Towards Robust Learning to Optimize with Theoretical Guarantees

Qingyu Song CUHK
Wei Lin CUHK
Juncheng Wang HKBU
Hong Xu CUHK

Abstract

Learning to optimize (L2O) is an emerging technique to solve mathematical optimization problems with learning-based methods. Although with great success in many real-world scenarios such as wireless communications, computer networks, and electronic design, existing L2O works lack theoretical demonstration of their performance and robustness in out-of-distribution (OOD) scenarios. We address this gap by providing comprehensive proofs. First, we prove a sufficient condition for a robust L2O model with homogeneous convergence rates over all In-Distribution (InD) instances. We assume an L2O model achieves robustness for an InD scenario. Based on our proposed methodology of aligning OOD problems to InD problems, we also demonstrate that the L2O model's convergence rate in OOD scenarios will deteriorate by an equation of the L2O model's input features. Moreover, we propose an L2O model with a concise gradient-only feature construction and a novel gradient-based history modeling method. Numerical simulation demonstrates that our proposed model outperforms the state-of-the-art baseline in both InD and OOD scenarios and achieves up to 10 $\times$ convergence speedup. The code of our method can be found from https://github.com/NetX-lab/GoMathL2O-Official.

1. Introduction

Learning to Optimize (L2O) is a promising new approach in applying learning-based methods to tackle optimization problems. In particular, L2O concentrates on problems with well-defined objective functions and constraints [7]. Thus, black-box optimization strategies, such as Bayesian Optimization [24], typically fall outside its scope. L2O has shown benefits in problems from various domains, including LASSO regression in sparse coding using multilayer perceptrons [8], and utility maximization in resource allocation wherein neural networks (NN) serve to approximate the expensive matrix inversion [11].

L2O can be categorized into three main types: black-box [6, 22, 26, 31], algorithm-unrolling [11, 21, 33], and math-inspired [9, 14]. Black-box L2O approaches the optimization problem as a traditional pattern recognition task, approximating a mapping function from manually constructed features to the solutions [26]. Algorithm-unrolling L2O leverages well-defined algorithms, such as gradient descent [19], to approximate the solutions of complex calculations. Besides, much research has gone into explainable and trustworthy L2O. For example, Heaton et al. [9] employ an existing algorithm to prevent the L2O model from entering irrecoverable areas. Liu et al. [14] introduce a mathematics-driven L2O (Math-L2O) framework for convex optimization, offering a general workflow for formulating an L2O model. Despite empirical results, a theoretical analysis on the robustness of L2O models under out-of-distribution (OOD) conditions is still missing in [14].

OOD generalization for L2O has emerged as a vital issue, often considered more critical in L2O than in other deep learning applications [23]. For L2O, OOD's challenge involves resolving previously unseen problems, potentially involving novel optimization problems with unique objectives [30]. Guaranteeing convergence in OOD scenarios remains elusive. For instance, a model's output in an OOD scenario could potentially veer into unpredictable areas when the domain changes significantly to an InD scenario.

Numerous efforts have been made to enhance the robustness of L2O models in training. Lv et al. [16] employ data augmentation to prevent L2O models from overfitting to specific tasks. Almeida et al. [2] transform the L2O model into a hyperparameter tuner for existing optimization algorithms. Wichrowska et al. [29] focus on minimizing parameters in NNs and assembling heterogeneous optimization tasks. Liu et al. [14] try to regularize L2O models with inspirations from existing algorithms. However, these studies predominantly aim to mitigate the limitations inherent in existing L2O methods, with no comprehensive analysis conducted on the impact of OOD on the deterioration of convergence. This gap in the literature motivates us to quantify this deterioration with rigorous analysis.

The central thesis of this paper is to propose a general and robust L2O model for both InD and OOD scenarios. Chiefly, we first investigate L2O's convergence behavior in InD contexts and derive the criteria for a uniformly robust

model applicable to all InD instances. Then, we characterize L2O's degradation of convergence under OOD conditions, presenting our findings as a series of corollaries. The main contributions of this paper are as follows.

1. We propose a methodology to link the L2O model's performances in InD and OOD situations based on the Math-L2O approach from Liu et al. [14]. First, we construct a virtual feature by subtracting the L2O model's input feature in InD from that in OOD. We then compute the corresponding difference in the model's outputs by applying this virtual feature. To depict a comprehensive deviation of OOD from InD, we align the variable sequence from an OOD situation with that from InD and construct a trajectory of virtual features. We use this trajectory to illustrate OOD's divergence from InD and then conduct theoretical analyses.
2. We establish the criteria for a robust L2O model in an InD setting and examine its response to OOD. First, we present a sufficient condition to guarantee a homogeneous convergence improvement in each iteration, confirming robustness in InD scenarios. Then, we derive the equations describing convergence gain in a single iteration and the overall convergence rate of the entire sequence relative to our proposed virtual feature. A collection of theorems and observations underscore that the magnitude of virtual features inherently exacerbates the deterioration of convergence in OOD situations.
3. Based on our theoretical insights, we propose a robust L2O model, GO-Math-L2O, that exclusively employs gradients as input features. This gradient-only approach enables a more concise virtual feature in OOD settings. We introduce a new gradient-only history modeling technique to model the optimization process's historical sequence. This method employs gradient (and subgradient) values as status indicators to modulate updates provided by the L2O model. We propose to recover the historical subgradient from an inversible model definition, thus eliminating the ambiguity of subgradient selection.
4. Through numerical experiments, we show that GO-Math-L2O outperforms state-of-the-art (SOTA) L2O models on convergence and optimality across both InD and OOD scenarios. Following training with a synthetic dataset, we deploy various OOD test cases with identical optimal values. Our proposed model's convergence speed is up to $10\times$ faster than SOTA L2O models in OOD scenarios.

The rest of this paper is organized as follows. In Sec. 2, we define OOD problems for L2O. In Sec. 3, we propose a method to quantify the solutions given by an L2O model in OOD scenarios. Then, in Sec. 4, we derive the convergence rate of an L2O model in OOD scenarios. Based on this, we propose our robust GO-Math-L2O model in Sec. 5. We empirically verify the proposed model with simulations in Sec. 6, and conclude the work in Sec. 7.

Notations: A smooth convex function and a non-smooth convex function are denoted by f and r, respectively. NNs' input vectors are denoted by z and z'. Variables of an optimization problem are denoted by x and x'. The optimal solution is denoted by x^*. An iteration and stopping iteration are denoted by k and K, respectively. A smooth gradient at x_k and a set of subgradients at x_k are denoted by $\nabla f(x_k)$ and $\partial r(x_k)$, respectively. A subgradient value of $\partial r(x_k)$ is denoted by g_k. Frobenius norm for a matrix and $L2$-norm for a vector is denoted by $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|$ respectively. Transpose is defined by $^\top$. The maximum length of history modeling is denoted by T. The Jacobian matrix of a vector-to-vector function is denoted by $\mathbf{J}$. An L2O model is denoted by d. A NN is denoted by operator $\mathbf{N}$.

2. Definitions

In this section, we first introduce the objective of the L2O problem. We then introduce the Math-L2O model in [14], whose iterative updates are defined by NNs. Last, we define the domains for both InD and OOD scenarios, which leads to the definitions of InD L2O and OOD L2O problems.

2.1. Optimizee (Optimization Objective)

Consider function $F(x) = f(x) + r(x)$. Here, $f(x)$ is a L-smooth function, and $r(x)$ is a non-smooth function. They are defined within the following function spaces:

$$\begin{aligned}\mathcal{F}_L\left(\mathbb{R}^n\right) =& \{f : \mathbb{R}^n \to \mathbb{R} | f \text{ is convex, differentiable, and} \\ & \|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|, \forall x, y \in \mathbb{R}^n\}, \\ \mathcal{F}\left(\mathbb{R}^n\right) =& \{r : \mathbb{R}^n \to \mathbb{R} | r \text{ is proper, closed, and convex}\}.\end{aligned}$$

We assume $r(x)$ is sub-differentiable, with its subgradient set at any point x defined below:

$$\partial r(x) = \{g \in \mathbb{R}^n \mid r(y) - r(x) \geq g^\top (y-x), \forall x, y \in \mathbb{R}^n\}.$$

We note here that the above optimization objective applies to both the InD and the OOD scenarios.

2.2. Optimizor (L2O Model)

Denote the L2O model as $d(z)$, where the input vector space is designated as $\mathcal{Z}$ such that $z \in \mathcal{Z} \subseteq \mathbb{R}^m$. We define $d(z)$ as a function mapping within the given function space [14]:

$$\begin{aligned}\mathcal{D}_C(\mathcal{Z}) = \{d : \mathcal{Z} \to \mathbb{R}^n \mid & d \text{ is differentiable}, \\ & \|\mathbf{J}_{d(z)}\|_{\mathrm{F}} \leq C, \forall z \in \mathcal{Z}, C \in \mathbb{R}^+\}.\end{aligned} \tag{1}$$

We choose features from x and $F(x)$ to define z, offering a wide range of feasible options. For instance, z could be defined with the optimization variable and its gradient as $\left[x^\top, \nabla f(x)^\top\right]^\top$ in [14]. Different from [14], we propose to define z solely as $\nabla f(x)$ to improve convergence

in OOD scenarios. From our experimental results, our approach achieves near-optimal solutions in some OOD cases and more robust performance than SOTA baselines in all OOD scenarios. Moreover, Corollaries 2 and 3 theoretically demonstrate the outperformance over the method in [14].

$d(z)$ iteratively updates the optimization variable. At each iteration k, given the previous variable $x_{k-1} \in \mathbb{R}^n$ and the input vector z_{k-1} for the L2O model, $d(z_{k-1})$ updates x_k as follows:

$$x_k = x_{k-1} - d(z_{k-1}). \tag{2}$$

2.3. InD and OOD Problems

The InD and OOD problems share the same space of optimization objective defined in Sec. 2.1 but with different optimization objectives or variable domains. Consider a convex and compact set, $\mathcal{S}_P \subseteq \mathbb{R}^n$. The complementary set of $\mathcal{S}_P$ is denoted as $\mathcal{S}_O$, such that $\mathcal{S}_O := \mathbb{R}^n \backslash \mathcal{S}_P$. We also suppose the existence of two function sets: $\mathcal{F}_{L,P} \subseteq \mathcal{F}_L(\mathbb{R}^n)$ and $\mathcal{F}_P \subseteq \mathcal{F}(\mathbb{R}^n)$. We define InD optimization problems as follows:

$$\min_x F(x), \tag{P}$$

where $x \in \mathcal{S}_P$, $F(x) = f(x) + r(x)$, $f \in \mathcal{F}_{L,P}$, and $r \in \mathcal{F}_P$. The dataset employed for training an L2O model is derived from a specific domain of x, f, and r. Consider an L2O model $d(z)$ that has undergone training with a domain of x, f, and r, sampled from Problem P. We then define the *InD L2O Problem* as: Given any initial point $x_0 \in \mathcal{S}_P$, using $d(z)$ to iteratively update x_0 in order to find a solution for any arbitrary InD problem as depicted in Problem P.

Note that instances outside this domain potentially yield more erroneous $d(z)$ outputs. Furthermore, non-learning algorithms, such as gradient descent, have demonstrated robustness across all domains [19]. One of the main goals of this paper is to propose an L2O model that is robust to OOD.

We characterize OOD in the context of L2O in the optimization objective's domain. We define the *OOD L2O Problem* as: Consider an L2O model $d(z)$ that has undergone training with a domain of x, f, and r, sampled from Problem P, using $d(z)$ to iteratively update $x'_0 \in \mathcal{S}_O$ in order to a solution for any following problem:

$$\min_{x'} F'(x'). \tag{O}$$

where $F'(x') = f'(x') + r'(x')$, $f' \notin \mathcal{F}_{L,P}$, and $r' \notin \mathcal{F}_P$.

We delineate the InD and OOD input vector spaces of $d(z)$. We denote the input vector spaces for an L2O model in the context of *InD L2O Problem* and *OOD L2O Problem* as $\mathcal{Z}_P$ and $\mathcal{Z}_O$, respectively. Then, we choose features of the variables and the objective functions to construct the input feature of $d(z)$. Specifically, we define $\mathcal{Z}_P$ and $\mathcal{Z}_O$ as the ensuing sets:

$$\begin{aligned}\mathcal{Z}_P =& \big\{ \big[x\text{-feature}^\top, f(x)\text{-feature}^\top, r(x)\text{-feature}^\top, \dots \big]^\top \\ & \mid \forall x \in \mathcal{S}_P, \forall f' \in \mathcal{F}_{L,P}, \forall r' \in \mathcal{F}_P \big\}, \\ \mathcal{Z}_O =& \big\{ \big[x'\text{-feature}^\top, f'(x')\text{-feature}^\top, r'(x')\text{-feature}^\top, \dots \big]^\top \\ & \mid \exists x' \in \mathcal{S}_O \text{ or } \exists f' \notin \mathcal{F}_{L,P} \text{ or } \exists r' \notin \mathcal{F}_P \big\},\end{aligned}$$

where "..." represents other feasible features such as the history of x. Some feasible feature constructions for x, $f(x)$, and $r(x)$ include x itself, $\nabla f(x)$, and $\partial r(x)$. Later in Sec. 5, we show how to contruct the input features of the L2O model $d(z)$ based only on $\nabla f(x)$ and $\partial r(x)$.

