PROVABLE LEARNING-BASED ALGORITHM FOR SPARSE RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering sparse parameters from observational data is a fundamental problem in machine learning with wide applications. Many classic algorithms can solve this problem with theoretical guarantees, but the theoretical performances depend on choosing the correct hyperparameters. Besides, they do not fully exploit the particular problem distribution of interest. In this work, we propose PLISA (Provable Learning-based Iterative Sparse recovery Algorithm) to learn algorithms automatically from data. PLISA is designed by unrolling a classic path-following algorithm, with some components being more flexible and learnable. With this structure, we theoretically show the improved recovery accuracy achievable by PLISA. Furthermore, we analyze the empirical Rademacher complexity of PLISA to characterize its generalization ability to solve new problems outside the training set. This paper contains novel theoretical contributions to the area of learning-based algorithms in the sense that (i) PLISA is generically applicable to a broad class of sparse estimation problems, (ii) generalization analysis has received less attention so far, and (iii) our analysis makes novel connections between the generalization ability and algorithmic properties such as stability and convergence, which leads to a tighter bound that can explain the empirical observations. The techniques could potentially be applied to analyze other learning-based algorithms in the literature.

1 INTRODUCTION

The problem of recovering a sparse vector $\boldsymbol{\beta}^*$ from finite observations $Z_{1:n} \sim (\mathbb{P}_{\boldsymbol{\beta}^*})^n$ is fundamental in machine learning, covering a broad family of problems including compressed sensing, sparse regression analysis, graphical model estimation, etc. It has also found applications in various domains. For example, in magnetic resonance imaging, sparse signals need to be reconstructed from measurements taken by a scanner. In computational biology, estimating a sparse graph structure from gene expression data is important for understanding gene regulatory networks.

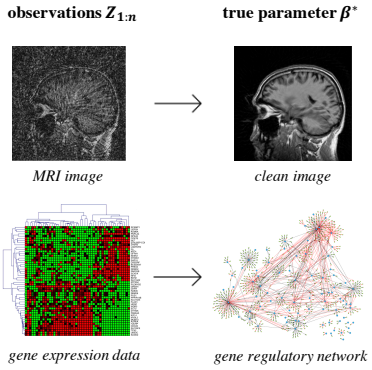

Figure 1: Sparse recovery problems.

Various classic algorithms are available for solving sparse recovery problems, and many of them come with theoretical guarantees for the recovery accuracy. However, the theoretical performance often relies on choosing the correct hyperparameters, such as regularization parameters and the learning rate, which may depend on unknown constants. Furthermore, in practice, similar problems may need to be solved repeatedly, but it is hard for classic algorithms to fully utilize the information of the particular problem distribution of interest.

To alleviate these limitations, we consider the approach of learning-to-learn and propose a neural algorithm, called PLISA (Provable Learning-based Iterative Sparse recovery Algorithm). PLISA is a deep learning model that takes the observations $Z_{1:n}$ as the input and outputs an estimation for $\boldsymbol{\beta}^*$. To make use of classic techniques developed by domain experts, we design the architecture of PLISA by *unrolling and modifying a classic path-following algorithm* proposed by Wang et al. (2014). To benefit from learning, some components in this classic algorithm are made more flexible

with careful design and treated as learnable parameters in PLISA. These parameters can be learned by optimizing the performances on a set of training problems. The learned PLISA can then be used for solving other problems in the target distribution. With the algorithm design problem converted to a deep learning problem, we naturally ask the two important questions:

1. *Capacity:* What's the recovery accuracy achievable by PLISA? Can the flexible components in PLISA lead to an algorithm which effectively improves the recovery performance?
2. *Generalization:* How well can the learned PLISA solve new problems outside the training set? Is the generalization behavior related to the algorithmic properties of PLISA?

Aiming at supplying rigorous answers to these questions, we conduct theoretical analysis of PLISA to provide guarantees for its recovery performance, measure its empirical Rademachar complexity, and characterize its generalization ability. The results and the techniques in our analysis can distinguish our work from existing studies on algorithm learning and deep unrolling, which has become an active research direction recently. To be more specific, we summarize our new and novel contributions into the following three aspects.

Theoretical understanding. In contrast to the plethora of empirical studies on algorithm learning, there have been relatively few studies devoted to the theoretical understanding. Existing theoretical efforts primarily focus on analyzing the convergence rate achievable by the neural algorithm (Chen et al., 2018; Liu et al., 2019a; Zhang & Ghanem, 2018; Wu et al., 2020), but the generalization error bound has received less attention so far. A substantial body of works only argue intuitively that algorithm unrolling architectures can generalize well because they contain a small number of parameters. In comparison, we provide theoretical guarantees for both the capacity and the generalization ability of PLISA, which are more solid arguments.

Novel connection. The algorithmic structure in the architecture of PLISA can make it behaves differently from conventional neural networks. Therefore, we largely utilize the analysis techniques in classic algorithms to derive its generalization bound. By combining the analysis tools of deep learning theory and optimization algorithms, our result reveals a novel connection between the generalization ability of PLISA and its algorithmic properties including the convergence rate and stability. Benefit from this connection, our generalization bound is tight in the sense that it matches the interesting behavior of PLISA observed in experiments - the generalization gap could decrease in the number of layers, which is rarely observed in conventional neural networks.

General setting. The problem setting in this paper is new and more challenging. Existing works mainly focus on a specific problem. For example, the compressed sensing problem with a fixed design matrix is the mostly investigated one. PLISA, however, is generic and is applicable to various sparse recovery problems as long as they satisfy certain conditions in Assumption 3.1. This is complementary to the literature of algorithm learning.

The remainder of the paper is organized as follows. In Section 2, we elaborate on the proposed architecture of PLISA, and present the learning-to-learn setting under which we learn the parameters. Section 3 states our first theoretical result on the achievable recovery performance of PLISA. Section 4 is devoted to the generalization analysis of PLISA. In Section 5, we summarize the related works. Section 6 supplies results of numerical simulations. Proofs are contained in the Appendix.

2 PLISA: Learning To Solve Sparse Estimation Problems

A sparse estimation problem is to recover $\boldsymbol{\beta}^*$ from finite observations $Z_{1:n}$ sampled from $\mathbb{P}_{\boldsymbol{\beta}^*}$. As a concrete example, in a sparse linear regression problem, n observations $\{Z_i = (\boldsymbol{x}_i, y_i)\}_{i=1}^n$ are sampled from a linear model $y = \boldsymbol{x}^\top \boldsymbol{\beta}^* + \epsilon$, and an algorithm needs to estimate the vector $\boldsymbol{\beta}^*$ in this linear model from the n observations. Therefore, an oracle algorithm that perfectly recovers the true parameter can be view as a function that maps the observations $Z_{1:n}$ to $\boldsymbol{\beta}^*$:

$$\text{Oracle algorithm (function): } Z_{1:n} \mapsto \boldsymbol{\beta}^*.$$

Based on the equivalent view of algorithms and functions, we propose to approximate the oracle algorithm by a deep learning model, called PLISA_θ. In this section, we will first describe the architecture of PLISA_θ and explain its design principles. After that, we will describe how to optimize the parameters in PLISA_θ under the learning-to-learn setting.

2.1 ARCHITECTURE OF $\texttt{PLISA}_\theta$

A naive architecture could be a feed-forward or recurrent neural network that takes $Z_{1:n}$ as the input and outputs a d-dimensional vector. However, a more favorable design choice in the literature of algorithm learning is *algorithm unrolling*, which unrolls and truncates a classic iterative algorithm to design the architecture. Based on this idea, we use a classic path-following algorithm proposed by Wang et al. (2014) as the basis to design $\texttt{PLISA}_\theta$.

The key idea of path-following algorithms is creating a sequence of surrogate objective functions to gradually approach the target objective function, which is supposed to be more difficult to solve. Following the similar design-logic of the path-following algorithm in Wang et al. (2014), $\texttt{PLISA}_\theta$ is designed to sequentially approximate the local minimizers of a sequence of objectives,

$$\text{for } t = 1, \cdots, T:$$
$$\boldsymbol{\beta}_t(Z_{1:n}; \theta) \approx \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t} \in \arg\min L_n(Z_{1:n}, \boldsymbol{\beta}) + P_{\boldsymbol{w}}(\boldsymbol{\lambda}_t, \boldsymbol{\beta}) \quad (1)$$

with regularization parameters $\boldsymbol{\lambda}_1, \cdots, \boldsymbol{\lambda}_T$. In this objective, L_n is an empirical loss that measures the "fit" between a parameter $\boldsymbol{\beta}$ and the observations $Z_{1:n}$. As an example, it can be the least squared loss for solving linear regression problems. The second term in the objective, $P_{\boldsymbol{w}}$, is a sparsity penalty function, which is set to be $\lambda\|\boldsymbol{\beta}\|_1$ in many classic algorithms.

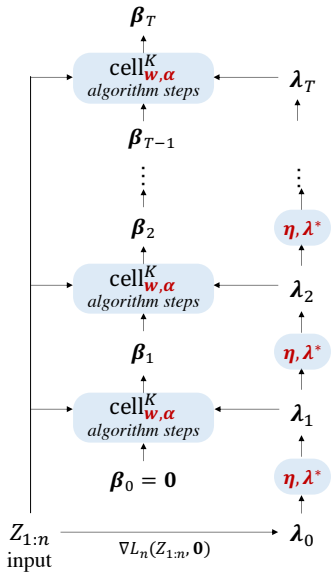

Figure 2: Overall architecture.

Consequently, as illustrated by Fig. 2 and Algorithm 1, the architecture of $\texttt{PLISA}_\theta$ contains T recurrent cells. The t-th cell corresponds to the t-th objective in Eq. 1, and it contains a K-step algorithm that iteratively minimize the cell-specific objective.

Algorithm 1: Architecture of $\texttt{PLISA}_\theta$

#cells: T, *#layers per cell:* K
Parameters: $\theta = \{\boldsymbol{\eta}, \boldsymbol{\lambda}^*, \boldsymbol{w}, \alpha\}$
Input: samples $Z_{1:n}$
$\boldsymbol{\beta}_0 \leftarrow \mathbf{0}, \quad \boldsymbol{\lambda}_0 \leftarrow \nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \mathbf{0})$
For $t = 1, \ldots, T$ **do**
  $\boldsymbol{\lambda}_t \leftarrow \max\{\sigma(\boldsymbol{\eta}) \circ \boldsymbol{\lambda}_{t-1}, \boldsymbol{\lambda}^*\}$
  $\boldsymbol{\beta}_t \leftarrow \texttt{Cell}^K_{\boldsymbol{w},\alpha}(Z_{1:n}, \boldsymbol{\beta}_{t-1}, \boldsymbol{\lambda}_t)$ by K many update steps in Eq. 5
return $\boldsymbol{\beta}_T$

While maintaining the overall structure of the path-following algorithm, $\texttt{PLISA}_\theta$ is more flexible, as it contains learnable parameters $\theta = \{\boldsymbol{\eta}, \boldsymbol{\lambda}^*, \boldsymbol{w}, \alpha\}$ that consist of (i) learnable decrease ratios $\boldsymbol{\eta}$ and learnable target regularization parameters $\boldsymbol{\lambda}^*$ that jointly determine the sequence of regularization parameters $\boldsymbol{\lambda}_1, \cdots, \boldsymbol{\lambda}_T$; (ii) a learnable sparsity penalty function $P_{\boldsymbol{w}}$ with parameters $\boldsymbol{w}$; and (iii) a learnable step size α for the optimization steps in the cells.

We explain the details in the following. Red-colored symbols are used to indicate learnable parameters in $\texttt{PLISA}_\theta$. The concrete computations in $\texttt{PLISA}_\theta$ are presented in Algorithm 1.

(i) Entry-wise regularzation parameters. Given input samples $Z_{1:n}$, the regularization parameters are initialized by a vector $\boldsymbol{\lambda}_0 := \nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \mathbf{0})$. Then $\texttt{PLISA}_\theta$ will update it sequentially by

$$\boldsymbol{\lambda}_t \leftarrow \max\{\sigma(\boldsymbol{\eta}) \circ \boldsymbol{\lambda}_{t-1}, \boldsymbol{\lambda}^*\}, \quad \text{for } t = 1, \cdots, T, \quad (2)$$

where $\sigma(\cdot)$ is the element-wise sigmoid function, $\circ$ is element-wise multiplication, $\max\{\cdot, \cdot\}$ is element-wise maximization, and $\{\boldsymbol{\eta}, \boldsymbol{\lambda}^*\}$ are learnable parameters. This operation will create a decreasing sequence $\boldsymbol{\lambda}_1, \cdots, \boldsymbol{\lambda}_T$ through the entry-wise decrease ratio $\sigma(\boldsymbol{\eta})$, until they reach the entry values of the learnable target vector $\boldsymbol{\lambda}^*$. It is worth mentioning that different from classic algorithms which employ a uniform regularization parameter λ across all entries of $\boldsymbol{\beta}$, $\texttt{PLISA}_\theta$ always maintains a d-dimensional vector $\boldsymbol{\lambda}_t = [\lambda_{t,1}, \cdots, \lambda_{t,d}]^\top$ to enforce different levels of sparse penalty to different entries.

(ii) Penalty function. Unlike classic algorithms that use a pre-defined penalty function such as the ℓ_1 regularizer, $\texttt{PLISA}_\theta$ parameterizes the penalty function as follows,

$$P_{\boldsymbol{w}}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \textstyle\sum_{i=1}^q \widetilde{w_i} \cdot P^{(i)}(\boldsymbol{\lambda}, \boldsymbol{\beta}), \quad \text{where } \widetilde{w_i} = \frac{\exp(w_i)}{\sum_{i'=1}^q \exp(w_{i'})}. \quad (3)$$

In other words, $P_{\boldsymbol{w}}$ is a convex combination of q penalty functions $(P^{(1)}, \cdots, P^{(q)})$, in which the weights of these functions are determined by a learnable vector $\boldsymbol{w} = [w_1, \cdots, w_q]$. In this paper, we focus on learning the combination of three well-known penalty functions:

$$P^{(1)}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1, \;\; P^{(2)}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \textstyle\sum_{j=1}^{p} \mathrm{MCP}(\lambda_j, \beta_j), \;\; P^{(3)}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \textstyle\sum_{j=1}^{p} \mathrm{SCAD}(\lambda_j, \beta_j),$$

where $P^{(1)}$ is convex, and MCP (Zhang, 2010a) and SCAD (Fan & Li, 2001) are *nonconvex penalties* whose analytical forms are given in Appendix B. In fact, one can include any other penalty functions as long as they satisfy a set of conditions specified in Appendix B.

(iii) $\texttt{Cell}^K_{\boldsymbol{w},\alpha}$ **and the step size.** As mentioned earlier, each cell in $\texttt{PLISA}_\theta$ contains a K-step algorithm that iteratively minimizes the cell-specific objective in Eq. 1. Mathematically, given the inputs $Z_{1:n}$, $\boldsymbol{\beta}_{t-1}$, and $\boldsymbol{\lambda}_t$, the t-th cell $\texttt{Cell}^K_{\boldsymbol{w},\alpha}$ initialize $\widetilde{\boldsymbol{\beta}}^0_t$ by the output of last cell $\boldsymbol{\beta}_{t-1}$, and computes the following *modified proximal gradient* steps:

$$\text{for } k = 1, \cdots, K, \quad \widetilde{\boldsymbol{\beta}}^k_t \leftarrow \mathcal{T}_{\alpha \cdot \boldsymbol{\lambda}_t} \left(\widetilde{\boldsymbol{\beta}}^{k-1}_t - \alpha \big(\nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \widetilde{\boldsymbol{\beta}}^{k-1}_t) + \nabla_{\boldsymbol{\beta}} Q_{\boldsymbol{w}}(\boldsymbol{\lambda}_t, \widetilde{\boldsymbol{\beta}}^{k-1}_t) \big) \right). \tag{4}$$

$$\text{cell output:} \qquad \texttt{Cell}^K_{\boldsymbol{w},\alpha}(Z_{1:n}, \boldsymbol{\beta}_{t-1}, \boldsymbol{\lambda}_t) = \boldsymbol{\beta}_t = \widetilde{\boldsymbol{\beta}}^K_t \tag{5}$$

In the classic algorithm, the step size α in Eq. 4 is obtained by performing line-search. In $\texttt{PLISA}_\theta$, α is a learnable parameter. Experimentally, we find this is much more efficient and effective.

For the notations in Eq. 4, $\mathcal{T}_{\alpha \cdot \boldsymbol{\lambda}_t}$ is the entry-wise soft-thresholding function, defined as $[\mathcal{T}_{\boldsymbol{\delta}}(\boldsymbol{\beta})]_j := \mathrm{sign}(\beta_j) \max\{|\beta_j| - \delta_j, 0\}$. The function $Q_{\boldsymbol{w}}$ represents the concave component of the penalty function $P_{\boldsymbol{w}}$, defined as $Q_{\boldsymbol{w}}(\boldsymbol{\lambda}, \boldsymbol{\beta}) := P_{\boldsymbol{w}}(\boldsymbol{\lambda}, \boldsymbol{\beta}) - \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1$. The analytical form of its gradient $\nabla_{\boldsymbol{\beta}} Q_{\boldsymbol{w}}(\boldsymbol{\lambda}_t, \boldsymbol{\beta})$ can be found in Appendix B.

Discussion. We choose the path-following algorithm in (Wang et al., 2014) as the basis to design $\texttt{PLISA}_\theta$ because this algorithm is applicable to nonconvex losses L_n and nonconvex penalty functions $P_{\boldsymbol{w}}$. This allows $\texttt{PLISA}_\theta$ to be applicable to a broader class of problems (e.g., nonlinear sparse regression problems may involve nonconvex loss). Furthermore, employing nonconvex L_n and $P_{\boldsymbol{w}}$ can potentially lead to better statistical properties (Fan & Li, 2001; Fan et al., 2009; Loh & Wainwright, 2015), for which we will explain more in Section 3 and in experiments.

2.2 Learning-to-learn Setting

Now we describe how to train the parameters θ in $\texttt{PLISA}_\theta$ under the learning-to-learn setting.

Training set. Similar to other works in this domain, we assume the access to m problems from the target problem-space $\mathcal{P}$, and use them as the training set:

$$\mathcal{D}_m = \{(Z^{(1)}_{1:n_1}, \boldsymbol{\beta}^{*(1)}), \cdots, (Z^{(m)}_{1:n_m}, \boldsymbol{\beta}^{*(m)})\} \quad \text{with} \quad (Z^{(i)}_{1:n_i}, \boldsymbol{\beta}^{*(i)}) \in \mathcal{P}.$$

Here each estimation problem is represented by a pair of observations and the corresponding true parameter to be recovered. A different problem i can contain a different number n_i of observations.

Training loss. Since the intermediate outputs $\boldsymbol{\beta}_t(Z_{1:n}; \theta)$ of $\texttt{PLISA}_\theta$ are also estimates of $\boldsymbol{\beta}^*$, a common design of the training loss is the weighted sum of the intermediate estimation errors (Chen et al., 2021). More specifically, we employ the following training loss:

$$\mathcal{L}^{\gamma}_{train}(\mathcal{D}_m; \theta) := \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{T} \gamma^{T-t} \left\| \boldsymbol{\beta}_t(Z^{(i)}_{1:n_i}; \theta) - \boldsymbol{\beta}^{*(i)} \right\|_2^2, \tag{6}$$

where $\gamma < 1$ is a discounting factor If $\gamma = 0$ then the loss is only estimated at the last layer.

Generalization error. The ultimate goal of algorithm learning is to minimize the estimation error on expectation over all problems in the target problem distribution:

$$\mathcal{L}_{gen}(\mathbb{P}(\mathcal{P}); \theta) := \mathbb{E}_{(Z_{1:n}, \boldsymbol{\beta}^*) \sim \mathbb{P}(\mathcal{P})} \|\boldsymbol{\beta}_T(Z_{1:n}; \theta) - \boldsymbol{\beta}^*\|_2^2, \tag{7}$$

where $\mathbb{P}(\mathcal{P})$ is some distribution in the target problem-space $\mathcal{P}$. Similar to other machine learning problems, this generalization error depends on both the *(i) empirical error (i.e., training loss)* and the *(ii) generalization gap*. We will theoretically characterize them in Section 3 and Section 4.

3 CAPACITY OF PLISA

Can $\texttt{PLISA}_\theta$ recover the true parameters accurately within a small number of layers (cells)? In this section, we answer this question by theoretically showing that $\texttt{PLISA}_\theta$ can achieve a smaller training error compared to the classic algorithm. Moreover, $\texttt{PLISA}_\theta$ enjoys a fast convergence rate, and therefore the small training error can be obtained without using too many layers.

3.1 PROBLEM SPACE ASSUMPTIONS

Before stating the theorem, we follow the notations in Wang et al. (2014); Loh & Wainwright (2015) to describe some classic assumptions on the estimation problems.

Assumption 3.1 (Problem Space)**.** *Let $s^*, \tilde{s}$ be positive integers and ρ_-, ρ_+ be positive constants such that $\tilde{s} > (121(\rho_+/\rho_-) + 144(\rho_+/\rho_-)^2)s^*$. Assume for every estimation problem $(Z_{1:n}, \boldsymbol{\beta}^*)$ in the space $\mathcal{P}$, the following conditions are satisfied.*

(a) $\|\boldsymbol{\beta}^*\|_0 \leq s^*$ *and* $\|\boldsymbol{\beta}^*\|_\infty \leq B_1$*;*

(b) For any nonzero $\boldsymbol{v} \in \mathbb{R}^d$ *with sparsity* $\|\boldsymbol{v}\|_0 \leq s^* + 2\tilde{s}$*, it holds* $\frac{\boldsymbol{v}^\top \nabla^2_{\boldsymbol{\beta}} L_n(Z_{1:n}, \boldsymbol{\beta})\boldsymbol{v}}{\|\boldsymbol{v}\|_2^2} \in [\rho_-, \rho_+]$*;*

(c) $8\,|[\nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \boldsymbol{\beta}^*)]_j| \leq |[\nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \mathbf{0})]_j| \leq B_2, \forall j = 1, \cdots, d.$

Condition (a) assumes $\boldsymbol{\beta}^*$ is s^*-sparse and B_1-bounded. Condition (b) is commonly referred to as 'sparse eigenvalue condition' (Zhang, 2010b; Wang et al., 2014), which is weaker than the well-known restricted isometry property (RIP) in compressed sensing (Candes & Tao, 2005). Note that the class of functions satisfying conditions of this type is much larger than the class of convex losses. In the special case when $L_n(Z_{1:n}, \boldsymbol{\beta})$ is strongly convex in $\boldsymbol{\beta}$, condition (b) holds with $\tilde{s} \to \infty$. The last condition bounds the gradient of the empirical loss L_n at the true parameter $\boldsymbol{\beta}^*$ and $\mathbf{0}$.

3.2 FIRST MAIN RESULT: CAPACITY

Let $\boldsymbol{\beta}_t(Z_{1:n}; \theta)$ be the output of the t-th cell in $\texttt{PLISA}_\theta$. Let $\boldsymbol{x} \vee a$ denote entry-wise maximal value $\max\{\boldsymbol{x}, a\}$. Let $(\boldsymbol{x})_S$ denote the sub-vector of $\boldsymbol{x}$ with entries indexed by the set S.

Theorem 3.1 (Capacity)**.** *Assume the problem space $\mathcal{P}$ satisfies Assumption 3.1 and $\mathcal{D}_m \subseteq \mathcal{P}$. Let T be the number of cells in $\texttt{PLISA}_\theta$ and let K be the number of layers in each cell. For any $\varepsilon > 0$, there exists a set of parameters $\theta = \{\boldsymbol{\eta}, \boldsymbol{\lambda}^*, \boldsymbol{w}, \alpha\}$ such that the estimation error of every problem $(Z_{1:n}, \boldsymbol{\beta}^*) \in \mathcal{D}_m$ is bounded as follows, $\forall T > t_0$,*

$$\|\boldsymbol{\beta}_T(Z_{1:n}; \theta) - \boldsymbol{\beta}^*\|_2 \leq \varepsilon^{-1} c_\theta s^* \exp(-C_\theta K(T - t_0)) \qquad \textbf{\textit{optimization error}} \tag{8}$$

$$+ c'_\theta \kappa_m \| \left(\nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \boldsymbol{\beta}^*) \vee \varepsilon\right)_{S^*} \|_2, \qquad \textbf{\textit{statistical error}} \tag{9}$$

where $S^* := \operatorname{supp}(\boldsymbol{\beta}^*)$ is the support indices of $\boldsymbol{\beta}^*$, c_θ, c'_θ, and C_θ are some positive values depending on the chosen θ, and κ_m is a condition number which reveals the similarity of the problems in $\mathcal{D}_m$. Note that K and t_0 are required to be larger than certain values, but we will elaborate in Appendix D that the required lower bounds are small. See Appendix D for the proof of this theorem.

