymbol{\xi}}^{k+1} \in \mathrm{Prox}_{\beta\lambda J(\cdot)}(\boldsymbol{\xi}^{k+1} - \beta\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^{k+1})). \tag{64}$$

We can also estimate

$$\|\overline{\mathbf{w}}^{k+1} - \mathbf{w}^{k+1}\| = \|[\alpha \nabla_{T_{k+1}} g_k(\mathbf{u}^{k+1}); \mathbf{w}^{k+1}_{\overline{T}_{k+1}}]\| \leq \max\{c_1\alpha, c_1\}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|$$
$$\|\overline{\boldsymbol{\xi}}^{k+1} - \boldsymbol{\xi}^{k+1}\| = \|[\beta \nabla_{\Gamma_{k+1}} g_k(\mathbf{u}^{k+1}); \boldsymbol{\xi}^{k+1}_{\overline{\Gamma}_{k+1}}]\| \leq \max\{c_2\beta, c_2\}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2.$$

Considering (26), $\lim_{k\to\infty} \|\overline{\mathbf{w}}^{k+1} - \mathbf{w}^{k+1}\| = \lim_{k\to\infty} \|\overline{\boldsymbol{\xi}}^{k+1} - \boldsymbol{\xi}^{k+1}\| = 0$ hold. This together with $\lim_{k\to\infty, k\in\mathcal{K}} \|\mathbf{u}^{k+1} - \mathbf{u}^*\| = 0$ leads to

$$\lim_{k\to\infty, k\in\mathcal{K}} \overline{\mathbf{w}}^{k+1} = \mathbf{w}^* \text{ and } \lim_{k\to\infty, k\in\mathcal{K}} \overline{\boldsymbol{\xi}}^{k+1} = \boldsymbol{\xi}^*. \tag{65}$$

Besides, passing $k \to \infty$ for $k \in \mathcal{K}$ on both sides of (56), (57) and (23) leads to

$$\begin{aligned} &\lim_{k\in, k\to\infty} \nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}) = \mathbf{w}^* + A^\top \mathbf{z}^* \\ &\lim_{k\in\mathcal{K}, k\to\infty} \nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^{k+1}) = -\mathbf{z}^*. \\ &A\mathbf{w}^* + \mathbf{1} - \boldsymbol{\xi}^* = 0. \end{aligned} \tag{66}$$

Since (64), (65) and (66) hold, it follows from (Rockafellar, 1976, Theorem 1.25) that $(\mathbf{u}^*, \mathbf{z}^*)$ will be a P-stationary pair satisfying (16). Finally, using Theorem 5, we can conclude that $\mathbf{u}^*$ is also a strict local minimizer of (12). ■

Proof of Theorem 9 Let us first prove that $\{\mathbf{u}^k\}_{k\in\mathbb{N}}$ converges to a P-stationary point of (12). Let $\mathbf{u}^* = (\mathbf{w}^*, \boldsymbol{\xi}^*)$ be an accumulation point. Lemma 16 and the proof of Theorem 5 indicate that $\mathbf{w}^*$ is the unique solution of the following problem

$$\min_{\mathbf{w},\boldsymbol{\xi}} \quad \frac{1}{2}\|\mathbf{w}\|^2, \quad \text{s.t.} \quad \mathbf{w}_{\overline{T}^*} = 0, \quad (A\mathbf{w} + \mathbf{1})_{\mathcal{I}^*_-} \leq 0,$$

because the objective $\|\mathbf{w}\|^2/2$ is strongly convex. Considering $T^* \subseteq [n]$ and $\mathcal{I}^*_- \subseteq [m]$, there are only finite accumulation points in sequence $\{\mathbf{u}^k\}_{k\in\mathbb{N}}$, and thus each accumulation point is isolated. Moreover, taking (26) into account, (Kanzow and Qi, 1999, Proposition 7) implies $\lim_{k\to\infty} \mathbf{u}^k = \mathbf{u}^*$. We further estimate

$$\|\mathbf{w}^*_{\overline{T}_{k+1}}\| \leq \|[\mathbf{w}^{k+1} - \mathbf{w}^*]_{\overline{T}_{k+1}}\| + \|\mathbf{w}^{k+1}_{\overline{T}_{k+1}}\| \overset{(22)}{\leq} \|\mathbf{w}^{k+1} - \mathbf{w}^*\| + c_1\|\mathbf{w}^{k+1} - \mathbf{w}^k\|$$

Taking limit as $k \to \infty$ on both sides of above inequality leads to $\lim_{k\to\infty} \|\mathbf{w}^*_{\overline{T}_{k+1}}\| = 0$, which means that $\mathbf{w}^*_{\overline{T}_{k+1}} = 0$ when k is sufficiently large. We then have

$$T_{k+1} \supseteq \mathcal{S}^* := \{i \in [n] : w^*_i \neq 0\}. \tag{67}$$

Suppose that $\mathbf{z}^*$ is a P-stationary multiplier associated with $\mathbf{u}^*$. Now let us prove $\lim_{k\to\infty} \mathbf{z}^k = \mathbf{z}^*$. To achieve this goal, we need to give an upper bound for $\|\mathbf{z}^{k+1} - \mathbf{z}^*\|$. We claim that the following equation holds when k is sufficiently large

$$(\mathbf{w}^* + A^\top \mathbf{z}^*)_{T_{k+1}} = 0.$$

Indeed, if $\|\mathbf{w}^*\|_0 = s$, then the $T_{k+1} = \mathcal{S}^*$ follows from (67) and $|T_{k+1}| = s$. This and (6) further leads to the above equation. If $\|\mathbf{w}^*\|_0 < s$, then we have $\mathbf{w}^* + A^\top \mathbf{z}^* = 0$ by (6). Moreover, considering that $|T_{k+1}| = s$ holds, we can use Assumption 1 and (56) to derive

$$\begin{aligned}\gamma\|\mathbf{z}^{k+1} - \mathbf{z}^*\| \le& \|[A^\top(\mathbf{z}^{k+1} - \mathbf{z}^*)]_{T_{k+1}}\| \\ \le& \|[\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}) - (\mathbf{w}^{k+1} - \mathbf{w}^*) - \mu(\mathbf{w}^{k+1} - \mathbf{w}^k)]_{T_{k+1}}\| \\ \le& \|\nabla_{T_{k+1}} g_k(\mathbf{u}^{k+1})\| + \|\mathbf{w}^{k+1} - \mathbf{w}^*\| + \mu\|\mathbf{w}^{k+1} - \mathbf{w}^k\| \\ \overset{(22)}{\le}& (c_1 + \mu)\|\mathbf{w}^{k+1} - \mathbf{w}^k\| + \|\mathbf{w}^{k+1} - \mathbf{w}^*\|. \end{aligned} \tag{68}$$

Considering that we have proved $\lim_{k\to\infty} \mathbf{u}^k = \mathbf{u}^*$, taking limit on both sides of the above inequality yields $\lim_{k\to\infty} \mathbf{z}^k = \mathbf{z}^*$. Overall, we have verified $\lim_{k\to\infty}(\mathbf{u}^k, \mathbf{z}^k) = (\mathbf{u}^*, \mathbf{z}^*)$. Using Lemma 16 and Theorem 5, we can arrive at the desired conclusion. ■

Appendix G. Corollary from global convergence

Corollary 17 *Under the premise of Theorem 9, the following holds.*

(i) For k is sufficiently large, it holds

$$\|\mathbf{w}^*_{\overline{T}_{k+1}}\| = 0, \ \begin{cases} T_{k+1} \supseteq \mathcal{S}^*, \ \text{if } \|\mathbf{w}^*\|_0 < s, \\ T_{k+1} = \mathcal{S}^*, \ \text{if } \|\mathbf{w}^*\|_0 = s. \end{cases} \tag{69}$$

$$\|\boldsymbol{\xi}^*_{\overline{\Gamma}_{k+1}}\| = 0, \ \|\mathbf{z}^*_{\Gamma_{k+1}}\| = 0, \ J(\boldsymbol{\xi}^{k+1}) = J(\boldsymbol{\xi}^*) \tag{70}$$

$$\|\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1})\| \le c_5\|\mathbf{w}^{k+1} - \mathbf{w}^k\| + c_6\|\mathbf{w}^{k+1} - \mathbf{w}^*\|, \ \text{if } \|\mathbf{w}^*\|_0 < s \tag{71}$$

(ii) It holds

$$\lim_{k\to\infty} \mathcal{M}_k = \mathcal{M}_* := \mathcal{M}_{\rho,\eta}(\mathbf{u}^*, \mathbf{z}^*, \mathbf{u}^*) = \frac{1}{2}\|\mathbf{w}^*\|^2 + \lambda J(\boldsymbol{\xi}^*).$$