3. Virtual Feature and Trajectory

In this section, we introduce a virtual feature methodology to correlate any arbitrary variable yielded by the L2O model in the OOD scenario (x'_k) to a corresponding variable x_k in the InD scenario. The virtual features are generated as a linear combination of the OOD and InD features and serve as a bridge to connect each L2O model's OOD outcome to its InD outcome. We then leverage the virtual-feature method to connect OOD and InD variable trajectories generated by the L2O model. Since the convergence of InD trajectories is deterministic, such a method facilitates the convergence and robustness analysis for OOD scenarios in Sec. 4.

3.1. Virtual Feature

Consider an arbitrary OOD variable $x' \in \mathcal{S}_O$ and a InD variable $x \in \mathcal{S}_P$ yielded by the L2O model. Let $s \in \mathbb{R}^n$ such that $s = x' - x$. In that case, we define the difference s' between the L2O model's features in the OOD scenario z' and the features in the InD scenario z. From the Mean Value Theorem [20], there exists a virtual Jacobian matrix $\mathbf{J}_d$, $\|\mathbf{J}_d\| \leq C\sqrt{n}$ such that the following equality holds:

$$d(z') = d(z) + \mathbf{J}_d(z' - z) = d(z) + \mathbf{J}_d s'. \tag{3}$$

The demonstrations are in Sec. 8.1. From equation 3, we can relate any variable of the L2O model in the OOD scenario to the InD scenario. Although the virtual Jacobian matrix $\mathbf{J}_d$ is non-deterministic, it is upper bounded from the definition of $d(z)$ in equation 1. This suffices for a quantitative analysis of the impact of the "shift" s' on convergence. For instance, our proposed Theorem 1 in Sec. 4 provides an upper bound on the convergence gain for a single iteration.

3.2. Trajectory

For the OOD Problem O, denote the initial variable as $x'_0 \in \mathcal{S}_O$. In the optimization process, we have two trajectories for the variable x' and the features of the L2O model z':

$$\{x'_0, x'_1, x'_2, \dots, x'_K\}, \{z'_0, z'_1, z'_2, \dots, z'_K\}.$$

where $x'_k \in \mathcal{S}_O, z'_k \in \mathcal{Z}_O, k = 0, 1, 2, \ldots, K$. Similarly, for the InD Problem P, denote the initial variable as $x_0 \in \mathcal{S}_P$. We have also have two trajectories for the variables x and the features of the L2O model z:

$$\{x_0, x_1, x_2, \ldots, x_K\}, \{z_0, z_1, z_2, \ldots, z_K\},$$

where $x_k \in \mathcal{S}_P, z_k \in \mathcal{Z}_P, k = 0, 1, 2, \ldots, K$. Utilizing the definitions in Sec. 3.1, we compute the differences between the variables and the features of the OOD trajectories and the InD trajectories as follows:

$$\{s_0, s_1, s_2, \ldots, s_K\}, \{s'_0, s'_1, s'_2, \ldots, s'_K\},$$

where $s_k := x'_k - x_k$ and $s'_k := z'_k - z_k$. Thus, we can represent the OOD trajectory by $\{x_k + s_k\}$ and $\{z_k + s'_k\}$. Furthermore, utilizing the virtual-feature method in Sec. 3.1, we have:

$$d(z'_{k-1}) = d(z_{k-1}) + \mathbf{J}_{d,k-1} s'_{k-1}, \tag{4}$$

where $\mathbf{J}_{d,k-1}$ is a virtual Jacobian matrix of $d(\tilde{z}_{k-1})$. Due to equation 2 in Sec. 2.2, x'_k is updated by $x'_{k-1} - d(z'_{k-1})$ and x_k is updated by $x_{k-1} - d(z_{k-1})$. Based on equation 4, we have:

$$s_k = s_{k-1} - \mathbf{J}_{d,k-1} s'_{k-1}. \tag{5}$$

4. White-Box OOD Generalization Analysis

In this section, we rigorously demonstrate that the robustness of the L2O model is limited by its input features of NNs. We prove that increased features adversely impact the L2O model's generalization ability in OOD scenarios.

4.1. The Smooth Case

Building upon the state-of-the-art Math-L2O [14], we systematically detail our conclusions through a series of theorems and lemmas.

We analyze the convergence rate of the OOD scenario when the objective function $F(x)$ is smooth, i.e., $r(x) = 0$ and $F(x) = f(x)$. Leveraging Theorem 1 from [14], the update of the variable at the k-th iteration can be expressed as $x_k = x_{k-1} - \mathbf{P}_{k-1} \nabla f(x_{k-1}) - b_{k-1}$, where $\mathbf{P}_{k-1} \in \mathbb{R}^{n \times n}$ and $b_{k-1} \in \mathbb{R}^n$ are parameters learned by NNs.

Let $\mathbf{P}_{k-1}$ and b_{k-1} be $\mathbf{N}_1(\mathcal{Z}) \in \mathcal{D}_{C_1}(\mathcal{Z})$ and $\mathbf{N}_2(\mathcal{Z}) \in \mathcal{D}_{C_2}(\mathcal{Z})$ respectively, for some positive constants $C_1, C_2 \in \mathbb{R}^+$. As suggested in [14], we assign $\mathbf{P}_k$ as a diagonal matrix. Without loss of generality, for any given variable x_{k-1}, where $x_{k-1} \in \mathbb{R}^n$, and any given function $f \in \mathcal{F}_L(\mathbb{R}^n)$, we define $z_{k-1} = [x_{k-1}^\top, \nabla f(x_{k-1})^\top]^\top$ [14]. The update of variable x_k at each iteration k can then be expressed as:

$$x_k = x_{k-1} - \mathrm{diag}(\mathbf{N}_1(z_{k-1})) \nabla f(x_{k-1}) - \mathbf{N}_2(z_{k-1}). \tag{6}$$

The OOD shift applied to the variable and its gradient yields the definition of virtual feature (Sec. 3):

$$s'_{k-1} := [s_{k-1}^\top, (\nabla f'(x'_{k-1}) - \nabla f(x_{k-1}))^\top]^\top. \tag{7}$$

We present the following lemma for $\mathbf{N}_1(z)$ and $\mathbf{N}_2(z)$ to yield a variable x_k that is no worse than the previous variable x_{k-1} at each iteration k.

Lemma 1. *Denote the angle between $\mathbf{N}_2(z_{k-1})$ and corresponding $\nabla f(x_{k-1})$ as θ_{k-1}. For $\forall z_{k-1} \in \mathcal{Z}_P, \forall x_{k-1} \in \mathcal{S}_P$, if $\mathbf{N}_1(z_{k-1})$ and $\mathbf{N}_2(z_{k-1})$ are respectively bounded by following compact sets:*

$$\mathbf{N}_1(z_{k-1}) := \lambda_{k-1}\mathbf{1}, \lambda_{k-1} \in \left[0, \frac{1}{L}\right],$$
$$\mathbf{N}_2(z_{k-1}) \in \left[\mathbf{0}, \frac{\|\nabla f(x_{k-1})\| \cos(\theta_{k-1})}{L}\mathbf{1}\right], \theta \in \left[0, \frac{\pi}{2}\right],$$

then, for x_k generated by L2O model in equation 6, we have:

$$F(x_k) - F(x_{k-1}) \leq 0.$$

Proof. See Sec. 8.2 in Appendix. □

As stated in Lemma 1, to maintain homogeneous improvement on the convergence, it is sufficient to set $\mathbf{N}_1(z)$ as an input-invariant constant, and limit $\mathbf{N}_2(z)$ according to the gradient $\nabla F(x_{k-1})$. Moreover, we can utilize some bounded activation functions in training an L2O model to fulfill the conditions to ensure convergence, such as Sigmoid [17] and Tanh [12].

The proof for Lemma 1 establishes that improvement is characterized by a quadratic relation to each element in $\mathbf{N}_1(z_{k-1})$ and $\|\mathbf{N}_2(z_{k-1})\|$. We can identify the optimal upper bound for convergence improvement in the InD L2O model by optimizing this quadratic relation, leading us to Corollary 1.

Corollary 1. *For any $z_{k-1} \in \mathcal{Z}_P$, we let:*

$$\mathbf{N}_1(z_{k-1}) := \frac{1}{2L}\mathbf{1}, \mathbf{N}_2(z_{k-1}) := \frac{\nabla f(x_{k-1})}{2L},$$

the Math-L2O model in equation 6 is exactly gradient descent update with convergence rate:

$$F(x_K) - F(x^*) \leq \frac{L}{2K}\|x_0 - x^*\|^2.$$

Proof. See Sec. 8.3 in Appendix. □

Corollary 1 implies that the L2O model can achieve gradient descent's convergence rate by particular settings. The $\mathbf{N}_1(z_{k-1})$ is set to be a homogeneous constant across all elements. The $\mathbf{N}_2(z_{k-1})$ is set to in correspondence with the gradient $\nabla f(x_{k-1})$. Moreover, Corollary 1 also provides the most robust L2O model with an identical per iteration convergence gain among all InD instances.

Per-Iteration Convergence Gain

To ascertain the convergence rate of OOD, following Corollary 1, we suppose that after training, the following assumption holds for the *InD L2O Problem* (not for the *OOD L2O Problem*) to ensure best robustness for the InD scenario:

Assumption 1. *After training,* $\forall x_{k-1} \in \mathcal{S}_P, \forall z_{k-1} \in \mathcal{Z}_P$, $\mathbf{N}_1(z_{k-1}) := \frac{1}{2L}\mathbf{1}$ *and* $\mathbf{N}_2(z_{k-1}) := \frac{\nabla f(x_{k-1})}{2L}$.

Based on the Lemma 1 and Corollary 1, Assumption 1 leads to an L2O model with best robustness on all InD instances. In the following theorem, we quantify the diminution in convergence rate instigated by the virtual feature s' defined in Sec. 3.

Theorem 1. *Under Assumption 1, there exists virtual Jacobian matrices* $\mathbf{J}_{1,k-1}, \mathbf{J}_{2,k-1}, k = 1, 2, \ldots, K$ *that the per iteration convergence improvement in the OOD scenario is upper bounded by:*

$$\begin{aligned} & F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\ \leq & -\frac{\|\nabla f'(x_{k-1} + s_{k-1})\|^2}{2L} \\ & + L\|\operatorname{diag}(\mathbf{J}_{1,k-1}s')\nabla f'(x_{k-1} + s_{k-1})\|^2 \\ & + L\|\frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} - \mathbf{J}_{2,k-1}s'\|^2. \end{aligned}$$

Proof. See Sec. 8.4 in Appendix. □

Theorem 1 discloses that for a single iteration, the convergence improvement of OOD is bounded by the gradient descent with a step size of $1/L$, resulting in $-|\nabla f|^2/2L$ convergence improvement. Hence, when Math-L2O is adequately trained, any OOD will dampen convergence. Additionally, given that the expression on the right-hand side is not strictly non-positive, we cannot unequivocally affirm that convergence will transpire within a single iteration. Further investigation also intimates that, even in the context of convex optimization problems, scenarios may arise where the value of the objective function deteriorates.

While the existence of virtual Jacobian matrices in Theorem 1 is assured, their specific values remain unknown. Given that boundedness is a defined characteristic of these matrices, we relax this constraint in Theorem 1 and introduce Corollary 2.

Corollary 2. *Under Assumption 1, the per iteration convergence improvement in the OOD scenario can be upper bounded w.r.t.* $\|s'_{k-1}\|$ *by:*

$$\begin{aligned} & F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\ \leq & -\frac{\|\nabla f'(x_{k-1} + s_{k-1})\|^2}{2L} \\ & + \frac{\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x)\|^2}{2L} \\ & + \left(LC_1^2 n\|\nabla f'(x_{k-1} + s_{k-1})\|^2 + 2LC_2^2 n\right)\|s'\|^2. \end{aligned}$$

Proof. See Sec. 8.5 in Appendix. □

Corollary 2 further elucidates that the decline in the convergence improvement of OOD is determined by the magnitude of the input (virtual) feature s' of the L2O model, as outlined in equation 7. This magnitude is intrinsically related to the vector's dimensionality, which relies on the feature construction of the L2O model. For example, to reduce its magnitude, we can eliminate s_{k-1} in equation 7. We achieve this feature shrinking and propose a novel gradient-only L2O model in Sec. 5.