This estimation error can be interpreted as a combination of the optimization error (in Eq. 8) and the statistical error (in Eq. 9). Eq. 8 shows the optimization error decreases linearly in both K and T. This fast convergence rate allows us to use fewer layers and cells in $\texttt{PLISA}_\theta$, which is important because a deeper architecture could be more difficult and less efficient to train. The statistical error in Eq. 9 occurs because of the randomness in $Z_{1:n}$. The gradient at the true parameter $\nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \boldsymbol{\beta}^*)$ characterizes how well the finite samples $Z_{1:n}$ can represent the distribution $\mathbb{P}_{\boldsymbol{\beta}^*}$.

In the following, we will elaborate on how the design of entry-wise regularization and learnable penalty function have improved the statistical error.

(i) Impact of entry-wise regularization. In classic algorithms that use a uniform regularization parameter, λ^* is usually taken to be proportional to $\|\nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \boldsymbol{\beta}^*)\|_\infty$ in order to achieve the optimal statistical rate (Wang et al., 2014; Loh & Wainwright, 2015). In our case, restricting the regularization to be uniform across entries will lead to an error bound that replaces the norm $\| \left(\nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \boldsymbol{\beta}^*) \vee \varepsilon\right)_{S^*} \|_2$ in Eq. 9 by $\sqrt{s^*}\,(\|\nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \boldsymbol{\beta}^*)\|_\infty \vee \varepsilon)$. To understand how the former has improved the latter, we can consider the sparse linear regression problem introduced

in Section 2 with L_n being the least square loss. If the design matrix is normalized such that $\max_{1\le j\le d}\|([\boldsymbol{x}_1]_j,\cdots,[\boldsymbol{x}_n]_j)\|_2\le\sqrt{n}$, then $\|\left(\nabla_{\boldsymbol{\beta}}L_n(Z_{1:n},\boldsymbol{\beta}^*)\vee\varepsilon\right)_{S^*}\|_2\le C\sqrt{s^*/n}$ with high probability. In comparison, $\sqrt{s^*}\|\nabla_{\boldsymbol{\beta}}L_n(Z_{1:n},\boldsymbol{\beta}^*)\|_\infty\le C\sqrt{s^*\log d/n}$ with high probability is a slower statistical rate due to the term $\log d$.

(ii) Impact of learnable penalty function. To explain the the benefit of using learnable penalty function, we give a more refined bound for the statistical error in Eq. 9 in the following lemma.

Lemma 3.1 (Refined bound)**.** *Assume the same conditions and parameters θ in Theorem 3.1. Assume $T\to\infty$ so that the optimization error can be ignored. For simplicity, assume $\widetilde{w_3}=0$ and only consider the weights $\widetilde{w_1}$ and $\widetilde{w_2}$ for ℓ_1 penalty and MCP. Then for every problem $(Z_{1:n},\boldsymbol{\beta}^*)\in\mathcal{D}_m$:*

$$\|\boldsymbol{\beta}_\infty(Z_{1:n};\theta)-\boldsymbol{\beta}^*\|_2\le\frac{1+8(1+\widetilde{w_2})\kappa_m}{\rho_--\widetilde{w_2}/b}\|(\nabla_{\boldsymbol{\beta}}L_n(Z_{1:n},\boldsymbol{\beta}^*)\vee\varepsilon)_{S_1^*}\|_2\quad\left(S_1^*\text{: Small }\left|\beta_j^*\right|\text{'s}\right)\tag{10}$$

$$+\frac{1+8(1-\widetilde{w_2})\kappa_m}{\rho_--\widetilde{w_2}/b}\|(\nabla_{\boldsymbol{\beta}}L_n(Z_{1:n},\boldsymbol{\beta}^*)\vee\varepsilon)_{S_2^*}\|_2\quad\left(S_2^*\text{: Large }\left|\beta_j^*\right|\text{'s}\right),\tag{11}$$

where $b>1$ is a hyperparameter in MCP*, and the index sets S_1^* and S_2^* are defined as $S_1^*:=\{j\in S^*:\left|\beta_j^*\right|\le b\lambda_j^*\}$ and $S_2^*:=\{j\in S^*:\left|\beta_j^*\right|>b\lambda_j^*\}$. See Appendix D for the proof.*

This refined bound reveals the benefit of learning the penalty function because:

1. According to Lemma 3.1, which penalty can minimize the error bound is problem-dependent. More specifically, if $(8(b\rho_-+1)\kappa_m+1)\|(\nabla_{\boldsymbol{\beta}}L_n(Z_{1:n},\boldsymbol{\beta}^*)\vee\varepsilon)_{S_1^*}\|_2$ is larger than $(8(b\rho_--1)\kappa_m-1)\|(\nabla_{\boldsymbol{\beta}}L_n(Z_{1:n},\boldsymbol{\beta}^*)\vee\varepsilon)_{S_2^*}\|_2$, choosing $\widetilde{w_2}=0$ can induce a smaller error bound. Otherwise, $\widetilde{w_2}=1$ is better. Therefore, learning is a more suitable way of choosing the penalty function.
2. Apart from the statistical error, the convergence speed C_θ in Eq. 8 is also affected by the weights, monotonely decreasing in $\widetilde{w_2}$. Through gradient-based training, we can automatically find the optimal combination of penalty functions to strike a nice balance between the statistical error and convergence speed.

4 Generalization Analysis

How well can the learned $\texttt{PLISA}_\theta$ solve new problems outside the training set? In this section, we analyze the generalization gap between the expected loss $\mathcal{L}_{gen}$ in Eq. 7 and the training loss $\mathcal{L}_{train}^{\gamma=0}$ in Eq. 6. In particular, our analysis is conducted in a novel way to focus on answering the questions:

How is the *generalization bound* of $\texttt{PLISA}_\theta$ related to its *algorithmic properties*?
And how is it *different from* conventional neural networks?

4.1 Second Main Result: Generalization Bound

To analyze the generalization properties of neural networks, many works have adopted the analysis framework of Bartlett & Mendelson (2002) to bound the Rademacher complexity via Dudley's integral (Bartlett et al., 2017; Chen et al., 2019; Garg et al., 2020; Joukovsky et al., 2021). A key step in this analysis framework is deriving the robustness of the training loss to the small perturbation in model parameters θ. In our case, we can view $\texttt{PLISA}_\theta$ as an iterative algorithm and therefore borrow the analysis tools of classic optimization algorithms to derive its robustness in θ. The following lemma states this key intermediate result, which clearly connects the Lipschitz constant to algorithmic properties of $\texttt{PLISA}_\theta$.

Lemma 4.1 (Robustness to θ)**.** *Assume $\mathcal{P}$ satisfies Assumption 3.1 and $\mathcal{D}_m\sim\mathbb{P}(\mathcal{P})^m$. Assume $\texttt{PLISA}_\theta$ contains $T>t_0$ cells and K layers. Consider a parameter space Θ in which the parameters satisfy (i) $\alpha\in[\alpha_{\min},\frac{1}{\rho_+}]$, (ii) $\eta_j\in[\sigma^{-1}(0.9),\eta_{\max}]$, (iii) $\widetilde{w_2}\frac{1}{b}+\widetilde{w_3}\frac{1}{a-1}\le\xi_{\max}<\rho_-$, and (iv) $\lambda_j^*\in[8\sup_{(Z_{1:n},\boldsymbol{\beta}^*)\in\mathcal{D}_m}|[\nabla_{\boldsymbol{\beta}}L_n(Z_{1:n},\boldsymbol{\beta}^*)]_j|\vee\varepsilon,\lambda_{\max}]$ with some positive constants $\alpha_{\min}$, $\eta_{\max}$, $\xi_{\max}$, and $\lambda_{\max}$. Then for any $\theta=\{\boldsymbol{\eta},\boldsymbol{\lambda}^*,\boldsymbol{w},\alpha\}$ and $\theta'=\{\boldsymbol{\eta}',\boldsymbol{\lambda}^{*\prime},\boldsymbol{w}',\alpha'\}$ in Θ, and for*

any recovery problem $(Z_{1:n}, \boldsymbol{\beta}^*) \in \mathcal{D}_m$*, the following inequality holds,*

$$\|\boldsymbol{\beta}_T(Z_{1:n};\theta) - \boldsymbol{\beta}_T(Z_{1:n};\theta')\|_2 \leq c_1 K(T-t_0)\sqrt{s^*}\,|\alpha-\alpha'|\underbrace{\exp(-C_\Theta K(T-t_0))}_{\text{convergence rate}} \tag{12}$$

$$+\left(c_2\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + c_3\|\boldsymbol{\lambda}^*-\boldsymbol{\lambda}^{*\prime}\|_2 + c_4\sqrt{d}\|\boldsymbol{w}-\boldsymbol{w}'\|_2\right)\underbrace{(1-\exp(-C_\Theta KT))}_{\text{stability rate}}, \tag{13}$$

where c_1, c_2, c_3, c_4 *and* C_Θ *are some positive constants. Note that similar to Theorem 3.1,* K *and* t_0 *are required to be larger than certain small values. See Appendix E.1 for the proof.*

Convergence rate & step size perturbation. As indicated in Eq. 12, the Lipschitz constant in the step size α scales at the same rate as the convergence rate of $\texttt{PLISA}_\theta$, decreasing exponentially in T and K (See Fig. 3 for a visualization). To understand this, consider when both step sizes α and α' are within the convergence region (i.e., $(0, \rho_+^{-1}]$). After infinitely many steps, their induced outputs will both converge to the same optimal point. This intuitively explains why the output perturbation caused by α-perturbation has the same decrease rate as the optimization error. In the proof, we exploit the techniques for analyzing algorithm convergence to obtain this result.

Stability rate & regularization perturbation. In the literature of optimization, stability of an algorithm expresses its robustness to small perturbation in the optimization objective. This is clearly related to the robustness of $\texttt{PLISA}_\theta$ to the perturbation in $\boldsymbol{\eta}, \boldsymbol{\lambda}^*, \boldsymbol{w}$, because these parameters jointly play the role of learning the regularization $P_{\boldsymbol{w}}(\boldsymbol{\lambda}_t, \boldsymbol{\beta})$, which is a part of the optimization objective in Eq. 1. Therefore, we exploit the analysis techniques for algorithmic stability to derive the robustness in $(\boldsymbol{\eta}, \boldsymbol{\lambda}^*, \boldsymbol{w})$-perturbation and obtain the Lipschitz constant in Eq. 13, which is bounded but increasing in T and K (See Fig. 3 for a visualization).

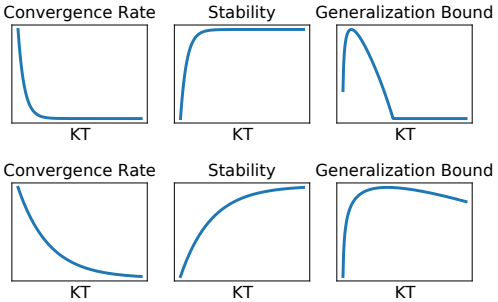

Figure 3: Visualization of convergence, stability, and generalization bound in Theorem 4.1. The two sets of visualizations are obtained by choosing different speeds C_Θ in the convergence rate and stability.

Based on the key result in Lemma 4.1, we can apply Dudley's integral to measure the empirical Rademacher complexity which immediately yields the following generalization bound.

Theorem 4.1 (Generalization gap)**.** *Assume the assumptions in Lemma 4.1. For any* $\epsilon > 0$*, with probability at least* $1-\epsilon$*, the generalization gap is bounded by*

$$\mathcal{L}_{gen}(\mathbb{P}(\mathcal{P});\theta) - \mathcal{L}_{train}^{\gamma=0}(\mathcal{D}_m;\theta) \leq c_1\sqrt{m^{-1}\log(4\epsilon^{-1})} + \tag{14}$$

$$\sqrt{c_2 m^{-1}\log\left(\sqrt{m}KT\underbrace{\exp(-C_\Theta K(T-t_0))}_{\text{convergence rate}} \vee 1\right) + c_3 d m^{-1}\log\left(\sqrt{m}\underbrace{(1-\exp(-C_\Theta KT))}_{\text{stability}}\right)},$$

where c_1, c_2, c_3, C_Θ *are constants independent of* d, m, K *and* T*. See Appendix E for the proof.*

Fig. 3 visualizes how the generalization bound in Theorem 4.1 grows when KT increases. The two sets of plots look slightly different by picking different constants C_Θ. We have also tried varying the values of c_2, c_3, d, m in Theorem 4.1. Overall, they will lead to the two types of behaviors in Fig. 3. It will increase first and then decrease to a constant, but the speed can vary.

An important observation in Theorem 4.1 and Figure 3 is that the generalization gap could *decrease in the number of layers*, and this matches the empirical observations of PLISA as we reported in Section 6. It also distinguishes algorithm-unrolling based architectures from conventional neural networks, whose generalization gaps rarely decrease in the number of layers.

Remark. The above generalization results are conducted on a constrained parameter space (as described in Lemma 4.1) so that we can utilize the algorithmic properties of $\texttt{PLISA}_\theta$. We focus on this space because the analysis contains more interesting and new ingredients. For parameters outside this space, the analysis procedure is similar to other conventional recurrent networks. Since the bound in Theorem 4.1 has matched the empirical observations, it is reasonable to believe that after training, the learned parameters are likely to be in this 'nice' constrained space.

5 RELATED WORK

Learning-to-learn and deep unrolling has become an active research direction in recent years (Franceschi et al., 2017; Niculae et al., 2018; Denevi et al., 2018; Pogančić et al., 2019; Liu et al., 2019b; Berthet et al., 2020). Many works share the idea of unrolling an iterative algorithm to design the architecture (Yang et al., 2017; Borgerding et al., 2017; Corbineau et al., 2019; Xie et al., 2019; Shrivastava et al., 2020; Chen et al., 2020a; Wei et al., 2020). A well-known example is LISTA (Gregor & LeCun, 2010) which interprets ISTA (Daubechies et al., 2004) as layers of neural networks and has since then been an active research topic (Zhang & Ghanem, 2018; Kamilov & Mansour, 2016; Chen et al., 2018; Liu et al., 2019a; Wu et al., 2020; Kim & Park, 2020).

However, not many works have been conducted on the theoretical understanding of algorithm learning, especially the generalization ability. The only exceptions are several recent works of Chen et al. (2020b); Wang et al. (2021); Behboodi et al. (2020); Joukovsky et al. (2021). However, Chen et al. (2020b) and Wang et al. (2021) only consider learning to optimize quadratic losses. Behboodi et al. (2020); Joukovsky et al. (2021) do not connect the generalization analysis with algorithmic properties to provide tighter bounds as in our work.

We will refer the audience to Shlezinger et al. (2020); Chen et al. (2021) for a more comprehensive summary of related works in this area.

6 EXPERIMENTS

In this section, we report the experimental results to validate the effectiveness of PLISA and our theoretical results. We create the synthetic data by sampling a set of linear sparse recovery problems $\{((X^{(i)}, Y^{(i)}), \boldsymbol{\beta}^*(i),)\}$. In each problem, the design matrix $X^{(i)} \in \mathbb{R}^{n\times d}$ contains $n = 64$ independent realizations of a random vector $\mathbf{x} \in \mathbb{R}^d$ with $d = 256, 1024$ in easy or difficult setting, respectively. $\mathbf{x}$ follows a zero mean Gaussian distribution with covariance matrix $(\Sigma)_{i,j} = 0.9 \cdot 1_{\{i\neq j\}} + 1 \cdot 1_{\{i=j\}}$. The true parameter vector $\boldsymbol{\beta}^{*(i)}$ has a sparsity $\|\boldsymbol{\beta}^{*(i)}\|_0 = 16$ and its nonzero entries take values uniformly sampled from $(-2, -\frac{1}{2}) \cup (\frac{1}{2}, 2)$. The support set of each $\boldsymbol{\beta}^{*(i)}$ is independently sampled from a union support set D, with $|D| = 128$ to allow some similarity among the problems. The observation $Y^{(i)}$ is sampled such that $Y^{(i)} - X^{(i)}\boldsymbol{\beta}^{*(i)}$ is a n-dimensional Gaussian random vector with zero mean and covariance matrix $\mathbf{I}_n$. In all the experiments, $m = 2000$ such problems are used for training.

Recovery performance. In the first experiment, we compare PLISA with alternative methods including APF (Wang et al., 2014), ALISTA (Liu et al., 2019a), RNN (Andrychowicz et al., 2016), and RNN-ℓ_1. APF is the classic approximate path-following algorithm that is used as the basis of our architecture. ALISTA is a representative of the algorithm unrolling based deep architecture, which is an advanced variant of LISTA. We have tried the vanilla LISTA, but it performs worse than ALISTA on our tasks so it is not reported. RNN refers to the LSTM-based model in Andrychowicz et al. (2016) which is applied to learning-to-learn. Besides, we add a soft-thresholding operator to this model to enforce sparsity, and include this variant as a baseline, called RNN-ℓ_1, which reveals to perform better than RNN. Except for APF, all methods are trained on the same set of training problems and selected by the validation problems. For APF, we perform grid-search to choose its hyperparameters, which is also selected by the validation set. The detailed specification of each model can be found in Appendix G.

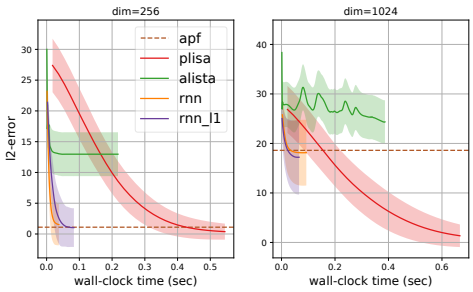

Figure 4: Convergence of recovery error. Since APF takes a long time to converge, its curve are outside the range of these plots. We use a dash-line to represent the final ℓ_2 error it achieves.

The comparison results are reported in Fig. 4, which shows the convergence of the recovery ℓ_2 error, $\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2$, in which the x-axis indicates the wall-clock time. It can be seen that in terms of the final recovery accuracy, PLISA outperforms all baseline methods. In the more difficult setting (i.e, when $d = 1024$), its advantage is very obvious. Although PLISA is slightly slower than other

deep learning based models due to the computations of MCP and SCAD, PLISA achieves a better accuracy and it has been converging much faster than the classic algorithm APF.

Table 1: Ablation study of PLISA ($p = 1024$). TPR is the true positive rate of recovering the nonzero entries of $\boldsymbol{\beta}^*$. FPS is the cardinality of false positive entries. Note that the true sparsity level is $s^* = 16$. Standard deviations over 100 test problems are present in the parantheses.

	PLISA	PLISA-single	PLISA-ℓ_1
ℓ_2 error	1.34 (2.28)	18.25 (6.06)	2.20 (2.76)
TPR	0.99 (0.01)	0.62 (0.19)	0.99 (0.02)
FPS	16.65 (13.60)	51.07 (6.67)	25.11 (13.30)

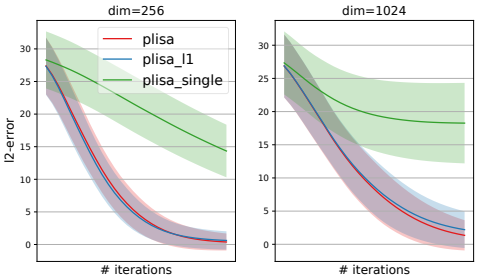

Figure 5: Ablation study.

Ablation study. In the second experiment, we perform an ablation study to verify the effectiveness of the designs in PLISA. We consider two variants of PLISA. One is PLISA-single which employs a single regularization parameter across different entries, i.e., $\eta_1 = \cdots = \eta_d$ and $\lambda_1^* = \cdots = \lambda_d^*$. The other is PLISA-$\ell_1$ which does not learn the penalty function but uses the ℓ_1 norm as the penalty function, i.e., $P_{\boldsymbol{w}}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1$. As reported in Fig. 5 and Table 1, the vanilla PLISA performs better than the alternatives. Especially, it has a much better accuracy than PLISA-single which does not use entry-wise regularization parameters. Therefore, this ablation study has validated the effectiveness of using entry-wise regularization parameters and learning the penalty function.

Generalization gap. We are interested in seeing what is the generalization behavior of PLISA experimentally. As this experiment is conducted for theoretical interest, we do not use the validation set to select the model. We vary the number of layers (K) and cells (T) in PLISA to create a set of models with different depths. For each depth, we train the model with 2000 training problems, and then test it on a separate set of 100 problems to approximate the generalization gap. In Fig. 6, we observe the interesting behavior of the generalization gap, where the left one increases in KT at the beginning and then decrease to a constant, and the right one increases fast and then decrease very slowly. This surprisingly matches the two different visualizations in Fig. 3 of the predicted bound given by Theorem 4.1.

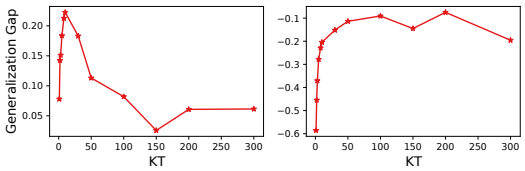

Figure 6: Generalization gap of PLISA with varying KT, for two different experimental settings.

7 Discussion

We proposed an algorithm-unrolling based architecture called PLISA for learning to solve sparse parameter recovery problems. We presented a theoretical analysis for both the capacity and generalization ability of PLISA. The results can shed new light on the relation between generalization and the algorithmic properties, and the techniques could be used to derive guarantees for other algorithm-unrolling based architectures. The current work can be extended in the future from several aspects. The model PLISA can be improved, for example, by using a more flexible penalty function (e.g., a conventional neural network) as long as it satisfies Assumption B.1. Theoretically, it will be interesting to extend the current result to the unsupervised learning scenario when the true parameters are unobserved.

REFERENCES

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, pp. 3981–3989, 2016.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pp. 6240–6249, 2017.

Arash Behboodi, Holger Rauhut, and Ekkehard Schnoor. Compressive sensing and neural networks from a statistical learning perspective. *arXiv preprint arXiv:2010.15658*, 2020.

Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable perturbed optimizers. *arXiv preprint arXiv:2002.08676*, 2020.

Mark Borgerding, Philip Schniter, and Sundeep Rangan. Amp-inspired deep networks for sparse linear inverse problems. *IEEE Transactions on Signal Processing*, 65(16):4293–4308, 2017.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *arXiv preprint arXiv:1405.4980*, 2014.

Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

Minshuo Chen, Xingguo Li, and Tuo Zhao. On generalization bounds of a family of recurrent neural networks. *arXiv preprint arXiv:1910.12947*, 2019.

Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to optimize: A primer and a benchmark. *arXiv preprint arXiv:2103.12828*, 2021.

Xiaohan Chen, Jialin Liu, Zhangyang Wang, and Wotao Yin. Theoretical linear convergence of unfolded ista and its practical weights and thresholds. In *Advances in Neural Information Processing Systems*, pp. 9061–9071, 2018.

Xinshi Chen, Yu Li, Ramzan Umarov, Xin Gao, and Le Song. Rna secondary structure prediction by learning unrolled algorithms. *arXiv preprint arXiv:2002.05810*, 2020a.

Xinshi Chen, Yufei Zhang, Christoph Reisinger, and Le Song. Understanding deep architecture with reasoning layer. *Advances in Neural Information Processing Systems*, 33, 2020b.

M-C Corbineau, Carla Bertocchi, Emilie Chouzenoux, Marco Prato, and J-C Pesquet. Learned image deblurring by unfolding a proximal interior point algorithm. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4664–4668. IEEE, 2019.

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

Giulia Denevi, Carlo Ciliberto, Dimitris Stamos, and Massimiliano Pontil. Learning to learn around a common mean. *Advances in Neural Information Processing Systems*, 31:10169–10179, 2018.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan, Yang Feng, and Yichao Wu. Network exploration via the adaptive lasso and scad penalties. *The annals of applied statistics*, 3(2):521, 2009.

Luca Franceschi, Paolo Frasconi, Michele Donini, and Massimiliano Pontil. A bridge between hyperparameter optimization and larning-to-learn. *stat*, 1050:18, 2017.