Proof. (i) Formulas (69) has been proved in Theorem 9. Moreover, $\|\boldsymbol{\xi}^*_{\overline{\Gamma}_{k+1}}\| = 0$ and $\|\mathbf{z}^*_{\Gamma_{k+1}}\| = 0$ can be derived from $\mathcal{R}_2(\mathbf{u}^{k+1}) \le c_2\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2$ by a similar procedure as that of $\|\mathbf{w}^*_{\overline{T}_{k+1}}\| = 0$. We will first prove $J(\boldsymbol{\xi}^{k+1}) = J(\boldsymbol{\xi}^*)$ when k is large enough. From the last line of (22) and the definition of Moreau envelop, we have

$$\begin{aligned}(\beta/2)\|\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^{k+1})\|^2 + \lambda J(\boldsymbol{\xi}^{k+1}) \le& \Phi^{\beta}_{\lambda J(\cdot)}(\boldsymbol{\xi}^{k+1} - \beta\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^{k+1})) + \vartheta_k \\ \le& \frac{1}{2\beta}\|\mathbf{w}^* - (\mathbf{w}^{k+1} - \beta\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^{k+1}))\|^2 + \lambda J(\boldsymbol{\xi}^*) + \vartheta_k. \end{aligned}$$

Taking the superior limits on both sides of the above inequality implies

$$\limsup_{k\to\infty} J(\boldsymbol{\xi}^{k+1}) \le J(\boldsymbol{\xi}^*).$$

Combining this with the lower semi-continuity of $J(\cdot)$ leads to $\lim_{k\to\infty} J(\boldsymbol{\xi}^{k+1}) = J(\boldsymbol{\xi}^*)$. Since the values of $J(\cdot)$ can only be taken from $[m]$, we can derive (70).

Now we will prove (71). If $\|\mathbf{w}^*\|_0 < s$, then from (67), $T_{k+1} \cap \overline{\mathcal{S}}^* \neq \varnothing$ holds. By the definition of T_{k+1}, we have the following chain of inequalities

$$\begin{aligned}\|[\mathbf{w}^{k+1} - \alpha\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}))]_{\overline{T}^{k+1}}\| \leq & |\overline{T}_{k+1}||[\mathbf{w}^{k+1} - \alpha\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}))]_i| \text{ for any } i \in T_{k+1} \cap \overline{\mathcal{S}}^* \\ \leq & (n-s)\|[\mathbf{w}^{k+1} - \alpha\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}))]_{T_{k+1}\cap\overline{\mathcal{S}}^*}\| \\ \leq & (n-s)(\|[\mathbf{w}^{k+1} - \mathbf{w}^*]_{T_{k+1}\cap\overline{\mathcal{S}}^*}\| + \alpha\|[\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1})]_{T_{k+1}\cap\overline{\mathcal{S}}^*}\|) \\ \overset{(22)}{\leq} & (n-s)(\|\mathbf{w}^{k+1} - \mathbf{w}^*\| + \alpha c_1\|\mathbf{w}^{k+1} - \mathbf{w}^k\|), \end{aligned} \tag{72}$$

where the first inequality follows from the fact that T_{k+1} contains the best s largest elements of $\mathbf{w}^{k+1} - \alpha\nabla_{\mathbf{w}} g(\mathbf{u}^{k+1})$ in absolute value. Then we can estimate

$$\begin{aligned}\|\nabla_{\overline{T}_{k+1}} g_k(\mathbf{u}^{k+1})\| = & \|[\mathbf{w}^{k+1} - (\mathbf{w}^{k+1} - \alpha\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}))]_{\overline{T}^{k+1}}\|/\alpha \\ \leq & (\|\mathbf{w}^{k+1}_{\overline{T}^{k+1}}\| + \|[\mathbf{w}^{k+1} - \alpha\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}))]_{\overline{T}^{k+1}}\|)/\alpha \\ \overset{(72)}{\leq} & (c_1\|\mathbf{w}^{k+1} - \mathbf{w}^k\| + (n-s)\|\mathbf{w}^{k+1} - \mathbf{w}^*\| + \alpha c_1(n-s)\|\mathbf{w}^{k+1} - \mathbf{w}^k\|)/\alpha. \end{aligned}$$

This result further leads to

$$\|\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1})\| \leq \|\nabla_{T_{k+1}} g_k(\mathbf{u}^{k+1})\| + \|\nabla_{\overline{T}_{k+1}} g_k(\mathbf{u}^{k+1})\| \overset{(22)}{\leq} c_5\|\mathbf{w}^{k+1} - \mathbf{w}^k\| + c_6\|\mathbf{w}^{k+1} - \mathbf{w}^*\|,$$

where $c_5 := c_1/\alpha + c_1(n+1-s)$ and $c_6 := (n-s)/\alpha$.

(ii) Applying the fact $\lim_{k\to\infty}(\mathbf{u}^k, \mathbf{z}^k) = (\mathbf{u}^*, \mathbf{z}^*)$, $\delta_{\mathbb{S}}(\mathbf{w}^{k+1}) = \delta_{\mathbb{S}}(\mathbf{w}^*) = 0$ and (70), we can derive $\lim_{k\to\infty} \mathcal{M}_k = \mathcal{M}_*$. ■

Appendix H. Proof of Theorem 10 on Convergence Rate of iPAL

The main steps for convergence rate analysis is as follows.

- To prove Theorem 10 (i), we will first estimate an upper bound of $\mathcal{M}_{k+1} - \mathcal{M}_*$ (see (76)). This, together with the sufficient descent property (25), leads to a recursion formula (77). This will give rise to the linear convergence rate of the Lyapunov function value sequence, see (27).

- For the linear convergence rate of iterate sequence, we will first investigate the relationship between $\|\mathbf{w}^{k+1} - \mathbf{w}^*\|$ and $\mathcal{M}_{k+1} - \mathcal{M}_*$ (see (78)). Then we will use (27) to derive linear convergence rate of $\|\mathbf{w}^{k+1} - \mathbf{w}^*\|$. Following a similar procedure, we can prove linear convergence rate of $\|\boldsymbol{\xi}^{k+1} - \boldsymbol{\xi}^*\|$ and $\|\mathbf{z}^{k+1} - \mathbf{z}^*\|$.

Proof of Theorem 10. (i) We start with several inequalities. The first one is a direct computation

$$\frac{1}{2}\|\mathbf{w}^{k+1}\|^2 - \frac{1}{2}\|\mathbf{w}^*\|^2 + \langle \mathbf{w}^{k+1}, \mathbf{w}^* - \mathbf{w}^{k+1}\rangle = -\frac{1}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^*\|^2. \tag{73}$$

We now estimate an upper bound for $|\langle\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^{k+1}), \boldsymbol{\xi}^{k+1}-\boldsymbol{\xi}^k\rangle|$. Since the sequence boundedness has been proved in Theorem 8 (ii), we can assume $\|(\mathbf{u}^{k+1};\mathbf{z}^{k+1})\| \le \tau$ for $\tau > 0$. Using the (57) and $\boldsymbol{\xi}^*_{\overline{\Gamma}_{k+1}} = 0$ for sufficiently large k, we obtain

$$\begin{aligned}
|\langle\nabla_{\overline{\Gamma}_{k+1}} g_k(\mathbf{u}^{k+1}), [\boldsymbol{\xi}^{k+1}-\boldsymbol{\xi}^*]_{\overline{\Gamma}_{k+1}}\rangle| =& |\langle\nabla_{\overline{\Gamma}_{k+1}} g_k(\mathbf{u}^{k+1}), \boldsymbol{\xi}^{k+1}_{\overline{\Gamma}_{k+1}}\rangle| \le \|\nabla_{\overline{\Gamma}_{k+1}} g_k(\mathbf{u}^{k+1})\|\|\boldsymbol{\xi}^{k+1}_{\overline{\Gamma}_{k+1}}\| \\
\le& \tau c_2\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2 \\
|\langle\nabla_{\Gamma_{k+1}} g_k(\mathbf{u}^{k+1}), [\boldsymbol{\xi}^{k+1}-\boldsymbol{\xi}^*]_{\Gamma_{k+1}}\rangle| \le& \|\nabla_{\Gamma_{k+1}} g_k(\mathbf{u}^{k+1})\|(\|\boldsymbol{\xi}^{k+1}_{\Gamma_{k+1}}\| + \|\boldsymbol{\xi}^*_{\Gamma_{k+1}}\|) \\
\le& 2\tau c_2\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2.
\end{aligned}$$

Adding the above inequalities implies

$$|\langle\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^{k+1}), \boldsymbol{\xi}^{k+1}-\boldsymbol{\xi}^k\rangle| \le 3\tau c_2\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2 \tag{74}$$

We shall also derive an upper bound for $|\langle\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}), \mathbf{w}^{k+1}-\mathbf{w}^*\rangle|$. If $\|\mathbf{w}^*\|_0 = s$, using $T_{k+1} = \mathcal{S}^*$ and (22), we can derive

$$\begin{aligned}
&|\langle\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}), \mathbf{w}^{k+1}-\mathbf{w}^*\rangle| \\
=&|\langle\nabla_{T_{k+1}} g_k(\mathbf{u}^{k+1}), [\mathbf{w}^{k+1}-\mathbf{w}^*]_{T_{k+1}}\rangle| \\
\le&\|\nabla_{T_{k+1}} g_k(\mathbf{u}^{k+1})\|\|[\mathbf{w}^{k+1}-\mathbf{w}^*]_{T_{k+1}}\| \le c_1\|\mathbf{w}^{k+1}-\mathbf{w}^k\|\|\mathbf{w}^{k+1}-\mathbf{w}^*\|.
\end{aligned}$$