Multi-Iteration Convergence Rate

Building upon Theorem 1, we extrapolate the convergence rate across numerous iterations, as delineated in Theorem 2.

Theorem 2. *Under Assumption 1, the* K *iterations' convergence rate in the OOD scenario is upper bounded by:*

$$\begin{aligned} & \min_{k=1,\ldots,K} F'(x_k + s_k) - F'(x^* + s^*) \\ \leq & \frac{L}{2}\|x_0 - x^* + s_0 - s^*\|^2 - \frac{L}{2}\|x_K - x^* + s_K - s^*\|^2 \\ & + \frac{L}{K}\sum_{k=1}^{K}(x_k + s_k - x^* - s^*)^\top \\ & \left(x_k + s_k - \left(x_{k-1} + s_{k-1} - \frac{\nabla f'(x_{k-1} + s_{k-1})}{L}\right)\right). \end{aligned}$$

Proof. See Sec. 8.6 in Appendix. □

The first two terms on the right-hand side of the above inequality represent the gradient descent convergence rate characterized by a step size of $1/L$. However, the third term is unbounded and could be either non-positive or positive. This suggests that there is no guaranteed global convergence in OOD situations, even with homogeneous robustness in InD scenarios.

The inequation above offers a direct approach to analyzing distinct cases of convergence. Included in the concluding line of Theorem 1 is a gradient descent equation, $x_{k-1} + s_{k-1} - \nabla f'(x_{k-1} + s_{k-1})/L$. Moreover, $x_k + s_k$ represents the updated solution by the L2O model. The subtraction of the two terms reveals the discrepancy between the updates made by L2O and gradient descent on the objective variable $x_{k-1} + s_{k-1}$, thereby creating a vector directed towards $x_k + s_k$. Similarly, $x_k + s_k - x^* - s^*$ signifies the relative position to the optimal solution, generating another vector directed towards $x_k + s_k$. The resulting inner product will be non-positive if the angle between these two vectors is $\pi/2$ or more. Moreover, if the trajectory of $x_k + s_k - x^* - s^*$ can be extrapolated from domain knowledge, a "trust region" surrounding $x_k + s_k$ can be established to augment the efficacy of gradient descent.

From Theorem 1, we develop a stringent formulation to illustrate the potential uncertainty of convergence in OOD

scenarios. If we know the relative position of the optimal solution, we can fine-tune an L2O model to outperform gradient descent. Based on Theorem 1, we establish an upper bound w.r.t. s'. This mirrors the approach in Corollary 2.

Corollary 3. *Under Assumption 1, L2O model $d(z)$'s OOD convergence rate is upper bounded w.r.t. $\|s'_{k-1}\|$ by:*

$$\begin{aligned}
&\min_{k=1,\dots,K} F'(x_k+s_k)-F'(x^*+s^*)\\
\leq&\frac{L}{2}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2}\|x_K+s_K-x^*-s^*\|^2\\
&+\frac{1}{2K}\sum_{k=1}^{K}\left(\nabla f'(x_{k-1}+s_{k-1})-\nabla f(x_{k-1})\right)^\top\\
&\qquad(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}\left(C_1\sqrt{n}\|\nabla f'(x_{k-1}+s_{k-1})\|\right.\\
&\qquad\left.+C_2\sqrt{n}\|x_k+s_k-x^*-s^*\|\right)\|s'_{k-1}\|.
\end{aligned}$$

Proof. See Sec. 8.7 in Appendix. □

Corollary 3 posits that the overall convergence rate is consistently upper bounded by the magnitude of s'. Based on Corollaries 2 and 3, we endeavor to reduce the magnitude of s' by eliminating variable, leading to the approach of a gradient-only Math-L2O framework in the next section.

4.2. Other Three Cases

We have developed several additional theorems and lemmas for non-smooth, incremental historical modeling, and integrated smooth-non-smooth cases.. Our approach mirrors that employed in the smooth case demonstration. The backbone algorithms of math-inspired L2O fundamentally limit their convergences. For example, the Gradient Descent [19] and Proximal Point [18] algorithms in the smooth case and the non-smooth case, respectively.

We extend the theorems and lemmas in the smooth case to derive formulas for convergence improvement of a single iteration and convergence rate across a sequence. These demonstrate the diminishing effect of OOD on convergence. Our findings conclude that constructing fewer features can mitigate this negative impact. More extensive demonstrations and complete proofs can be found in Appendix.

5. Gradient-Only L2O Model

Informed by the theorems and lemmas posited in Sec. 4, we introduce a gradient-only L2O model, GO-Math-L2O, which aims to enhance robustness in OOD scenarios by eliminate variable-related input features for the L2O model.

To derive the formulation of GO-Math-L2O, we employ the workflow delineated in [14]. Let T denote the history length. At the k-th iteration, suppose there exists an operator $d_k \in \mathcal{D}_C(\mathbb{R}^{3n})$, we formulate the input of our GO-Math-L2O as follows:

$$x_k = x_{k-1} - d_k(\nabla f(x_{k-1}), g_k, v_{k-1}), \tag{8}$$

where g_k denotes the implicit subgradient vector of x_k to invoke the proximal gradient method [14]. Moreover, we eliminate all variable-related features and define v_k as the result of historical modeling [14]. Different from the variable approach in [14], we propose to utilize gradient (and subgradient) to model the historical information of the optimization process since gradient sufficiently and necessarily indicates optimality in convex optimization scenarios. Such an approach reduces the magnitude of L2O's input feature (defined in Sec. 3) by $1/3$, which facilitates convergence based on our proposed corollaries in Sec. 4.

Suppose there exists an operator $u_k \in \mathcal{D}_C(\mathbb{R}^{Tn})$, we define the following model to generate v_k from the gradient and subgradient of T historical iterations:

$$v_k = d_k(\nabla f(x_{k-1})+g_{k-1},\dots,\nabla f(x_{k-T})+g_{k-T}). \tag{9}$$

where each g represents a subgradient vector. For subgradient selection, we should carefully choose an instance from the subgradient set of each non-smooth point since an arbitrary selection may lead to poor convergence [25].

We achieve a lightweight subgradient selection based on the gradient map method [28] and our following model constructions. From the objective definition in Sec. 2.1, the non-smooth objective r is trivially solvable by $\arg\min$. Thus, at k-th iteration, we can recover an implicit subgradient vector g_k of $\arg\min$ by k-th solution x_k and k_{-1}-th solution x_{k-1} if the L2O operator d_k in equation 8 is inversible. Next, we achieve an inversible d_k based on the workflow proposed in [14].

With the above feature and component constructions, we start to define the structures and learnable parameters of our L2O operator d_k in equation 8. We formulate d_k as the necessary condition of convergence [14], which means the formulation that d_k should follow if convergence is achieved. First, denote a candidate optimal solution as x^*, we construct two sufficient conditions (Asymptotic **F**ixed **P**oint and **G**lobal **C**onvergence) of convergence for our L2O operator d_k in equation 8:

$$\lim_{k\to\infty} d_k(\nabla f(x^*), -\nabla f(x^*), 0) = \mathbf{0}, \tag{FP}$$

$$\lim_{k\to\infty} x_k = x^*. \tag{GC}$$

As discussed in [14], such two conditions are essential for optimization algorithms.

Then, we present the following Theorem 3 to construct d_k's parameters. Theorem 3 shows that if d_k converges, it should be in the form of equation 10. Then, if we add a further assumption on some of the parameters, the solution on each iteration can be uniquely obtained by equation 11.

Theorem 3. *Suppose $T = 2$, given $f \in \mathcal{F}_L(\mathbb{R}^n)$ and $r \in \mathcal{F}(\mathbb{R}^n)$, we pick an operators from $\mathcal{D}_C(\mathbb{R}^{3n})$ and $\mathcal{D}_C(\mathbb{R}^{2n})$. If Condition FP and Condition GC hold, there exist $\mathbf{R}_k \succ 0, \mathbf{Q}_k, \mathbf{B}_k \in \mathbb{R}^{n\times n}$ and $b_{1,k}, b_{2,k} \in \mathbb{R}^n$ and satisfying:*

$$\begin{aligned} x_k =& x_{k-1} - \mathbf{R}_k \nabla f(x_{k-1}) - \mathbf{R}_k g_k - \mathbf{Q}_k v_{k-1} - b_{1,k}, \\ v_k =& (\mathbf{I} - \mathbf{B}_k) G_k + \mathbf{B}_k G_{k-1} - b_{2,k}, \\ G_k :=& \mathbf{R}_k^{-1}(x_{k-1} - x_k - \mathbf{Q}_k v_{k-1} - b_{1,k}), \end{aligned} \tag{10}$$

where for $k = 0, 1, 2, \ldots$, $g_{k+1} \in \partial r(x_{k+1})$ represents implicit subgradient vector, $\mathbf{R}_k$, $\mathbf{Q}_k$, and $\mathbf{B}_k$ are bounded parameter matrices and $b_{1,k} \to 0, b_{2,k} \to 0$ as $k \to \infty$. Since $\mathbf{R}_k$ is symmetric positive definite, x_{k+1} is uniquely determined through:

$$\arg\min_{x\in\mathbb{R}^n} r(x) + \frac{1}{2}\|x - \mathbf{R}_k \nabla f(x_k) - \mathbf{Q}_k v_k - b_{1,k}\|^2_{\mathbf{R}_k^{-1}}, \tag{11}$$

where $\|\cdot\|_{\mathbf{R}_k^{-1}}$ is defined as $\|x\|_{\mathbf{R}_k^{-1}} = \sqrt{x^\top \mathbf{R}_k^{-1} x}$.

Proof. See Sec. 8.8 in Appendix. □

As a necessary condition for convergence, Theorem 3 suggests that our gradient-only L2O model should construct parameters $\mathbf{R}$, $\mathbf{Q}$, $\mathbf{B}$, b_1, and b_2. It is worth noting that this model does not guarantee satisfaction of conditions FP and GC. The convergence is promoted by training.

We learn to construct the parameters in Theorem 3. First, the proof elucidates that the bias terms approach zero upon convergence. Thus, we set $b_1, b_2 := 0$ and learn to construct $\mathbf{R}$, $\mathbf{Q}$, and $\mathbf{B}$. We take the construction in [14] to implement our GO-Math-L2O model with a two-layer LSTM cell. Then, we utilize three one-layer linear neural network models with Sigmoid activation function [17] to generate $\mathbf{R}$, $\mathbf{Q}$, and $\mathbf{B}$ at each iteration, respectively, which ensures that all the matrices are bounded.

6. Experiments

We perform experiments with Python 3.9 and PyTorch 1.12 on an Ubuntu 18.04 system equipped with 128GB of memory, an Intel Xeon Gold 5320 CPU, and a pair of NVIDIA RTX 3090 GPUs. We strictly follow the experimental setup presented in [14] for constructing InD evaluations. Due to the page limit, the implementation details are in Sec. 12.

We use the Adam optimizer [13] to train our proposed model and learning-based baselines on datasets of 32,000 optimization problems with randomly sampled parameters and optimal solutions. We generate a test dataset of 1,000 iterations' objective values, averaging over 1,024 pre-generated optimization problems. We evaluate different training configurations and loss functions to select the best setting. Details are in Sec. 12.5, Appendix.

Baselines. We compare our GD-Math-L2O (Section 5) against both learning-based methods and non-learning algorithms. Our main competater is the state-of-the-art (SOTA) math-inspired L2O model in [14]. Specifically, we select the best variant from this study, L2O-PA. Consistent with the outlined methodology, we also compare our approach with several hand-crafted algorithms: ISTA, FISTA [5], Adam [13], and AdamHD [4], which is Adam complemented by an adaptive learning rate. Moreover, we assess our model against two black-box L2O models, namely L2O-DM[3] and L2O-RNNprop [15], and one Ada-LISTA [1] that unrolls the gradient descent algorithm with learning.

Optimization Objective. We choose the two regression problems in [14]: *LASSO Regression* and *Logistic Regression*, defined as follows:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} F(x) =& \frac{1}{2}\|\mathbf{A}x - b\|^2 + \lambda\|x\|_1, \\ \min_{x\in\mathbb{R}^n} F(x) =& -\frac{1}{m}\sum_{i=1}^{m}\big[b_i \log(h(a_i^\top x)) \\ &+ (1 - b_i)\log(1 - h(a_i^\top x))\big] + \lambda\|x\|_1, \end{aligned}$$

where $m := 1000$. $\mathbf{A} \in \mathbb{R}^{250\times 500}$ and $b \in \mathbb{R}^{500}$, $\{(a_i, b_i) \in \mathbb{R}^{50} \times \{0, 1\}\}_{i=1}^m$ are given parameters. $h(x) := 1/(1 + e^{-x})$ is sigmoid function. We utilize the standard normal distribution to generate samples and set $\lambda := 0.1$ for both scenarios [14].