Vikas K Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. *arXiv preprint arXiv:2002.06157*, 2020.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 399–406. Omnipress, 2010.

Boris Joseph Joukovsky, Tanmoy Mukherjee, Nikos Deligiannis, et al. Generalization error bounds for deep unfolding rnns. In *Proceedings of Machine Learning Research*. Journal of Machine Learning Research, 2021.

Ulugbek S Kamilov and Hassan Mansour. Learning optimal nonlinearities for iterative thresholding algorithms. *IEEE Signal Processing Letters*, 23(5):747–751, 2016.

Dohyun Kim and Daeyoung Park. Element-wise adaptive thresholds for learned iterative shrinkage thresholding algorithms. *IEEE Access*, 8:45874–45886, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jialin Liu, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. ALISTA: Analytic weights are as good as learned weights in LISTA. In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=B1lnzn0ctQ.

Risheng Liu, Shichao Cheng, Yi He, Xin Fan, Zhouchen Lin, and Zhongxuan Luo. On the convergence of learning-based iterative methods for nonconvex inverse problems. *IEEE transactions on pattern analysis and machine intelligence*, 42(12):3027–3039, 2019b.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.

Vlad Niculae, Andre Martins, Mathieu Blondel, and Claire Cardie. Sparsemap: Differentiable sparse structured inference. In *International Conference on Machine Learning*, pp. 3799–3808, 2018.

Marin Vlastelica Pogančić, Anselm Paulus, Vit Musil, Georg Martius, and Michal Rolinek. Differentiation of blackbox combinatorial solvers. In *International Conference on Learning Representations*, 2019.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Nir Shlezinger, Jay Whang, Yonina C Eldar, and Alexandros G Dimakis. Model-based deep learning. *arXiv preprint arXiv:2012.08405*, 2020.

Harsh Shrivastava, Xinshi Chen, Binghong Chen, Guanghui Lan, Srinivas Aluru, Han Liu, and Le Song. GLAD: Learning sparse graph recovery. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BkxpMTEtPB.

Xiang Wang, Shuai Yuan, Chenwei Wu, and Rong Ge. Guarantees for tuning the step size using a learning-to-learn approach. In *International Conference on Machine Learning*, pp. 10981–10990. PMLR, 2021.

Zhaoran Wang, Han Liu, and Tong Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of statistics*, 42(6):2164, 2014.

Kaixuan Wei, Angelica Aviles-Rivero, Jingwei Liang, Ying Fu, Carola-Bibiane Schönlieb, and Hua Huang. Tuning-free plug-and-play proximal algorithm for inverse imaging problems. In *International Conference on Machine Learning*, pp. 10158–10169. PMLR, 2020.

Kailun Wu, Yiwen Guo, Ziang Li, and Changshui Zhang. Sparse coding with gated learned ista. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BygPO2VKPH.

Xingyu Xie, Jianlong Wu, Guangcan Liu, Zhisheng Zhong, and Zhouchen Lin. Differentiable linearized admm. In *International Conference on Machine Learning*, pp. 6902–6911. PMLR, 2019.

Y Yang, J Sun, H Li, and Z Xu. Admm-net: A deep learning approach for compressive sensing mri. corr. *arXiv preprint arXiv:1705.06869*, 2017.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942, 2010a.

Jian Zhang and Bernard Ghanem. Ista-net: Interpretable optimization-inspired deep network for image compressive sensing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1828–1837, 2018.

Tong Zhang. Analysis of multi-stage convex relaxation for sparse regularization. *Journal of Machine Learning Research*, 11(3), 2010b.

A LIST OF DEFINITIONS AND NOTATIONS

For the convenience of the reader, we summarize a list of notations blow.

1. $\|\boldsymbol{\beta}\|_{\boldsymbol{\lambda},1} := \sum_{j=1}^{d} \lambda_j |\beta_j|$ denotes element-wise ℓ_1 norm.
2. $(\boldsymbol{x})_S$ is the sub-vector of $\boldsymbol{x}$ with entries in the index set S.
3. $\boldsymbol{x} \vee \epsilon = \max\{\boldsymbol{x}, \epsilon\} = [\max\{x_1, \epsilon\}, \cdots, \max\{x_d, \epsilon\}]^\top$.
4. Element-wise soft-threshold:

$$\mathcal{T}_{s\boldsymbol{\lambda}}(\boldsymbol{\beta}) = (|\boldsymbol{\beta}| - s\boldsymbol{\lambda})_+ \circ \operatorname{sign}(\boldsymbol{\beta}) \tag{15}$$

$$= \begin{bmatrix} (|\beta_1| - s\lambda_1)_+ \operatorname{sign}(\beta_1) \\ \vdots \\ (|\beta_d| - s\lambda_d)_+ \operatorname{sign}(\beta_d) \end{bmatrix} \tag{16}$$

5. Denote a single iteration of the modified proximal gradient step as

$$\begin{aligned} \boldsymbol{\beta}^k &= \texttt{MPG}(\boldsymbol{\beta}^{k-1}; \boldsymbol{\lambda}, \boldsymbol{w}, \alpha) \\ &:= \mathcal{T}_{\alpha\cdot\boldsymbol{\lambda}}\left(\boldsymbol{\beta}^{k-1} - \alpha\left(\nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^{k-1})\right)\right) \end{aligned}$$

B PENALTY FUNCTIONS

We provide detailed descriptions of the penalty functions used in PLISA. We focus on learning the combination of three well-known penalty functions:

$$P^{(1)}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1, \;\; P^{(2)}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \sum_{j=1}^{p} \mathrm{MCP}(\lambda_j, \beta_j), \;\; P^{(3)}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \sum_{j=1}^{p} \mathrm{SCAD}(\lambda_j, \beta_j),$$

where MCP (Zhang, 2010a) and SCAD (Fan & Li, 2001) are nonconvex penalties whose analytical forms are given below.

B.1 ANALYTICAL FORMS

The MCP penalty can be written as

$$\mathrm{MCP}(\lambda, \beta) = \left(\lambda |\beta| - \frac{\beta^2}{2b}\right) \cdot \mathbb{1}\left(|\beta| \le b\lambda\right) + \frac{b\lambda^2}{2} \cdot \mathbb{1}\left(|\beta| > b\lambda\right),$$

where $b > 0$ is some hyperparameter.

The SCAD penalty can be written as

$$\begin{aligned} \mathrm{SCAD}(\lambda, \beta) = \lambda |\beta| \cdot \mathbb{1}\left(|\beta| \le \lambda\right) - \frac{\beta^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)} \cdot \mathbb{1}\left(\lambda < |\beta| \le a\lambda\right) \\ + \frac{(a+1)\lambda^2}{2} \cdot \mathbb{1}\left(|\beta| > a\lambda\right), \end{aligned}$$

where $a > 2$ is some hyperparameter.

PLISA uses the concave components of these penalty functions. Their analytical forms are as follows.

$$q^{(2)}(\lambda, \boldsymbol{\beta}) = \mathrm{MCP}(\lambda, \beta) - \lambda|\beta| = -\frac{\beta^2}{2b} \cdot \mathbb{1}\left(|\beta| \le b\lambda\right) + \left(\frac{b\lambda^2}{2} - \lambda|\beta|\right) \cdot \mathbb{1}\left(|\beta| > b\lambda\right)$$

$$\begin{aligned} q^{(3)}(\lambda, \boldsymbol{\beta}) = \mathrm{SCAD}(\lambda, \beta) - \lambda|\beta| = \frac{2\lambda|\beta| - \beta^2 - \lambda^2}{2(a-1)} \cdot \mathbb{1}\left(\lambda < |\beta| \le a\lambda\right) \\ + \frac{(a+1)\lambda^2 - 2\lambda|\beta|}{2} \cdot \mathbb{1}\left(|\beta| > a\lambda\right) \end{aligned}$$

B.2 IMPLICATIONS

Based on the definitions of these penalty functions, it is easy to check that their convex combination satisfies a set of conditions as stated in the following Assumption B.1.

Assumption B.1 (Regularization)**.** *Let $P(\boldsymbol{\lambda}, \boldsymbol{\beta}) = Q(\boldsymbol{\lambda}, \boldsymbol{\beta}) + \|\boldsymbol{\beta}\|_{\boldsymbol{\lambda},1}$ where Q is concave in $\boldsymbol{\beta}$, and that $Q(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \sum_{j=1}^{d} q(\lambda_j, \beta_j)$. Denote the partial gradient of q as $q'_\lambda(\beta) := \frac{\partial q(\lambda,\beta)}{\partial \beta}$. Assume the following regularization conditions.*

(a) There exists constants $\xi \geq 0$ such that for any $\beta' > \beta$,

$$-\xi \leq \frac{q'_\lambda(\beta') - q'_\lambda(\beta)}{\beta' - \beta} \leq 0.$$

(b) $q(\lambda, -\beta) = q(\lambda, \beta)$.

(c) $q'_\lambda(0) = q(\lambda, 0) = 0$.

(d) $|q'_\lambda(\beta)| \leq \lambda$.

(e) $|q'_{\lambda_1}(\beta) - q'_{\lambda_2}(\beta)| \leq |\lambda_1 - \lambda_2|$.

Note that the penalty function $P_{\boldsymbol{w}}(\boldsymbol{\lambda}, \boldsymbol{\beta})$ in `PLISA`$_\theta$ *satisfies all these conditions, with the constant ξ in condition (a) being*

$$\xi_{\boldsymbol{w}} = \widetilde{w_2}\frac{1}{b} + \widetilde{w_3}\frac{1}{a-1},$$

where a and b are the hyperparameters in SCAD *and* MCP.

Next, we present an important lemma which states the restricted strongly convexity and smoothness of the modified loss

$$L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}) := L_n(Z_{1:n}, \boldsymbol{\beta}) + Q_{\boldsymbol{w}}(\boldsymbol{\lambda}, \boldsymbol{\beta}).$$

Lemma B.1 (Restricted Strongly Convex and Restricted Strongly Smooth)**.** *Assume $\mathcal{P}$ satisfies Assumption 3.1 and $(Z_n, \boldsymbol{\beta}^*) \in \mathcal{P}$. If $Q(\boldsymbol{\lambda}, \boldsymbol{\beta})$ satisfies the regularization conditions in Assumption B.1 with a constant $\xi < \rho_-$. Then*

$$L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) := L_n(Z_{1:n}, \boldsymbol{\beta}) + Q(\boldsymbol{\lambda}, \boldsymbol{\beta})$$

is strongly convex and smooth in the subspace $\{\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_0 \leq s^* + 2\tilde{s}\}$. That is, for any $\boldsymbol{\beta}, \boldsymbol{\beta}'$ such that $\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_0 \leq s^* + 2\tilde{s}$, the following inequalities hold

$$L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}') \geq L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta})^\top(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{\rho_- - \xi}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2 \tag{17}$$

$$L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}') \leq L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta})^\top(\boldsymbol{\beta}' - \boldsymbol{\beta}) + \frac{\rho_+}{2}\|\boldsymbol{\beta}' - \boldsymbol{\beta}\|_2^2 \tag{18}$$

Proof. The proof is straightforward by using the condition (a) in Assumption B.1 and the sparse-eigenvalue condition in Assumption 3.1. □

C RECOVERY GUARANTEE FOR PARAMETERS IN CONSTRAINED SPACE

Our generalization analysis is conducted on a constrained parameter space. In this section, we provide theoretical guarantees for the estimation error of $\texttt{PLISA}_\theta$ when θ is in the constrained space. We first formally define the parameter space as follows.

Definition C.1 (Constrained parameter space)**.** *Assume $\mathcal{P}$ satisfies Assumption 3.1 and $\mathcal{D}_m \subseteq \mathcal{P}$. Given some positive values $\alpha_{\min}$, $\eta_{\max}$, $\xi_{\max}$, and $\lambda_{\max}$. We definite the constrained parameter space $\Theta(\mathcal{D}_m)$ as $\Theta(\mathcal{D}_m) := \{\theta = \{\boldsymbol{\eta}, \boldsymbol{\lambda}^*, \boldsymbol{w}, \alpha\} :$ conditions (i) (ii) (iii) (iv) are satisfied$\}$, where the conditions are:*

(i) $\alpha \in [\alpha_{\min}, \frac{1}{\rho_+}]$.

(ii) $\eta_j \in [\eta_{\min} := \sigma^{-1}(0.9), \eta_{\max}]$, *for all* $j = 1, \cdots, d$.

(iii) $\widetilde{w}_2 \frac{1}{b} + \widetilde{w}_3 \frac{1}{a-1} \leq \xi_{\max} < \rho_-$, *where b is the hyperparameter in* MCP *and a is the hyperparameter in* SCAD.

(iv) $\lambda_j^* \in [8 \sup_{(Z_{1:n}, \boldsymbol{\beta}^*) \in \mathcal{D}_m} |[\nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \boldsymbol{\beta}^*)]_j| \vee \varepsilon, \lambda_{\max}]$, *for all* $j = 1, \cdots, d$.

For parameters in this constrained space, we can show the performance guarantee for the outputs of $\texttt{PLISA}_\theta$, as stated in the following Theorem C.1.

Theorem C.1 (Recovery guarantee)**.** *Assume $\mathcal{P}$ satisfies Assumption 3.1 and $\mathcal{D}_m \subseteq \mathcal{P}$. Let $\Theta(\mathcal{D}_m)$ be the constrained parameter space defined in Definition C.1 and assume $\theta = \{\boldsymbol{\eta}, \boldsymbol{\lambda}^*, \boldsymbol{w}, \alpha\} \in \Theta(\mathcal{D}_m)$. Assume*

$$K \geq \log\left(4\sqrt{\frac{29\rho_+}{\rho_- - \xi_{\max}}}(1 + \frac{1}{\alpha_{\min}\rho_+})\frac{\|(B_2 \vee \boldsymbol{\lambda}^*)_{S^*}\|_2}{\varepsilon}\right) / \log\left(\delta^{-1/2}\right) + 1,$$

where $\delta := 1 - \alpha_{\min}(\rho_- - \xi_{\max})$. Then for any problem $(Z_{1:n}, \boldsymbol{\beta}^*)$ in the training set $\mathcal{D}_m$, all intermediate outputs of $\texttt{PLISA}_\theta$ have the following error bounds for all $t = 1, \cdots, T$.

$$\|\widetilde{\boldsymbol{\beta}}_t^k(Z_{1:n}; \theta) - \boldsymbol{\beta}^*\|_2 \leq \frac{15/2}{\rho_- - \xi_{\max}}\|(\boldsymbol{\lambda}_t(Z_{1:n}; \theta))_{S^*}\|_2, \quad \forall k = 1, \cdots, K,$$
$$\|\boldsymbol{\beta}_t(Z_{1:n}; \theta) - \boldsymbol{\beta}^*\|_2 \leq \frac{19/8}{\rho_- - \xi_{\max}}\|(\boldsymbol{\lambda}_t(Z_{1:n}; \theta))_{S^*}\|_2.$$

Proof. The proof of this theorem is based on a series of key lemmas for the properties of the modified proximal gradient steps in each cell of PLISA. We state these key lemmas and their proofs in Section F. With these lemmas, the proof of this theorem is straightforward, and we state the proof below.

Given a fixed problem $(Z_{1:n}, \boldsymbol{\beta}^*) \in \mathcal{D}_m$ and a fixed set of parameters $\theta = \{\boldsymbol{\eta}, \boldsymbol{\lambda}^*, \boldsymbol{w}, \alpha\} \in \Theta(\mathcal{D}_m)$, we simplify some notations in this proof by removing the dependency on $Z_{1:n}$ and θ. For example, we denote $\boldsymbol{\beta}_t = \boldsymbol{\beta}_t(Z_{1:n}; \theta)$, $\widetilde{\boldsymbol{\beta}}_t^k = \widetilde{\boldsymbol{\beta}}_t^k(Z_{1:n}; \theta)$, $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}_t(Z_{1:n}; \theta)$, etc.

Notations. Furthermore, we will use the following simplified notations and definitions.

$$\begin{aligned}
&\textit{empirical loss} && L_n(\boldsymbol{\beta}) := L_n(Z_{1:n}, \boldsymbol{\beta}) \\
&\textit{modified loss} && L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) := L_n(Z_{1:n}, \boldsymbol{\beta}) + Q_{\boldsymbol{w}}(\boldsymbol{\lambda}, \boldsymbol{\beta}) \\
&\textit{regularized loss} && \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) := L_n(Z_{1:n}, \boldsymbol{\beta}) + P_{\boldsymbol{w}}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \|\boldsymbol{\beta}\|_{\boldsymbol{\lambda},1} \\
&\textit{local optimal} && \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} \in \underset{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \leq \tilde{s}}{\arg\min} \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) \\
&\textit{sub-optimality} && \omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) := \min_{\boldsymbol{\xi}' \in \partial\|\boldsymbol{\beta}\|_1} \max_{\boldsymbol{\beta}'} \left\{\frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda},1}}(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}')\right\}.
\end{aligned}$$

Now we are ready to state the proof. We first show the following statement holds true for all $t \leq T$ by mathematical induction.

$$\textbf{Statement}(t)\textbf{:} \quad \omega_{\boldsymbol{\lambda}_t}(\widetilde{\boldsymbol{\beta}}_t^0) \leq \frac{1}{2}, \quad \text{and} \quad \|(\widetilde{\boldsymbol{\beta}}_t^k)_{\overline{S^*}}\|_0 \leq \tilde{s}, \quad \forall k = 1, \cdots, K.$$

Statement(1)**.** First we verify that Statement(1) holds true. Recall that $\boldsymbol{\lambda}_0 = |\nabla L_n(\mathbf{0})|$. In the following, we prove that $\widetilde{\boldsymbol{\beta}}_1^0$ is a local solution of $\phi_{\boldsymbol{\lambda}_0}(\boldsymbol{\beta}) := L_{\boldsymbol{\lambda}_0}(\boldsymbol{\beta}) + \|\boldsymbol{\beta}\|_{\boldsymbol{\lambda}_0,1}$.

$$\begin{aligned}
&\nabla L_{\boldsymbol{\lambda}_0}(\widetilde{\boldsymbol{\beta}}_1^0) + \boldsymbol{\lambda}_0 \circ \partial\|\widetilde{\boldsymbol{\beta}}_1^0\|_1 \\
&= \nabla L_n(\mathbf{0}) + \nabla_{\boldsymbol{\beta}} Q_{\boldsymbol{w}}(\boldsymbol{\lambda}_0, \mathbf{0}) + \boldsymbol{\lambda}_0 \circ \partial\|\mathbf{0}\|_1 \\
&= \operatorname{sign}(\nabla L_n(\mathbf{0})) \circ |\nabla L_n(\mathbf{0})| + \mathbf{0} + \boldsymbol{\lambda}_0 \circ \partial\|\mathbf{0}\|_1 \\
&= \operatorname{sign}(\nabla L_n(\mathbf{0})) \circ \boldsymbol{\lambda}_0 + \boldsymbol{\lambda}_0 \circ \partial\|\mathbf{0}\|_1.
\end{aligned}$$

Since $-\operatorname{sign}(\nabla L_n(\mathbf{0})) \in \partial\|\mathbf{0}\|_1$, then we have $\mathbf{0} \in \nabla L_{\boldsymbol{\lambda}_0}(\widetilde{\boldsymbol{\beta}}_1^0) + \boldsymbol{\lambda}_0 \circ \partial\|\widetilde{\boldsymbol{\beta}}_1^0\|_1$. Therefore, $\omega_{\boldsymbol{\lambda}_0}(\widetilde{\boldsymbol{\beta}}_1^0) = 0$. Since $\boldsymbol{\lambda}_1 = \boldsymbol{\lambda}_0 \circ \boldsymbol{\eta}_1$, by Lemma F.6, we have

$$\omega_{\boldsymbol{\lambda}_1}(\widetilde{\boldsymbol{\beta}}_1^0) \le \frac{\omega_{\boldsymbol{\lambda}_0}(\widetilde{\boldsymbol{\beta}}_1^0) + 0.2}{0.9} = \frac{2}{9} \le 1/2.$$

By Lemma F.5, we have $\phi_{\boldsymbol{\lambda}_1}(\widetilde{\boldsymbol{\beta}}_1^0) - \phi_{\boldsymbol{\lambda}_1}(\boldsymbol{\beta}^*) \le \frac{21/2\|(\boldsymbol{\lambda}_1)_{S^*}\|_2^2}{\rho_- - \xi_{\max}}$, which implies that the conditions in Lemma F.4 are satisfied. Therefore, we have proved that

$$\|(\boldsymbol{\beta}_1^k)_{\overline{S^*}}\|_0 \le \tilde{s} \quad \forall k = 1, \ldots, K.$$

This finishes the proof of Statement(1).

Statement(t)**.** Now we assume Statement$(t-1)$ is true and prove Statement(t). First, we prove that $\omega_{\boldsymbol{\lambda}_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}^K) \le 1/4$. By Lemma F.2 and Lemma F.5,

$$\begin{aligned}
\omega_{\boldsymbol{\lambda}_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}^K) &\le \frac{\sqrt{2}(1 + \frac{1}{\alpha_{\min}\rho_+})\sqrt{\rho_+}}{\min(\boldsymbol{\lambda}_{t-1})} \sqrt{(1 - \alpha_{\min}(\rho_- - \xi_{\max}))^{K-1} \frac{29/2\|(\boldsymbol{\lambda}_{t-1})_{S^*}\|_2^2}{\rho_- - \xi_{\max}}} \\
&\le \sqrt{\frac{29\rho_+}{\rho_- - \xi_{\max}}}(1 + \frac{1}{\alpha_{\min}\rho_+}) \frac{\|(\boldsymbol{\lambda}_0 \vee \boldsymbol{\lambda}^*)_{S^*}\|_2}{\varepsilon} \sqrt{(1 - \alpha_{\min}(\rho_- - \xi_{\max}))}^{K-1}.
\end{aligned}$$

Assume $K \ge \log\left(4\sqrt{\frac{29\rho_+}{\rho_- - \xi_{\max}}}(1 + \frac{1}{\alpha_{\min}\rho_+})\frac{\|(\boldsymbol{\lambda}_0 \vee \boldsymbol{\lambda}^*)_{S^*}\|_2}{\varepsilon}\right) / \log\left(\delta^{-1/2}\right) + 1$, where

$$\delta := 1 - \alpha_{\min}(\rho_- - \xi_{\max}).$$

Then we have

$$\omega_{\boldsymbol{\lambda}_{t-1}}(\widetilde{\boldsymbol{\beta}}_{t-1}^K) \le 1/4. \tag{19}$$

Since $\widetilde{\boldsymbol{\beta}}_t^0 = \widetilde{\boldsymbol{\beta}}_{t-1}^K$, Lemma F.6 implies that

$$\omega_{\boldsymbol{\lambda}_t}(\widetilde{\boldsymbol{\beta}}_t^0) \le 1/2.$$

Furthermore, by Lemma F.5, $\omega_{\boldsymbol{\lambda}_t}(\widetilde{\boldsymbol{\beta}}_t^0) \le 1/2$ implies

$$\phi_{\boldsymbol{\lambda}_t}(\widetilde{\boldsymbol{\beta}}_t^0) - \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}^*) \le \frac{21/2\|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi_{\max}}, \tag{20}$$

which in turns implies that the conditions in Lemma F.4 are satisfied. Therefore, we have

$$\|(\widetilde{\boldsymbol{\beta}}_t^k)_{\overline{S^*}}\|_0 \le \tilde{s} \quad \forall k = 1, \ldots, K,$$

which completes the proof of Statement(t).