If $\|\mathbf{w}^*\|_0 < s$, using $T_{k+1} \supseteq \mathcal{S}^*$, we have

$$\begin{aligned}
&\langle\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}), \mathbf{w}^{k+1}-\mathbf{w}^*\rangle \\
=&\langle\nabla_{T_{k+1}} g_k(\mathbf{u}^{k+1}), [\mathbf{w}^{k+1}-\mathbf{w}^*]_{T_{k+1}}\rangle + \langle\nabla_{\overline{T}_{k+1}} g_k(\mathbf{u}^{k+1}), [\mathbf{w}^{k+1}-\mathbf{w}^*]_{\overline{T}_{k+1}}\rangle \\
=&\langle\nabla_{T_{k+1}} g_k(\mathbf{u}^{k+1}), [\mathbf{w}^{k+1}-\mathbf{w}^*]_{T_{k+1}}\rangle + \langle\nabla_{\overline{T}_{k+1}} g_k(\mathbf{u}^{k+1}), \mathbf{w}^{k+1}_{\overline{T}_{k+1}}\rangle.
\end{aligned}$$

Then from (71) and (22), the following chain of inequalities holds

$$\begin{aligned}
&|\langle\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}), \mathbf{w}^{k+1}-\mathbf{w}^*\rangle| \\
\le&\|\nabla_{T_{k+1}} g_k(\mathbf{u}^{k+1})\|\|[\mathbf{w}^{k+1}-\mathbf{w}^*]_{T_{k+1}}\| + \|\nabla_{\overline{T}_{k+1}} g_k(\mathbf{u}^{k+1})\|\|\mathbf{w}^{k+1}_{\overline{T}_{k+1}}\| \\
\le& c_1\|\mathbf{w}^{k+1}-\mathbf{w}^k\|\|\mathbf{w}^{k+1}-\mathbf{w}^*\| + c_1(c_5\|\mathbf{w}^{k+1}-\mathbf{w}^k\| + c_6\|\mathbf{w}^{k+1}-\mathbf{w}^*\|)\|\mathbf{w}^{k+1}-\mathbf{w}^k\| \\
\le& c_1c_5\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2 + (c_1c_6+c_1)\|\mathbf{w}^{k+1}-\mathbf{w}^k\|\|\mathbf{w}^{k+1}-\mathbf{w}^*\|.
\end{aligned}$$

These two cases lead to

$$\begin{aligned}
|\langle\nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}), \mathbf{w}^{k+1}-\mathbf{w}^*\rangle| \le& c_1c_5\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2 \\
&+ (c_1c_6+c_1)\|\mathbf{w}^{k+1}-\mathbf{w}^k\|\|\mathbf{w}^{k+1}-\mathbf{w}^*\|
\end{aligned} \tag{75}$$

Now let us consider $\mathcal{M}_{k+1} - \mathcal{M}_*$. For sufficiently large k, using definition of Lyapunov function, $J(\boldsymbol{\xi}^{k+1}) = J(\boldsymbol{\xi}^*)$ and $A\mathbf{w}^* + \mathbf{1} - \boldsymbol{\xi}^* = 0$, we have

$$\begin{aligned}
&\mathcal{M}_{k+1} - \mathcal{M}_* \\
=&\frac{1}{2}\|\mathbf{w}^{k+1}\|^2 - \frac{1}{2}\|\mathbf{w}^*\|^2 + \langle \mathbf{z}^{k+1}, A\mathbf{w}^{k+1} + \mathbf{1} - \boldsymbol{\xi}^{k+1}\rangle + \frac{1}{2\rho}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 + \frac{\eta}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 \\
=&\frac{1}{2}\|\mathbf{w}^{k+1}\|^2 - \frac{1}{2}\|\mathbf{w}^*\|^2 + \langle \mathbf{z}^{k+1}, A\mathbf{w}^{k+1} + \mathbf{1} - \boldsymbol{\xi}^{k+1} - (A\mathbf{w}^* + \mathbf{1} - \boldsymbol{\xi}^*)\rangle \\
&+ \frac{1}{2\rho}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 + \frac{\eta}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 \\
=&\frac{1}{2}\|\mathbf{w}^{k+1}\|^2 - \frac{1}{2}\|\mathbf{w}^*\|^2 + \langle A^\top \mathbf{z}^{k+1}, \mathbf{w}^{k+1} - \mathbf{w}^*\rangle - \langle \mathbf{z}^{k+1}, \boldsymbol{\xi}^{k+1} - \boldsymbol{\xi}^*\rangle \\
&+ \frac{1}{2\rho}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2 + \frac{\eta}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2.
\end{aligned}$$

Applying (56) and (57), we can further derive

$$\begin{aligned}
&\mathcal{M}_{k+1} - \mathcal{M}_* \\
&= \frac{1}{2}\|\mathbf{w}^{k+1}\|^2 - \frac{1}{2}\|\mathbf{w}^*\|^2 - \langle \mathbf{w}^{k+1}, \mathbf{w}^{k+1} - \mathbf{w}^*\rangle + \langle \nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^{k+1}), \boldsymbol{\xi}^{k+1} - \boldsymbol{\xi}^*\rangle + \frac{\eta}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 \\
&+ \langle \nabla_{\mathbf{w}} g_k(\mathbf{u}^{k+1}), \mathbf{w}^{k+1} - \mathbf{w}^*\rangle - \mu\langle \mathbf{w}^{k+1} - \mathbf{w}^k, \mathbf{w}^{k+1} - \mathbf{w}^*\rangle + \frac{1}{2\rho}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|^2
\end{aligned}$$

Then we can use the previous inequalities (61), (73), (74), (75) as well as the fact $-\mu\langle \mathbf{w}^{k+1} - \mathbf{w}^k, \mathbf{w}^{k+1} - \mathbf{w}^*\rangle \le \mu\|\mathbf{w}^{k+1} - \mathbf{w}^k\|\|\mathbf{w}^{k+1} - \mathbf{w}^*\|$ to obtain

$$\begin{aligned}
&\mathcal{M}_{k+1} - \mathcal{M}_* \\
\le& -\frac{1}{2}\|\mathbf{w}^{k+1} - \mathbf{w}^*\|^2 + \underbrace{(c_1c_6 + c_1 + \mu)}_{:=c_7}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|\|\mathbf{w}^{k+1} - \mathbf{w}^*\| \\
&+ \underbrace{(3\tau c_2 + c_1c_5 + \frac{\eta}{2} + \frac{c_3^2}{\rho})}_{:=c_8}\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 + \frac{c_4^2}{\rho}\|\mathbf{w}^k - \mathbf{w}^{k-1}\|^2 \\
=& -\frac{1}{2}(\|\mathbf{w}^{k+1} - \mathbf{w}^*\| - c_7\|\mathbf{w}^{k+1} - \mathbf{w}^k\|)^2 + (c_8 + \frac{c_7^2}{2})\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 + \frac{c_4^2}{\rho}\|\mathbf{w}^k - \mathbf{w}^{k-1}\|^2 \\
\le& \tau_1(\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 + \|\mathbf{w}^k - \mathbf{w}^{k-1}\|^2), \qquad (76)
\end{aligned}$$

where $\tau_1 := c_8 + c_7^2/2$. Now taking the descent property (25) into account, we can estimate

$$\begin{aligned}
(\mathcal{M}_{k-1} - \mathcal{M}_*) - (\mathcal{M}_{k+1} - \mathcal{M}_*) =& \mathcal{M}_{k-1} - \mathcal{M}_k + \mathcal{M}_k - \mathcal{M}_{k+1} \\
\ge& \frac{\mu}{4}(\|\mathbf{w}^{k+1} - \mathbf{w}^k\|^2 + \|\mathbf{w}^k - \mathbf{w}^{k-1}\|^2).
\end{aligned}$$

Combining this with (76) leads to

$$\mathcal{M}_{k+1} - \mathcal{M}_* \le \frac{1}{1 + \mu/(4\tau_1)}(\mathcal{M}_{k-1} - \mathcal{M}_*). \qquad (77)$$

This means that there exists a sufficiently large k^* such that (27) holds for constants

$$q := \sqrt{\frac{1}{1+\mu/(4\tau_1)}} \quad \text{and} \quad c_m := (1/q)^{k^*+1}(\mathcal{M}_0 - \mathcal{M}_*).$$

(ii) Suppose that index k is sufficiently large. It follows from (76) that

$$\begin{aligned}\sqrt{\mathcal{M}_{k+1}-\mathcal{M}_*} \leq& \sqrt{\tau_1(\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2 + \|\mathbf{w}^k-\mathbf{w}^{k-1}\|^2)}\\ \leq& \sqrt{\tau_1}(\|\mathbf{w}^{k+1}-\mathbf{w}^k\| + \|\mathbf{w}^k-\mathbf{w}^{k-1}\|).\end{aligned}$$