We implement the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [5], executing 5,000 iterations to generate labels (optimal objective values) [14]. Due to page limit, we confine our presentation to *LASSO Regression*. The results of *Logistic Regression* are in Sec. 12.8, Appendix.

OOD Scenarios. We aim to quantify the effect of OOD on convergence rates. We specifically formulate two types of OOD trajectories triggered by different actions. It is crucial to note that both OOD and InD scenarios maintain an identical optimality on both objective and solution.

1) $s_0 \neq 0, s_0 \in \mathbb{R}^n$. x_0 is altered by an adjustment factor s_0 that x_0' falls within the OOD set $\mathcal{S}_O$. Assuming the objective remains consistent, we expect x' to move from the OOD $\mathcal{S}_O$ to the InD $\mathcal{S}_P$.
2) $F'(x) = F(x + t), t \in \mathbb{R}^n$. The OOD perturbation introduces a translation t along the axes of the objective variable to the objective function. Thus, the optimal solution x'^* diverges from that obtained under the original InD domain, even though the optimal value remains. This illustrates a scenario where the domain translates in inference. If the starting point is unchanged, x' is expected to move from InD domain to OOD domain.

We derive the non-smooth function's proximal operator for the OOD scenario, specifically for the ℓ_1-norm. We define $r(x)$ as $\lambda|x|_1$, and define the OOD translation as t on

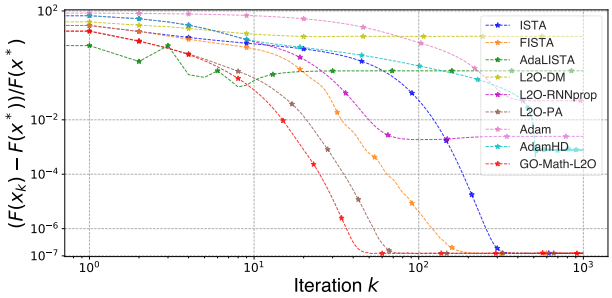

Figure 1. LASSO Regression: InD.

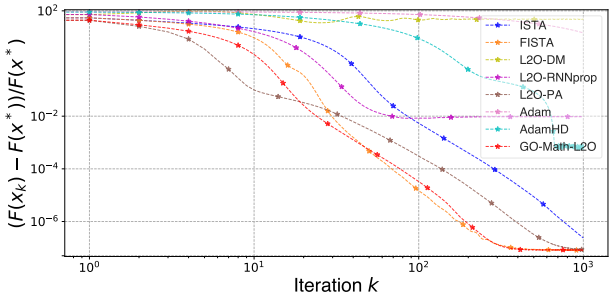

Figure 2. LASSO Regression: Real-World OOD.

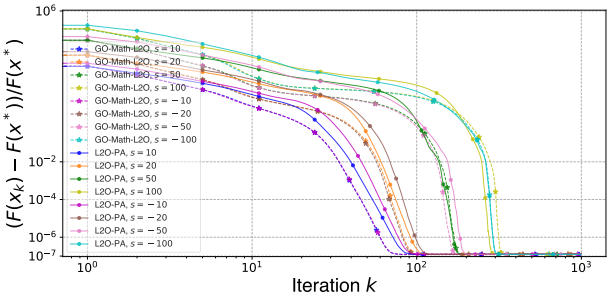

Figure 3. LASSO Regression: OOD by Trigger 1.

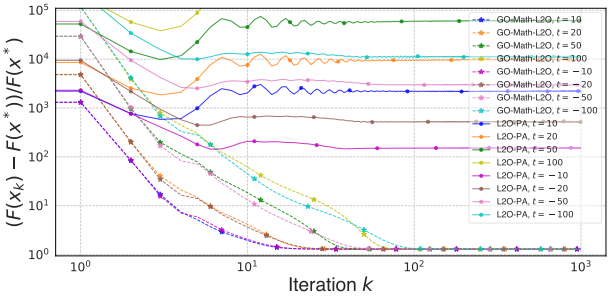

Figure 4. LASSO Regression: OOD by Trigger 2.

variable. The OOD proximal operator with t is given by:

$$(\operatorname{prox}_{r,p_k}(\bar{x}))_i := -t + \operatorname{sign}(\bar{x}_i)\max(0, |\bar{x}_i| - \lambda(p_k)_i + \operatorname{sign}(\bar{x}_i)t).$$

6.1. InD Comparison

The trajectories of solving the *LASSO Regression* problems are shown in Figure 1, where the vertical axis represents the normed objective value at a given iteration (indicated on the horizontal axis) with a label generated by FISTA [5]. Our proposed method (red line) surpasses all other methods, demonstrating better optimality and quicker convergence.

Furthermore, we utilize several ablation studies on model configuration, such as gradient map recovery strategies in Sec. 12.4 and hyperparameter settings for learned parameter matrices in Sec. 12.6, to determine the best model configuration. The details are in the Appendix.

6.2. OOD Comparison

The real-world results in Firgure 2 show that our GO-Math-L2O (converges at 400 iterations) outperforms all other baselines (1,000 iterations). Considering the lackluster performances of other baselines in Figures 1 and 2, we primarily compare our GO-Math-L2O model against SOTA L2O-PA [14]. We construct two synthetic OOD scenarios with the two trigger settings, where the optimal objectives align with those in Figure 1.

Figure 3 portrays the scenario wherein the initial point shifts such that $s_0 \neq 0$, with the legends denoting sixteen cases. Our GO-Math-L2O model (represented by dashed lines) outshines L2O-PA (solid lines) in all instances, asserting its superior robustness.

The observations in Figure 4 for the OOD scenario involve function shifting that $F'(x) = F(x + t)$. The optimal values achieved by both methods deteriorate from 10^{-7} (as seen in Figure 3) to 10^0. However, our GO-Math-L2O still outperforms L2O-PA in all cases. For example, when $t = \pm 10$, our model converges at around 20 steps, but L2O-PA fails to converge.

7. Conclusion

This paper aims to improve the robustness of L2O in OOD scenarios. We derive a general condition to ensure robustness in InD scenarios. We propose virtual features to connect the OOD L2O's outputs with InD L2O's outputs of a whole trajectory. Based on such connections, we prove formulations to demonstrate the convergence performances in OOD scenarios. Based on the observations, we establish that the magnitude of the L2O model's input features intrinsically limits the OOD's convergence. Furthermore, we propose a robust L2O model with concise gradient-only features and modeling historical features with gradient and subgradient. Experiments show our model significantly outperforms SOTA baselines.

Acknowledgements

This work is partly supported by funding from the Research Grants Council of Hong Kong (11209520, CRF C7004-22G) and CUHK (4055199).

References

[1] Aviad Aberdam, Alona Golts, and Michael Elad. Ada-lista: Learned solvers adaptive to varying models. *IEEE TPAMI*, 44(12):9222–9235, 2021. 7

[2] Diogo Almeida, Clemens Winter, Jie Tang, and Wojciech Zaremba. A generalizable approach to learning optimizers. *arXiv preprint arXiv:2106.00958*, 2021. 1

[3] 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 *NeurIPS*, 2016. 7

[4] Atilim Gunes Baydin, Robert Cornish, David Martinez Rubio, Mark Schmidt, and Frank Wood. Online learning rate adaptation with hypergradient descent. *arXiv preprint arXiv:1703.04782*, 2017. 7

[5] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009. 7, 8, 16, 29, 45, 49

[6] Yanmei Cao, Guomei Zhang, Guobing Li, and Jia Zhang. A Deep Q-Network Based-Resource Allocation Scheme for Massive MIMO-NOMA. *IEEE Communications Letters*, 25 (5):1544–1548, 2021. 1

[7] Tianlong Chen, Xiaohan Chen, Wuyang Chen, Zhangyang Wang, Howard Heaton, Jialin Liu, and Wotao Yin. Learning to optimize: A primer and a benchmark. *The Journal of Machine Learning Research*, 23(1):8562–8620, 2022. 1

[8] Karol Gregor and Yann LeCun. Learning Fast Approximations of Sparse Coding. In *ICML*, pages 399–406, 2010. 1

[9] Howard Heaton, Xiaohan Chen, Zhangyang Wang, and Wotao Yin. Safeguarded learned convex optimization. In *AAAI*, pages 7848–7855, 2023. 1

[10] Michael (https://math.stackexchange.com/users/155065/michael). Proof of convergence for the proximal point algorithm. Mathematics Stack Exchange, 2015. URL:https://math.stackexchange.com/q/1303325 (version: 2015-05-30). 14, 17

[11] Qiyu Hu, Yunlong Cai, Qingjiang Shi, Kaidi Xu, Guanding Yu, and Zhi Ding. Iterative Algorithm Induced Deep-Unfolding Neural Networks: Precoding Design for Multiuser MIMO Systems. *IEEE TWC*, 20(2):1394–1410, 2020. 1

[12] B.L. Kalman and S.C. Kwasny. Why tanh: choosing a sigmoidal function. In *IJCNN International Joint Conference on Neural Networks*, pages 578–581 vol.4, 1992. 4

[13] D Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *ICLR*, page 6. San Diego, California;, 2015. 7, 50

[14] Jialin Liu, Xiaohan Chen, Zhangyang Wang, Wotao Yin, and HanQin Cai. Towards Constituting Mathematical Structures for Learning to Optimize. In *ICML*, 2023. 1, 2, 3, 4, 6, 7, 8, 14, 15, 16, 17, 29, 30, 31, 48, 49, 50, 52, 53, 54, 55

[15] Kaifeng Lv, Shunhua Jiang, and Jian Li. Learning gradient descent: Better generalization and longer horizons. In *ICML*, pages 2247–2255. PMLR, 2017. 7, 29

[16] Kaifeng Lv, Shunhua Jiang, and Jian Li. Learning gradient descent: Better generalization and longer horizons. In *ICML*, pages 2247–2255. PMLR, 2017. 1

[17] Sridhar Narayan. The generalized sigmoid activation function: Competitive supervised learning. *Information Sciences*, 99(1):69–82, 1997. 4, 7

[18] R Tyrrell Rockafellar. Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization*, 14(5):877–898, 1976. 6, 17

[19] Sebastian Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016. 1, 3, 6

[20] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 1976. 3

[21] Lukas Schynol and Marius Pesavento. Coordinated Sum-Rate Maximization in Multicell MU-MIMO With Deep Unrolling. *IEEE JSAC*, 41(4):1120–1134, 2023. 1

[22] Yifei Shen, Yuanming Shi, Jun Zhang, and Khaled B. Letaief. Graph Neural Networks for Scalable Radio Resource Management: Architecture Design and Theoretical Analysis. *IEEE JSAC*, 39(1):101–115, 2021. 1

[23] Z Shen, J Liu, Y He, X Zhang, R Xu, H Yu, and P Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2023. 1

[24] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. 2012. 1

[25] Stephen Boyd, Lin Xiao, and Almir Mutapcic. Subgradient Methods. *Stanford EE392o Optimization Projects*, 2003. 6

[26] Haoran Sun, Xiangyi Chen, Qingjiang Shi, Mingyi Hong, Xiao Fu, and Nicholas D Sidiropoulos. Learning to optimize: Training deep neural networks for interference management. *IEEE TSP*, 66(20):5438–5453, 2018. 1

[27] Ryan Tibshirani. Lecture 6: September 12. *CMU 10-725: Optimization*, 2013. 2, 3

[28] L. Vandenberghe. Proximal gradient method. *ECE236C (Spring 2022)*, 2022. 6, 17

[29] Olga Wichrowska, Niru Maheswaranathan, Matthew W Hoffman, Sergio Gomez Colmenarejo, Misha Denil, Nando Freitas, and Jascha Sohl-Dickstein. Learned optimizers that scale and generalize. In *ICML*, pages 3751–3760. PMLR, 2017. 1

[30] Junjie Yang, Tianlong Chen, Mingkang Zhu, Fengxiang He, Dacheng Tao, Yingbin Liang, and Zhangyang Wang. Learning to Generalize Provably in Learning to Optimize. In *International Conference on Artificial Intelligence and Statistics*, pages 9807–9825, 2023. 1

[31] Yu Zhao, Ignas G. Niemegeers, and Sonia M.Heemstra De Groot. Dynamic Power Allocation for Cell-Free Massive MIMO: Deep Reinforcement Learning Methods. *IEEE Access*, 9:102953–102965, 2021. 1