Now we derive the error bounds. Similar to how we derive Eq. 19 and Eq. 20, it is easy to show that

$$\begin{aligned}
&\omega_{\boldsymbol{\lambda}_t}(\widetilde{\boldsymbol{\beta}}_t^K) \le 1/4, \quad \forall t = 1, \cdots, T, \\
&\text{and } \phi_{\boldsymbol{\lambda}_t}(\widetilde{\boldsymbol{\beta}}_t^0) - \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}^*) \le \frac{21/2\|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi_{\max}} \quad \forall t = 1, \cdots, T.
\end{aligned}$$

Combining $\phi_{\boldsymbol{\lambda}_t}(\widetilde{\boldsymbol{\beta}}_t^0) - \phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}^*) \le \frac{21/2\|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi_{\max}}$ with Lemma F.3, we have

$$\|\widetilde{\boldsymbol{\beta}}_t^k - \boldsymbol{\beta}^*\|_2 \le \frac{15/2\|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi_{\max}}, \quad \forall k = 1, \cdots, K. \tag{21}$$

Combining $\omega_{\boldsymbol{\lambda}_t}(\widetilde{\boldsymbol{\beta}}_t^K) \le 1/4$ with Lemma F.5, we have

$$\|\widetilde{\boldsymbol{\beta}}_t^K - \boldsymbol{\beta}^*\|_2 \le \left(\frac{1}{4} + \frac{17}{8}\right) \frac{\|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi_{\max}} = \frac{19/8\|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi_{\max}}.$$

□

D CAPACITY ANALYSIS: PROOF OF THEOREM 3.1

Proof. This theorem is a direct result of Theorem C.1, by taking a suitable set of parameters θ. Similar to the proof of Theorem C.1, some notations are simplified. Please refer to the proof of Theorem C.1 for detailed illustrations.

Denote the union support set of the true parameters $\boldsymbol{\beta}^*$ by $S_m^* := \{j : j \in \text{supp}(\boldsymbol{\beta}^*), (Z_{1:n}, \boldsymbol{\beta}^*) \in \mathcal{D}_m\}$. Then we specify a set of parameters $\theta = \{\boldsymbol{\eta}, \boldsymbol{\lambda}^*, \boldsymbol{w}, \alpha\}$ for $\texttt{PLISA}_\theta$ as follows.

- $\alpha = \frac{1}{\rho_+}$.
- $\boldsymbol{w}$: The weight $\boldsymbol{w}$ can take any values as long as it satisfies $\widetilde{w_2}\frac{1}{b} + \widetilde{w_3}\frac{1}{a-1} < \rho_-$. In particular, we denote this value by $\xi_{\boldsymbol{w}} := \widetilde{w_2}\frac{1}{b} + \widetilde{w_3}\frac{1}{a-1}$.
- $\boldsymbol{\lambda}^*$: For each $j \in S_m^*$, we take $\lambda_j^* := 8\max_{(Z_n,\boldsymbol{\beta}^*)\in\mathcal{D}_m} |\nabla L_n(Z_n, \boldsymbol{\beta}^*)_j| \vee \varepsilon$ as the target regularization parameter. For each $j \notin S_m^*$, we take $\lambda_j^* = B_2$, which is the upper bound of $\|\nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \mathbf{0})\|_\infty$ by Assumption 3.1.
- $\boldsymbol{\eta}$: For all j, take $\eta_j = \log 9$ so that the decrease ratio is $\sigma(\eta_j) = 0.9$.

In the following, we will show that, with this specification of the parameters, $\texttt{PLISA}_\theta$ can achieve the recovery accuracy stated in Theorem 3.1.

Let $t_0 = \min\{t : t \geq 0, \boldsymbol{\lambda}^* = \max\{0.9^t |\nabla_{\boldsymbol{\beta}} L_n(Z_{1:n}, \mathbf{0})|, \boldsymbol{\lambda}^*\}, Z_{1:n} \in \mathcal{D}_m, \boldsymbol{\lambda}^* \in \theta \in \Theta(\mathcal{D}_M)\}$ be the number of blocks after which $\boldsymbol{\lambda}_t(Z_{1:n}; \theta) = \boldsymbol{\lambda}^*$. We know t_0 is a small number because the decrease rate is linear. To be more clear, it is easy to show that $t_0 \leq \log\left(\frac{B_2}{\varepsilon}\right) / \log(10/9)$. Therefore, after t_0 cells, the regularization parameters do not change anymore. That is,

$$\boldsymbol{\lambda}_t(Z_{1:n}; \theta) = \boldsymbol{\lambda}^*, \quad \forall t \geq t_0.$$

Since the specified parameters satisfy the conditions in Theorem C.1, we can follow the same way as how we derive Eq. 19 to obtain that

$$\omega_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta}_{t-1}) \leq 1/4, \quad \forall t = 1, \cdots, T.$$

Since we have $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}^*$ for $t \geq t_0$. The last $(T - t_0)$ cells can be viewed as a single cell with $K(T - t_0)$ many steps. Therefore, we can apply Lemma F.2 and Lemma F.5 to obtain

$$\begin{aligned}
\omega_{\boldsymbol{\lambda}_t}(\widetilde{\boldsymbol{\beta}}_t^K) &\leq \frac{2\sqrt{2\rho_+}}{\epsilon}\sqrt{(1 - \alpha(\rho_- - \xi_{\boldsymbol{w}}))^{K(t-t_0)-1}\,\frac{29/2\|(\boldsymbol{\lambda}_{t-1})_{S^*}\|_2^2}{\rho_- - \xi_{\boldsymbol{w}}}} \\
&\leq \sqrt{\frac{29\rho_+}{\rho_- - \xi_{\boldsymbol{w}}}}(1 + \frac{1}{\alpha\rho_+})\frac{\|(\boldsymbol{\lambda}_0 \vee \boldsymbol{\lambda}^*)_{S^*}\|_2}{\varepsilon}\sqrt{(1 - \alpha(\rho_- - \xi_{\boldsymbol{w}}))}^{K(t-t_0)-1} \\
&\leq \sqrt{\frac{29\rho_+}{\rho_- - \xi_{\boldsymbol{w}}}}(1 + \frac{1}{\alpha\rho_+})\frac{\sqrt{s^*}B_2}{\varepsilon}\sqrt{(1 - \alpha(\rho_- - \xi_{\boldsymbol{w}}))}^{K(t-t_0)-1} \\
&= c_\theta\sqrt{s^*}\varepsilon^{-1}\exp(-C_\theta K(t - t_0)),
\end{aligned}$$

where $C_\theta = -\frac{1}{2}\log(1 - \alpha(\rho_- - \xi_{\boldsymbol{w}}))$, $c_\theta = \sqrt{\frac{29\rho_+}{\rho_- - \xi_{\boldsymbol{w}}}}(1 + \frac{1}{\alpha\rho_+})\frac{B_2}{1-\alpha(\rho_- - \xi_{\boldsymbol{w}})}$. By Lemma F.5,

$$\begin{aligned}
\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 &\leq \frac{\left(\omega_{\boldsymbol{\lambda}_t}(\widetilde{\boldsymbol{\beta}}_t^K) + \frac{17}{8}\right)\|\boldsymbol{\lambda}_{S^*}^*\|_2}{\rho_- - \xi_{\boldsymbol{w}}} \\
&\leq \frac{\|\boldsymbol{\lambda}_{S^*}^*\|_2}{\rho_- - \xi_{\boldsymbol{w}}}c_\theta\varepsilon^{-1}\sqrt{s^*}\exp(-C_\theta K(t - t_0)) + \frac{\frac{17}{8}\|\boldsymbol{\lambda}_{S^*}^*\|_2}{\rho_- - \xi_{\boldsymbol{w}}} \\
&\leq \tilde{c}_\theta\varepsilon^{-1}s^*\exp(-C_\theta K(t - t_0)) + c'_\theta\|\boldsymbol{\lambda}_{S^*}^*\|_2
\end{aligned} \tag{22}$$

where $\tilde{c}_\theta = \frac{\sqrt{s^*}B_2}{\rho_- - \xi_{\boldsymbol{w}}}c_\theta$ and $c'_\theta = \frac{17}{8(\rho_- - \xi_{\boldsymbol{w}})}$. What remains is bounding $\|\boldsymbol{\lambda}_{S^*}^*\|_2$. First, we define a condition number as

$$\kappa_m = \max_{j \in S_m^*}\left\{\frac{8\max_{(Z_n,\boldsymbol{\beta}^*)\in\mathcal{D}_m}|\nabla_{\boldsymbol{\beta}} L_n(Z_n, \boldsymbol{\beta}^*)_j|}{8\min_{(Z_n,\boldsymbol{\beta}^*)\in\mathcal{D}_m}|\nabla_{\boldsymbol{\beta}} L_n(Z_n, \boldsymbol{\beta}^*)_j| \vee \varepsilon}\right\}.$$

In fact, $\kappa_m \leq \frac{B_2}{\varepsilon}$. Then recall the specification of $\boldsymbol{\lambda}^*$ at the beginning of this proof. We have

$$\begin{aligned}\|\boldsymbol{\lambda}^*_{S^*}\|_2 &\leq \sum_{j \in S^*} 8\kappa_m \left|\nabla_{\boldsymbol{\beta}} L_n(Z_n, \boldsymbol{\beta}^*)_j\right| \vee \varepsilon \\ &\leq 8\kappa_m \|(\nabla_{\boldsymbol{\beta}} L_n(Z_n, \boldsymbol{\beta}^*) \vee \varepsilon)_{S^*}\|_2.\end{aligned}$$

Plugging this back to Eq. 22, we have

$$\|\boldsymbol{\beta}_t - \boldsymbol{\beta}^*\|_2 \leq \tilde{c}_\theta \varepsilon^{-1} s^* \exp(-C_\theta K(t - t_0)) + c'_\theta 8\kappa_m \|(\nabla_{\boldsymbol{\beta}} L_n(Z_n, \boldsymbol{\beta}^*) \vee \varepsilon)_{S^*}\|_2.$$

This is equivalent to the statement to be proved. □

E GENERALIZATION ANALYSIS: PROOF OF THEOREM 4.1

We first state a well-known inequality that bounds the generalization gap by *empirical Rademacher complexity*.

Theorem E.1 (Adapted from Theorem 26.5 (Shalev-Shwartz & Ben-David, 2014))**.** *Assume that for all $(Z_{1:n}, \boldsymbol{\beta}^*) \in \mathcal{P}$ and all $\theta \in \Theta$ we have that $\ell(\boldsymbol{\beta}_T(Z_{1:n};\theta), \boldsymbol{\beta}^*) \leq c$. Then with probability at least $1-\epsilon$, for all $\theta \in \Theta$,*

$$\mathcal{L}_{gen}(\mathbb{P}(\mathcal{P});\theta) - \mathcal{L}_{train}^{\gamma=0}(\mathcal{D}_m;\theta) \leq 2R_m(\ell_\Theta) + 4c\sqrt{\frac{2\log(4\epsilon^{-1})}{m}},$$

where $\ell_\Theta := \{(Z_{1:n}, \boldsymbol{\beta}^*) \mapsto \|\boldsymbol{\beta}_T(Z_{1:n};\theta) - \boldsymbol{\beta}^*\|_2^2 : \theta \in \Theta\}$ is the space of loss functions of our model $\texttt{PLISA}_\theta$, and $R_m(\ell_\Theta)$ is its empirical Rademacher complexity. It is defined as

$$R_m(\ell_\Theta) := \mathbb{E}_{\boldsymbol{\sigma}} \sup_{\theta\in\Theta} \tfrac{1}{m}\textstyle\sum_{i=1}^m \sigma_i \left\|\boldsymbol{\beta}_T(Z_{1:n_i}^{(i)};\theta) - \boldsymbol{\beta}^{*(i)}\right\|_2^2,$$

where $\{\sigma_i\}_{i=1}^m$ are m independent Rademacher random variables.

Therefore, it resorts to bound the empirical Rademacher complexity, which can be bounded via Dudley's integral. The following theorem states the bound for the empirical Rademacher complexity and its proof follows.

Theorem E.2 (Empirical Rademacher complexity bound)**.** *Assume the assumptions in Theorem 4.1. Let $\Theta = \Theta(\mathcal{D}_m)$ be the constrained space defined in Definition C.1. Let $\ell_\Theta := \{(Z_{1:n}, \boldsymbol{\beta}^*) \mapsto \|\boldsymbol{\beta}_T(Z_{1:n};\theta) - \boldsymbol{\beta}^*\|_2^2 : \theta \in \Theta\}$ be the space of loss functions of our model $\texttt{PLISA}_\theta$. Then the empirical Rademacher complexity is bounded by $R_m\ell_\Theta \leq$*

$$\frac{c_1}{\sqrt{m}}\sqrt{\log\left(c_2\sqrt{m}K(T-t_0)\exp(-C_\Theta K(T-t_0)) \vee 1\right) + c_3 d\log\left(c_4\sqrt{m}\left(1-\exp(-C_\Theta KT)\right)\right)}$$

where c_1, c_2, c_3, c_4 are some constants independent of d, m, K and T.

Proof. The classical Dudley's entropy integral bound gives us an upper bound for the empirical Rademacher complexity in terms of the covering number.

$$\begin{aligned}
R_m\ell_\Theta &\leq \inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{m}}\int_\alpha^{\|\ell_\Theta\|_{P_m,\infty}} \sqrt{\log\mathcal{N}\left(\epsilon, \ell_\Theta, L_2(P_m)\right)}\,d\epsilon\right) \\
&\leq \frac{4}{\sqrt{m}} + \frac{12}{\sqrt{m}}\int_{\frac{1}{\sqrt{m}}}^{\|\ell_\Theta\|_{P_m,\infty}} \sqrt{\log\mathcal{N}\left(\epsilon, \ell_\Theta, L_2(P_m)\right)}\,d\epsilon \\
&\leq \frac{4}{\sqrt{m}} + \frac{12\|\ell_\Theta\|_{P_m,\infty}}{\sqrt{m}}\sqrt{\log\mathcal{N}\left(\frac{1}{\sqrt{m}}, \ell_\Theta, L_2(P_m)\right)}.
\end{aligned}$$

The notation P_m means the empirical measure defined based on the samples in $\mathcal{D}_m$, and

$$\begin{aligned}
\|\ell_\Theta\|_{P_m,\infty} &:= \sup_{\theta\in\Theta} P_m(\ell_\theta) \\
&= \sup_{\theta\in\Theta}\frac{1}{m}\sum_{(Z_{1:n},\boldsymbol{\beta}^*)\in\mathcal{D}_m} \|\boldsymbol{\beta}_T(Z_{1:n};\theta) - \boldsymbol{\beta}^*\|_2^2 \\
\text{by Theorem C.1} \quad &\leq \sup_{\theta\in\Theta}\frac{1}{m}\sum_{(Z_{1:n},\boldsymbol{\beta}^*)\in\mathcal{D}_m} \frac{19/8}{\rho_- - \xi_{\max}}\|(\boldsymbol{\lambda}_t(Z_{1:n};\theta))_{S^*}\|_2 \\
&\leq \sup_{\theta\in\Theta}\frac{1}{m}\sum_{(Z_{1:n},\boldsymbol{\beta}^*)\in\mathcal{D}_m} \frac{19/8}{\rho_- - \xi_{\max}}\sqrt{s^*}\max\{B_2, \lambda_{\max}\} \leq C.
\end{aligned}$$

Here we use a constant C as the upper bound because we only care about the scales in d, m, K and T. Next, we bound the covering number $\mathcal{N}\left(\frac{1}{\sqrt{m}}, \ell_\Theta, L_2(P_m)\right)$. By the key result in Lemma 4.1, it

is easy to derive that

$$\begin{aligned}\|\ell_\theta - \ell_{\theta'}\|_{L_2(P_m)} &\le 2C\Big(c_1 K(T-t_0)\sqrt{s^*}\,|\alpha-\alpha'|\exp(-C_\Theta K(T-t_0)) \\ &\quad + \Big(c_2\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + c_3\|\boldsymbol{\lambda}^*-\boldsymbol{\lambda}^{*\prime}\|_2 + c_4\sqrt{d}\|\boldsymbol{w}-\boldsymbol{w}'\|_2\Big)\left(1-\exp(-C_\Theta KT)\right)\Big) \\ &\le cK(T-t_0)\,|\alpha-\alpha'|\exp(-C_\Theta K(T-t_0)) \\ &\quad + c\left(\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + \|\boldsymbol{\lambda}^*-\boldsymbol{\lambda}^{*\prime}\|_2 + \sqrt{d}\|\boldsymbol{w}-\boldsymbol{w}'\|_2\right)\left(1-\exp(-C_\Theta KT)\right)/\end{aligned}$$

From now on, we abuse the symbol c to generally represent some constant that does not depend on d, m, K, T.

$$\begin{aligned}\mathcal{N}\left(\epsilon, \ell_\Theta, L_2(P_m)\right) &\le \mathcal{N}\left(\frac{\epsilon}{cK(T-t_0)\exp(-C_\Theta K(T-t_0))}, \left[\alpha_{\min}, \frac{1}{\rho_+}\right], |\cdot|\right) \\ &\quad \times \mathcal{N}\left(\frac{\epsilon}{c\left(1-\exp(-C_\Theta KT)\right)}, [\eta_{\min}, \eta_{\max}]^d, \ell_2\right) \\ &\quad \times \mathcal{N}\left(\frac{\epsilon}{c\left(1-\exp(-C_\Theta KT)\right)}, \prod_{j=1}^d \times[\lambda_{j,\min}, \lambda_{j,\max}], \ell_2\right) \\ &\quad \times \mathcal{N}\left(\frac{\epsilon}{c\sqrt{d}\left(1-\exp(-C_\Theta KT)\right)}, [w_{\min}, w_{\max}]^3, \ell_2\right).\end{aligned}$$

For the covering number of a d-dimensional compact vector space $T \subseteq \mathbb{R}^d$, we can bound it by the ϵ-packing number. That is, $\mathcal{N}(T, \epsilon, \ell_2) \le \mathcal{M}(T, \epsilon, \ell_2) \le c\prod_{j=1}^d \left(\frac{T_{j,\max}-T_{j,\min}}{\epsilon}+1\right)$ for some constant c. Applying this fact, we have

$$\begin{aligned}\mathcal{N}\left(\epsilon, \ell_\Theta, L_2(P_m)\right) \le &\left(\frac{cK(T-t_0)\exp(-C_\Theta K(T-t_0))}{\epsilon} \vee 1\right) \\ &\times \prod_{j=1}^d \left(\frac{c\left(1-\exp(-C_\Theta KT)\right)}{\epsilon}+1\right) \\ &\times \prod_{j=1}^d \left(\frac{c\left(1-\exp(-C_\Theta KT)\right)}{\epsilon}+1\right) \\ &\times \left(\frac{3c\sqrt{d}\left(1-\exp(-C_\Theta KT)\right)}{\epsilon}\right)^3\end{aligned}$$

which implies

$$\begin{aligned}&\log \mathcal{N}\left(\frac{1}{\sqrt{m}}, \ell_\Theta, L_2(P_m)\right) \\ &\le \log\left(c\sqrt{m}K(T-t_0)\exp(-C_\Theta K(T-t_0)) \vee 1\right) \\ &+ (2d)\log\left(c\sqrt{m}\left(1-\exp(-C_\Theta KT)\right)+1\right) + 3\log\left(c\sqrt{dm}\left(1-\exp(-C_\Theta KT)\right)\right) \\ &\le \log\left(c\sqrt{m}K(T-t_0)\exp(-C_\Theta K(T-t_0)) \vee 1\right) + cd\log\left(c\sqrt{m}\left(1-\exp(-C_\Theta KT)\right)+1\right).\end{aligned}$$

Therefore, $R_m \ell_\Theta \le$

$$\frac{c}{\sqrt{m}}\sqrt{\log\left(c\sqrt{m}K(T-t_0)\exp(-C_\Theta K(T-t_0)) \vee 1\right) + cd\log\left(c\sqrt{m}\left(1-\exp(-C_\Theta KT)\right)\right)}.$$

□

E.1 Robustness to Parameter Perturbation: Proof of Lemma 4.1

We split the parameters $\theta = \{\boldsymbol{\lambda}^*, \boldsymbol{w}, \boldsymbol{\eta}, \alpha\}$ into two sets $\theta = \theta_1 \cup \theta_2$, where $\theta_1 = \{\alpha\}$ and $\theta_2 = \{\boldsymbol{\lambda}^*, \boldsymbol{w}, \boldsymbol{\eta}\}$. Clearly, the robustness can be bounded by

$$\|\boldsymbol{\beta}_T(Z_{1:n}; \theta_1 \cup \theta_2) - \boldsymbol{\beta}_T(Z_{1:n}; \theta_1' \cup \theta_2)\|_2 + \|\boldsymbol{\beta}_T(Z_{1:n}; \theta_1' \cup \theta_2) - \boldsymbol{\beta}_T(Z_{1:n}; \theta_1' \cup \theta_2')\|_2.$$

We derive the upper bounds of these two terms in Lemma E.1 and Lemma E.2, which imply

$$\begin{aligned}&\|\boldsymbol{\beta}_T(Z_{1:n}; \theta) - \boldsymbol{\beta}_T(Z_{1:n}; \theta')\|_2 \leq C\sqrt{s^*}K\left(T - t_0\right)\delta^{K(T-t_0)}\left|\alpha - \alpha'\right| \\ &+ \left(C_1\|\boldsymbol{\eta} - \boldsymbol{\eta}'\|_2 + C_2\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^{*\prime}\|_2 + C_4\sqrt{d}\|\boldsymbol{w} - \boldsymbol{w}'\|_2\right)\frac{1 - \delta^{KT}}{1 - \delta},\end{aligned}$$

where $\delta := \sqrt{1 - \alpha_{\min}(\rho_- - \xi_{\max})} < 1$. Lemma 4.1 is equivalent to this inequality by taking $C_\Theta = -\log(\delta) > 0$. Therefore, the key derivation steps are in the proof of Lemma E.1 and Lemma E.2, as stated below.