Using this relationship and the concavity of $\sqrt{(\cdot)}$, we can obtain

$$\begin{aligned}\varepsilon_k :=& \sqrt{\mathcal{M}_k-\mathcal{M}_*} - \sqrt{\mathcal{M}_{k+1}-\mathcal{M}_*}\\ \geq& \frac{\mathcal{M}_k-\mathcal{M}_{k+1}}{2\sqrt{\mathcal{M}_k-\mathcal{M}_*}} \geq \frac{\mu\|\mathbf{w}^{k+1}-\mathbf{w}^k\|^2}{8\sqrt{\tau_1}(\|\mathbf{w}^k-\mathbf{w}^{k-1}\| + \|\mathbf{w}^{k-1}-\mathbf{w}^k\|)},\end{aligned}$$

This further leads to

$$\begin{aligned}\|\mathbf{w}^{k+1}-\mathbf{w}^k\| \leq& \left(\frac{8\sqrt{\tau_1}\varepsilon_k}{\mu}(\|\mathbf{w}^k-\mathbf{w}^{k-1}\| + \|\mathbf{w}^{k-1}-\mathbf{w}^{k-2}\|)\right)^{\frac{1}{2}}\\ \leq& \frac{1}{4}(\|\mathbf{w}^k-\mathbf{w}^{k-1}\| + \|\mathbf{w}^{k-1}-\mathbf{w}^{k-2}\|) + \frac{8\sqrt{\tau_1}}{\mu}\varepsilon_k.\end{aligned}$$

Let us consider the sum of the above terms from $\ell = k+2$ to $\ell = \widetilde{k}$.

$$\begin{aligned}\sum_{\ell=k+2}^{\widetilde{k}}\|\mathbf{w}^{\ell+1}-\mathbf{w}^\ell\| \leq& \frac{1}{4}\sum_{\ell=k+2}^{\widetilde{k}}\|\mathbf{w}^{\ell}-\mathbf{w}^{\ell-1}\| + \frac{1}{4}\sum_{\ell=k+2}^{\widetilde{k}}\|\mathbf{w}^{\ell-1}-\mathbf{w}^{\ell-2}\| + \frac{8\sqrt{\tau_1}}{\mu}\sum_{\ell=k+2}^{\widetilde{k}}\varepsilon_\ell\\ \leq& \frac{1}{4}\sum_{\ell=k+2}^{\widetilde{k}}\|\mathbf{w}^{\ell+1}-\mathbf{w}^\ell\| + \frac{1}{4}\sum_{\ell=k+2}^{\widetilde{k}}\|\mathbf{w}^{\ell+1}-\mathbf{w}^\ell\| + \frac{8\sqrt{\tau_1}}{\mu}\sum_{\ell=k+2}^{\widetilde{k}}\varepsilon_\ell\\ &+ \frac{1}{2}\|\mathbf{w}^{k+2}-\mathbf{w}^{k+1}\| + \frac{1}{4}\|\mathbf{w}^{k+1}-\mathbf{w}^k\|\end{aligned}$$

After some algebraic manipulation, we have

$$\begin{aligned}\sum_{\ell=k}^{\widetilde{k}}\|\mathbf{w}^{\ell+1}-\mathbf{w}^\ell\| \leq& \frac{3}{2}\|\mathbf{w}^{k+1}-\mathbf{w}^k\| + 2\|\mathbf{w}^{k+2}-\mathbf{w}^{k+1}\| + \frac{16\sqrt{\tau_1}}{\mu}\sum_{\ell=k+2}^{\widetilde{k}}\varepsilon_\ell\\ \leq& \frac{3}{\sqrt{\mu}}\sqrt{\mathcal{M}_k-\mathcal{M}_{k+1}} + \frac{4}{\sqrt{\mu}}\sqrt{\mathcal{M}_{k+1}-\mathcal{M}_{k+2}} + \frac{16\sqrt{\tau_1}}{\mu}\sqrt{\mathcal{M}_{k+2}-\mathcal{M}_*}\\ \leq& (\frac{7}{\sqrt{\mu}} + \frac{16\sqrt{\tau_1}}{\mu})\sqrt{\mathcal{M}_k-\mathcal{M}_*}.\end{aligned}$$

The taking $\widetilde{k} \to \infty$ for above inequality yields

$$\|\mathbf{w}^k-\mathbf{w}^*\| \leq \sum_{\ell=k}^{\infty}\|\mathbf{w}^{\ell+1}-\mathbf{w}^\ell\| \leq \underbrace{(7/\sqrt{\mu} + 16\sqrt{\tau_1}/\mu)}_{:=\tau_2}\sqrt{\mathcal{M}_k-\mathcal{M}_*}. \tag{78}$$

Since (27) holds, we can derive the R-linear convergence rate for $\{\mathbf{w}^k\}_{k\in\mathbb{N}}$ in (28) with constant $c_w := \tau_2\sqrt{c_m}$.

Next, we will prove the R-liner convergence rate of $\{\mathbf{z}^k\}_{k\in\mathbb{N}}$. From (68), we can estimate

$$\begin{aligned}\gamma\|\mathbf{z}^k-\mathbf{z}^*\| \leq & (c_1+\mu)\|\mathbf{w}^k-\mathbf{w}^{k-1}\|+\|\mathbf{w}^k-\mathbf{w}^*\| \\ \overset{(25),(78)}{\leq} & (2(c_1+\mu)/\sqrt{\mu})\sqrt{\mathcal{M}_{k-1}-\mathcal{M}_k}+\tau_2\sqrt{\mathcal{M}_k-\mathcal{M}_*} \\ \leq & \underbrace{(2(c_1+\mu)/\sqrt{\mu}+\tau_2)}_{:=\tau_3}\sqrt{\mathcal{M}_{k-1}-\mathcal{M}_*}.\end{aligned}$$

By using (27), we can arrive at $\|\mathbf{z}^k-\mathbf{z}^*\| \leq c_z\sqrt{q}^k$ with constant $c_z := (\tau_3/\gamma)\sqrt{c_m/q}$. Finally, we prove the linear convergence rate of $\{\boldsymbol{\xi}^k\}_{k\in\mathbb{N}}$. Using (23) and $A\mathbf{w}^*+\mathbf{1}-\boldsymbol{\xi}^*=0$, we can estimate

$$\begin{aligned}\|\boldsymbol{\xi}^k-\boldsymbol{\xi}^*\| \leq & \|A\|\|\mathbf{w}^k-\mathbf{w}^*\|+\|\mathbf{z}^k-\mathbf{z}^{k-1}\|/\rho \overset{(60)}{\leq} \|A\|\|\mathbf{w}^k-\mathbf{w}^*\| \\ & +(c_3/\rho)\|\mathbf{w}^k-\mathbf{w}^{k-1}\|+(c_4/\rho)\|\mathbf{w}^{k-1}-\mathbf{w}^{k-2}\| \\ \overset{(78,25)}{\leq} & \|A\|\tau_2\sqrt{\mathcal{M}_k-\mathcal{M}_*}+\frac{2c_3}{\rho\sqrt{\mu}}\sqrt{\mathcal{M}_{k-1}-\mathcal{M}_k}+\frac{2c_4}{\rho\sqrt{\mu}}\sqrt{\mathcal{M}_{k-2}-\mathcal{M}_{k-1}} \\ \leq & \underbrace{\left(\|A\|\tau_2+\frac{2(c_3+c_4)}{\rho\sqrt{\mu}}\right)}_{:=\tau_4}\sqrt{\mathcal{M}_{k-2}-\mathcal{M}_*}.\end{aligned}$$

This means $\|\mathbf{z}^k-\mathbf{z}^*\| \leq c_z\sqrt{q}^k$ can be verified with $c_z = \tau_4\sqrt{c_m/q^2}$. ■

Appendix I. Proofs on Convergence Properties of PGN

First we explain the general ideas for the proof of Theorem 11.

- To prove Theorem 11 (i), we first show the objective function G enjoys sufficient descent (81) on the proximal gradient iterate $\mathbf{u}^{j+1/2}$. Then if Newton step is accepted, G also enjoys the sufficient descent (34). These results lead to (37). We then show the convergence of $\{G(\mathbf{u}^j)\}_{j\in\mathbb{N}}$, which further implies (38).
- The procedure to prove (ii) is similar to the global convergence of iPAL. First, we show that the sequence $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$ is bounded. Second, the boundedness ensures the existence of accumulated points and we will prove each of them is a P-stationary point of subproblem (29). Finally, we will utilize (Kanzow and Qi, 1999, Proposition 7) to show the whole sequence is convergent. Again, this proposition requires (38) and isolatedness of the P-stationary points, which we will show in the following proof.