[32] Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018. 17

[33] Minghe Zhu, Tsung-Hui Chang, and Mingyi Hong. Learning to beamform in heterogeneous massive MIMO networks. *IEEE TWC*, 2022. 1

Towards Robust Learning to Optimize with Theoretical Guarantees

Supplementary Material

8. Proofs

8.1. Preliminary

Demonstration of Equation 3

Proof. Based on demonstration for Lemma 1 in [14], since $d \in \mathcal{D}_C(m)$, the outcome of$d$ is an n-dimensional vector. Denote the i-th element as $d_i (1 \leq i \leq n)$ and convert d into a matrix form:

$$d(z') = \left[d_1(z'), \ldots, d_n(z')\right]^\top,$$
$$d(z) = \left[d_1(z), \ldots, d_n(z)\right]^\top.$$

Regarding each $d_i(z')$ as a multi-variable function and applying the Mean Value Theorem on it, for some $\xi_i \in (0, 1)$, we can construct following equality:

$$d_i(z') - d_i(z) = \left\langle \frac{\partial d_i}{\partial z_i}\big(\xi_i z' + (1-\xi_i)z\big), z' - z \right\rangle.$$

Stacking all partial derivatives into one matrix yields:

$$\mathbf{J}_d = \left[\frac{\partial d_1}{\partial z}\big(\xi_1 z' + (1-\xi_1)z\big), \ldots, \frac{\partial d_n}{\partial z}\big(\xi_n z' + (1-\xi_n)z\big)\right].$$

We can directly get equation 3. And the upper bound of $\|\mathbf{J}_d\|$ is given by:

$$\|\mathbf{J}_d\|^2 \leq \|\mathbf{J}_d\|_{\mathrm{F}}^2 = \sum_{i=1}^{n} \left\| \frac{\partial d_i}{\partial z}\big(\xi_i z' + (1-\xi_i)\big) \right\|^2 \leq nC^2.$$

□

A General Upper Bound from L-smoothness

Suppose $z, \tilde{z}, z' \in \mathcal{Z}$ are input feature vectors of the L2O model. There exists a $\xi \in [0, 1]$ that $z' = \xi z + (1-\xi)\tilde{z}, z' \in \mathcal{Z}$. Thus, we are able to represent OOD feature z' with InD feature z. Denote the virtual Jacobian matrix of $\mathbf{N}_1(z') \in \mathcal{D}_{C_1}$ and $\mathbf{N}_2(z') \in \mathcal{D}_{C_2}$ at point z' as $\mathbf{J}_1$ and $\mathbf{J}_2$. Since $\mathbf{N}_1(z)$ and $\mathbf{N}_1(z)$ are smooth, due to the Mean Value Theorem, we have the following equality between OOD and InD outputs of the L2O model:

$$\mathbf{N}_1(z) = \mathbf{N}_1(\tilde{z}) + \mathbf{J}_1(z - \tilde{z}), \quad \mathbf{N}_2(z) = \mathbf{N}_2(\tilde{z}) + \mathbf{J}_2(z - \tilde{z}).$$

As demonstrated above, we have $\|\mathbf{J}_1\| \leq \sqrt{n}C_1$, and $\|\mathbf{J}_2\| \leq \sqrt{n}C_2$.

As illustrated in Sec. 3, we represent OOD input feature vector z' as a combination of InD input feature vector z and virtual feature vector s'. For a $z' = z + s'$, we have the following equalities:

$$\begin{aligned} \mathbf{N}_1(z + s') &= \mathbf{N}_1(z) + \mathbf{J}_1(z + s' - z) = \mathbf{N}_1(z) + \mathbf{J}_1 s', \\ \mathbf{N}_2(z + s') &= \mathbf{N}_2(z) + \mathbf{J}_2(z + s' - z) = \mathbf{N}_2(z) + \mathbf{J}_2 s'. \end{aligned} \tag{12}$$

Following [14], in the smooth objective case, if we use variable and gradient to construct input features, we can further formulate s' as $s' := \left[s^\top, \big(\nabla f'(x+s) - \nabla f(x)\big)^\top\right]^\top$. For s', we have the following inequalities:

$$\|s'\|^2 = \|s\|^2 + \|\nabla f'(x+s) - \nabla f(x)\|^2 \geq \|\nabla f'(x+s) - \nabla f(x)\|^2. \tag{13}$$

The above inequation gives a theoretical lower bound of the input feature's magnitude in [14]. Our following results will demonstrate that the convergence rate of learning-to-optimize will be the upper bound with respect to the magnitude of the L2O model's input feature. Based on such a lower bound, we are able to improve the convergence rate by eliminating variable features.

For any $x, x^+ \in \mathbb{R}^n$, we use x and x^+ to denote the variables before and after the update. Based on the definition of L-smoothness [27], we have following upper bound on objective F:

$$F(x^+) \leq F(x) + \nabla f(x)^\top (x^+ - x) + \frac{L}{2}\|x^+ - x\|^2.$$

Note that in problem O (Sec. 2), we define that an objective $F(x)$ has two parts: smooth part $f(x)$ and non-smooth part $r(x)$. And in the smooth case, we set $r(x) := 0$.

Substituting $x^+ = x - \operatorname{diag}\big(\mathbf{N}_1(z)\big)^\top \nabla f(x) - \mathbf{N}_2(z)$, we have:

$$\begin{aligned} F(x^+) &\leq F(x) + \nabla f(x)^\top \big(x - \operatorname{diag}\big(\mathbf{N}_1(z)\big)\nabla f(x) - \mathbf{N}_2(z) - x\big) + \frac{1}{2}L\big\|x - \operatorname{diag}\big(\mathbf{N}_1(z)\big)\nabla f(x) - \mathbf{N}_2(z) - x\big\|^2, \\ &= F(x) - \nabla f(x)^\top \operatorname{diag}\big(\mathbf{N}_1(z)\big)\nabla f(x) - \nabla f(x)^\top \mathbf{N}_2(z) + \frac{L}{2}\big\|\operatorname{diag}\big(\mathbf{N}_1(z)\big)\nabla f(x) + \mathbf{N}_2(z)\big\|^2. \end{aligned} \tag{14}$$

If $\nabla f(x) := \mathbf{0}$, we have:

$$F(x^+) \leq F(x) + \frac{L}{2}\big\|\mathbf{N}_2(z)\big\|^2.$$

To ensure $F(x^+) \leq F(x)$, we should set $\big\|\mathbf{N}_2(z)\big\| := \mathbf{0}$. Thus, $\mathbf{N}_2(z) := \mathbf{0}$. Otherwise $\nabla f(x) \neq \mathbf{0}$, we can split $\mathbf{N}_1(z)$ and $\mathbf{N}_2(z)$ in equation 14 by:

$$\begin{aligned} F(x^+) &\leq F(x) - \nabla f(x)^\top \operatorname{diag}\big(\mathbf{N}_1(z)\big)\nabla f(x) - \nabla f(x)^\top \mathbf{N}_2(z) + \frac{L}{2}\big\|\operatorname{diag}\big(\mathbf{N}_1(z)\big)\nabla f(x) + \mathbf{N}_2(z)\big\|^2, \\ &\leq F(x) - \nabla f(x)^\top \operatorname{diag}\big(\mathbf{N}_1(z)\big)\nabla f(x) - \nabla f(x)^\top \mathbf{N}_2(z) + L\big\|\operatorname{diag}\big(\mathbf{N}_1(z)\big)\nabla f(x)\big\|^2 + L\big\|\mathbf{N}_2(z)\big\|^2, \\ &= F(x) - \nabla f(x)^\top \operatorname{diag}\big(\mathbf{N}_1(z)\big)\nabla f(x) - \nabla f(x)^\top \mathbf{N}_2(z) + L\nabla f(x)^\top \operatorname{diag}\big(\mathbf{N}_1(z)\big)^2\nabla f(x) + L\big\|\mathbf{N}_2(z)\big\|^2, \\ &= F(x) - \nabla f(x)^\top \Big(\operatorname{diag}\big(\mathbf{N}_1(z)\big) - L\operatorname{diag}\big(\mathbf{N}_1(z)\big)^2\Big)\nabla f(x) - \nabla f(x)^\top \mathbf{N}_2(z) + L\big\|\mathbf{N}_2(z)\big\|^2. \end{aligned}$$

We construct the following inequality to ensure a homogeneous decrease in objective:

$$-\nabla f(x)^\top \Big(\operatorname{diag}\big(\mathbf{N}_1(z)\big) - L\operatorname{diag}\big(\mathbf{N}_1(z)\big)^2\Big)\nabla f(x) - \nabla f(x)^\top \mathbf{N}_2(z) + L\big\|\mathbf{N}_2(z)\big\|^2 \leq 0,$$

where $()^2$ on a matrix or a vector represents the entry-wise square, respectively. We continue to use this denotation below.

We first demonstrate the convergence gain on each iteration by $-\nabla f(x)^\top \Big(\operatorname{diag}\big(\mathbf{N}_1(z)\big) - L\operatorname{diag}\big(\mathbf{N}_1(z)\big)^2\Big)\nabla f(x)$ and $-\nabla f(x)^\top \mathbf{N}_2(z) + L\big\|\mathbf{N}_2(z)\big\|^2$ respectively (Lemma 1) and overall convergence rate over K iterations. Then, we analyze the out-of-distribution effect on the convergence rate.

8.2. Proof of Lemma 1

Proof. This proof demonstrates a sufficient condition to ensure a robust L2O model with per iteration convergence guarantee.

Due to the proof in Section A.2. in [14], $\big\|\mathbf{N}_2(z)\big\| \to 0$ when $x^+ \to x^*$. We frist derive the conditions for $\mathbf{N}_2(z) = \mathbf{0}$ case. Based on the solutions, we derive the conditions for $\mathbf{N}_2(z)$ in $\mathbf{N}_2(z) \neq \mathbf{0}$ case.

Case 1 $\mathbf{N}_2(z) = \mathbf{0}$**.**

$$\begin{aligned} &-\nabla f(x)^\top \Big(\operatorname{diag}\big(\mathbf{N}_1(z)\big) - L\operatorname{diag}\big(\mathbf{N}_1(z)\big)^2\Big)\nabla f(x) - \nabla f(x)^\top \mathbf{N}_2(z) + L\big\|\mathbf{N}_2(z)\big\|^2 \\ =& -\nabla f(x)^\top \Big(\operatorname{diag}\big(\mathbf{N}_1(z)\big) - L\operatorname{diag}\big(\mathbf{N}_1(z)\big)^2\Big)\nabla f(x), \\ =& -\nabla f(x)^\top \Big(\operatorname{diag}(\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2)\Big)\nabla f(x), \end{aligned} \tag{15}$$

where $\mathbf{N}_1(z)^2$ represents coordinate-wise square over a vector. To keep $-\nabla f(x)^\top \Big(\operatorname{diag}\left(\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2\right)\Big)\nabla f(x) \leq 0, \forall \nabla f(x)$, we should have:

$$\operatorname{diag}\left(\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2\right) \succeq 0,$$
$$\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2 \geq 0,$$

where $L > 0$ is given by L-smoothness definition in equation 2.1. Solving the above quadratic inequality, we have the following range for $\mathbf{N}_1(z)$:

$$0 \leq \mathbf{N}_1(z) \leq \frac{1}{L}, \forall z \in \mathcal{Z}. \tag{16}$$

The left-hand side $\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2$ achieves maxima $\frac{1}{4L}$ at $\mathbf{N}_1(z) = \frac{1}{2L}$.