Lemma E.1. *Assume the assumptions in Lemma 4.1 are satisfied. Let*

$$\theta = \{\boldsymbol{\lambda}^*, \boldsymbol{w}, \boldsymbol{\eta}, \alpha\} \quad \textit{and} \quad \theta' = \{\boldsymbol{\lambda}^*, \boldsymbol{w}, \boldsymbol{\eta}, \alpha'\}$$

be parameters in the constrained space Θ. Let $\{\boldsymbol{\beta}_t\}_{t=1}^T = \texttt{PLISA}_\theta(Z_{1:n})$ and $\{\boldsymbol{\beta}_t{}'\}_{t=1}^T = \texttt{PLISA}_{\theta'}(Z_{1:n})$ be the intermediate outputs of $\texttt{PLISA}$ with different parameters θ and θ'. Assume

$$K \geq \log\left(4\sqrt{\frac{29\rho_+}{\rho_- - \xi_{\max}}}(1 + \frac{1}{\alpha_{\min}\rho_+})\frac{\sqrt{s^*}(B_2 + \lambda_{\max})}{\varepsilon}\right) / \log\left(\delta^{-1}\right) + 1,$$

where $\delta := \sqrt{1 - \alpha_{\min}(\rho_- - \xi_{\max})} < 1$. Then

$$\|\boldsymbol{\beta}_T - \boldsymbol{\beta}_T{}'\|_2 \leq C\sqrt{s^*}K\left(T - t_0\right)\delta^{K(T-t_0)}\left|\alpha - \alpha'\right|.$$

Proof. By triangular inequality,

$$\begin{aligned}&\left\|\widetilde{\boldsymbol{\beta}}_t^k - \widetilde{\boldsymbol{\beta}}_t^{k\prime}\right\|_2 = \left\|\texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1}; \boldsymbol{\lambda}_t, \boldsymbol{w}, \alpha) - \texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1\prime}; \boldsymbol{\lambda}_t, \boldsymbol{w}, \alpha')\right\|_2 \\ \leq &\left\|\texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1}; \boldsymbol{\lambda}_t, \boldsymbol{w}, \alpha) - \texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1}; \boldsymbol{\lambda}_t, \boldsymbol{w}, \alpha')\right\|_2 \\ &+ \left\|\texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1}; \boldsymbol{\lambda}_t, \boldsymbol{w}, \alpha') - \texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1\prime}; \boldsymbol{\lambda}_t, \boldsymbol{w}, \alpha')\right\|_2\end{aligned}$$

WLOG, assume $\alpha > \alpha'$. Applying Lemma E.3 and Lemma E.4 to the above two terms on the right hand side, we obtain

$$\begin{aligned}\left\|\widetilde{\boldsymbol{\beta}}_t^k - \widetilde{\boldsymbol{\beta}}_t^{k\prime}\right\|_2 &\leq \delta\|\widetilde{\boldsymbol{\beta}}_t^{k-1} - \widetilde{\boldsymbol{\beta}}_t^{k-1\prime}\|_2 + \sqrt{\frac{|\alpha - \alpha'|}{\alpha + \alpha'}}\left\|\widetilde{\boldsymbol{\beta}}_t^k - \widetilde{\boldsymbol{\beta}}_t^{k-1}\right\|_2 \\ &\leq \delta\|\widetilde{\boldsymbol{\beta}}_t^{k-1} - \widetilde{\boldsymbol{\beta}}_t^{k-1\prime}\|_2 + c\left|\alpha - \alpha'\right|\left\|\widetilde{\boldsymbol{\beta}}_t^k - \widetilde{\boldsymbol{\beta}}_t^{k-1}\right\|_2\end{aligned} \tag{23}$$

where c is the Lipschtiz constant. c is a finite number because α is bounded and positive. Applying it recursively, we have

$$\left\|\widetilde{\boldsymbol{\beta}}_t^K - \widetilde{\boldsymbol{\beta}}_t^{K\prime}\right\|_2 \leq \delta^K\|\widetilde{\boldsymbol{\beta}}_t^0 - \widetilde{\boldsymbol{\beta}}_t^{0\prime}\|_2 + c\left|\alpha - \alpha'\right|\sum_{k=1}^K \delta^{K-k}\left\|\widetilde{\boldsymbol{\beta}}_t^k - \widetilde{\boldsymbol{\beta}}_t^{k-1}\right\|_2. \tag{24}$$

Now we bound the term $\left\|\widetilde{\boldsymbol{\beta}}_t^k - \widetilde{\boldsymbol{\beta}}_t^{k-1}\right\|_2$. By Lemma F.1 and Lemma F.2,

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\beta}}_t^k - \widetilde{\boldsymbol{\beta}}_t^{k-1}\right\|_2 &\leq \sqrt{\frac{2\alpha_{\min}}{2-\alpha_{\min}\rho_+}\left(\phi_{\boldsymbol{\lambda}_t,\boldsymbol{w}}(\widetilde{\boldsymbol{\beta}}_t^{k-1}) - \phi_{\boldsymbol{\lambda}_t,\boldsymbol{w}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t})\right)} \\
&\leq \sqrt{\frac{2}{\rho_+}\left(\phi_{\boldsymbol{\lambda}_t,\boldsymbol{w}}(\widetilde{\boldsymbol{\beta}}_t^{0}) - \phi_{\boldsymbol{\lambda}_t,\boldsymbol{w}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t})\right)}\delta^{k-1} \\
\text{(By Lemma F.6)} \quad &\leq \sqrt{\frac{2}{\rho_+}}\sqrt{\frac{10}{\rho_- - \xi_{\max}}}\|(\boldsymbol{\lambda}_t)_{S^*}\|_2\delta^{k-1} \\
&\leq \sqrt{\frac{2}{\rho_+}}\sqrt{\frac{10}{\rho_- - \xi_{\max}}}B_2\sqrt{s^*}\delta^{k-1} \\
&= c'\sqrt{s^*}\delta^k,
\end{aligned}$$

where $c' := \sqrt{\frac{2}{\rho_+}}\sqrt{\frac{10}{\rho_- - \xi_{\max}}}B_2\delta^{-1}$. However, we can have a tighter bound for this term if t is larger than a certain integer. More precisely, recall the definition that $\boldsymbol{\lambda}_t = \max\{\boldsymbol{\lambda}_0 \circ \sigma(\boldsymbol{\eta})^t, \boldsymbol{\lambda}^*\}$. Since $\sigma(\eta_{\max}) < 1$, the entries of $\boldsymbol{\lambda}_0 \circ \sigma(\boldsymbol{\eta})^t$ will decrease exponentially to reach the entry values of $\boldsymbol{\lambda}^*$. Let $t_0 = \min\{t : t \geq 0, \boldsymbol{\lambda}^* = \max\{\boldsymbol{\lambda}_0 \circ \sigma(\eta_{\max})^t, \boldsymbol{\lambda}^*\}, \forall \boldsymbol{\lambda}^* \in \theta \in \Theta(\mathcal{D}_M)\}$ be the number of blocks after which $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}^*$. We know t_0 is a small number because the decrease rate is linear.

Now, the regularization parameters satisfy $\boldsymbol{\lambda}_t = \boldsymbol{\lambda}^*$, $\forall t \geq t_0$. Therefore, we can apply Lemma F.1 and Lemma F.2 again to obtain the following bound for all $t \geq t_0$:

$$\begin{aligned}
\left\|\widetilde{\boldsymbol{\beta}}_t^k - \widetilde{\boldsymbol{\beta}}_t^{k-1}\right\|_2 &\leq \sqrt{\frac{2\alpha_{\min}}{2-\alpha_{\min}\rho_+}\left(\phi_{\boldsymbol{\lambda}^*,\boldsymbol{w}}(\widetilde{\boldsymbol{\beta}}_t^{k-1}) - \phi_{\boldsymbol{\lambda}^*,\boldsymbol{w}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}^*})\right)} \\
&\leq \sqrt{\frac{2}{\rho_+}\left(\phi_{\boldsymbol{\lambda}^*,\boldsymbol{w}}(\widetilde{\boldsymbol{\beta}}_{t_0}^{k-1}) - \phi_{\boldsymbol{\lambda}^*,\boldsymbol{w}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}^*})\right)}\delta^{K(t-t_0)+k-1} \\
\text{(By Lemma F.6)} \quad &\leq \sqrt{\frac{2}{\rho_+}}\sqrt{\frac{10}{\rho_- - \xi_{\max}}}\|(\boldsymbol{\lambda}^*)_{S^*}\|_2\delta^{K(t-t_0)+k-1} \\
&= c'\sqrt{s^*}\delta^{K(t-t_0)+k}.
\end{aligned}$$

Combining the two different bounds for $\left\|\widetilde{\boldsymbol{\beta}}_t^k - \widetilde{\boldsymbol{\beta}}_t^{k-1}\right\|_2$ with Eq. 24, we have

$$\begin{aligned}
\|\boldsymbol{\beta}_t - \boldsymbol{\beta}_t'\|_2 &\leq \delta^K\|\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1}'\|_2 + \begin{cases} C\left|\alpha - \alpha'\right|\sqrt{s^*}\sum_{k=1}^K \delta^{K-k}\delta^k & \text{if } t < t_0 \\ C\left|\alpha - \alpha'\right|\sqrt{s^*}\sum_{k=1}^K \delta^{K-k}\delta^{K(t-t_0)+k} & \text{if } t \geq t_0 \end{cases} \\
&= \delta^K\|\boldsymbol{\beta}_{t-1} - \boldsymbol{\beta}_{t-1}'\|_2 + \begin{cases} C\left|\alpha - \alpha'\right|\sqrt{s^*}K\delta^K & \text{if } t < t_0 \\ C\left|\alpha - \alpha'\right|\sqrt{s^*}K\delta^{K(t-t_0+1)} & \text{if } t \geq t_0 \end{cases}
\end{aligned}$$

where $C = cc'$. We apply the above inequality recursively and obtain that

$$\begin{aligned}
\|\boldsymbol{\beta}_T - \boldsymbol{\beta}_T'\|_2 &\leq C\sqrt{s^*}K\delta^K\left(\sum_{t=1}^{t_0-1}\delta^{K(T-t)}\left|\alpha - \alpha'\right| + \sum_{t=t_0}^{T}\delta^{K(T-t)}\delta^{K(t-t_0)}\left|\alpha - \alpha'\right|\right) \\
&= C\sqrt{s^*}K\delta^K\left(\delta^{K(T-t_0+1)}\frac{1-\delta^{K(t_0-1)}}{1-\delta^K} + (T-t_0)\delta^{K(T-t_0)}\right)\left|\alpha - \alpha'\right| \\
&= C\sqrt{s^*}K\delta^K\left(\left(\frac{\delta^K(1-\delta^{K(t_0-1)})}{1-\delta^K} + (T-t_0)\right)\delta^{K(T-t_0)}\right)\left|\alpha - \alpha'\right| \\
&\leq C\sqrt{s^*}K\left(T-t_0+1\right)\delta^{K(T-t_0+1)}\left|\alpha - \alpha'\right| \\
&\leq C\sqrt{s^*}K\left(T-t_0\right)\delta^{K(T-t_0)}\left|\alpha - \alpha'\right|
\end{aligned}$$

□

Lemma E.2. *Assume the assumptions in Lemma 4.1 are satisfied. Let*

$$\theta = \{\boldsymbol{\lambda}^*, \boldsymbol{w}, \boldsymbol{\eta}, \alpha\} \quad \text{and} \quad \theta' = \{\boldsymbol{\lambda}^{*\prime}, \boldsymbol{w}', \boldsymbol{\eta}', \alpha\}$$

be parameters in the constrained space Θ. Let $\{\boldsymbol{\beta}_t\}_{t=1}^T = \texttt{PLISA}_\theta(Z_{1:n})$ and $\{\boldsymbol{\beta}_t{}'\}_{t=1}^T = \texttt{PLISA}_{\theta'}(Z_{1:n})$ be the intermediate outputs of $\texttt{PLISA}$ with different parameters θ and θ'. Then $\|\boldsymbol{\beta}_T(Z_{1:n};\theta) - \boldsymbol{\beta}_T(Z_{1:n},\theta')\|_2 \leq$

$$\left(C_1\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + C_2\|\boldsymbol{\lambda}^*-\boldsymbol{\lambda}^{*\prime}\|_2 + C_4\sqrt{d}\|\boldsymbol{w}-\boldsymbol{w}'\|_2\right)\frac{1-\delta^{KT}}{1-\delta},$$

where $\delta := \sqrt{1-\alpha_{\min}(\rho_- - \xi_{\max})} < 1$ and $C_{1,2,4}$ are some constants.

Proof. Let $\{\boldsymbol{\beta}_t\}_{t=1}^T \cup \{\widetilde{\boldsymbol{\beta}}_t^k\}_{t=1}^T{}_{k=1}^K$ and $\{\boldsymbol{\beta}_t{}'\}_{t=1}^T \cup \{\widetilde{\boldsymbol{\beta}}_t^{k\prime}\}_{t=1}^T{}_{k=1}^K$ be the intermediate outputs of $\texttt{PLISA}_\theta(Z_{1:n})$ and $\texttt{PLISA}_{\theta'}(Z_{1:n})$ respectively. By triangle inequality,

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{\beta}}_t^k - \widetilde{\boldsymbol{\beta}}_t^{k\prime}\right\|_2 = \left\|\texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1};\boldsymbol{\lambda}_t,\boldsymbol{w},\alpha) - \texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1\prime};\boldsymbol{\lambda}_t{}',\boldsymbol{w}',\alpha')\right\|_2 \\
&\leq \left\|\texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1};\boldsymbol{\lambda}_t,\boldsymbol{w},\alpha) - \texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1\prime};\boldsymbol{\lambda}_t,\boldsymbol{w},\alpha)\right\|_2 \\
&+ \left\|\texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1\prime};\boldsymbol{\lambda}_t,\boldsymbol{w},\alpha) - \texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1\prime};\boldsymbol{\lambda}_t{}',\boldsymbol{w},\alpha)\right\|_2 \\
&+ \left\|\texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1\prime};\boldsymbol{\lambda}_t{}',\boldsymbol{w},\alpha) - \texttt{MPG}(\widetilde{\boldsymbol{\beta}}_t^{k-1\prime};\boldsymbol{\lambda}_t{}',\boldsymbol{w}',\alpha)\right\|_2.
\end{aligned}$$

Applying Lemma E.3, Lemma E.5, and Lemma E.6 to the above 3 terms on the right hand side, we obtain

$$\left\|\widetilde{\boldsymbol{\beta}}_t^k - \widetilde{\boldsymbol{\beta}}_t^{k\prime}\right\|_2 \leq \delta\|\widetilde{\boldsymbol{\beta}}_t^{k-1} - \widetilde{\boldsymbol{\beta}}_t^{k-1\prime}\|_2 + \underbrace{\frac{2}{\rho_+}\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_t'\|_2}_{(i)} + \underbrace{C\|\boldsymbol{w}-\boldsymbol{w}'\|_2\|\boldsymbol{\lambda}_t\|_2}_{(ii)}, \tag{25}$$

where C is some absolute constant.

Term (i). By definition and triangular inequality,

$$\begin{aligned}
\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_t{}'\|_2 &\leq \|\max\left\{\boldsymbol{\lambda}_0 \circ \sigma(\boldsymbol{\eta})^t, \boldsymbol{\lambda}^*\right\} - \max\left\{\boldsymbol{\lambda}_0 \circ \sigma(\boldsymbol{\eta}')^t, \boldsymbol{\lambda}^*\right\}\|_2 \\
&\quad + \|\max\left\{\boldsymbol{\lambda}_0 \circ \sigma(\boldsymbol{\eta}')^t, \boldsymbol{\lambda}^*\right\} - \max\left\{\boldsymbol{\lambda}_0 \circ \sigma(\boldsymbol{\eta}')^t, \boldsymbol{\lambda}^{*\prime}\right\}\|_2 \\
&\leq \|\boldsymbol{\lambda}_0\|_\infty t\sigma(\eta_{\max})^{t-1}t\|\sigma(\boldsymbol{\eta}) - \sigma(\boldsymbol{\eta}')\|_2 + \|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^{*\prime}\|_2 \\
&\leq \frac{B_2}{4}\sigma(\eta_{\max})^t t\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + \|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^{*\prime}\|_2
\end{aligned}$$

The last inequality holds by Assumption 3.1 and the bounded first derivative of $\sigma(\cdot)$. Therefore, for some constants $C_2 = \frac{2}{\rho_+}$ and $\tilde{C}_1 = C_2\frac{B_2}{4}$,

$$\begin{aligned}
(i) = \frac{2}{\rho_+}\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_t'\|_2 &\leq \tilde{C}_1\sigma(\eta_{\max})^t t\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + C_2\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^{*\prime}\|_2 \\
&\leq C_1\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + C_2\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^{*\prime}\|_2.
\end{aligned}$$

The last inequality holds because the value $\sigma(\eta_{\max})^t t$ is bounded by some constant.

Term (ii). Since $\|\boldsymbol{\lambda}_t\|_2 \leq \|\max\{\sigma(\eta_{\max})^t\boldsymbol{\lambda}_0, \boldsymbol{\lambda}^{*\prime}\}\|_2 \leq \|\max\{\sigma(\eta_{\max})^t\boldsymbol{\lambda}_0, \boldsymbol{\lambda}_0\}\|_2 = \|\boldsymbol{\lambda}_0\|_2 \leq B\sqrt{d}$,

$$(ii) = C\|\boldsymbol{w}-\boldsymbol{w}'\|_2\|\boldsymbol{\lambda}_t\|_2 \leq C_4\sqrt{d}\|\boldsymbol{w}-\boldsymbol{w}'\|_2.$$

Plugging the bounds for (i)-(iii) into Eq. 25, we have

$$\left\|\widetilde{\boldsymbol{\beta}}_t^k - \widetilde{\boldsymbol{\beta}}_t^{k\prime}\right\|_2 \leq \delta\|\widetilde{\boldsymbol{\beta}}_t^{k-1} - \widetilde{\boldsymbol{\beta}}_t^{k-1\prime}\|_2 + C_1\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + C_2\|\boldsymbol{\lambda}^* - \boldsymbol{\lambda}^{*\prime}\|_2 + C_4\sqrt{d}\|\boldsymbol{w}-\boldsymbol{w}'\|_2.$$

By direct computation,

$$\begin{aligned}
&\left\|\widetilde{\boldsymbol{\beta}}_t^K - \widetilde{\boldsymbol{\beta}}_t^{K\prime}\right\|_2 \\
&\leq \delta^K \left\|\widetilde{\boldsymbol{\beta}}_t^0 - \widetilde{\boldsymbol{\beta}}_t^{0\prime}\right\|_2 + \left(C_1\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + C_2\|\boldsymbol{\lambda}^*-\boldsymbol{\lambda}^{*\prime}\|_2 + C_4\sqrt{d}\|\boldsymbol{w}-\boldsymbol{w}'\|_2\right)\sum_{k=0}^{K-1}\delta^k \\
&= \delta^K \left\|\widetilde{\boldsymbol{\beta}}_t^0 - \widetilde{\boldsymbol{\beta}}_t^{0\prime}\right\|_2 + \left(C_1\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + C_2\|\boldsymbol{\lambda}^*-\boldsymbol{\lambda}^{*\prime}\|_2 + C_4\sqrt{d}\|\boldsymbol{w}-\boldsymbol{w}'\|_2\right)\frac{1-\delta^K}{1-\delta}.
\end{aligned}$$

Recall that $\widetilde{\boldsymbol{\beta}}_t^K = \boldsymbol{\beta}_t$ and $\widetilde{\boldsymbol{\beta}}_t^0 = \boldsymbol{\beta}_{t-1}$. We can apply the above inequality recursively and obtain that

$$\begin{aligned}
&\|\boldsymbol{\beta}_T - \boldsymbol{\beta}_T{}'\|_2 \\
&\leq \delta^K \|\boldsymbol{\beta}_{T-1} - \boldsymbol{\beta}_{T-1}{}'\|_2 + \left(C_1\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + C_2\|\boldsymbol{\lambda}^*-\boldsymbol{\lambda}^{*\prime}\|_2 + C_4\sqrt{d}\|\boldsymbol{w}-\boldsymbol{w}'\|_2\right)\frac{1-\delta^K}{1-\delta} \\
&\leq \left(C_1\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + C_2\|\boldsymbol{\lambda}^*-\boldsymbol{\lambda}^{*\prime}\|_2 + C_4\sqrt{d}\|\boldsymbol{w}-\boldsymbol{w}'\|_2\right)\frac{1-\delta^K}{1-\delta}\sum_{t=0}^{T-1}(\delta^K)^t \\
&= \left(C_1\|\boldsymbol{\eta}-\boldsymbol{\eta}'\|_2 + C_2\|\boldsymbol{\lambda}^*-\boldsymbol{\lambda}^{*\prime}\|_2 + C_4\sqrt{d}\|\boldsymbol{w}-\boldsymbol{w}'\|_2\right)\frac{1-\delta^{KT}}{1-\delta}.
\end{aligned}$$

□

Lemma E.3 (Robustness to $\boldsymbol{\beta}$)**.** *Assume $(Z_{1:n}, \boldsymbol{\beta}^*) \in \mathcal{P}$ and $\mathcal{P}$ satisfies Assumption 3.1. Assume $\boldsymbol{\beta}, \boldsymbol{\beta}'$ satisfy $\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_0 \leq s = s^* + 2\tilde{s}$ and $\alpha \leq \frac{1}{\rho_+}$. Then*

$$\|\texttt{MPG}(\boldsymbol{\beta};\boldsymbol{\lambda},\boldsymbol{w},\alpha) - \texttt{MPG}(\boldsymbol{\beta}';\boldsymbol{\lambda},\boldsymbol{w},\alpha)\|_2 \leq \sqrt{1 - 2\alpha\frac{(\rho_- - \xi_{\boldsymbol{w}})\rho_+}{\rho_- - \xi_{\boldsymbol{w}} + \rho_+}}\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2 \tag{26}$$

$$\leq \sqrt{1 - \alpha\left(\rho_- - \xi_{\boldsymbol{w}}\right)}\,\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2 \tag{27}$$

Proof.

$$\begin{aligned}
&\|\texttt{MPG}(\boldsymbol{\beta};\boldsymbol{\lambda},\boldsymbol{w},\alpha) - \texttt{MPG}(\boldsymbol{\beta}';\boldsymbol{\lambda},\boldsymbol{w},\alpha)\|_2^2 \\
&= \|\mathcal{T}_{\alpha\cdot\boldsymbol{\lambda}}\left(\boldsymbol{\beta} - \alpha\left(\nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta})\right)\right) - \mathcal{T}_{\alpha\cdot\boldsymbol{\lambda}}\left(\boldsymbol{\beta}' - \alpha\left(\nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}')\right)\right)\|_2^2 \\
&\leq \|\boldsymbol{\beta} - \alpha\left(\nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta})\right) - \boldsymbol{\beta}' + \alpha\left(\nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}')\right)\|_2^2 \\
&= \|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2^2 + \alpha^2\|\nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}) - \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}')\|_2^2 - 2\alpha\langle\boldsymbol{\beta}-\boldsymbol{\beta}', \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}) - \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}')\rangle.
\end{aligned} \tag{28}$$

Since on the restricted subspace $\{\boldsymbol{\beta} : \|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_0 \leq s^* + 2\tilde{s}\}$, $L_{\boldsymbol{\lambda}}$ is $(\rho_- - \xi_{\boldsymbol{w}})$-strongly convex and ρ_+-smooth, by Lemma F.7,

$$\begin{aligned}
&\langle\boldsymbol{\beta}-\boldsymbol{\beta}', \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}) - \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}')\rangle \\
&\geq \frac{(\rho_- - \xi_{\boldsymbol{w}})\rho_+}{\rho_- - \xi_{\boldsymbol{w}} + \rho_+}\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2^2 + \frac{\|\nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}) - \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}')\|_2^2}{\rho_- - \xi_{\boldsymbol{w}} + \rho_+}.
\end{aligned}$$

Combining this inequality with Eq. 28, we have

$$\begin{aligned}
&\|\texttt{MPG}(\boldsymbol{\beta};\boldsymbol{\lambda},\boldsymbol{w},\alpha) - \texttt{MPG}(\boldsymbol{\beta}';\boldsymbol{\lambda},\boldsymbol{w},\alpha)\|_2^2 \\
&\leq \left(1 - 2\alpha\frac{(\rho_- - \xi_{\boldsymbol{w}})\rho_+}{\rho_- - \xi_{\boldsymbol{w}} + \rho_+}\right)\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2^2 \\
&+ \alpha\left(\alpha - \frac{2}{\rho_- - \xi_{\boldsymbol{w}} + \rho_+}\right)\|\nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}) - \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}')\|_2^2
\end{aligned}$$

By the assumption that $\alpha \leq \frac{1}{\rho_+}$, the second term in the above inequality is non-positive. Furthermore, this assumption also implies that $1 - 2\alpha\frac{(\rho_- - \xi_{\boldsymbol{w}})\rho_+}{\rho_- - \xi_{\boldsymbol{w}} + \rho_+} \geq 0$. Therefore,

$$\|\texttt{MPG}(\boldsymbol{\beta};\boldsymbol{\lambda},\boldsymbol{w},\alpha) - \texttt{MPG}(\boldsymbol{\beta}';\boldsymbol{\lambda},\boldsymbol{w},\alpha)\|_2 \leq \sqrt{1 - 2\alpha\frac{(\rho_- - \xi_{\boldsymbol{w}})\rho_+}{\rho_- - \xi_{\boldsymbol{w}} + \rho_+}}\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2$$

□

Lemma E.4 (Robustness to α)**.** *With out loss of generalization, assume $\alpha \geq \alpha'$. Denote*

$$\boldsymbol{\beta}^+ := \texttt{MPG}(\boldsymbol{\beta}; \boldsymbol{\lambda}, \boldsymbol{w}, \alpha) \quad \textit{and} \quad \boldsymbol{\beta}^{+\prime} := \texttt{MPG}(\boldsymbol{\beta}; \boldsymbol{\lambda}, \boldsymbol{w}, \alpha').$$

Then

$$\|\boldsymbol{\beta}^{+\prime} - \boldsymbol{\beta}^+\|_2 \leq \sqrt{\frac{\alpha - \alpha'}{\alpha + \alpha'}} \left\|\boldsymbol{\beta}^+ - \boldsymbol{\beta}\right\|_2$$

Proof. Define the quadratic approximation function as

$$\psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{z}, \boldsymbol{\beta}) := L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}) + \langle \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}), \boldsymbol{z} - \boldsymbol{\beta} \rangle + \frac{1}{2\alpha} \|\boldsymbol{z} - \boldsymbol{\beta}\|_2^2 + \|\boldsymbol{z}\|_{\boldsymbol{\lambda},1}. \tag{29}$$

Clearly,

$$\texttt{MPG}(\boldsymbol{\beta}; \boldsymbol{\lambda}, \boldsymbol{w}, \alpha) = \arg\min_{\boldsymbol{z}} \psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{z}, \boldsymbol{\beta}).$$

Since $\psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{z}, \boldsymbol{\beta})$ is $\frac{1}{\alpha}$-strongly convex in $\boldsymbol{z}$, and that $\boldsymbol{\beta}^+ := \texttt{MPG}(\boldsymbol{\beta}; \boldsymbol{\lambda}, \boldsymbol{w}, \alpha)$ is its optimal point,

$$\psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{z}, \boldsymbol{\beta}) \geq \psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^+, \boldsymbol{\beta}) + \frac{1}{2\alpha} \left\|\boldsymbol{z} - \boldsymbol{\beta}^+\right\|_2^2 \quad \forall \boldsymbol{z}. \tag{30}$$

Similarly,

$$\psi_{\alpha',\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{z}, \boldsymbol{\beta}) \geq \psi_{\alpha',\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^{+\prime}, \boldsymbol{\beta}) + \frac{1}{2\alpha'} \left\|\boldsymbol{z} - \boldsymbol{\beta}^{+\prime}\right\|_2^2 \quad \forall \boldsymbol{z}. \tag{31}$$

Taking $\boldsymbol{z} = \boldsymbol{\beta}^{+\prime}$ in Eq. 30 and $\boldsymbol{z} = \boldsymbol{\beta}^+$ in Eq. 31 yields

$$\begin{aligned}
\frac{1}{2\alpha} \left\|\boldsymbol{\beta}^+ - \boldsymbol{\beta}^{+\prime}\right\|_2^2 &\leq \psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^{+\prime}, \boldsymbol{\beta}) - \psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^+, \boldsymbol{\beta}), \\
\frac{1}{2\alpha'} \left\|\boldsymbol{\beta}^+ - \boldsymbol{\beta}^{+\prime}\right\|_2^2 &\leq \psi_{\alpha',\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^+, \boldsymbol{\beta}) - \psi_{\alpha',\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^{+\prime}, \boldsymbol{\beta}).
\end{aligned}$$

Summing the above two inequalities, we have

$$\begin{aligned}
\frac{1}{2}\left(\frac{1}{\alpha} + \frac{1}{\alpha'}\right) \left\|\boldsymbol{\beta}^+ - \boldsymbol{\beta}^{+\prime}\right\|_2^2 &\leq \left(\psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^{+\prime}, \boldsymbol{\beta}) - \psi_{\alpha',\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^{+\prime}, \boldsymbol{\beta})\right) \\
&\quad + \left(\psi_{\alpha',\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^+, \boldsymbol{\beta}) - \psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^+, \boldsymbol{\beta})\right) \\
\text{(by } \alpha \geq \alpha') \quad &\leq \psi_{\alpha',\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^+, \boldsymbol{\beta}) - \psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^+, \boldsymbol{\beta}) \\
&= \frac{1}{2}\left(\frac{1}{\alpha'} - \frac{1}{\alpha}\right) \left\|\boldsymbol{\beta}^+ - \boldsymbol{\beta}\right\|_2^2.
\end{aligned}$$

□

Lemma E.5 (Robustness to $\boldsymbol{\lambda}$)**.** *Let $\boldsymbol{\beta}^+ = \texttt{MPG}(\boldsymbol{\beta}; \boldsymbol{\lambda}, \boldsymbol{w}, \alpha)$ and $\boldsymbol{\beta}^{+\prime} = \texttt{MPG}(\boldsymbol{\beta}; \boldsymbol{\lambda}', \boldsymbol{w}, \alpha)$. Then*

$$\|\boldsymbol{\beta}^+ - \boldsymbol{\beta}^{+\prime}\|_2 \leq 2\alpha \|(\boldsymbol{\lambda} - \boldsymbol{\lambda}')_S\|_2, \tag{32}$$

where $S := \operatorname{supp}(\boldsymbol{\beta}^+) \cup \operatorname{supp}(\boldsymbol{\beta}^{+\prime})$.