Proof of Theorem 11 (i) Let us first prove the descent property of G. From (31), and the definition of projection and proximal operator, we have

$$\begin{cases}\frac{1}{2\alpha}\|\mathbf{w}^{j+1/2}-(\mathbf{w}^j-\alpha\nabla_{\mathbf{w}}g(\mathbf{u}^j))\|^2 \leq \frac{\alpha}{2}\|\nabla_{\mathbf{w}}g(\mathbf{u}^j)\|^2 \\ \frac{1}{2\beta}\|\boldsymbol{\xi}^{j+1/2}-(\boldsymbol{\xi}^j-\beta\nabla_{\boldsymbol{\xi}}g(\mathbf{u}^j))\|^2+\lambda J(\boldsymbol{\xi}^{j+1/2}) \leq \frac{\beta}{2}\|\nabla_{\boldsymbol{\xi}}g(\mathbf{u}^j)\|^2+\lambda J(\boldsymbol{\xi}^j).\end{cases}$$

By some simple algebraic manipulation, the following inequalities can be deduced

$$\begin{cases} \langle \nabla_{\mathbf{w}} g(\mathbf{u}^j), \mathbf{w}^{j+1/2} - \mathbf{w}^j \rangle \le -\dfrac{1}{2\alpha} \|\mathbf{w}^{j+1/2} - \mathbf{w}^j\|^2 \\ \langle \nabla_{\boldsymbol{\xi}} g(\mathbf{u}^j), \boldsymbol{\xi}^{j+1/2} - \boldsymbol{\xi}^j \rangle + \lambda J(\boldsymbol{\xi}^{j+1/2}) - \lambda J(\boldsymbol{\xi}^j) \le -\dfrac{1}{2\beta} \|\boldsymbol{\xi}^{j+1/2} - \boldsymbol{\xi}^j\|^2. \end{cases} \tag{79}$$

Using the descent lemma (Beck, 2017, Lemma 5.7) on function G yields

$$g(\mathbf{u}^{j+1/2}) \le g(\mathbf{u}^j) + \langle \nabla g(\mathbf{u}^j), \mathbf{u}^{j+1/2} - \mathbf{u}^j \rangle + \frac{\ell_g}{2} \|\mathbf{u}^{j+1/2} - \mathbf{u}^j\|^2. \tag{80}$$

Taking $\delta_{\mathbb{S}}(\mathbf{w}^{j+1/2}) = \delta_{\mathbb{S}}(\mathbf{w}^j)$ into account and adding (79) and (80), we obtain

$$G(\mathbf{u}^j) - G(\mathbf{u}^{j+1/2}) \ge \zeta \|\mathbf{u}^{j+1/2} - \mathbf{u}^j\|^2. \tag{81}$$

Then in each case of the update step (34), we have

$$G(\mathbf{u}^{j+1/2}) - G(\mathbf{u}^{j+1}) \ge (\sigma_g/4) \|\mathbf{u}^{j+1/2} - \mathbf{u}^{j+1}\|^2 \tag{82}$$

Adding the above two inequalities directly leads to (37).

Since g is strongly convex, and $\delta_{\mathbb{S}}(\cdot)$ and $J(\cdot)$ are lower bounded, we can conclude G is also bounded below. Then $\{G(\mathbf{u}^j)\}_{j\in\mathbb{N}}$ is a nonincreasing and bounded sequence, which implies the sequence is convergent. This result together with (37) yields (38).

(ii) We will prove the global convergence of $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$ according to the three steps mentioned at the beginning of this section.

Step 1. Since g is strongly convex, it is also coercive, i.e. $\lim_{\|\mathbf{u}\|\to\infty} g(\mathbf{u}) = \infty$. Combining this with the lower boundedness and lower semi-continuity of $\delta_{\mathbb{S}}(\cdot)$ and $\lambda J(\cdot)$ implies that G is lower semi-continuous and coercive. It follows from (Mordukhovich and Nam, 2013, Theorem 4.10) that $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$ is bounded.

Step 2. Suppose that $\widehat{\mathbf{u}}$ is an accumulation point of $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$. Then there exists a subsequence $\{\mathbf{u}^j\}_{j\in\mathcal{J}}$ converging to $\widehat{\mathbf{u}}$. It follows from the continuous differentiability of g that

$$\lim_{j\in\mathcal{J}} \mathbf{w}^j - \alpha \nabla_{\mathbf{w}} g_k(\mathbf{u}^j) = \widehat{\mathbf{w}} - \alpha \nabla_{\mathbf{w}} g_k(\widehat{\mathbf{u}}) \text{ and } \lim_{j\in\mathcal{J}} \boldsymbol{\xi}^j - \beta \nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^j) = \widehat{\boldsymbol{\xi}} - \alpha \nabla_{\mathbf{w}} g_k(\widehat{\mathbf{u}}).$$

Since $\|\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}\| \le \|\mathbf{u}^{j+1/2} - \mathbf{u}^j\| + \|\mathbf{u}^j - \widehat{\mathbf{u}}\|$, by using $\lim_{j\to\infty, j\in\mathcal{J}} \mathbf{u}^j = \widehat{\mathbf{u}}$ and (38), we have

$$\lim_{j\to\infty, j\in\mathcal{J}} \mathbf{u}^{j+1/2} = \widehat{\mathbf{u}}$$

Finally, it follows from (Rockafellar, 1976, Theorem 1.25) that $\widehat{\mathbf{u}}$ must satisfy system (39).

Step 3. Let $\widehat{\mathbf{u}}$ be an accumulation point of $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$. We define $\widehat{\mathcal{I}}_- := \{i \in [m] : \widehat{\boldsymbol{\xi}}_i \le 0\}$ and select $\widehat{T} \in \mathbb{T} := \{T : T \supseteq \widehat{\mathcal{S}},\ |T| = s\}$. We consider the following convex programming.

$$\min_{\mathbf{u}:=(\mathbf{w},\boldsymbol{\xi})} g(\mathbf{u}) \quad s.t. \quad \boldsymbol{\xi}_{\widehat{\mathcal{I}}_-} \le 0, \quad \mathbf{w}_i = 0,\ i \notin \widehat{T}. \tag{83}$$

Since the objective function g is strongly convex and the constraints are linear, if a point $\mathbf{u}$ satisfies the following KKT system, then it must be the unique global minimizer of the above convex programming

$$\begin{cases} [\nabla_{\mathbf{w}} g(\mathbf{u})]_{\widehat{T}} = 0, \quad \mathbf{w}_i = 0,\ i \notin \widehat{T} \\ \boldsymbol{\xi}_{\widehat{\mathcal{I}}_-} \le 0, \quad -[\nabla_{\boldsymbol{\xi}} g(\mathbf{u})]_{\widehat{\mathcal{I}}_-} \ge 0, \quad \langle \boldsymbol{\xi}_{\widehat{\mathcal{I}}_-}, [\nabla_{\boldsymbol{\xi}} g(\mathbf{u})]_{\widehat{\mathcal{I}}_-} \rangle = 0 \\ \quad [\nabla_{\boldsymbol{\xi}} g(\mathbf{u})]_i = 0,\ i \notin \widehat{\mathcal{I}}_- \end{cases}$$

Considering that the accumulation point $\widehat{\mathbf{u}}$ satisfies (39), Lemmas 1 and 2 imply that $\widehat{\mathbf{u}}$ must satisfy the above KKT system, and thus it is the unique global minimizer of (83). Since the numbers of the choices of $\widehat{\mathcal{I}}_-$ and $\widehat{T}$ are finite, the number of accumulation for $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$ is also finite. Therefore, each accumulation point must be isolated. Finally, taking (38) into account, it follows from (Kanzow and Qi, 1999, Proposition 7) that the whole sequence $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$ must converge to $\widehat{\mathbf{u}}$. ■

Since we have proved that the sequence $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$ converges to a P-stationary point of (29) and the inexact criteria is just an approximation of the P-stationary system. Then the iterate can satisfy the inexact criteria after finite steps. This is what we will prove in Corollary 12.

Proof of Corollary 12. By Theorem 11 (ii), $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$ must converge to a P-stationary point $\widehat{\mathbf{u}}$ of (29). Let us first prove

$$\lim_{j\to\infty} J(\boldsymbol{\xi}^{j+1/2}) = \lim_{j\to\infty} J(\boldsymbol{\xi}^{j+1}) = J(\widehat{\boldsymbol{\xi}}). \tag{84}$$

From (31) and the definition of proximal operator, we have

$$\frac{1}{2\beta}\|\boldsymbol{\xi}^{j+1/2} - \boldsymbol{\xi}^j + \beta\nabla_{\boldsymbol{\xi}} g(\mathbf{u}^j)\|^2 + \lambda J(\boldsymbol{\xi}^{j+1/2}) \le \frac{1}{2\beta}\|\widehat{\boldsymbol{\xi}} - \boldsymbol{\xi}^j + \beta\nabla_{\boldsymbol{\xi}} g(\mathbf{u}^j)\|^2 + \lambda J(\widehat{\boldsymbol{\xi}}).$$

Taking the superior limits on both sides of above inequality implies $\limsup_{j\to\infty} J(\boldsymbol{\xi}^{j+1/2}) \le J(\widehat{\boldsymbol{\xi}})$. Combining this with the lower semi-continuity of $J(\cdot)$ leads to $\lim_{j\to\infty} J(\boldsymbol{\xi}^{j+1/2}) = J(\widehat{\boldsymbol{\xi}})$.