Hence, due to the L-smoothness of $f(x)$, InD's convergence gain in one iteration has the following lower bound:

$$-\frac{L}{4} \leq -\frac{\|\nabla f(x)\|^2}{4L} \leq -\nabla f(x)^\top \Big(\operatorname{diag}\left(\mathbf{N}_1(z)\right) - L \operatorname{diag}\left(\mathbf{N}_1(z)\right)^2\Big)\nabla f(x). \tag{17}$$

Case 2 $\mathbf{N}_2(z) \neq \mathbf{0}$. We first freeze $-\nabla f(x)^\top \mathbf{N}_2(z) + L\big\|\mathbf{N}_2(z)\big\|^2$ and apply the derivation in Case 1 to keep a non-positive $-\nabla f(x)^\top \Big(\operatorname{diag}\left(\mathbf{N}_1(z)\right) - L \operatorname{diag}\left(\mathbf{N}_1(z)\right)^2\Big)\nabla f(x)$. Similar to the demonstration of $\mathbf{N}_1(z)$, we construct the following inequation for $\mathbf{N}_2(z)$:

$$-\nabla f(x)^\top \mathbf{N}_2(z) + L\big\|\mathbf{N}_2(z)\big\|^2 \leq 0.$$

Suppose the angle between $\mathbf{N}_2(z)$ and $\nabla f(x)$ is θ. The left-hand side of the above equation can be represented by:

$$-\|\nabla f(x)\| * \big\|\mathbf{N}_2(z)\big\| \cos(\theta) + L\big\|\mathbf{N}_2(z)\big\|^2 = \big\|\mathbf{N}_2(z)\big\|\Big(-\|\nabla f(x)\| \cos(\theta) + L\big\|\mathbf{N}_2(z)\big\|\Big) \leq 0.$$

Note that θ should follow $\theta \in [0, \frac{\pi}{2}]$ to avoid inherent non-negative of left-hand side in the above inequalities. Solve the above inequality, we have:

$$0 \leq \big\|\mathbf{N}_2(z)\big\| \leq \frac{\|\nabla f(x)\| \cos(\theta)}{L} \leq \frac{\|\nabla f(x)\|}{L}. \tag{18}$$

Substituting it back, we get minima at $\big\|\mathbf{N}_2(z)\big\| = \frac{\|\nabla f(x)\|}{2L}$ as $-\frac{\|\nabla f(x)\|^2}{4L}$. Note that if $\theta = 0$, the above equation achieves maxima at $-\frac{1}{4L}\|\nabla f(x)\|^2$, which means $\mathbf{N}_2(z)$ is in the same direction with $\nabla f(x)$, i.e., $\mathbf{N}_2(z) = \frac{\nabla f(x)}{2L}$.

□

8.3. Proof of Corollary 1

Proof. From the proof of Lemma 1, we construct separate conditions for the output of neural networks $\mathbf{N}_1$ and $\mathbf{N}_2$ by decomposing the per iteration convergence rate into two quadratic formulas with respect to the neural network models in the L2O model. This proof demonstrates the most robust L2O model in the InD scenario, which achieves the per iteration convergence gain of gradient descent.

The quadratic formular $\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2$ with respect to $\mathbf{N}_1(z)$ achieves maxima $\frac{1}{4L}$ at $\mathbf{N}_1(z) = \frac{1}{2L}$. The quadratic formular $\big\|\mathbf{N}_2(z)\big\|\Big(-\|\nabla f(x)\| \cos(\theta) + L\big\|\mathbf{N}_2(z)\big\|\Big)$ with respect to $\mathbf{N}_2(z)$ achieves minima $-\frac{\|\nabla f(x)\|^2}{4L}$ at $\big\|\mathbf{N}_2(z)\big\| = \frac{\|\nabla f(x)\|}{2L}$. We derive the best convergence rate after K iterations in this part for InD and OOD cases, respectively. The convexity of $f(x)$ yields $f(x) \leq f(x^*) + \nabla f(x)^\top (x - x^*)$ [27].

Due to equation 16 and equation 18, when $\mathbf{N}_1(z) = \frac{1}{2L}$ and $\mathbf{N}_1(z) = \frac{\nabla f(x)}{2L}$, the update formula for one iteration is as following, which is precisely gradient descent with $\frac{1}{L}$ step size.

$$x^+ = x - \frac{1}{2L}\nabla f(x) - \frac{\nabla f(x)}{2L} = x - \frac{1}{L}\nabla f(x). \tag{19}$$

We note that this corollary demonstrates that the L2O model can achieve the best upper-bound convergence rate of gradient descent. However, its lower bound is non-deterministic and relies on training.

Based on the definition of L-smoothness on f, we have:

$$F(x^+) \leq F(x^*) + \nabla f(x)^\top (x - x^*) - \frac{\|\nabla f(x)\|^2}{2L}.$$

In k-th iteration, $k \geq 1$, we have:

$$\begin{aligned}
F(x_k) - F(x^*) &\leq \nabla f(x_{k-1})^\top (x_{k-1} - x^*) - \frac{\left\|\nabla f(x_{k-1})\right\|^2}{2L}, \\
&= \frac{1}{2L}\Big(2L\nabla f(x_{k-1})^\top (x_{k-1} - x^*) - \left\|\nabla f(x_{k-1})\right\|^2\Big), \\
&= \frac{1}{2L}\Big(2L\nabla f(x_{k-1})^\top (x_{k-1} - x^*) - \left\|\nabla f(x_{k-1})\right\|^2 - L^2\|x_{k-1} - x^*\|^2 + L^2\|x_{k-1} - x^*\|^2\Big), \\
&= \frac{1}{2L}\Big(L^2\|x_{k-1} - x^*\|^2 - \left\|L(x_{k-1} - x^*) - \nabla f(x_{k-1})\right\|^2\Big), \\
&= \frac{1}{2L}\Big(L^2\|x_{k-1} - x^*\|^2 - L^2\left\|x_{k-1} - \frac{1}{L}\nabla f(x_{k-1}) - x^*\right\|^2\Big), \\
\text{(Due to equation 19)} &= \frac{L}{2}\Big(\|x_{k-1} - x^*\|^2 - \|x_k - x^*\|^2\Big).
\end{aligned}$$

Sum over all K iterations, we have:

$$\sum_{k=1}^{K} F(x_k) - F(x^*) \leq \frac{L}{2}\sum_{k=1}^{K}\left(\|x_{k-1} - x^*\|^2 - L^2\|x_k - x^*\|^2\right) = \frac{L}{2}\|x_0 - x^*\|^2.$$

Since $F(x_K) - F(x^*)$ is the minimum of the left-hand side of the above, we have:

$$F(x_K) - F(x^*) \leq \frac{L}{2K}\|x_0 - x^*\|^2.$$

□

8.4. Proof of Theorem 1

Proof. In the proof of Lemma 1, we have demonstrated that to ensure a robust L2O with a homogeneous convergence gain for any x and F, we should bound the neural networks $\mathbf{N}_1(z)$ and $\mathbf{N}_2(z)$ in the L2O model into those compact sets respectively. In Corollary 1, we give one sufficient condition to ensure the best robustness for the L2O model. In this proof, upon Corollary 1, we formulate the L2O model's convergence in the OOD scenario in this proof.

Convergence Gain by $\mathbf{N}_1(z)$. Due to equation 12, we have the following equation:

$$\begin{aligned}
&\mathbf{N}_1(z + s') - L\mathbf{N}_1(z + s')^2 \\
=&\mathbf{N}_1(z) + \mathbf{J}_1 s' - L\big(\mathbf{N}_1(z) + v\big)^2, \\
=&\mathbf{N}_1(z) + \mathbf{J}_1 s' - L\big(\mathbf{N}_1(z)^2 + (\mathbf{J}_1 s')^2 + 2\mathbf{N}_1(z)\mathbf{J}_1 s'\big), \\
=&\underbrace{\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2}_{①} + \underbrace{\big(1 - 2L\mathbf{N}_1(z)\big)\mathbf{J}_1 s' - L(\mathbf{J}_1 s')^2}_{②}.
\end{aligned} \tag{20}$$

Term ② in equation 20 is a quadratic formula of v with the following range thanks to v:

$$\begin{aligned}
0 \leq ② \leq \frac{\big(1 - 2L\mathbf{N}_1(z)\big)^2}{4L}&, \text{ if } 0 \leq \mathbf{J}_1 s' \leq \frac{1}{L} - 2\mathbf{N}_1(z), \\
② < 0&, \text{ if } \mathbf{J}_1 s' < 0 \text{ or } \mathbf{J}_1 s' > \frac{1}{L} - 2\mathbf{N}_1(z).
\end{aligned} \tag{21}$$

Due to equation 20, we have the following OOD convergence gain on shifted data $z + s'$:

$$
\begin{aligned}
&- \nabla f'(x+s)^\top \Big(\operatorname{diag}\big(\mathbf{N}_1(z+s')\big) - L \operatorname{diag}\big(\mathbf{N}_1(z+s')\big)^2 \Big) \nabla f'(x+s) \\
&= - \nabla f'(x+s)^\top \operatorname{diag}\Big(\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2 + \big(1 - 2L\mathbf{N}_1(z)\big)\mathbf{J}_1 s' - L(\mathbf{J}_1 s')^2\Big) \nabla f'(x+s), \\
&= - \nabla f'(x+s)^\top \operatorname{diag}\big(\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2\big) \nabla f'(x+s) - \nabla f'(x+s)^\top \operatorname{diag}\Big(\big(1 - 2L\mathbf{N}_1(z)\big)\mathbf{J}_1 s' - L(\mathbf{J}_1 s')^2\Big) \nabla f'(x+s), \\
&= - \nabla f'(x+s)^\top \operatorname{diag}\big(\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2\big) \nabla f'(x+s) - \nabla f'(x+s)^\top \operatorname{diag}\Big(\big(1 - 2L\mathbf{N}_1(z)\big)\mathbf{J}_1 s' - L(\mathbf{J}_1 s')^2\Big) \nabla f'(x+s).
\end{aligned} \tag{22}
$$

Moreover, we derive the following equation of convergence gain with respect to the L2O model's virtual feature s' (the difference between OOD and InD features):

$$
\begin{aligned}
&- \nabla f'(x+s)^\top \operatorname{diag}\big(\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2\big) \nabla f'(x+s) - \nabla f'(x+s)^\top \operatorname{diag}\Big(\big(1 - 2L\mathbf{N}_1(z)\big)\mathbf{J}_1 s' - L\big(\mathbf{J}_1 s'\big)^2\Big) \nabla f'(x+s) \\
&= - \nabla f'(x+s)^\top \operatorname{diag}\big(\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2\big) \nabla f'(x+s) \\
&\quad - \nabla f'(x+s)^\top \Big(\operatorname{diag}\big(1 - 2L\mathbf{N}_1(z)\big) \operatorname{diag}(\mathbf{J}_1 s') - L \operatorname{diag}\big((\mathbf{J}_1 s')^\top (\mathbf{J}_1 s')\big) \Big) \nabla f'(x+s).
\end{aligned} \tag{23}
$$

The above equation is lower bounded by $-\nabla f'(x+s)^\top \operatorname{diag}\big(\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2\big) \nabla f'(x+s) + \frac{1}{4L} \nabla f'(x+s)^\top \operatorname{diag}\Big(\big(1 - 2L\mathbf{N}_1(z)\big)^2\Big) \nabla f'(x+s)$ when $\mathbf{J}_1 s' := \frac{1}{2L} - \mathbf{N}_1(z)$.

Moreover, if $\mathbf{N}_1(z) := \frac{1}{2L}, \forall z \in \mathcal{Z}_P$, which means the best robutstness is achieved on the training set, the convergence gain compared with InD decreases to:

$$
\begin{aligned}
&- \nabla f'(x+s)^\top \operatorname{diag}\left(\frac{1}{2L} - L\frac{1}{2L}^2\right) \nabla f'(x+s) \\
&- \nabla f'(x+s)^\top \big(\operatorname{diag}(1 - 2L\frac{1}{2L}) \operatorname{diag}(\mathbf{J}_1 s') - L \operatorname{diag}(\mathbf{J}_1 s')^\top \operatorname{diag}(\mathbf{J}_1 s') \big) \nabla f'(x+s) \\
&= - \frac{\|\nabla f'(x+s)\|^2}{4L} + L \nabla f'(x+s)^\top \operatorname{diag}(\mathbf{J}_1 s')^\top \operatorname{diag}(\mathbf{J}_1 s') \nabla f'(x+s), \\
&= - \frac{\|\nabla f'(x+s)\|^2}{4L} + L \big\| \operatorname{diag}(\mathbf{J}_1 s') \nabla f'(x+s) \big\|^2.
\end{aligned} \tag{24}
$$

The above result demonstrates that any non-zero $\mathbf{J}_1 s'$ leads to worse convergence gain, which implies that a more well-training L2O model leads to less generalization ability. Only inadequately trained L2O models can achieve increments of convergence.