Proof. Recall the quadratic approximation in Eq. 44 and the definition that $\texttt{MPG}(\boldsymbol{\beta}; \boldsymbol{\lambda}, \boldsymbol{w}, \alpha) = \arg\min_{\boldsymbol{z}} \psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{z}, \boldsymbol{\beta})$. By the $\frac{1}{\alpha}$-strongly convexity of $\psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}$ in $\boldsymbol{z}$, it holds true for all $\boldsymbol{z}$ that

$$\begin{aligned}
&\frac{1}{2\alpha} \left\|\boldsymbol{z} - \boldsymbol{\beta}^+\right\|_2^2 \leq \psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{z}, \boldsymbol{\beta}) - \psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^+, \boldsymbol{\beta}) \\
&= \langle \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}), \boldsymbol{z} - \boldsymbol{\beta}^+ \rangle + \frac{1}{2\alpha}\left(\|\boldsymbol{z} - \boldsymbol{\beta}\|_2^2 - \left\|\boldsymbol{\beta}^+ - \boldsymbol{\beta}\right\|_2^2\right) + (\|\boldsymbol{z}\|_{\boldsymbol{\lambda},1} - \left\|\boldsymbol{\beta}^+\right\|_{\boldsymbol{\lambda},1})
\end{aligned} \tag{33}$$

Similarly,

$$\frac{1}{2\alpha} \left\|\boldsymbol{z} - \boldsymbol{\beta}^{+\prime}\right\|_2^2 \tag{34}$$

$$\leq \langle \nabla L_{\boldsymbol{\lambda}',\boldsymbol{w}}(\boldsymbol{\beta}), \boldsymbol{z} - \boldsymbol{\beta}^{+\prime} \rangle + \frac{1}{2\alpha}\left(\|\boldsymbol{z} - \boldsymbol{\beta}\|_2^2 - \left\|\boldsymbol{\beta}^{+\prime} - \boldsymbol{\beta}\right\|_2^2\right) + (\|\boldsymbol{z}\|_{\boldsymbol{\lambda}',1} - \left\|\boldsymbol{\beta}^{+\prime}\right\|_{\boldsymbol{\lambda}',1}). \tag{35}$$

Taking $\boldsymbol{z} = \boldsymbol{\beta}^{+\prime}$ in Eq. 33, taking $\boldsymbol{z} = \boldsymbol{\beta}^{+}$ in Eq. 35, and summing up the two inequalities, we have

$$\begin{aligned}
\frac{1}{\alpha}\left\|\boldsymbol{\beta}^{+}-\boldsymbol{\beta}^{+\prime}\right\|_2^2 &\leq \langle \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}) - \nabla L_{\boldsymbol{\lambda}',\boldsymbol{w}}(\boldsymbol{\beta}), \boldsymbol{\beta}^{+\prime} - \boldsymbol{\beta}^{+}\rangle \\
&\quad + \|\boldsymbol{\beta}^{+}\circ(\boldsymbol{\lambda}'-\boldsymbol{\lambda})\|_1 - \|\boldsymbol{\beta}^{+\prime}\circ(\boldsymbol{\lambda}'-\boldsymbol{\lambda})\|_1 \\
&= \langle \nabla \mathcal{Q}_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}) - \nabla \mathcal{Q}_{\boldsymbol{\lambda}',\boldsymbol{w}}(\boldsymbol{\beta}), \boldsymbol{\beta}^{+\prime} - \boldsymbol{\beta}^{+}\rangle + \|\boldsymbol{\beta}^{+}-\boldsymbol{\beta}^{+\prime}\|_{|\boldsymbol{\lambda}-\boldsymbol{\lambda}'|,1} \\
&\leq \sum_{j=1}^{d} |q'_{\boldsymbol{w}}(\lambda_j,\beta_j) - q'_{\boldsymbol{w}}(\lambda'_j,\beta_j)||\beta_j^{+\prime}-\beta_j^{+}| + \|\boldsymbol{\beta}^{+}-\boldsymbol{\beta}^{+\prime}\|_{|\boldsymbol{\lambda}-\boldsymbol{\lambda}'|,1} \\
\text{(by Assumption B.1)}\quad &\leq \sum_{j=1}^{d} |\lambda_j-\lambda'_j||\beta_j^{+\prime}-\beta_j^{+}| + \|\boldsymbol{\beta}^{+}-\boldsymbol{\beta}^{+\prime}\|_{|\boldsymbol{\lambda}-\boldsymbol{\lambda}'|,1} \\
&= 2\|\boldsymbol{\beta}^{+}-\boldsymbol{\beta}^{+\prime}\|_{|\boldsymbol{\lambda}-\boldsymbol{\lambda}'|,1} \\
&\leq 2\|(\boldsymbol{\lambda}-\boldsymbol{\lambda}')_S\|_2 \left\|\boldsymbol{\beta}^{+\prime}-\boldsymbol{\beta}^{+}\right\|_2
\end{aligned}$$

where $S := \mathrm{supp}(\boldsymbol{\beta}^{+}) \cup \mathrm{supp}(\boldsymbol{\beta}^{+\prime})$. Dividing both side by $\|\boldsymbol{\beta}^{+\prime}-\boldsymbol{\beta}^{+}\|_2$ draws the conclusion. □

Lemma E.6 (Robustness to $\boldsymbol{w}$)**.** *Let $\boldsymbol{\beta}^{+} = \mathtt{MPG}(\boldsymbol{\beta};\boldsymbol{\lambda},\boldsymbol{w},\alpha)$ and $\boldsymbol{\beta}^{+\prime} = \mathtt{MPG}(\boldsymbol{\beta};\boldsymbol{\lambda},\boldsymbol{w}',\alpha)$. Then*

$$\|\boldsymbol{\beta}^{+}-\boldsymbol{\beta}^{+\prime}\|_2 \leq C\|\boldsymbol{w}-\boldsymbol{w}'\|_2\|(\boldsymbol{\lambda})_S\|_2, \tag{36}$$

where $S := \mathrm{supp}(\boldsymbol{\beta}^{+}) \cup \mathrm{supp}(\boldsymbol{\beta}^{+\prime})$ and C is some absolute constant.

Proof. Recall the quadratic approximation in Eq. 44 and the definition that $\mathtt{MPG}(\boldsymbol{\beta};\boldsymbol{\lambda},\boldsymbol{w},\alpha) = \arg\min_{\boldsymbol{z}} \psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{z},\boldsymbol{\beta})$. By the $\frac{1}{\alpha}$-strongly convexity of $\psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}$ in $\boldsymbol{\beta}$, it holds true for all $\boldsymbol{z}$ in the restricted subspace that

$$\begin{aligned}
&\frac{1}{2\alpha}\left\|\boldsymbol{z}-\boldsymbol{\beta}^{+}\right\|_2^2 \leq \psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{z},\boldsymbol{\beta}) - \psi_{\alpha,\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}^{+},\boldsymbol{\beta}) \\
&= \langle \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}), \boldsymbol{z}-\boldsymbol{\beta}^{+}\rangle + \frac{1}{2\alpha}\left(\|\boldsymbol{z}-\boldsymbol{\beta}\|_2^2 - \left\|\boldsymbol{\beta}^{+}-\boldsymbol{\beta}\right\|_2^2\right) + (\|\boldsymbol{z}\|_{\boldsymbol{\lambda},1} - \left\|\boldsymbol{\beta}^{+}\right\|_{\boldsymbol{\lambda},1})
\end{aligned} \tag{37}$$

Similarly,

$$\begin{aligned}
&\frac{1}{2\alpha}\left\|\boldsymbol{z}-\boldsymbol{\beta}^{+\prime}\right\|_2^2 \\
&\leq \langle \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}'}(\boldsymbol{\beta}), \boldsymbol{z}-\boldsymbol{\beta}^{+\prime}\rangle + \frac{1}{2\alpha}\left(\|\boldsymbol{z}-\boldsymbol{\beta}\|_2^2 - \left\|\boldsymbol{\beta}^{+\prime}-\boldsymbol{\beta}\right\|_2^2\right) + (\|\boldsymbol{z}\|_{\boldsymbol{\lambda},1} - \left\|\boldsymbol{\beta}^{+\prime}\right\|_{\boldsymbol{\lambda},1}).
\end{aligned} \tag{38}$$

Taking $\boldsymbol{z} = \boldsymbol{\beta}^{+\prime}$ in Eq. 37, taking $\boldsymbol{z} = \boldsymbol{\beta}^{+}$ in Eq. 38, and summing up the two inequalities, we have

$$\begin{aligned}
\frac{1}{\alpha}\left\|\boldsymbol{\beta}^{+}-\boldsymbol{\beta}^{+\prime}\right\|_2^2 &\leq \langle \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}) - \nabla L_{\boldsymbol{\lambda},\boldsymbol{w}'}(\boldsymbol{\beta}), \boldsymbol{\beta}^{+\prime} - \boldsymbol{\beta}^{+}\rangle \\
&= \langle \nabla \mathcal{Q}_{\boldsymbol{\lambda},\boldsymbol{w}}(\boldsymbol{\beta}) - \nabla \mathcal{Q}_{\boldsymbol{\lambda},\boldsymbol{w}'}(\boldsymbol{\beta}), \boldsymbol{\beta}^{+\prime} - \boldsymbol{\beta}^{+}\rangle \\
&\leq \sum_{j=1}^{d} |q'_{\boldsymbol{w}}(\lambda_j,\beta_j) - q'_{\boldsymbol{w}'}(\lambda_j,\beta_j)||\beta_j^{+\prime}-\beta_j^{+}|.
\end{aligned}$$

Note that $q'_{\boldsymbol{w}}(\lambda_j,\beta_j) = \sum_{i=1}^{3} \frac{\exp(w_i)}{Z(\boldsymbol{w})} q^{(i)\prime}(\lambda_j,\boldsymbol{\beta}_j)$ where $Z(\boldsymbol{w}) = \sum_{i=1}^{3}\exp(w_i)$. Then

$$\begin{aligned}
|q'_{\boldsymbol{w}}(\lambda_j,\beta_j) - q'_{\boldsymbol{w}'}(\lambda_j,\beta_j)| &\leq \sum_{i=1}^{3}\left|\frac{\exp(w_i)}{Z(\boldsymbol{w})} - \frac{\exp(w'_i)}{Z(\boldsymbol{w}')}\right|\left|q^{(i)\prime}(\lambda_j,\boldsymbol{\beta}_j)\right| \\
\text{(by Assumption B.1)}\quad &\leq \sum_{i=1}^{3}\left|\frac{\exp(w_i)}{Z(\boldsymbol{w})} - \frac{\exp(w'_i)}{Z(\boldsymbol{w}')}\right|\lambda_j.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{1}{\alpha}\left\|\boldsymbol{\beta}^{+}-\boldsymbol{\beta}^{+\prime}\right\|_2^2 &\leq \sum_{i=1}^{3}\left|\frac{\exp(w_i)}{Z(\boldsymbol{w})}-\frac{\exp(w_i')}{Z(\boldsymbol{w}')}\right|\sum_{j=1}^{d}\lambda_j|\beta_j^{+\prime}-\beta_j^{+}| \\
&\leq \sum_{i=1}^{3}\left|\frac{\exp(w_i)}{Z(\boldsymbol{w})}-\frac{\exp(w_i')}{Z(\boldsymbol{w}')}\right|\|(\boldsymbol{\lambda})_S\|_2\left\|\boldsymbol{\beta}^{+\prime}-\boldsymbol{\beta}^{+}\right\|_2 \\
\text{(By Lipschitz continuity)} \quad &\leq C\|\boldsymbol{w}-\boldsymbol{w}'\|_2\|(\boldsymbol{\lambda})_S\|_2\left\|\boldsymbol{\beta}^{+\prime}-\boldsymbol{\beta}^{+}\right\|_2,
\end{aligned}$$

for some constant C. □

F KEY LEMMAS

This section supplies some fundamental results for the modified proximal gradient algorithm. All the results in this section are based on the following assumption and notations.

Assumption F.1. *Assume $\mathcal{P}$ satisfies Assumption 3.1 and consider a problem $(Z_n, \boldsymbol{\beta}^*) \in \mathcal{P}$. Assume we have a penalty function $P(\boldsymbol{\lambda}, \boldsymbol{\beta})$ whose concave component $Q(\boldsymbol{\lambda}, \boldsymbol{\beta}) = P(\boldsymbol{\lambda}, \boldsymbol{\beta}) - \|\boldsymbol{\lambda} \circ \boldsymbol{\beta}\|_1$ satisfies the regularization conditions in Assumption B.1 with a Lipschitz constant ξ for the condition (a). We adopt the following notations for the statements in this sections.*

$$
\begin{aligned}
\textit{empirical loss} \quad & L_n(\boldsymbol{\beta}) := L_n(Z_{1:n}, \boldsymbol{\beta}) \\
\textit{modified loss} \quad & L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) := L_n(Z_{1:n}, \boldsymbol{\beta}) + Q(\boldsymbol{\lambda}, \boldsymbol{\beta}) \\
\textit{regularized loss} \quad & \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) := L_n(Z_{1:n}, \boldsymbol{\beta}) + P(\boldsymbol{\lambda}, \boldsymbol{\beta}) = L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \|\boldsymbol{\beta}\|_{\boldsymbol{\lambda},1} \\
\textit{local optimal} \quad & \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} \in \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta}_{\overline{S^*}}\|_0 \le \tilde{s}} \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) \\
\textit{quadratic approximation} \quad & \psi_{\alpha,\boldsymbol{\lambda}}(\boldsymbol{z}, \boldsymbol{\beta}) := L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \langle \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}), \boldsymbol{z} - \boldsymbol{\beta} \rangle + \frac{1}{2\alpha} \|\boldsymbol{z} - \boldsymbol{\beta}\|_2^2 + \|\boldsymbol{z}\|_{\boldsymbol{\lambda},1} \\
\textit{sub-optimality} \quad & \omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) := \min_{\boldsymbol{\xi}' \in \partial \|\boldsymbol{\beta}\|_1} \max_{\boldsymbol{\beta}'} \left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda},1}} \left(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}' \right) \right\}.
\end{aligned}
$$

Lemma F.1 (Contraction of MPG)**.** *Assume Assumption F.1. Let $\boldsymbol{\beta}^0, \cdots, \boldsymbol{\beta}^k$ be a sequence of vectors obtained by the modified proximal gradient updates:*

$$
\begin{aligned}
\boldsymbol{\beta}^k &= \textit{MPG}(\boldsymbol{\beta}^{k-1}; \boldsymbol{\lambda}, \boldsymbol{w}, \alpha) \\
&:= \mathcal{T}_{\alpha \cdot \boldsymbol{\lambda}} \left(\boldsymbol{\beta}^{k-1} - \alpha \left(\nabla L_n(\boldsymbol{\beta}^{k-1}) + \nabla_{\boldsymbol{\beta}} Q(\boldsymbol{\lambda}, \boldsymbol{\beta}^{k-1}) \right) \right).
\end{aligned}
$$

Assume $\|(\boldsymbol{\beta}^k)_{\overline{S^*}}\|_0 \le \tilde{s}$. Then

$$
\left(\frac{1}{\alpha} - \frac{\rho_+}{2} \right) \left\| \boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1} \right\|_2^2 \le \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) - \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k). \tag{39}
$$

Proof. Note that

$$
\boldsymbol{\beta}^k = \boldsymbol{\beta}^{k-1} - \alpha \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1}), \quad \text{where } \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}) := \alpha^{-1} \left(\boldsymbol{\beta} - \mathcal{T}_{\alpha\boldsymbol{\lambda}} \left(\boldsymbol{\beta} - \alpha \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) \right) \right).
$$

It is easy to observe that $\boldsymbol{\delta}_\alpha(\boldsymbol{\beta}) \in \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \boldsymbol{\lambda} \circ \partial \|\boldsymbol{\beta} - \alpha \boldsymbol{\delta}_\alpha(\boldsymbol{\beta})\|_1$. By RSS in Lemma B.1, it holds that

$$
L_{\boldsymbol{\lambda}}\left(\boldsymbol{\beta}^k\right) \le L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) - \alpha \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})^\top \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1}) + \frac{\rho_+}{2} \alpha^2 \left\| \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1}) \right\|_2^2.
$$

Furthermore, by RSC in Lemma B.1, for any $\boldsymbol{z}$ in the space $\{\left\|\boldsymbol{z} - \boldsymbol{\beta}^{k-1}\right\|_0 \le s^* + 2\tilde{s}\}$,

$$
L_{\boldsymbol{\lambda}}\left(\boldsymbol{\beta}^k\right) \le L_{\boldsymbol{\lambda}}(\boldsymbol{z}) + \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})^\top (\boldsymbol{\beta} - \boldsymbol{z}) - \alpha \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})^\top \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1}) \tag{40}
$$

$$
+ \frac{\rho_+}{2} \alpha^2 \left\| \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1}) \right\|_2^2 - \frac{\rho_- - \xi}{2} \left\| \boldsymbol{\beta}^{k-1} - \boldsymbol{z} \right\|_2^2. \tag{41}
$$

Denote $\boldsymbol{v} = \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1}) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})$. We have $\boldsymbol{v} \in \boldsymbol{\lambda} \circ \partial \left\| \boldsymbol{\beta}^{k-1} - \alpha \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1}) \right\|_1$ since $\boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1}) \in \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) + \boldsymbol{\lambda} \circ \partial \left\| \boldsymbol{\beta}^{k-1} - \alpha \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1}) \right\|_1$. By the convexity of $\|\cdot\|_{\boldsymbol{\lambda},1}$,

$$
\left\| \boldsymbol{\beta}^{k-1} - \alpha \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1}) \right\|_{\boldsymbol{\lambda},1} \le \|\boldsymbol{z}\|_{\boldsymbol{\lambda},1} + \boldsymbol{v}^\top \left(\boldsymbol{\beta}^{k-1} - \alpha \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1}) - \boldsymbol{z} \right).
$$

Combining this inequality with Eq. 41, we have

$$
L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) + \|\boldsymbol{\beta}^k\|_{\boldsymbol{\lambda},1} \le L_{\boldsymbol{\lambda}}(\boldsymbol{z}) + \lambda \|\boldsymbol{z}\|_{\boldsymbol{\lambda},1} + \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1})^\top (\boldsymbol{\beta}^{k-1} - \boldsymbol{z}) \tag{42}
$$

$$
- \alpha \left(1 - \frac{\alpha(\rho_+)}{2} \right) \left\| \boldsymbol{\delta}_\alpha(\boldsymbol{\beta}^{k-1}) \right\|_2^2 - \frac{\rho_- - \xi}{2} \left\| \boldsymbol{\beta}^{k-1} - \boldsymbol{z} \right\|_2^2. \tag{43}
$$

Taking $\boldsymbol{z} = \boldsymbol{\beta}^{k-1}$ in Eq. 43 implies that

$$
L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) + \|\boldsymbol{\beta}^k\|_{\boldsymbol{\lambda},1} \le L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) + \left\| \boldsymbol{\beta}^{k-1} \right\|_{\boldsymbol{\lambda},1} - \frac{1}{\alpha} \left(1 - \frac{\alpha(\rho_+)}{2} \right) \left\| \boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1} \right\|_2^2.
$$

Therefore,

$$\left(\frac{1}{\alpha}-\frac{\rho_+}{2}\right)\left\|\boldsymbol{\beta}^k-\boldsymbol{\beta}^{k-1}\right\|_2^2 \leq \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})-\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k).$$

□

Lemma F.2 (Convergence of modified proximal gradient)**.** *Assume Assumption F.1. Let $\boldsymbol{\beta}^0, \cdots, \boldsymbol{\beta}^k$ be a sequence of vectors obtained by the modified proximal gradient updates:*

$$\begin{aligned}\boldsymbol{\beta}^k &= \texttt{MPG}(\boldsymbol{\beta}^{k-1};\boldsymbol{\lambda},\boldsymbol{w},\alpha)\\ &:= \mathcal{T}_{\alpha\cdot\boldsymbol{\lambda}}\left(\boldsymbol{\beta}^{k-1}-\alpha\left(\nabla L_n(\boldsymbol{\beta}^{k-1})+\nabla_{\boldsymbol{\beta}}Q(\boldsymbol{\lambda},\boldsymbol{\beta}^{k-1})\right)\right)\end{aligned}$$

with a step size α in $(0, \frac{1}{\rho_+}]$. Assume $\|(\boldsymbol{\beta}^k)_{\overline{S^*}}\|_0 \leq \tilde{s}$. Then

$$\begin{aligned}\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k)-\phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) &\leq (1-\alpha(\rho_- - \xi_{\boldsymbol{w}}))^k\left(\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^0)-\phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})\right),\\ \|\boldsymbol{\beta}^k-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|_2^2 &\leq (1-\alpha(\rho_- - \xi_{\boldsymbol{w}}))^k\|\boldsymbol{\beta}^0-\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|_2^2,\\ \omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) &\leq \frac{\sqrt{2}(1+\frac{1}{\alpha\rho_+})\sqrt{\rho_+}}{\min(\boldsymbol{\lambda})}\sqrt{(1-\alpha(\rho_- - \xi_{\boldsymbol{w}}))^{k-1}\left(\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^0)-\phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})\right)}.\end{aligned}$$