From (82) and the fact $\delta_{\mathbb{S}}(\mathbf{w}^{j+1}) = \delta_{\mathbb{S}}(\mathbf{w}^{j+1/2}) = 0$, we have

$$g(\mathbf{u}^{j+1}) + \lambda J(\boldsymbol{\xi}^{j+1}) + (\sigma_g/4)\|\mathbf{u}^{j+1/2} - \mathbf{u}^{j+1}\| \le g(\mathbf{u}^{j+1/2}) + \lambda J(\boldsymbol{\xi}^{j+1/2}).$$

Taking the superior limits on both sides of the above inequality, we have

$$\limsup_{j\to\infty} J(\boldsymbol{\xi}^{j+1}) \le \limsup_{j\to\infty} J(\boldsymbol{\xi}^{j+1/2}) = J(\widehat{\boldsymbol{\xi}}).$$

This together with lower semi-continuity of $J(\cdot)$ leads to $\lim_{j\to\infty} J(\boldsymbol{\xi}^{j+1}) = J(\widehat{\boldsymbol{\xi}})$.

We now show $\lim_{j\to\infty} \mathcal{R}_i(\mathbf{u}^j) = 0$ for $i = 1, 2, 3$. The first line of (22) directly follows from (37) and (31). Furthermore, we have

$$\begin{aligned} \mathcal{R}_1(\mathbf{u}^j) &= \|[\nabla_{T_j} g_k(\mathbf{u}^j); \mathbf{w}^j_{\overline{T}_j}]\| \le \max\{1/\alpha, 1\}\|\mathbf{w}^{j+1/2} - \mathbf{w}^j\| \\ \mathcal{R}_2(\mathbf{u}^j) &= \|[\nabla_{\Gamma_j} g_k(\mathbf{u}^j); \boldsymbol{\xi}^j_{\overline{\Gamma}_j}]\| \le \max\{1/\beta, 1\}\|\boldsymbol{\xi}^{j+1/2} - \boldsymbol{\xi}^j\|, \end{aligned} \tag{85}$$

where T_j and Γ_j are corresponding index sets for the j-th identification step. Then we derive $\lim_{j\to\infty}\mathcal{R}_1(\mathbf{u}^j) = \lim_{j\to\infty}\mathcal{R}_2(\mathbf{u}^j) = 0$ by using (38). Applying the definition of Moreau envelop and (31) yields

$$\begin{aligned}
&\lim_{j\to\infty}\mathcal{R}_3(\mathbf{u}^j) = \lim_{j\to\infty}\frac{\beta}{2}\|\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^j)\|^2 + \lambda J(\boldsymbol{\xi}^j) - \Phi^{\beta}_{\lambda J(\cdot)}(\boldsymbol{\xi}^j - \beta\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^j))\\
&= \lim_{j\to\infty}\frac{\beta}{2}\|\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^j)\|^2 + \lambda J(\boldsymbol{\xi}^j) - \frac{1}{2\beta}\|\boldsymbol{\xi}^{j+1/2} - \boldsymbol{\xi}^j + \beta\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^j)\|^2 - \lambda J(\boldsymbol{\xi}^{j+1/2})\\
&= \lim_{j\to\infty} -\frac{1}{2\beta}\|\boldsymbol{\xi}^{j+1/2} - \boldsymbol{\xi}^j\|^2 - \langle\nabla_{\boldsymbol{\xi}} g_k(\mathbf{u}^j), \boldsymbol{\xi}^{j+1/2} - \boldsymbol{\xi}^j\rangle + \lambda J(\boldsymbol{\xi}^j) - \lambda J(\boldsymbol{\xi}^{j+1/2}) \overset{(84)}{=} 0. \quad (86)
\end{aligned}$$

Meanwhile, we can derive $\lim_{j\to\infty}\|\mathbf{w}^{k,j} - \mathbf{w}^k\| = \|\widehat{\mathbf{w}} - \mathbf{w}^k\| = \|\widehat{\mathbf{w}} - \mathbf{w}^{k,0}\| \neq 0$. Combing this with (85) and (86), we arrive at the desired conclusion. ■

Next we will prove the local quadratic convergence rate of PGN. The main ideas for this proof are presented as follows.

- The changeable index sets T_j and Γ_j in (33) brings difficulties for convergence rate analysis. We will show that they exactly contain nonzero elements of the solution (active sets) after finite iterate (see Lemma 18 below).
- Newton step plays a crucial role to ensure quadratic convergence rate. We will prove the update condition (34) always holds after finite iterations. Thus Newton step is accepted (see the first part of Theorem 13).
- Once the above two points prove to be true, the Newton iteration will be always performed on a fixed subspace. Then considering the strong convexity of g, the local quadratic convergence of PGN just follows from classical theory. It is also noteworthy that there is a gradient step before Newton step and both of them are performed on the same subspace

  $$\mathbf{u}^j \longrightarrow \mathbf{u}^{j+1/2} \text{ (gradient step)} \longrightarrow \mathbf{u}^{j+1} \text{ (Newton step)}.$$

  We shall show the gradient iteration will not influence the quadratic convergence rate (see, the second part of Theorem 13).

Lemma 18 *(Finite Identification) Let $\{\mathbf{u}^j\}_{j\in\mathbb{N}}$ be a sequence converging to a* P*-stationary point $\widehat{\mathbf{u}}$ of* (29)*. Suppose that $\widehat{\boldsymbol{\xi}}$ and $\nabla_{\boldsymbol{\xi}} g(\widehat{\mathbf{u}})$ satisfy strictly complementary condition* (40)*, then there exists sufficiently large integer $\widehat{j}$ such that*

$$\Gamma_j = \mathcal{S}(\boldsymbol{\xi}^{j+1/2}) = \mathcal{S}(\widehat{\boldsymbol{\xi}}), \quad \begin{cases} T_j = \mathcal{S}(\mathbf{w}^{j+1/2}) = \mathcal{S}(\widehat{\mathbf{w}}), & \text{if } \|\widehat{\mathbf{w}}\|_0 = s,\\ T_j \supseteq \mathcal{S}(\mathbf{w}^{j+1/2}) \supseteq \mathcal{S}(\widehat{\mathbf{w}}), & \text{if } \|\widehat{\mathbf{w}}\|_0 < s, \end{cases} \quad \forall j \geq \widehat{j}, \qquad (87)$$

where $\mathcal{S}(\cdot)$ includes indices of nonzero elements for a given vector.

Proof Let us first prove the relationship involving T_j. Since $\lim_{j\to\infty}\mathbf{u}^j = \widehat{\mathbf{u}}$ and (38), we know that $\lim_{j\to\infty}\mathbf{u}^{j+1/2} = \widehat{\mathbf{u}}$ also holds. Then when j is sufficiently large, we have $\mathcal{S}(\mathbf{w}^{j+1/2}) \supseteq \mathcal{S}(\widehat{\mathbf{w}})$. The relationship $T_j \supseteq \mathcal{S}(\mathbf{w}^{j+1/2})$ directly follows from (35), and thus

$T_j \supseteq \mathcal{S}(\mathbf{w}^{j+1/2}) \supseteq \mathcal{S}(\widehat{\mathbf{w}})$ always holds when j is sufficiently large. When $\|\widehat{\mathbf{w}}\|_0 = s$, $|T_j| = s$ indicates $T_j = \mathcal{S}(\mathbf{w}^{j+1/2}) = \mathcal{S}(\widehat{\mathbf{w}})$.

We now prove $\Gamma_j = \mathcal{S}(\widehat{\boldsymbol{\xi}})$. From $\lim_{j\to\infty} \mathbf{u}^j = \widehat{\mathbf{u}}$, $\mathcal{S}(\boldsymbol{\xi}^{j+1/2}) \supseteq \mathcal{S}(\widehat{\boldsymbol{\xi}})$ holds when j is large enough. Noticing that (31) and (32) lead to $\mathcal{S}(\boldsymbol{\xi}^{j+1/2}) = \Gamma_j$, we can obtain $\Gamma_j \supseteq \mathcal{S}(\widehat{\boldsymbol{\xi}})$.