Convergence Gain by $\mathbf{N}_2(z)$. Assume equation 18 holds in training:

$$
\begin{aligned}
&- \nabla f'(x+s)^\top \mathbf{N}_2(z+s') + L \big\| \mathbf{N}_2(z+s') \big\|^2 \\
&= - \nabla f'(x+s)^\top \big(\mathbf{N}_2(z) + \mathbf{J}_2 s' \big) + L \big\| \mathbf{N}_2(z) + \mathbf{J}_2 s' \big\|^2, \\
&= - \nabla f'(x+s)^\top \mathbf{N}_2(z) + L \big\| \mathbf{N}_2(z) \big\|^2 - \nabla f'(x+s)^\top \mathbf{J}_2 s' + L \|\mathbf{J}_2 s'\|^2 + 2L \mathbf{N}_2(z)^\top \mathbf{J}_2 s', \\
&= - \nabla f'(x+s)^\top \mathbf{N}_2(z) + L \big\| \mathbf{N}_2(z) \big\|^2 - \nabla f'(x+s)^\top \mathbf{J}_2 s' + L \|\mathbf{J}_2 s'\|^2 + 2L \mathbf{N}_2(z)^\top \mathbf{J}_2 s', \\
&= - \nabla f'(x+s)^\top \mathbf{N}_2(z) + L \big\| \mathbf{N}_2(z) \big\|^2 - \big(\nabla f'(x+s) - 2L \mathbf{N}_2(z) \big)^\top \mathbf{J}_2 s' + L \|\mathbf{J}_2 s'\|^2.
\end{aligned} \tag{25}
$$

If we further assume training leads to best convergence gain, i.e., $\mathbf{N}_2(z) = \frac{\nabla f(x)}{2L}, \forall z \in \mathcal{Z}_P$, which means best convergence gain achieved on a training set, we have the following per iteration convergence gain in the OOD scenario from $\mathbf{N}_2(z)$:

$$
\begin{aligned}
&-\nabla f'(x+s)^\top \mathbf{N}_2(z) + L\left\|\mathbf{N}_2(z)\right\|^2 - \left(\nabla f'(x+s) - 2L\mathbf{N}_2(z)\right)^\top \mathbf{J}_2 s' + L\|\mathbf{J}_2 s'\|^2 \\
=&-\frac{\nabla f'(x+s)^\top \nabla f(x)}{2L} + \frac{\|\nabla f(x)\|^2}{4L} - \left(\nabla f'(x+s) - 2L\frac{\nabla f(x)}{2L}\right)^\top \mathbf{J}_2 s' + L\|\mathbf{J}_2 s'\|^2, \\
=&-\frac{\nabla f'(x+s)^\top \nabla f(x)}{2L} + \frac{\|\nabla f(x)\|^2}{4L} \\
&+ L\left(-\frac{\left(\nabla f'(x+s) - \nabla f(x)\right)^\top}{L}\mathbf{J}_2 s' + \|\mathbf{J}_2 s'\|^2 + \frac{\left\|\nabla f'(x+s) - \nabla f(x)\right\|^2}{4L^2} - \frac{\left\|\nabla f'(x+s) - \nabla f(x)\right\|^2}{4L^2}\right), \\
=&-\frac{\nabla f'(x+s)^\top \nabla f(x)}{2L} + \frac{\|\nabla f(x)\|^2}{4L} - \frac{\|\nabla f'(x+s) - \nabla f(x)\|^2}{4L} + L\left\|\frac{\nabla f'(x+s) - \nabla f(x)}{2L} - \mathbf{J}_2 s'\right\|^2, \\
=&-\frac{\|\nabla f'(x+s)\|^2}{4L} + L\left\|\frac{\nabla f'(x+s) - \nabla f(x)}{2L} - \mathbf{J}_2 s'\right\|^2.
\end{aligned} \tag{26}
$$

Overall Convergence Gain of One Iteration. Sum up equation 24 and equation 26, we have the following OOD's integrated convergence gain of one iteration:

$$
-\frac{\left\|\nabla f'(x+s)\right\|^2}{2L} + \underbrace{L\left\|\operatorname{diag}(\mathbf{J}_1 s')\nabla f'(x+s)\right\|^2 + L\left\|\frac{\nabla f'(x+s) - \nabla f(x)}{2L} - \mathbf{J}_2 s'\right\|^2}_{③}. \tag{27}
$$

$-\frac{\left\|\nabla f'(x+s)\right\|^2}{2L}$ is equivalent to the convergence rate of gradient descent, which is also the most robust convergence rate that the L2O model could achieve in the InD scenario. Moreover, we would like the value of the above equation to be as small as possible. However, the non-negativity of ③ shows that the convergence gain deteriorates as long as OOD happens. □

8.5. Proof of Corollary 2

Proof. In this proof, we formulate the upper bound of per iteration convergence gain with respect to the L2O model input feature vectors.

Based on Triangle and Cauchy-Schwarz inequalities, we have:

$$
\begin{aligned}
&-\frac{\left\|\nabla f'(x+s)\right\|^2}{2L} + L\left\|\operatorname{diag}(\mathbf{J}_1 s')\nabla f'(x+s)\right\|^2 + L\left\|\frac{\nabla f'(x+s) - \nabla f(x)}{2L} - \mathbf{J}_2 s'\right\|^2 \\
=&-\frac{\left\|\nabla f'(x+s)\right\|^2}{2L} + L\left\|\operatorname{diag}(\mathbf{J}_1 s')\nabla f'(x+s)\right\|^2 + L\left\|\frac{\nabla f'(x+s) - \nabla f(x)}{2L} - \mathbf{J}_2 s'\right\|^2, \\
\leq&-\frac{\left\|\nabla f'(x+s)\right\|^2}{2L} + L\left\|\operatorname{diag}(\mathbf{J}_1 s')\nabla f'(x+s)\right\|^2 + 2L\left\|\frac{\nabla f'(x+s) - \nabla f(x)}{2L}\right\|^2 + 2L\|\mathbf{J}_2 s'\|^2, \\
=&-\frac{\left\|\nabla f'(x+s)\right\|^2}{2L} + \frac{\|\nabla f'(x+s) - \nabla f(x)\|^2}{2L} + L\left\|\operatorname{diag}(\mathbf{J}_1 s')\nabla f'(x+s)\right\|^2 + 2L\|\mathbf{J}_2 s'\|^2, \\
\leq&-\frac{\left\|\nabla f'(x+s)\right\|^2}{2L} + \frac{\|\nabla f'(x+s) - \nabla f(x)\|^2}{2L} + L\left\|\operatorname{diag}(\mathbf{J}_1 s')\nabla f'(x+s)\right\|^2 + 2L\|\mathbf{J}_2\|^2\|s'\|^2.
\end{aligned}
$$

Due to $\|\mathbf{J}_1\| \leq C_1\sqrt{n}$ and $\|\mathbf{J}_2\| \leq C_2\sqrt{n}$, the above inequality is upper bounded by:

$$
\begin{aligned}
&-\frac{\left\|\nabla f'(x+s)\right\|^2}{2L} + \frac{\|\nabla f'(x+s) - \nabla f(x)\|^2}{2L} + L\left\|\operatorname{diag}(\mathbf{J}_1 s')\nabla f'(x+s)\right\|^2 + 2L\|\mathbf{J}_2\|^2\|s'\|^2, \\
\leq&-\frac{\left\|\nabla f'(x+s)\right\|^2}{2L} + \frac{\|\nabla f'(x+s) - \nabla f(x)\|^2}{2L} + LC_1^2 n\|s'\|^2\nabla f'(x+s)^\top \mathbf{I}\nabla f'(x+s) + 2LC_2^2 n\|s'\|^2, \\
=&-\frac{\left\|\nabla f'(x+s)\right\|^2}{2L} + \frac{\|\nabla f'(x+s) - \nabla f(x)\|^2}{2L} + \left(LC_1^2 n\|\nabla f'(x+s)\|^2 + 2LC_2^2 n\right)\|s'\|^2.
\end{aligned} \tag{28}
$$

Due to the formulation of $\|s'\|$ in equation 13, if we set $s' := \nabla f'(x+s) - \nabla f(x)$, which means we remove the variable feature and only use gradient as the input feature of the L2O model, we can decrease such convergence gain's upper bound from the right-hand side of inequation 27 to:

$$-\frac{\left\|\nabla f'(x+s)\right\|^2}{2L} + \frac{\|\nabla f'(x+s) - \nabla f(x)\|^2}{2L} + \left(LC_1^2 n\|\nabla f'(x+s)\|^2 + 2LC_2^2\right)\|\nabla f'(x+s) - \nabla f(x)\|^2, \tag{29}$$

where we just replace $\|s'\|$ in inequation 27 with $\|\nabla f'(x+s) - \nabla f(x)\|$.

□

8.6. Proof of Theorem 2

Proof. In Sec. 3, we split the trajectory of OOD variable x' into a trajectory of InD variable x and a trajectory of the virtual variable s. x is updated with "well-trained" L2O with robustness guarantee (Corollary 2), which is independent of OOD and deterministically performs as gradient descent. Thus, the uncertainty of the OOD scenario can be formulated with respect to the virtual variable s. This proof constructs the formulation of multi-iteration convergence rate with respect to s.

First, we assume Corollary 2 hold, which ensures the L2O model's robust performance on InD variable x. Upon equation 12, we have following equalities between the L2O model's outputs of ODD and InD scenarios:

$$\begin{aligned}\mathbf{N}_1(z+s') &= \frac{1}{2L} + \mathbf{J}_1 s',\\ \mathbf{N}_2(z+s') &= \frac{\nabla f(x)}{2L} + \mathbf{J}_2 s',\end{aligned}$$

where as defined in Sec. 3, $z+s'$ and z represent L2O model's input features of OOD and InD scenarios respectively. And s' formulates the difference between them.

In k-th iteration, $k \geq 1$, based on the above equations, we can represent the L2O model's update formula in the OOD scenario as follows:

$$\begin{aligned}x_k + s_k &= x_{k-1} + s_{k-1} - \mathbf{N}_1(z_{k-1} + s'_{k-1})\nabla f'(x_{k-1} + s_{k-1}) - \mathbf{N}_2(z_{k-1} + s'_{k-1}),\\ &= x_{k-1} + s_{k-1} - \left(\frac{1}{2L} - \operatorname{diag}(\mathbf{J}_{1,k-1}s'_{k-1})\right)\nabla f'(x_{k-1} + s_{k-1}) - \left(\frac{\nabla f(x_{k-1})}{2L} + \mathbf{J}_{2,k-1}s'_{k-1}\right).\end{aligned} \tag{30}$$

From the above equation, we want to split the InD part on the InD variable x and the OOD part on the virtual variable s. By adding an entra term $\frac{\nabla f(x_{k-1})}{L}$, we have the following reformulations:

$$\begin{aligned}x_k + s_k =& x_{k-1} - \frac{\nabla f(x_{k-1})}{L} + s_{k-1} - \frac{\nabla f'(x_{k-1} + s_{k-1})}{2L} + \frac{\nabla f(x_{k-1})}{2L} - \operatorname{diag}(\mathbf{J}_{1,k-1}s'_{k-1})\nabla f'(x_{k-1} + s_{k-1})\\ &- \mathbf{J}_{2,k-1}s'_{k-1},\\ =& x_{k-1} - \frac{\nabla f(x_{k-1})}{L} + s_{k-1} - \frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} - \operatorname{diag}(\mathbf{J}_{1,k-1}s'_{k-1})\nabla f'(x_{k-1} + s_{k-1})\\ &- \mathbf{J}_{2,k-1}s'_{k-1}.\end{aligned} \tag{31}$$

Then, we can split the OOD trajectory of $x' = x + s$ into two parts: InD and OOD parts on x and s. First, we assume the InD variable x is updated with the following update formula with the gradient of InD objective $f(x_{k-1})$:

$$x_k = x_{k-1} - \frac{\nabla f(x_{k-1})}{L}, \tag{32}$$

For the x part, with an unchanged InD initial point $x_0 \in \mathcal{S}_P$ and an InD objective $f \in \mathcal{F}_{L,P}$ is solutions given by the L2O model are always in the InD scenario. The optimal solution x^* of f is deterministic as well.

Moreover, removing equation 32, the remaining terms of equation 31 constitutes the update formula on virtual variable s:

$$s_k = s_{k-1} - \frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} - \operatorname{diag}(\mathbf{J}_1 s'_{k-1})\nabla f'(x_{k-1} + s_{k-1}) - \mathbf{J}_2 s'_{k-1}.$$

Now is the time to derive the convergence rate. Based on the definition of L-smoothness on f', in k-th iteration, we have the following upper bound of OOD objective $f'(x_k + s_k)$:

$$
\begin{aligned}
&F'(x_k + s_k) \\
\leq &F'(x_{k-1} + s_{k-1}) + \nabla f'(x_{k-1} + s_{k-1})^\top \big(x_k + s_k - (x_{k-1} + s_{k-1})\big) + \frac{L}{2} \left\| x_k + s_k - (x_{k-1} + s_{k-1}) \right\|^2, \\
= &F'(x_{k-1} + s_{k-1}) \\
&+ \nabla f'(x_{k-1} + s_{k-1})^\top \left(-\frac{\nabla f(x_{k-1})}{L} - \frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} - \operatorname{diag}(\mathbf{J}_1 s'_{k-1}) \nabla f'(x_{k-1} + s_{k-1}) - \mathbf{J}_2 s'_{k-1} \right) \\
&+ \frac{L}{2} \left\| -\frac{\nabla f(x_{k-1})}{L} - \frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} - \operatorname{diag}(\mathbf{J}_1 s'_{k-1}) \nabla f'(x_{k-1} + s_{k-1}) - \mathbf{J}_2 s'_{k-1} \right\|^2, \\
\leq &F'(x^* + s^*) + \nabla f'(x_{k-1} + s_{k-1})^\top \big(x_{k-1} + s_{k-1} - (x^* + s^*)\big) \\
&+ \nabla f'(x_{k-1} + s_{k-1})^\top \left(-\frac{\nabla f(x_{k-1})}{L} - \frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} - \operatorname{diag}(\mathbf{J}_1 s'_{k-1}) \nabla f'(x_{k-1} + s_{k-1}) - \mathbf{J}_2 s'_{k-1} \right) \\
&+ \frac{L}{2} \left\| -\frac{\nabla f(x_{k-1})}{L} - \frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} - \operatorname{diag}(\mathbf{J}_1 s'_{k-1}) \nabla f'(x_{k-1} + s_{k-1}) - \mathbf{J}_2 s'_{k-1} \right\|^2,
\end{aligned}
$$

where in the second step, we import the L2O model's in OOD update process defined in equation 31. In the third step, we apply the definition of convexity on F'.