Proof. Consider the quadratic approximation function

$$\psi_{\alpha,\boldsymbol{\lambda}}(\boldsymbol{z},\boldsymbol{\beta}) := L_{\boldsymbol{\lambda}}(\boldsymbol{\beta})+\langle\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}),\boldsymbol{z}-\boldsymbol{\beta}\rangle+\frac{1}{2\alpha}\|\boldsymbol{z}-\boldsymbol{\beta}\|_2^2+\|\boldsymbol{z}\|_{\boldsymbol{\lambda},1}. \tag{44}$$

By definition,

$$\texttt{MPG}(\boldsymbol{\beta}^{k-1};\boldsymbol{\lambda},\boldsymbol{w},\alpha) = \arg\min_{\boldsymbol{z}}\psi_{\alpha,\boldsymbol{\lambda}}(\boldsymbol{z},\boldsymbol{\beta}^{k-1}).$$

Since $\psi_{\alpha,\boldsymbol{\lambda}}(\boldsymbol{z},\boldsymbol{\beta}^{k-1})$ is $\frac{1}{\alpha}$-strongly convex with respect to $\boldsymbol{z}$, and that $\boldsymbol{\beta}^k$ is its optimal point, we can obtain

$$\psi_{\alpha,\boldsymbol{\lambda}}(\boldsymbol{z},\boldsymbol{\beta}^{k-1}) \geq \psi_{\alpha,\boldsymbol{\lambda}}(\boldsymbol{\beta}^k,\boldsymbol{\beta}^{k-1})+\frac{1}{2\alpha}\left\|\boldsymbol{z}-\boldsymbol{\beta}^k\right\|_2^2, \quad \forall \boldsymbol{z}. \tag{45}$$

Note that by Lemma B.1,

$$\begin{aligned}\psi_{\alpha,\boldsymbol{\lambda}}(\boldsymbol{\beta}^k,\boldsymbol{\beta}^{k-1}) &= L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})+\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})^\top\left(\boldsymbol{\beta}^k-\boldsymbol{\beta}^{k-1}\right)+\frac{1}{2\alpha}\|\boldsymbol{\beta}^k-\boldsymbol{\beta}^{k-1}\|_2^2\\ &\quad+\|\boldsymbol{\beta}^k\|_{\boldsymbol{\lambda},1}\\ &\geq L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k)+\frac{\frac{1}{\alpha}-\rho_+}{2}\|\boldsymbol{\beta}^k-\boldsymbol{\beta}^{k-1}\|_2^2+\|\boldsymbol{\beta}^k\|_{\boldsymbol{\lambda},1}\\ &= \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k)+\frac{\frac{1}{\alpha}-\rho_+}{2}\|\boldsymbol{\beta}^k-\boldsymbol{\beta}^{k-1}\|_2^2.\end{aligned}$$

Similarly, by Lemma B.1,

$$\begin{aligned}\psi_{\alpha,\boldsymbol{\lambda}}(\boldsymbol{z},\boldsymbol{\beta}^{k-1}) &= L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})+\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})^\top\left(\boldsymbol{z}-\boldsymbol{\beta}^{k-1}\right)+\frac{1}{2\alpha}\|\boldsymbol{z}-\boldsymbol{\beta}^{k-1}\|_2^2\\ &\quad+\|\boldsymbol{z}\|_{\boldsymbol{\lambda},1}\\ &\leq \phi_{\boldsymbol{\lambda}}(\boldsymbol{z})+\frac{\frac{1}{\alpha}-(\rho_- - \xi)}{2}\|\boldsymbol{z}-\boldsymbol{\beta}^{k-1}\|_2^2.\end{aligned}$$

Combining the above two results with Eq. 45, we have

$$\begin{aligned}\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) &\leq \phi_{\boldsymbol{\lambda}}(\boldsymbol{z})+\frac{\frac{1}{\alpha}-(\rho_- - \xi)}{2}\|\boldsymbol{z}-\boldsymbol{\beta}^{k-1}\|_2^2-\frac{\frac{1}{\alpha}-\rho_+}{2}\|\boldsymbol{\beta}^k-\boldsymbol{\beta}^{k-1}\|_2^2\\ &\quad-\frac{1}{2\alpha}\left\|\boldsymbol{z}-\boldsymbol{\beta}^k\right\|_2^2\\ &\leq \phi_{\boldsymbol{\lambda}}(\boldsymbol{z})+\frac{\frac{1}{\alpha}-(\rho_- - \xi)}{2}\|\boldsymbol{z}-\boldsymbol{\beta}^{k-1}\|_2^2-\frac{\frac{1}{\alpha}-\rho_+}{2}\|\boldsymbol{\beta}^k-\boldsymbol{\beta}^{k-1}\|_2^2.\end{aligned} \tag{46}$$

Taking $\boldsymbol{z} = c\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} + (1-c)\boldsymbol{\beta}^{k-1}$ for some $c \in [0,1]$, then

$$
\begin{aligned}
&\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) \\
&\leq c\phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) + (1-c)\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) + \frac{\left(\frac{1}{\alpha} - (\rho_- - \xi)\right)c^2}{2}\|\boldsymbol{\beta}^{k-1} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|_2^2 - \frac{\frac{1}{\alpha} - \rho_+}{2}\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}\|_2^2 \\
&\leq c\phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) + (1-c)\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) + \frac{c\left(c\frac{1}{\alpha} - (\rho_- - \xi)\right)}{2}\|\boldsymbol{\beta}^{k-1} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|_2^2 - \frac{\frac{1}{\alpha} - \rho_+}{2}\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}\|_2^2
\end{aligned}
$$

Taking $c = \alpha\left(\rho_- - \xi\right)$, it implies

$$
\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) - \phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) \leq \left(1 - \alpha\left(\rho_- - \xi\right)\right)\left(\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) - \phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})\right) - \frac{\frac{1}{\alpha} - \rho_+}{2}\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}\|_2^2.
$$

Assume $\alpha \leq \frac{1}{\rho_+}$. Then

$$
\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) - \phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) \leq \left(1 - \alpha\left(\rho_- - \xi\right)\right)\left(\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) - \phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})\right).
$$

Taking $\boldsymbol{z} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ in Eq. 46 implies

$$
\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) \leq \phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) + \frac{\frac{1}{\alpha} - (\rho_- - \xi)}{2}\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \boldsymbol{\beta}^{k-1}\|_2^2 - \frac{1}{2\alpha}\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \boldsymbol{\beta}^k\|_2^2.
$$

By the optimality of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$, we have

$$
\frac{\frac{1}{\alpha} - (\rho_- - \xi)}{2}\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \boldsymbol{\beta}^{k-1}\|_2^2 - \frac{1}{2\alpha}\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \boldsymbol{\beta}^k\|_2^2 \geq 0,
$$

which implies

$$
\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \boldsymbol{\beta}^k\|_2^2 \leq \left(1 - \alpha\left(\rho_- - \xi\right)\right)\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \boldsymbol{\beta}^{k-1}\|_2^2.
$$

Finally, we derive the upper bound for $\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k)$. Since $\boldsymbol{\beta}^k = \arg\min_{\boldsymbol{z}} \psi_{\alpha,\boldsymbol{\lambda}}(\boldsymbol{z}, \boldsymbol{\beta}^{k-1})$, by the optimality condition, for each $j = 1, \cdots, d$, there exists a subgradient $\xi_j \in \partial\left|\beta_j^k\right|$ such that

$$
\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) + \frac{1}{\alpha}(\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi} = 0.
$$

By the definition of $\omega_{\boldsymbol{\lambda}}$ and combining it with the above equality,

$$
\begin{aligned}
\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) &= \min_{\boldsymbol{\xi}' \in \partial\|\boldsymbol{\beta}^k\|_1} \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\boldsymbol{\beta}^k - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta}^k - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda},1}} (\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}') \right\} \\
&\leq \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\boldsymbol{\beta}^k - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta}^k - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda},1}} (\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}) \right\} \\
&= \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\boldsymbol{\beta}^k - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta}^k - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda},1}} \left(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) + \frac{1}{\alpha}(\boldsymbol{\beta}^{k-1} - \boldsymbol{\beta}^k) \right) \right\} \\
&= \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \sum_{j=1}^d \frac{(\beta_j^k - \beta_j')\lambda_j}{\|\boldsymbol{\beta}^k - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda},1}} \frac{\left(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) + \frac{1}{\alpha}(\boldsymbol{\beta}^{k-1} - \boldsymbol{\beta}^k)\right)_j}{\lambda_j} \right\} \\
&\leq \frac{1}{\min(\boldsymbol{\lambda})}\|\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) + \frac{1}{\alpha}(\boldsymbol{\beta}^{k-1} - \boldsymbol{\beta}^k)\|_\infty \\
&\leq \frac{1}{\min(\boldsymbol{\lambda})}\|\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) + \frac{1}{\alpha}(\boldsymbol{\beta}^{k-1} - \boldsymbol{\beta}^k)\|_2 \\
&\leq \frac{1}{\min(\boldsymbol{\lambda})}\|\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})\|_2 + \frac{1}{\alpha}\|(\boldsymbol{\beta}^{k-1} - \boldsymbol{\beta}^k)\|_2 \\
\text{(by RSS)} \quad &\leq \frac{1}{\min(\boldsymbol{\lambda})}(\rho_+ + \frac{1}{\alpha})\|(\boldsymbol{\beta}^{k-1} - \boldsymbol{\beta}^k)\|_2.
\end{aligned}
$$

By Lemma F.1, $\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^k)\|_2 \le \sqrt{\frac{2}{2\frac{1}{\alpha}-(\rho_+)}\left(\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})-\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k)\right)}$. Therefore,

$$\begin{aligned}
\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) &\le \frac{1}{\min(\boldsymbol{\lambda})}(\rho_+ + \frac{1}{\alpha})\sqrt{\frac{2}{2\frac{1}{\alpha}-(\rho_+)}\left(\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})-\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k)\right)} \\
&\le \frac{1}{\min(\boldsymbol{\lambda})}(\rho_+ + \frac{1}{\alpha})\sqrt{\frac{2}{\rho_+}\left(\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})-\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k)\right)} \\
&\le \frac{1}{\min(\boldsymbol{\lambda})}(\rho_+ + \frac{1}{\alpha})\sqrt{\frac{2}{\rho_+}\left(\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})-\phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})\right)} \\
&\le \frac{1}{\min(\boldsymbol{\lambda})}(\rho_+ + \frac{1}{\alpha})\sqrt{\frac{2}{\rho_+}\left(1-\alpha\left(\rho_- - \xi\right)\right)^{k-1}\left(\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{0})-\phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})\right)}.
\end{aligned}$$

□

Lemma F.3 (Statistical ℓ_2 error)**.** *Assume Assumption F.1. If the following conditions are satisfied*

$$\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_0 \le s^* + 2\tilde{s}, \tag{47}$$

$$\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) \le \frac{21/2}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2, \tag{48}$$

$$\lambda_j \ge 8\left|\left[\nabla L_n(\boldsymbol{\beta}^*)\right]_j\right|, \tag{49}$$

then $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \le \min_{a\ge 0}\max\left\{\sqrt{21+\frac{17}{4}a}, \frac{21}{a}+\frac{17}{4}\right\}\frac{\|\boldsymbol{\lambda}_{S^*}\|_2}{\rho_- - \xi} \le \frac{15/2\|\boldsymbol{\lambda}_{S^*}\|_2}{\rho_- - \xi}$.

Proof. By the restricted strong convexity of $L_{\boldsymbol{\lambda}}$ in Lemma B.1,

$$\begin{aligned}
&\frac{\rho_- - \xi}{2}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 \\
&\le L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) - L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*)^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*) \\
&= \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) + (\|\boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1} - \|\boldsymbol{\beta}\|_{\boldsymbol{\lambda},1}) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*)^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*) \\
&\le \frac{21/2}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2 + \underbrace{(\|\boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1} - \|\boldsymbol{\beta}\|_{\boldsymbol{\lambda},1})}_{\text{(i)}}\underbrace{-\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*)^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)}_{\text{(ii)}}.
\end{aligned} \tag{50}$$

For the term (i),

$$\begin{aligned}
\|\boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1} - \|\boldsymbol{\beta}\|_{\boldsymbol{\lambda},1} &= \|\boldsymbol{\beta}^*_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} - \|\boldsymbol{\beta}_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} + (\|\boldsymbol{\beta}^*_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} - \|\boldsymbol{\beta}_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1}) \\
&= \|\boldsymbol{\beta}^*_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} - \|\boldsymbol{\beta}_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} - \|\boldsymbol{\beta}_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} \\
&\le \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} - \|\boldsymbol{\beta}_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} \\
&= \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} - \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1}.
\end{aligned}$$

For the term (ii),

$$\begin{aligned}
-\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*)^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*) &= -\nabla L_n(\boldsymbol{\beta}^*)^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*) - \nabla Q_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*)^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*) \\
\text{(By Eq. 49)} \quad &\le \frac{1}{8}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1} - \nabla Q_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*)^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*) \\
\text{(since } q'_{\lambda_j}(0)=0) \quad &= \frac{1}{8}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1} - \nabla Q_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*_{S^*})^\top(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*} \\
\text{(since } \left|q'_{\lambda_j}(\beta)\right| \le \lambda_j) \quad &\le \frac{1}{8}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1} + \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} \\
&= \frac{9}{8}\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} + \frac{1}{8}\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1}.
\end{aligned}$$

Combining the upper bounds of term (i) and term (ii) with Eq. 50, we have

$$\begin{aligned}
&\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 \\
&\leq \frac{21}{(\rho_- - \xi)^2}\|\boldsymbol{\lambda}_{S^*}\|_2^2 + \frac{17}{4(\rho_- - \xi)}\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} - \frac{7}{4(\rho_- - \xi)}\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S}^*}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1}, \quad (51)\\
&\leq \frac{1}{\rho_- - \xi}\left(\frac{21}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2 + \frac{17}{4}\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1}\right).
\end{aligned}$$

Let $a \geq 0$ be a constant. If $\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} \leq \frac{a}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2$, then

$$\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \leq \sqrt{21+\frac{17}{4}a}\frac{\|\boldsymbol{\lambda}_{S^*}\|_2}{\rho_- - \xi}.$$

If $\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} > \frac{a}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2$, then

$$\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2^2 \leq \frac{1}{\rho_- - \xi}\left(\frac{21}{a}+\frac{17}{4}\right)\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} \leq \left(\frac{21}{a}+\frac{17}{4}\right)\frac{\|\boldsymbol{\lambda}_{S^*}\|_2}{\rho_- - \xi}\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\|_2,$$

which implies $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \leq \left(\frac{21}{a}+\frac{17}{4}\right)\frac{\|\boldsymbol{\lambda}_{S^*}\|_2}{\rho_- - \xi}$. Combining the two cases, we know that

$$\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \leq \min_{a\geq 0}\max\left\{\sqrt{21+\frac{17}{4}a}, \frac{21}{a}+\frac{17}{4}\right\}\frac{\|\boldsymbol{\lambda}_{S^*}\|_2}{\rho_- - \xi}.$$

Taking $a = 7$ it is easy to derive that

$$\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 \leq \frac{15/2\|\boldsymbol{\lambda}_{S^*}\|_2}{\rho_- - \xi}.$$

□

Lemma F.4 (Retain in the restricted space)**.** *Assume Assumption F.1. Let $\boldsymbol{\beta}^0, \cdots, \boldsymbol{\beta}^k$ be a sequence of vectors obtained by the following update:*

$$\boldsymbol{\beta}^k = \texttt{MPG}(\boldsymbol{\beta}^{k-1}; \boldsymbol{\lambda}, \boldsymbol{w}, \alpha)$$

with a step sizes α in $[\alpha_{\min}, \frac{1}{\rho_+}]$. Assume the following conditions:

$$\begin{aligned}
\|(\boldsymbol{\beta}^0)_{\overline{S^*}}\|_0 &\leq \tilde{s},\\
\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^0) - \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) &\leq \frac{21/2}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2,\\
\lambda_j &\geq 8\left|[\nabla L_n(\boldsymbol{\beta}^*)]_j\right|,\\
\left(\frac{121}{\alpha_{\min}(\rho_- - \xi)} + 144\left(\frac{\rho_+}{\rho_- - \xi}\right)^2\right)\left(\frac{\|\boldsymbol{\lambda}_{S^*}\|_2}{\min(\boldsymbol{\lambda}_{\overline{S^*}})}\right)^2 &\leq \tilde{s},
\end{aligned}$$

where $\kappa := \frac{\rho_+}{\rho_- - \xi}$. Then for all $k = 1, \cdots, K$,

$$\|(\boldsymbol{\beta}^k)_{\overline{S^*}}\|_0 \leq \tilde{s},$$

and

$$\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) - \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) \leq \frac{21/2}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2.$$

Proof. Recall that $\boldsymbol{\beta}^k = \mathcal{T}_{\alpha\boldsymbol{\lambda}}\left(\bar{\boldsymbol{\beta}}^k\right)$ where

$$\begin{aligned}
\bar{\boldsymbol{\beta}}^k &= \boldsymbol{\beta}^{k-1} - \alpha\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})\\
&= \boldsymbol{\beta}^{k-1} - \alpha\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) + \alpha\left(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})\right).
\end{aligned}$$

To show $\left\|(\boldsymbol{\beta}^k)_{\overline{S^*}}\right\|_0 \leq \tilde{s}$, we need to prove that, for $j \in \overline{S^*}$, the number of j's such that $|\bar{\boldsymbol{\beta}}_j^k| > \alpha\lambda_j$ is no more than $\tilde{s}$. Note that

$$\begin{aligned}
|\bar{\boldsymbol{\beta}}_j^k| &\leq |\boldsymbol{\beta}_j^{k-1}| + \alpha\left|\left(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})\right)_j\right| + \alpha\left|[\nabla L_n(\boldsymbol{\beta}^*)]_j\right| \\
&\leq |\boldsymbol{\beta}_j^{k-1}| + \alpha\left|\left(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})\right)_j\right| + \frac{1}{8}\cdot\alpha\lambda_j.
\end{aligned}$$

Therefore, $\forall c \in [0,1]$,

$$\left\|(\boldsymbol{\beta}^k)_{\overline{S^*}}\right\|_0 \leq \mathrm{Card}\left(\{j \in \overline{S^*} : |\boldsymbol{\beta}_j^{k-1}| > c\frac{7}{8}\alpha\lambda_j\}\right) \tag{52}$$

$$+ \mathrm{Card}\left(\{j \in \overline{S^*} : \left|\left(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})\right)_j\right| > (1-c)\cdot\frac{7}{8}\lambda_j\}\right), \tag{53}$$

where $\mathrm{Card}\,()$ represents the size of the set. For the term in Eq. 52,

$$\begin{aligned}
\mathrm{Card}\left(\{j \in \overline{S^*} : |\boldsymbol{\beta}_j^{k-1}| > c\frac{7}{8}\alpha\lambda_j\}\right) &\leq \sum_{j\in\overline{S^*}} \frac{|\boldsymbol{\beta}_j^{k-1}|}{c\frac{7}{8}\alpha\lambda_j} = \frac{8}{c\cdot 7\alpha}\sum_{j\in\overline{S^*}}\frac{|\boldsymbol{\beta}_j^{k-1}|}{\lambda_j} \\
&\leq \frac{8}{c\cdot 7\alpha\min(\boldsymbol{\lambda}_{\overline{S^*}})^2}\|(\boldsymbol{\beta}^{k-1})_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} \\
&= \frac{8}{c\cdot 7\alpha\min(\boldsymbol{\lambda}_{\overline{S^*}})^2}\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1}
\end{aligned}$$

Bounding $\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_1$. Following the derivation of Eq. 51, we can obtain

$$\begin{aligned}
&(\rho_- - \xi)\|\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*\|_2^2 \\
&\leq \frac{21}{(\rho_- - \xi)}\|\boldsymbol{\lambda}_{S^*}\|_2^2 + \frac{17}{4}\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} - \frac{7}{4}\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)_{\overline{S}^*}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1},
\end{aligned}$$

which implies

$$\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)_{\overline{S}^*}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} \leq \frac{12}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2 + \frac{17}{7}\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1}.$$

Let $a \geq 0$ be a constant. If $\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} \leq \frac{a}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2$, then

$$\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)_{\overline{S}^*}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} \leq \left(12 + \frac{17a}{7}\right)\frac{\|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi}.$$

If $\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} > \frac{a}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2$, then

$$\begin{aligned}
\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)_{\overline{S}^*}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} &\leq \left(\frac{12}{a} + \frac{17}{7}\right)\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} \\
&\leq \left(\frac{12}{a} + \frac{17}{7}\right)\|\boldsymbol{\lambda}_{S^*}\|_2\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)\|_2 \\
\text{(Lemma F.3)} \quad &\leq \frac{15}{2}\left(\frac{12}{a} + \frac{17}{7}\right)\frac{\|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi}.
\end{aligned}$$

Combining the above two cases yields

$$\begin{aligned}
\|(\boldsymbol{\beta}^{k-1}-\boldsymbol{\beta}^*)_{\overline{S}^*}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} &\leq \min_{a\geq 0}\max\left\{12 + \frac{17a}{7}, \frac{15}{2}\left(\frac{12}{a} + \frac{17}{7}\right)\right\}\frac{\|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi} \\
&\leq \left(\frac{423}{14}\right)\frac{\|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi},
\end{aligned}$$

where the last inequality is obtained by taking $a = \frac{255}{34}$. Therefore, the term in Eq. 52 can be bounded by

$$\begin{aligned}
\text{Card}\left(\{j \in \overline{S^*} : |\boldsymbol{\beta}_j^{k-1}| > c\frac{7}{8}\alpha\lambda_j\}\right) &\leq \frac{8\|(\boldsymbol{\beta}^{k-1} - \boldsymbol{\beta}^*)_{\overline{S}^*}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1}}{c \cdot 7\alpha \min(\boldsymbol{\lambda}_{\overline{S^*}})^2} \\
&\leq \left(\frac{1692}{49c\alpha \min(\boldsymbol{\lambda}_{\overline{S^*}})^2}\right) \frac{\|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi} \\
&\leq \frac{1692}{49c\alpha(\rho_- - \xi)} \left(\frac{\|\boldsymbol{\lambda}_{S^*}\|_2)}{\min(\boldsymbol{\lambda}_{\overline{S^*}})}\right)^2.
\end{aligned}$$

Now we bound the term in Eq. 53 as follows. Denote the set as $S' := \{j \in \overline{S^*} : \left|\left(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})\right)_j\right| > (1-c) \cdot \frac{7}{8}\lambda_j\}$ and its size as $s' := |S'|$. Then

$$\begin{aligned}
s' &= \text{Card}\left(\left\{j \in \overline{S^*} : \left|\left(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})\right)_j\right| > (1-c) \cdot \frac{7}{8}\lambda_j\right\}\right) \\
&\leq \frac{1}{(1-c) \cdot \frac{7}{8}} \sum_{j \in S'} \frac{\left|\left(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})\right)_j\right|}{\lambda_j} \\
&\leq \frac{1}{(1-c) \cdot \frac{7}{8}} \sum_{j \in S'} \frac{\left|\left(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})\right)_j\right|}{\min(\boldsymbol{\lambda}_{\overline{S^*}})} \\
&\leq \frac{1}{(1-c) \cdot \frac{7}{8}} \frac{\sqrt{s'}}{\min(\boldsymbol{\lambda}_{\overline{S^*}})} \left\|\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) - \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1})\right\|_2 \\
&\leq \frac{1}{(1-c) \cdot \frac{7}{8}} \frac{\sqrt{s'}}{\min(\boldsymbol{\lambda}_{\overline{S^*}})} (\rho_+) \|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{k-1}\|_2.
\end{aligned}$$

The last inequality holds because $L_{\boldsymbol{\lambda}}$ is (ρ_+)-smooth in the restricted subspace by Lemma B.1. Applying Lemma F.3, we have

$$\begin{aligned}
s' &\leq \frac{64(\rho_+)^2}{49(1-c)^2 \min(\boldsymbol{\lambda}_{\overline{S^*}})^2} \|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{k-1}\|_2^2 \\
&\leq \frac{64(\rho_+)^2}{49(1-c)^2 \min(\boldsymbol{\lambda}_{\overline{S^*}})^2} \left(\frac{15/2\|\boldsymbol{\lambda}_{S^*}\|_2}{\rho_- - \xi}\right)^2 \leq \frac{3600}{49(1-c)^2} \left(\frac{\rho_+}{\rho_- - \xi}\right)^2 \left(\frac{\|\boldsymbol{\lambda}_{S^*}\|_2}{\min(\boldsymbol{\lambda}_{\overline{S^*}})}\right)^2.
\end{aligned}$$

Combining the above bounds for the terms in Eq. 52 and Eq. 53, we have

$$\begin{aligned}
\left\|(\boldsymbol{\beta}^k)_{\overline{S^*}}\right\|_0 &\leq \min_{c \in (0,1)} \left(\left(\frac{1692}{49c\alpha}\right) \frac{1}{\rho_- - \xi} + \frac{3600}{49(1-c)^2} \left(\frac{\rho_+}{\rho_- - \xi}\right)^2\right) \left(\frac{\|\boldsymbol{\lambda}_{S^*}\|_2}{\min(\boldsymbol{\lambda}_{\overline{S^*}})}\right)^2 \\
&\leq \left(\frac{121}{\alpha_{\min}(\rho_- - \xi)} + 144\left(\frac{\rho_+}{\rho_- - \xi}\right)^2\right) \left(\frac{\|\boldsymbol{\lambda}_{S^*}\|_2}{\min(\boldsymbol{\lambda}_{\overline{S^*}})}\right)^2 \\
&\leq \tilde{s}
\end{aligned}$$

The second last inequality is obtained by taking $c = \frac{5}{7}$.