Finally, we need to prove $\Gamma_j \subseteq \mathcal{S}(\widehat{\boldsymbol{\xi}})$. Suppose for the contradiction that there exists an infinite index set $\widehat{\mathcal{J}}$ such that $\Gamma_j \not\subseteq \mathcal{S}(\widehat{\boldsymbol{\xi}})$ for any $j \in \widehat{\mathcal{J}}$. Then considering $|\Gamma_j| \subseteq [m]$ is finite, without loss of generality, we can assume that there exists a fixed index $\widehat{i} \in \Gamma_j$ but $\widehat{i} \notin \mathcal{S}(\widehat{\boldsymbol{\xi}})$ for all $j \in \widehat{\mathcal{J}}$. From (35), we have $[\nabla_{\boldsymbol{\xi}} g(\mathbf{u}^j)]_{\widehat{i}} = -[\boldsymbol{\xi}^{j+1/2} - \boldsymbol{\xi}^j]_{\widehat{i}}/\beta$. Passing limit $j \to \infty$ on both sides of this equality leads to $[\nabla_{\boldsymbol{\xi}} g(\widehat{\mathbf{u}})]_{\widehat{i}} = 0$. Since $\widehat{i} \notin \mathcal{S}(\widehat{\boldsymbol{\xi}})$, $\widehat{\boldsymbol{\xi}}_{\widehat{i}} = 0$ must hold. This contradicts to the strictly complementary assumption. Therefore, $\Gamma_j = \mathcal{S}(\widehat{\boldsymbol{\xi}})$ holds. ■

Proof of the first part in Theorem 13. Let us first prove $[\nabla g(\widehat{\mathbf{u}})]_{\Upsilon_j} = 0$. Indeed, $[\nabla_{\mathbf{w}} g(\widehat{\mathbf{u}})]_{T_j} = 0$ follows from (87) and (6). $[\nabla_{\boldsymbol{\xi}} g(\widehat{\mathbf{u}})]_{\Gamma_j} = 0$ can be verified by (87) and Lemma 2.

Denote $H(t) := [\nabla^2 g(\widehat{\mathbf{u}} + t(\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}))]_{\Upsilon_j,\Upsilon_j}$. We analyze the relationship between $\|\widetilde{\mathbf{u}}^{j+1} - \widehat{\mathbf{u}}\|$ and $\|\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}\|$ below.

$$\begin{aligned}
\|\widetilde{\mathbf{u}}^{j+1} - \widehat{\mathbf{u}}\| \overset{(87)}{=} &\|[\widetilde{\mathbf{u}}^{j+1} - \widehat{\mathbf{u}}]_{\Upsilon_j}\| = \|[\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}]_{\Upsilon_j} - (H^{j+1/2})^{-1}[\nabla g(\mathbf{u}^{j+1/2})]_{\Upsilon_j}\| \\
\leq &\frac{1}{\sigma_g}\|H^{j+1/2}[\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}]_{\Upsilon_j} - [\nabla g(\mathbf{u}^{j+1/2})]_{\Upsilon_j}\| \\
\leq &\frac{1}{\sigma_g}\|H^{j+1/2}[\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}]_{\Upsilon_j} - [\nabla g(\mathbf{u}^{j+1/2}) - \nabla g(\widehat{\mathbf{u}})]_{\Upsilon_j}\| \\
\leq &\frac{1}{\sigma_g}\|\int_0^1 (H^{j+1/2} - H(t))[\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}]_{\Upsilon_j} \mathrm{d}t\| \leq \frac{1}{\sigma_g}\int_0^1 L_g(1-t)\|\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}\|^2 \mathrm{d}t \\
\leq &\frac{L_g}{2\sigma_g}\|\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}\|^2, \qquad (88)
\end{aligned}$$

where the first inequality is derived by using σ_g-strong convexity of g and the fourth inequality follows from the Lipschitz continuity of $\nabla^2 g$. Then $\lim_{j\to\infty} \widetilde{\mathbf{u}}^{j+1} = \widehat{\mathbf{u}}$ directly follows from $\lim_{j\to\infty} \mathbf{u}^{j+1/2} = \widehat{\mathbf{u}}$. We also have the following equations by using (87) and (34) when j is sufficiently large

$$J(\widetilde{\boldsymbol{\xi}}^{j+1}) = J(\widetilde{\boldsymbol{\xi}}^{j+1}_{\Gamma_j}) = J(\widehat{\boldsymbol{\xi}}_{\Gamma_j}), \quad J(\boldsymbol{\xi}^{j+1/2}) = J(\boldsymbol{\xi}^{j+1/2}_{\Gamma_j}) = J(\widehat{\boldsymbol{\xi}}_{\Gamma_j}).$$

Finally, we prove that the descent property in (34) when j is sufficiently large, and thereby the Newton step will always be adopted.

$$\begin{aligned}
&G(\widetilde{\mathbf{u}}^{j+1}) - G(\mathbf{u}^{j+1/2})\\
=&g(\widetilde{\mathbf{u}}^{j+1}) - g(\mathbf{u}^{j+1/2}) + \lambda J(\widetilde{\boldsymbol{\xi}}^{j+1}) - \lambda J(\boldsymbol{\xi}^{j+1/2})\\
\overset{(87)}{=}&\langle [\nabla g(\mathbf{u}^{j+1/2})]_{\Upsilon_j}, [\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}]_{\Upsilon_j}\rangle + \frac{1}{2}[\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}]_{\Upsilon_j}^\top H^{j+1/2}[\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}]_{\Upsilon_j}\\
&+ o(\|\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}\|^2)\\
\overset{(34)}{\leq}&-\frac{1}{2}[\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}]_{\Upsilon_j}^\top H^{j+1/2}[\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}]_{\Upsilon_j} + o(\|\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}\|^2)\\
\leq&-\frac{\sigma_g}{2}\|[\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}]_{\Upsilon_j}\|^2 + o(\|\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}\|^2)\\
\overset{(87)}{=}&-\frac{\sigma_g}{2}\|\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}\|^2 + o(\|\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}\|^2) \leq -\frac{\sigma_g}{4}\|\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}\|^2,
\end{aligned}$$

where the second inequality follows from the σ_g-strong convexity of g and the last inequality is derived form $\lim_{j\to\infty}\|\widetilde{\mathbf{u}}^{j+1} - \mathbf{u}^{j+1/2}\| = 0$.

Proof of the second part in Theorem 13. Notice that (88) has indicated the relationship between $\|\mathbf{u}^{j+1} - \widehat{\mathbf{u}}\|$ and $\|\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}\|$. To prove the quadratic convergence, we just need to analyze the relationship between $\|\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}\|$ and $\|\mathbf{u}^{j} - \widehat{\mathbf{u}}\|$. By using (35), (87) and $[\nabla g(\widehat{\mathbf{u}})]_{\Upsilon_j} = 0$, we have the following estimation:

$$\begin{aligned}
\|\mathbf{w}^{j+1/2} - \widehat{\mathbf{w}}\| =&\|[\mathbf{w}^{j+1/2} - \widehat{\mathbf{w}}]_{T_j}\| = \|[\mathbf{w}^{j} - \widehat{\mathbf{w}} - \alpha\nabla_{\mathbf{w}} g(\mathbf{u}^j)]_{T_j}\|\\
=&\|[\mathbf{w}^{j} - \widehat{\mathbf{w}} - \alpha(\nabla_{\mathbf{w}} g(\mathbf{u}^j) - \nabla_{\mathbf{w}} g(\widehat{\mathbf{u}}))]_{T_j}\|\\
\leq&\|\mathbf{w}^{j} - \widehat{\mathbf{w}}\| + \alpha\ell_g\|\mathbf{u}^j - \widehat{\mathbf{u}}\| \leq \underbrace{(1+\alpha\ell_g)}_{:=\zeta_2}\|\mathbf{u}^j - \widehat{\mathbf{u}}\|,\\
\|\boldsymbol{\xi}^{j+1/2} - \widehat{\boldsymbol{\xi}}\| =&\|[\boldsymbol{\xi}^{j+1/2} - \widehat{\boldsymbol{\xi}}]_{\Gamma_j}\| = \|[\boldsymbol{\xi}^{j} - \widehat{\boldsymbol{\xi}} - \beta\nabla_{\boldsymbol{\xi}} g(\mathbf{u}^j)]_{\Gamma_j}\|\\
=&\|[\boldsymbol{\xi}^{j} - \widehat{\boldsymbol{\xi}}]_{\Gamma_j} - \beta[\nabla_{\boldsymbol{\xi}} g(\mathbf{u}^j) - \nabla_{\boldsymbol{\xi}} g(\widehat{\mathbf{u}})]_{\Gamma_j}\|\\
\leq&\|\boldsymbol{\xi}^{j} - \widehat{\boldsymbol{\xi}}\| + \beta\ell_g\|\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}\| \leq \underbrace{(1+\beta\ell_g)}_{:=\zeta_3}\|\mathbf{u}^j - \widehat{\mathbf{u}}\|.
\end{aligned}$$

These two results lead to

$$\|\mathbf{u}^{j+1/2} - \widehat{\mathbf{u}}\|^2 = \|\mathbf{w}^{j+1/2} - \widehat{\mathbf{w}}\|^2 + \|\boldsymbol{\xi}^{j+1/2} - \widehat{\boldsymbol{\xi}}\|^2 \leq (\zeta_2^2 + \zeta_3^2)\|\mathbf{u}^j - \widehat{\mathbf{u}}\|^2,$$

which combining with (88) implies local quadratic rate. ■

References

Steve M Bajgier and Arthur V Hill. An experimental comparison of statistical and linear programming approaches to the discriminant problem. *Decision Sciences*, 13(4):604–618, 1982.

Liqun Ban, Boris S Mordukhovich, and Wen Song. Lipschitzian stability of parametric variational inequalities over generalized polyhedra in banach spaces. *Nonlinear Analysis: Theory, Methods & Applications*, 74(2):441–461, 2011.

Amir Beck. *First-Order Methods in Optimization*. MOS-SIAM Series on Optimization. Society for Industrial and Applied Mathematics, Philadelphia, 2017.

Amir Beck and Yonina C Eldar. Sparsity constrained nonlinear optimization: Optimality conditions and algorithms. *SIAM Journal on Optimization*, 23(3):1480–1509, 2013.

Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena scientific optimization and computation series. Athena Scientific, Nashua, 1996.

Ernesto G Birgin and José Mario Martínez. *Practical augmented Lagrangian methods for constrained optimization*. SIAM, 2014.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Nonconvex Lagrangian-based optimization: monitoring schemes and global convergence. *Mathematics of Operations Research*, 43(4):1210–1232, 2018.

Radu Ioan Boţ and Dang-Khoa Nguyen. The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. *Mathematics of Operations Research*, 45(2):682–712, 2020.

J Paul Brooks. Support vector machines with the ramp loss and the hard margin loss. *Operations research*, 59(2):467–479, 2011.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology*, 2(3):1–27, 2011.

Xiaojun Chen, Lei Guo, Zhaosong Lu, and Jane J Ye. An augmented Lagrangian method for non-lipschitz nonconvex programming. *SIAM Journal on Numerical Analysis*, 55(1): 168–193, 2017.

Nello Cristianini, John Shawe-Taylor, et al. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

Ying Cui and Jong-Shi Pang. *Modern nonconvex nondifferentiable optimization*. SIAM, 2021.

Ying Cui, Junyi Liu, and Jong-Shi Pang. Nonconvex and nonsmooth approaches for affine chance-constrained stochastic programs. *Set-Valued and Variational Analysis*, 30(3): 1149–1211, 2022.

Ying Cui, Junyi Liu, and Jong-Shi Pang. The minimization of piecewise functions: Pseudo stationarity. *arXiv preprint arXiv:2305.14798*, 2023.

Sheng Dai. Variable selection in convex quantile regression: ℓ_1-norm or ℓ_0-norm regularization? *European Journal of Operational Research*, 305(1):338–355, 2023.

Alberto De Marchi, Xiaoxi Jia, Christian Kanzow, and Patrick Mehlitz. Constrained composite optimization and augmented Lagrangian methods. *Mathematical Programming*, pages 1–34, 2023.

Antoine Dedieu, Hussein Hazimeh, and Rahul Mazumder. Learning sparse classifiers: Continuous and mixed integer optimization perspectives. *The Journal of Machine Learning Research*, 22(1):6008–6054, 2021.

Antoine Dedieu, Rahul Mazumder, and Haoyue Wang. Solving ℓ_1-regularized svms and related linear programs: Revisiting the effectiveness of column and constraint generation. *Journal of Machine Learning Research*, 23(164):1–41, 2022.

Mingbin Feng, John E Mitchell, Jong-Shi Pang, Xin Shen, and Andreas Wächter. Complementarity formulations of ℓ_0-norm optimization problems. *Pacific Journal of Optimization*, 14:273–305, 2018.

Glenn Fung and Olvi L Mangasarian. Proximal support vector machine classifiers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 77–86, 2001.

Glenn M Fung and Olvi L Mangasarian. A feature selection Newton method for support vector machine classification. *Computational Optimization and Applications*, 28:185–202, 2004.

Andrés Gómez, Ziyu He, and Jong-Shi Pang. Linear-step solvability of some folded concave and singly-parametric sparse optimization problems. *Mathematical Programming*, 198(2): 1339–1380, 2023.

Shaoning Han, Ying Cui, and Jong-Shi Pang. Analysis of a class of minimization problems lacking lower semicontinuity. *Mathematics of Operations Research*, 2024. URL `https://doi.org/10.1287/moor.2023.0295`.

Magnus R Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, 1969.

Xiaoxi Jia, Christian Kanzow, Patrick Mehlitz, and Gerd Wachsmuth. An augmented Lagrangian method for optimization problems with structured geometric constraints. *Mathematical Programming*, 199(11):1365–1415, 2023.

Kory D Johnson, Dongyu Lin, Lyle H Ungar, Dean P Foster, and Robert A Stine. A risk ratio comparison of ℓ_0 and ℓ_1 penalized regression. *arXiv preprint arXiv:1510.06319*, 2015.

Christian Kanzow and Hou-Duo Qi. A QP-free constrained Newton-type method for variational inequality problems. *Mathematical Programming*, 85(1):81–106, 1999.

Christian Kanzow, Andreas B Raharja, and Alexandra Schwartz. An augmented Lagrangian method for cardinality-constrained optimization problems. *Journal of Optimization Theory and Applications*, 83:793–813, 2021.

Christian Kanzow, Alexandra Schwarz, and Felix Weiß. The sparse (st) optimization problem: Reformulations, optimality, stationarity, and numerical results. *arXiv preprint arXiv:2210.09589*, 2022.

Yuh-Jye Lee and Olvi L Mangasarian. SSVM: A smooth support vector machine for classification. *Computational Optimization and Applications*, 20(1):5–22, 2001.

Guoyin Li and Ting Kei Pong. Global convergence of splitting methods for nonconvex composite optimization. *SIAM Journal on Optimization*, 25(4):2434–2460, 2015.

Huan Li and Zhouchen Lin. Accelerated proximal gradient methods for nonconvex programming. *Advances in neural information processing systems*, 28, 2015.

JM Liittschwager and C Wang. Integer programming solution of a classification problem. *Management Science*, 24(14):1515–1525, 1978.

Olvi L Mangasarian and David R Musicant. Lagrangian support vector machines. *Journal of Machine Learning Research*, 1(Mar):161–177, 2001.

Boris S Mordukhovich and Nguyen Mau Nam. *An easy path to convex analysis and applications*. Synthesis Lectures on Mathematics and Statistics. Morgan & Claypool Publishers, California, 2013.

Jorge Nocedal and Stephen J. Wright. *Numerical optimization*. Springer series in operations research and financial engineering. Springer, New York, 2006.

Li-Li Pan, Nai-Hua Xiu, and Sheng-Long Zhou. On solutions of sparsity constrained optimization. *Journal of the Operations Research Society of China*, 3(4):421–439, 2015.

Michael JD Powell. A method for nonlinear constraints in minimization problems. *Optimization*, pages 283–298, 1969.

R Tyrrell Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1(2):97–116, 1976.

Yuan-Hai Shao, Chun-Na Li, Ling-Wei Huang, Zhen Wang, Nai-Yang Deng, and Yan Xu. Joint sample and feature selection via sparse primal and dual LSSVM. *Knowledge-Based Systems*, 185:104915, 2019.

Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14:199–222, 2004.

Jiebo Song, Jia Li, Zhengan Yao, Kaisheng Ma, and Chenglong Bao. Zero norm based analysis model for image smoothing and reconstruction. *Inverse Problems*, 36(11):115009, 2020.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Andreas Themelis, Lorenzo Stella, and Panagiotis Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM Journal on Optimization*, 28(3):2274–2303, 2018.

Berk Ustun and Cynthia Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102:349–391, 2016.

Vladimir N. Vapnik. *Statistical Learning Theory*. John-Wiley and Sons, INC, 1998.

Fenghui Wang, Wenfei Cao, and Zongben Xu. Convergence of multi-block Bregman ADMM for nonconvex composite problems. *Science China Information Sciences*, 61(12):1–12, 2018.

Huajun Wang, Yuanhai Shao, Shenglong Zhou, Ce Zhang, and Naihua Xiu. Support vector machine classifier via $L_{0/1}$ soft-margin loss. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7253–7265, 2021.

Yue Xie and Stephen J Wright. Complexity of proximal augmented Lagrangian for nonconvex optimization with nonlinear equality constraints. *Journal of Scientific Computing*, 86:1–30, 2021.

Yangyang Xu. Iteration complexity of inexact augmented Lagrangian methods for constrained convex programming. *Mathematical Programming*, 185(1):199–244, 2021.

Guo-Xun Yuan, Kai-Wei Chang, Cho-Jui Hsieh, and Chih-Jen Lin. A comparison of optimization methods and software for large-scale ℓ_1-regularized linear classification. *The Journal of Machine Learning Research*, 11:3183–3234, 2010.

Penghe Zhang, Naihua Xiu, and Hou-Duo Qi. iNALM: An inexact Newton augmented Lagrangian method for zero-one composite optimization. *arXiv preprint arXiv:2306.08991*, 2023.

Shenglong Zhou, Lili Pan, Naihua Xiu, and Hou-Duo Qi. Quadratic convergence of smoothing Newton's method for 0/1 loss optimization. *SIAM Journal on Optimization*, 31(4): 3184–3211, 2021a.

Shenglong Zhou, Naihua Xiu, and Hou-Duo Qi. Global and quadratic convergence of Newton hard-thresholding pursuit. *The Journal of Machine Learning Research*, 22(12):1–45, 2021b.

Ji Zhu, Saharon Rosset, Robert Tibshirani, and Trevor Hastie. 1-norm support vector machines. *Advances in neural information processing systems*, 16, 2003.