Finally, if we can show $\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) - \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) \leq \frac{21/2}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2$, then we can show by math induction that $\left\|(\boldsymbol{\beta}^k)_{\overline{S^*}}\right\|_0 \leq \tilde{s}$ for all k.

Taking $\boldsymbol{z} = \boldsymbol{\beta}^{k-1}$ in Eq. 46, we have

$$\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) \leq \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) - \frac{2\frac{1}{s} - \rho_+}{2} \left\|\boldsymbol{\beta}^k - \boldsymbol{\beta}^{k-1}\right\|_2^2 \leq \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}).$$

Therefore,

$$\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^k) - \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) \leq \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^{k-1}) - \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) \leq \frac{21/2}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2.$$

□

Lemma F.5 (Statistical error)**.** *Assume Assumption F.1. If*

$$\begin{aligned}&\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) \le \epsilon, \quad \textit{with } \epsilon \le 1/2,\\ &\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_0 \le s^* + 2\tilde{s},\\ &\lambda_j \ge 8\left|[L_n(\boldsymbol{\beta}^*)]_j\right|,\end{aligned}$$

then the following relations hold true.

$$\begin{aligned}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2 &\le \frac{\left(\epsilon+\frac{17}{8}\right)\|\boldsymbol{\lambda}_{S^*}\|_2}{\rho_- - \xi} \le \frac{21/8}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2,\\ \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) &\le \frac{3\epsilon(\epsilon+17/8)}{7/8-\epsilon}\frac{\|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi} \le \frac{21/2}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2.\\ \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) &\le \epsilon\left(\frac{3\left(\epsilon+17/8\right)}{7/8-\epsilon}+51/7\right)\frac{\|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi} \le \frac{29/2}{\rho_- - \xi}\|\boldsymbol{\lambda}_{S^*}\|_2^2.\end{aligned}$$

Proof. Since $\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_0 \le s^* + 2\tilde{s}$, then by the RSC and RSS in Lemma B.1,

$$(\rho_- - \xi)\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_2^2 \le (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) - (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*). \tag{54}$$

Let $\boldsymbol{\xi} \in \partial\|\boldsymbol{\beta}\|_1$ be the subgradient that attains the minimum in $\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$. Adding and subtracting $(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top(\boldsymbol{\lambda}\circ\boldsymbol{\xi})$ to the right-hand side of Eq. 54 yields

$$\begin{aligned}(\rho_- - \xi)\|\boldsymbol{\beta}^*-\boldsymbol{\beta}\|_2^2 \le &\underbrace{(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \boldsymbol{\lambda}\circ\boldsymbol{\xi})}_{\text{(i)}} - \underbrace{(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*)}_{\text{(ii)}}\\ &- \underbrace{(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top(\boldsymbol{\lambda}\circ\boldsymbol{\xi})}_{\text{(iii)}}.\end{aligned}$$

Now we bound the terms (i)-(iii).

Bound (i) using $\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$. Since $\boldsymbol{\xi}$ attains the minimum in $\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta})$, by definition,

$$\omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) = \max_{\boldsymbol{\beta}'\in\Omega}\left\{\frac{(\boldsymbol{\beta}-\boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_{\boldsymbol{\lambda},1}}(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi})\right\}.$$

Therefore,

$$\text{(i)} = (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top(\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \lambda\boldsymbol{\xi}) \le \omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta})\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1} \le \epsilon\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1}.$$

Bound (ii). Note that $L_{\boldsymbol{\lambda}} = L_n + Q_{\boldsymbol{\lambda}}$. Thus,

$$\begin{aligned}|\text{(ii)}| &= \left|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \nabla L_n(\boldsymbol{\beta}^*) + (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \nabla Q_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*)\right|\\ &\le \left|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \nabla L_n(\boldsymbol{\beta}^*)\right| + \left|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \nabla Q_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*)\right|\end{aligned}$$

Since $\lambda_j \ge 8\left|[L_n(\boldsymbol{\beta}^*)]_j\right|$, we have

$$|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \nabla L_n(\boldsymbol{\beta}^*)| \le \frac{1}{8}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1}.$$

Since $q'_{\lambda_j}(0) = 0$ and $q'_{\lambda_j}(\beta) \le \lambda_j$ (Assumption B.1), it holds true that

$$\begin{aligned}|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top \nabla Q_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*)| &= |(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}^\top(\nabla Q_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*))_{S^*}|\\ &\le \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1}.\end{aligned}$$

Therefore,

$$|\text{(ii)}| \le \frac{1}{8}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1} + \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1}.$$

Bound (iii) by the definition of subgradient.

$$\begin{aligned}\text{(iii)} &= (\boldsymbol{\beta}-\boldsymbol{\beta}^*)^\top(\boldsymbol{\lambda}\circ\boldsymbol{\xi})\\ &= \langle\boldsymbol{\lambda}_{S^*}\circ(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}, \boldsymbol{\xi}_{S^*}\rangle + \langle\boldsymbol{\lambda}_{\overline{S^*}}\circ(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}, \boldsymbol{\xi}_{\overline{S^*}}\rangle.\end{aligned}$$

By Holder's inequality and the fact $\|\boldsymbol{\xi}\|_\infty \leq 1$,

$$\langle \boldsymbol{\lambda}_{S^*} \circ (\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}, \boldsymbol{\xi}_{S^*} \rangle \geq -\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1}.$$

Since $\boldsymbol{\xi} \in \partial\|\boldsymbol{\beta}\|_1$ and that $\boldsymbol{\beta}^*_{\overline{S^*}} = \mathbf{0}$,

$$\langle \boldsymbol{\lambda}_{\overline{S^*}} \circ (\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}, \boldsymbol{\xi}_{\overline{S^*}} \rangle = \langle \boldsymbol{\lambda}_{\overline{S^*}} \circ \boldsymbol{\beta}_{\overline{S^*}}, \boldsymbol{\xi}_{\overline{S^*}} \rangle = \|\boldsymbol{\beta}_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} = \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1}. \tag{55}$$

Combining the above three equations, it holds true that

$$\text{(iii)} \geq -\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} + \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1}.$$

Combine (i)-(iii). Combining the bounds for (i)-(iii), it holds true that

$$\begin{aligned}
&(\rho_- - \xi)\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_2^2 \\
&\leq \left(\epsilon + \frac{1}{8} + 1 + 1\right) \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} - \left(1 - \epsilon - \frac{1}{8}\right) \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} \\
&= \left(\epsilon + \frac{17}{8}\right) \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} - \underbrace{\left(\frac{7}{8} - \epsilon\right)}_{\geq 0} \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} \\
&\leq \left(\epsilon + \frac{17}{8}\right) \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1}.
\end{aligned} \tag{56}$$

Using the fact that $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} \leq \|\boldsymbol{\lambda}_{S^*}\|_2 \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_2$, it implies the first conclusion:

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|_2 \leq \frac{\left(\epsilon + \frac{17}{8}\right) \|\boldsymbol{\lambda}_{S^*}\|_2}{\rho_- - \xi}. \tag{57}$$

Now we prove the objective value. Since $L_{\boldsymbol{\lambda}}$ is convex on $\boldsymbol{\beta}, \boldsymbol{\beta}^*$ and $\|\cdot\|_1$ is convex,

$$\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) \geq \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + (\boldsymbol{\beta}^* - \boldsymbol{\beta})^\top (\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}),$$

which implies

$$\begin{aligned}
\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) &\leq (\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top (\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi}) \\
&\leq \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1} \omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) \leq \epsilon \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1}
\end{aligned} \tag{58}$$

Now we bound the norm $\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1}$ as follows. By triangle inequality,

$$\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1} \leq \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} + \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1}. \tag{59}$$

To bound $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1}$, moving the term $\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1}$ in Eq. 56 to the left hand side yields

$$(7/8 - \epsilon)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} \leq (\epsilon + 17/8)\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1},$$

which implies

$$\|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{\overline{S^*}}\|_{\boldsymbol{\lambda}_{\overline{S^*}},1} \leq \frac{\epsilon + 17/8}{7/8 - \epsilon} \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1}. \tag{60}$$

Plugging it into Eq. 59, we have

$$\begin{aligned}
\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1} &\leq \left(1 + \frac{\epsilon + 17/8}{7/8 - \epsilon}\right) \|(\boldsymbol{\beta} - \boldsymbol{\beta}^*)_{S^*}\|_{\boldsymbol{\lambda}_{S^*},1} \leq \left(\frac{3}{7/8 - \epsilon}\right) \|\boldsymbol{\lambda}_{S^*}\|_2 \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \\
\text{(by Eq. 57)} \ &\leq \left(\frac{3}{7/8 - \epsilon}\right) \|\boldsymbol{\lambda}_{S^*}\|_2 \frac{\left(\epsilon + \frac{17}{8}\right) \|\boldsymbol{\lambda}_{S^*}\|_2}{\rho_- - \xi} \\
&= \frac{3(\epsilon + 17/8)}{7/8 - \epsilon} \frac{\|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi}.
\end{aligned} \tag{61}$$

Combining it with Eq. 58, we have

$$\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}^*) \leq \frac{3\epsilon(\epsilon + 17/8)}{7/8 - \epsilon} \frac{\|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi}.$$

Finally, we derive the bound for the term $\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})$. Since $\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|_0 \leq s^* + 2\tilde{s}$ and $\phi_{\boldsymbol{\lambda}}$ is convex on the restricted space, then

$$\begin{aligned}
&\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) \leq (\nabla L_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) + \boldsymbol{\lambda} \circ \boldsymbol{\xi})^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) \\
&\leq \omega_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|_{\boldsymbol{\lambda},1} \\
&\leq \epsilon \left(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\lambda},1} + \|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|_{\boldsymbol{\lambda},1} \right) \\
\text{(by Eq. 61)} \quad &\leq \epsilon \left(\frac{3(\epsilon + 17/8)}{7/8 - \epsilon} \frac{\|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi} + \|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|_{\boldsymbol{\lambda},1} \right). \qquad (62)
\end{aligned}$$

Now we only need to bound the term $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|_{\boldsymbol{\lambda},1}$. We can derive its bound following the same steps as we derive Eq. 61 and using the fact that $\omega_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) = 0$, which will imply that

$$\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} - \boldsymbol{\beta}^*\|_1 \leq \frac{3\,(0 + 17/8)}{7/8 - 0} \frac{\|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi} = \frac{51/7 \|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi}.$$

Plugging it into Eq. 62, we have

$$\begin{aligned}
&\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) \\
&\leq \epsilon \left(\frac{3\,(\epsilon + 17/8)}{7/8 - \epsilon} + 51/7 \right) \frac{\|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi}.
\end{aligned}$$

Assuming $\epsilon \leq 1/2$, then

$$\phi_{\boldsymbol{\lambda}}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) \leq \frac{29/2 \|\boldsymbol{\lambda}_{S^*}\|_2^2}{\rho_- - \xi}.$$

□

Lemma F.6 (Optimality when λ decreases)**.** *Assume Assumption F.1. If*

$$\omega_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta}) \leq \epsilon_{t-1}, \qquad (63)$$

$$\boldsymbol{\lambda}_t = \boldsymbol{\eta} \circ \boldsymbol{\lambda}_{t-1} \quad \textit{with} \quad \eta_j \in [0.9, 1], \qquad (64)$$

$$\|\boldsymbol{\beta}\|_0 \leq \tilde{s}, \qquad (65)$$

then

$$\begin{aligned}
&\omega_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) \leq \frac{\epsilon_{t-1} + 0.2}{0.9} \\
&\phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}_t}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t}) \leq \left(\frac{3\,(\epsilon_{t-1} + 17/8)}{0.9(7/8 - \epsilon_{t-1})} + 51/7 \right) \frac{\epsilon_{t-1} + 0.2}{0.9} \frac{\|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi}.
\end{aligned}$$

If $\epsilon_{t-1} \leq 1/4$, then

$$\begin{aligned}
&\omega_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) \leq \frac{1}{2} \\
&\phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}_t}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t}) \leq \frac{10 \|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi}.
\end{aligned}$$

Proof. Recall the definition of $\omega_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta})$.

$$\omega_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta}) = \min_{\boldsymbol{\xi}' \in \partial \|\boldsymbol{\beta}\|_1} \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_{t-1},1}} (\nabla L_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta}) + \boldsymbol{\lambda}_{t-1} \boldsymbol{\xi}') \right\}$$

Let $\boldsymbol{\xi} \in \partial \|\boldsymbol{\beta}\|_1$ be the subgradient that attains the minimum in $\omega_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta})$. Then

$$\omega_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta}) = \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_{t-1},1}} (\nabla L_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta}) + \boldsymbol{\lambda}_{t-1} \boldsymbol{\xi}) \right\}.$$

Now we consider $\boldsymbol{\lambda}_t = \boldsymbol{\eta} \circ \boldsymbol{\lambda}_{t-1}$. By definition,

$$\begin{aligned}\omega_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) &= \min_{\boldsymbol{\xi}' \in \partial \|\boldsymbol{\beta}\|_1} \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_t,1}} (\nabla L_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) + \boldsymbol{\lambda}_t \boldsymbol{\xi}') \right\} \\ &\leq \max_{\boldsymbol{\beta}' \in \Omega} \left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_t,1}} (\nabla L_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) + \boldsymbol{\lambda}_t \boldsymbol{\xi}) \right\}.\end{aligned}$$

Note that

$$\begin{aligned}\nabla L_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) + \boldsymbol{\lambda}_t \boldsymbol{\xi} = &\left(\nabla L_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta}) + \boldsymbol{\lambda}_{t-1} \boldsymbol{\xi}\right) \\ &+ (\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}) \circ \boldsymbol{\xi} \\ &+ \left(\nabla Q_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) - \nabla Q_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta})\right),\end{aligned}$$

which implies

$$\begin{aligned}\omega_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) \leq &\max_{\boldsymbol{\beta}' \in \Omega} \underbrace{\left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_t,1}} \left(\nabla L_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta}) + \boldsymbol{\lambda}_{t-1} \boldsymbol{\xi}\right) \right\}}_{\text{(i)}} \\ &+ \max_{\boldsymbol{\beta}' \in \Omega} \underbrace{\left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_t,1}} (\boldsymbol{\lambda}_t - \boldsymbol{\lambda}_{t-1}) \circ \boldsymbol{\xi} \right\}}_{\text{(ii)}} \\ &+ \max_{\boldsymbol{\beta}' \in \Omega} \underbrace{\left\{ \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_t,1}} \left(\nabla Q_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) - \nabla Q_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta})\right) \right\}}_{\text{(iii)}}.\end{aligned}$$

Using the fact that $\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_t,1} \geq 0.9 \|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_{t-1},1}$, term (i) can be bounded as follows.

$$\text{(i)} \leq \frac{1}{0.9} \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}')^\top}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_{t-1},1}} \left(\nabla L_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta}) + \boldsymbol{\lambda}_{t-1} \boldsymbol{\xi}\right) \leq \frac{1}{0.9} \omega_{\boldsymbol{\lambda}_{t-1}}(\boldsymbol{\beta}) \leq \frac{\epsilon_{t-1}}{0.9}.$$

Using the fact $|\xi_j| \leq 1$ and the fact that $[\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t]_j = \left(\frac{1}{\eta_j} - 1\right) [\boldsymbol{\lambda}_t]_j \leq \left(\frac{1}{0.9} - 1\right) [\boldsymbol{\lambda}_t]_j$, term (ii) can be bounded as follows.

$$\text{(ii)} \leq \sum_{j=1}^{d} \frac{|\boldsymbol{\beta}_j - \boldsymbol{\beta}'_j|}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_t,1}} \left(\frac{1}{0.9} - 1\right) [\boldsymbol{\lambda}_t]_j = \frac{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_t,1}}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_t,1}} \left(\frac{1}{0.9} - 1\right) = \frac{1}{0.9} - 1.$$

Term (iii) is bounded similarly as term (ii). Since $\left|q_{\lambda_j}(\beta) - q_{\lambda'_j}(\beta)\right| \leq |\lambda_j - \lambda'_j|$, we have

$$\text{(iii)} \leq \sum_{j=1}^{d} \frac{|\boldsymbol{\beta}_j - \boldsymbol{\beta}'_j|}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_t,1}} [\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}_t]_j \leq \sum_{j=1}^{d} \frac{|\boldsymbol{\beta}_j - \boldsymbol{\beta}'_j|}{\|\boldsymbol{\beta} - \boldsymbol{\beta}'\|_{\boldsymbol{\lambda}_t,1}} \left(\frac{1}{0.9} - 1\right) [\boldsymbol{\lambda}_t]_j = \frac{1}{0.9} - 1.$$

Combining the bounds for (i)-(iii), we have

$$\omega_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) \leq \frac{\epsilon_{t-1} + 2}{0.9} - 2 = \frac{\epsilon_{t-1} + 0.2}{0.9}$$

The first part of the proof is finished. Now we bound the term $\phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}_t}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t})$. Since $\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}\|_0 \leq s^* + 2\tilde{s}$ and $\phi_{\boldsymbol{\lambda}}$ is convex on the restricted space, then

$$\begin{aligned}&\phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}_t}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t}) \\ &\leq (\nabla L_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) + \boldsymbol{\lambda}_t \circ \boldsymbol{\xi})^\top (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t}) \\ &\leq \omega_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t}\|_{\boldsymbol{\lambda}_t,1} \\ &\leq \frac{\epsilon_{t-1} + 0.2}{0.9} \left(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_{\boldsymbol{\lambda}_t,1} + \|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t}\|_{\boldsymbol{\lambda}_t,1}\right) \qquad (66)\end{aligned}$$

Following the same way we derive Eq. 60, we can obtain that

$$\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_{(\boldsymbol{\lambda}_t)_{\overline{S^*}},1} \leq \frac{\epsilon_{t-1}+17/8}{7/8-\epsilon_{t-1}}\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{(\boldsymbol{\lambda}_t)_{S^*},1}.$$

Therefore,

$$\begin{aligned}
\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_{\boldsymbol{\lambda}_t,1} &= \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\overline{S^*}}\|_{(\boldsymbol{\lambda}_t)_{\overline{S^*}},1} + \|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{(\boldsymbol{\lambda}_t)_{S^*},1} \qquad (67)\\
&\leq \frac{3}{7/8-\epsilon_{t-1}}\|(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{S^*}\|_{(\boldsymbol{\lambda}_t)_{S^*},1}\\
&\leq \frac{3\|(\boldsymbol{\lambda}_t)_{S^*}\|_2}{7/8-\epsilon_{t-1}}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|_2\\
\text{(By Lemma F.5)} \quad &\leq \frac{3\|(\boldsymbol{\lambda}_t)_{S^*}\|_2}{7/8-\epsilon_{t-1}}\frac{\left(\epsilon_{t-1}+\frac{17}{8}\right)\|(\boldsymbol{\lambda}_{t-1})_{S^*}\|_2}{\rho_- - \xi}\\
&= \frac{3\left(\epsilon_{t-1}+17/8\right)}{0.9(7/8-\epsilon_{t-1})}\frac{\|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi}. \qquad (68)
\end{aligned}$$

Using the same arguments for the term $\|\boldsymbol{\beta}^* - \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t}\|_{\boldsymbol{\lambda}_t,1}$ and using the fact that $\omega_{\boldsymbol{\lambda}_t}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t}) \leq 0 \cdot \boldsymbol{\lambda}_t$, we obtain that

$$\|\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t} - \boldsymbol{\beta}^*\|_1 \leq \frac{3\left(0+17/8\right)}{7/8-0}\frac{\|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi} = \frac{51/7\|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi}.$$

Combining this inequality with Eq. 66 and Eq. 68, we have

$$\begin{aligned}
&\phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}_t}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t})\\
&\leq \left(\frac{3\left(\epsilon_{t-1}+17/8\right)}{0.9(7/8-\epsilon_{t-1})} + 51/7\right)\frac{\epsilon_{t-1}+0.2}{0.9}\frac{\|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi}.
\end{aligned}$$

Assuming $\epsilon_{t-1} \leq 1/4$, then

$$\phi_{\boldsymbol{\lambda}_t}(\boldsymbol{\beta}) - \phi_{\boldsymbol{\lambda}_t}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}_t}) \leq \frac{10\|(\boldsymbol{\lambda}_t)_{S^*}\|_2^2}{\rho_- - \xi}.$$

□

Lemma F.7 (Coercivity of the gradient)**.** *[Lemma 3.11 in Bubeck (2014)] Let f be a-strongly convex b-smooth and on $\mathcal{B} \subseteq \mathbb{R}^d$. Then for all $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \mathcal{B}$, one has*

$$\langle \nabla f(\boldsymbol{\beta}) - \nabla f(\boldsymbol{\beta}'), \boldsymbol{\beta} - \boldsymbol{\beta}\rangle \geq \frac{ab}{a+b}\|\boldsymbol{\beta}-\boldsymbol{\beta}'\|_2^2 + \frac{1}{a+b}\|\nabla f(\boldsymbol{\beta}) - \nabla f(\boldsymbol{\beta}')\|_2^2.$$

G EXPERIMENT DETAILS

G.1 IMPLEMENTATION

ALISTA: We follow the implementation in Liu et al. (2019a).

RNN: The RNN is designed following Andrychowicz et al. (2016). Initialize the β_0 as 0. At every step, we compute the gradient $\nabla_\beta L(\beta_t)$ w.r.t. to objective as the input to a LSTM with hidden dimension equals to 128. Then we use the output of the LSTM as the increment:

$$\beta_{t+1} = \beta_t + \text{LSTM}(\nabla_\beta L(\beta_t)) \tag{69}$$

We repeat this iteration 20 steps and use β_{20} as our final output.

RNN-ℓ_1: RNN-ℓ_1 use the same architecture as RNN. The only difference is we add an extra learnable parameter λ_t for soft thresholding at each step:

$$\beta_{t+1} = \beta_t + \text{LSTM}(\nabla_\beta L(\beta_t)) \tag{70}$$

$$\beta_{t+1} = \eta_{\lambda_t}(\beta_{t+1}) \tag{71}$$

where the softthresholding η_λ is an elementwise operator maps $\eta_\lambda(x) = 0$ is $|x| < \lambda$ and $\eta_\lambda(x) = \text{sign}(x)(|x| - \lambda)$.

APF: We follow the implementation in Wang et al. (2014).

G.2 TRAINING AND EVALUATION

We trained the neural networks to minimize the weighted l2 loss. That's to say, given the sequence of output $(\beta_1, ..., \beta_T)$, the loss is computed as:

$$L = \sum_{i=1}^{T} \gamma^{T-i} \|\beta_i - \beta^*\|_2^2 \tag{72}$$

where $\gamma = 0.5$. We use training set containing 2000 problems and valid set containing 200 instances. We use optimizer Adam (Kingma & Ba, 2014) to train 500 epochs with learning rate 1e-4 and select the model based on the l2 loss on 200 valid problems.

The evaluation is performed on a server with CPU: Intel(R) Xeon(R) Silver 4116 CPU @ 2.10GHz, GPU: Nvidia GTX 2080TI, Memory 264G, in single thread.