We make the following reformulation and denote the update on virtual variable s as Δs_{k-1}:

$$
\begin{aligned}
&F'(x_k + s_k) - F'(x^* + s^*) \\
\leq &\nabla f'(x_{k-1} + s_{k-1})^\top \big(x_{k-1} + s_{k-1} - (x^* + s^*)\big) \\
&+ \nabla f'(x_{k-1} + s_{k-1})^\top \left(-\frac{\nabla f(x_{k-1})}{L} - \frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} - \operatorname{diag}(\mathbf{J}_1 s'_{k-1}) \nabla f'(x_{k-1} + s_{k-1}) - \mathbf{J}_2 s'_{k-1} \right) \\
&+ \frac{L}{2} \left\| -\frac{\nabla f(x_{k-1})}{L} - \frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} - \operatorname{diag}(\mathbf{J}_1 s'_{k-1}) \nabla f'(x_{k-1} + s_{k-1}) - \mathbf{J}_2 s'_{k-1} \right\|^2, \\
= &\nabla f'(s_{k-1} + x_{k-1})^\top (s_{k-1} + x_{k-1} - (s^* + x^*)) \\
&+ \nabla f'(s_{k-1} + x_{k-1})^\top \\
&\quad \Bigg(x_{k-1} - \frac{\nabla f(x_{k-1})}{L} - x^* - \big((s_{k-1} + x_{k-1}) - (s^* + x^*)\big) \\
&\qquad + s_{k-1} - \underbrace{\Big(\frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} + \operatorname{diag}(\mathbf{J}_1 s'_{k-1}) \nabla f'(x_{k-1} + s_{k-1}) + \mathbf{J}_2 s'_{k-1}\Big)}_{\Delta s_{k-1}} - s^* \Bigg) \\
&+ \frac{L}{2} \Bigg\| x_{k-1} - \frac{\nabla f(x_{k-1})}{L} - x^* - \big((s_{k-1} + x_{k-1}) - (s^* + x^*)\big) \\
&\qquad + s_{k-1} - \underbrace{\frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} + \operatorname{diag}(\mathbf{J}_1 s'_{k-1}) \nabla f'(x_{k-1} + s_{k-1}) + \mathbf{J}_2 s'_{k-1}}_{\Delta s_{k-1}} - s^* \Bigg\|^2,
\end{aligned}
$$

where we place right-hand side items according to whether they are updates for x_{k-1} or s_{k-1}. Then, we can simplify the

above inequation with Δs_{k-1} as follows:

$$
\begin{aligned}
&F'(x_k+s_k)-F'(x^*+s^*)\\
\leq&\nabla f'(s_{k-1}+x_{k-1})^\top(s_{k-1}+x_{k-1}-(s^*+x^*))\\
&+\nabla f'(s_{k-1}+x_{k-1})^\top\Big(x_{k-1}-\frac{\nabla f(x_{k-1})}{L}-x^*+s_{k-1}-\Delta s_{k-1}-s^*-\big((s_{k-1}+x_{k-1})-(s^*+x^*)\big)\Big)\\
&+\frac{L}{2}\Big\|x_{k-1}-\frac{\nabla f(x_{k-1})}{L}-x^*+s_{k-1}-\Delta s_{k-1}-s^*-\big((s_{k-1}+x_{k-1})-(s^*+x^*)\big)\Big\|^2,\\
=&\nabla f'(s_{k-1}+x_{k-1})^\top(s_{k-1}+x_{k-1}-(s^*+x^*))-\nabla f'(s_{k-1}+x_{k-1})^\top\big(s_{k-1}+x_{k-1}-(s^*+x^*)\big)\\
&+\nabla f'(s_{k-1}+x_{k-1})^\top\Big(x_{k-1}-\frac{\nabla f(x_{k-1})}{L}-x^*+s_{k-1}-\Delta s_{k-1}-s^*\Big)\\
&+\frac{L}{2}\Big\|x_{k-1}-\frac{\nabla f(x_{k-1})}{L}-x^*+s_{k-1}-\Delta s_{k-1}-s^*-\big((s_{k-1}+x_{k-1})-(s^*+x^*)\big)\Big\|^2,\\
=&\nabla f'(s_{k-1}+x_{k-1})^\top\Big(x_{k-1}-\frac{\nabla f(x_{k-1})}{L}-x^*+s_{k-1}-\Delta s_{k-1}-s^*\Big)\\
&+\frac{L}{2}\Big\|x_{k-1}-\frac{\nabla f(x_{k-1})}{L}-x^*+s_{k-1}-\Delta s_{k-1}-s^*-\big((s_{k-1}+x_{k-1})-(s^*+x^*)\big)\Big\|^2,
\end{aligned}
$$

where in the second and third steps, we combine the similar terms of the right-hand side.

We continue by expanding the quadratic formula of the right-hand side and making up new ones as follows:

$$
\begin{aligned}
&\nabla f'(s_{k-1}+x_{k-1})^\top\left(x_{k-1}-\frac{\nabla f(x_{k-1})}{L}-x^*\right)+\nabla f'(s_{k-1}+x_{k-1})^\top(s_{k-1}-\Delta s_{k-1}-s^*)\\
&+\frac{L}{2}\Big\|x_{k-1}-\frac{\nabla f(x_{k-1})}{L}-x^*+s_{k-1}-\Delta s_{k-1}-s^*-\big((s_{k-1}+x_{k-1})-(s^*+x^*)\big)\Big\|^2\\
=&\Big(\nabla f'(s_{k-1}+x_{k-1})-L\big((s_{k-1}+x_{k-1})-(s^*+x^*)\big)\Big)^\top\left(x_{k-1}-\frac{\nabla f(x_{k-1})}{L}-x^*+s_{k-1}-\Delta s_{k-1}-s^*\right)\\
&+\frac{L}{2}\left\|x_{k-1}-\frac{\nabla f(x_{k-1})}{L}-x^*+s_{k-1}-\Delta s_{k-1}-s^*\right\|^2+\frac{L}{2}\|(s_{k-1}+x_{k-1})-(s^*+x^*)\|^2,\\
=&\frac{L}{2}\left(2\Big(\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\big((s_{k-1}+x_{k-1})-(s^*+x^*)\big)\Big)^\top\left(x_{k-1}-\frac{\nabla f(x_{k-1})}{L}-x^*+s_{k-1}-\Delta s_{k-1}-s^*\right)\right.\\
&\qquad\left.+\|x_{k-1}-\frac{\nabla f(x_{k-1})}{L}-x^*+s_{k-1}-\Delta s_{k-1}-s^*\|^2\right)+\frac{L}{2}\|(s_{k-1}+x_{k-1})-(s^*+x^*)\|^2,\\
=&\frac{L}{2}\left(\left\|\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\frac{\nabla f(x_{k-1})}{L}-\Delta s_{k-1}\right\|^2-\left\|\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\big((s_{k-1}+x_{k-1})-(s^*+x^*)\big)\right\|^2\right)\\
&+\frac{L}{2}\|(s_{k-1}+x_{k-1})-(s^*+x^*)\|^2,
\end{aligned}
$$

where we expand a quadratic formula in the first step and make up a new quadratic formula in the second and thrid steps.

We add an extra term $\frac{L}{2}\|(s_k+x_k)-(s^*+x^*)\|^2-\frac{L}{2}\|(s_k+x_k)-(s^*+x^*)\|^2$ to expand the second quadratic formular:

$$
\begin{aligned}
&\frac{L}{2}\left(\left\|\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\frac{\nabla f(x_{k-1})}{L}-\Delta s_{k-1}\right\|^2-\left\|\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\big((s_{k-1}+x_{k-1})-(s^*+x^*)\big)\right\|^2\right)\\
&+\frac{L}{2}\|(s_{k-1}+x_{k-1})-(s^*+x^*)\|^2+\frac{L}{2}\|(s_k+x_k)-(s^*+x^*)\|^2-\frac{L}{2}\|(s_k+x_k)-(s^*+x^*)\|^2\\
=&\frac{L}{2}\Bigg(\left\|\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\frac{\nabla f(x_{k-1})}{L}-\Delta s_{k-1}\right\|^2\\
&\quad-\left\|\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\big((s_{k-1}+x_{k-1})-(s^*+x^*)\big)\right\|^2+\|(s_k+x_k)-(s^*+x^*)\|^2\Bigg)\\
&+\frac{L}{2}\|(s_{k-1}+x_{k-1})-(s^*+x^*)\|^2-\frac{L}{2}\|(s_k+x_k)-(s^*+x^*)\|^2,\\
=&\frac{L}{2}\Bigg(\left\|\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\frac{\nabla f(x_{k-1})}{L}-\Delta s_{k-1}\right\|^2+\frac{L}{2}\|(s_{k-1}+x_{k-1})-(s^*+x^*)\|^2-\frac{L}{2}\|(s_k+x_k)-(s^*+x^*)\|^2\\
&+\left((s_k+x_k)-(s_{k-1}+x_{k-1})+\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}\right)^\top\\
&\left((s_k+s_{k-1}+x_k+x_{k-1})-2(s^*+x^*)-\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}\right)\Bigg).
\end{aligned}
$$

Based on the definition of x and s updates, we can further combine the first and second terms in the above formulation after expanding the first quadratic formula:

$$
\begin{aligned}
&\frac{L}{2}\Bigg(\left\|\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\frac{\nabla f(x_{k-1})}{L}-\Delta s_{k-1}\right\|^2\\
&+\left((s_k+x_k)-(s_{k-1}+x_{k-1})+\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}\right)^\top\\
&\quad\left((s_k+s_{k-1}+x_k+x_{k-1})-2(s^*+x^*)-\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}\right)\Bigg)\\
&+\frac{L}{2}\|(s_{k-1}+x_{k-1})-(s^*+x^*)\|^2-\frac{L}{2}\|(s_k+x_k)-(s^*+x^*)\|^2,\\
=&\frac{L}{2}\Bigg(\left(\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\frac{\nabla f(x_{k-1})}{L}-\Delta s_{k-1}\right)^\top\Big(2(s_{k-1}+x_{k-1})-2(s^*+x^*)\\
&\quad-\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\frac{\nabla f(x_{k-1})}{L}-\Delta s_{k-1}+\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\frac{\nabla f(x_{k-1})}{L}-\Delta s_{k-1}\Big)\Bigg)\\
&+\frac{L}{2}\|(s_{k-1}+x_{k-1})-(s^*+x^*)\|^2-\frac{L}{2}\|(s_k+x_k)-(s^*+x^*)\|^2,\\
=&\frac{L}{2}\left(\left(\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}-\frac{\nabla f(x_{k-1})}{L}-\Delta s_{k-1}\right)^\top\left(2(s_{k-1}+x_{k-1})-2(s^*+x^*)-2\frac{\nabla f(x_{k-1})}{L}-2\Delta s_{k-1}\right)\right)\\
&+\frac{L}{2}\|(s_{k-1}+x_{k-1})-(s^*+x^*)\|^2-\frac{L}{2}\|(s_k+x_k)-(s^*+x^*)\|^2,\\
=&L\left(\frac{\nabla f'(s_{k-1}+x_{k-1})}{L}+x_k-x_{k-1}+s_k-s_{k-1}\right)^\top\big((s_k+x_k)-(s^*+x^*)\big)\\
&+\frac{L}{2}\|(s_{k-1}+x_{k-1})-(s^*+x^*)\|^2-\frac{L}{2}\|(s_k+x_k)-(s^*+x^*)\|^2.
\end{aligned}
$$

We subsitute Δs_{k-1} back and sum over K iterations to get final multi-iteration convergence rate:

$$
\begin{aligned}
&\sum_{k=1}^{K} F'(x_k + s_k) - F'(x^* + s^*) \\
\leq & L \sum_{k=1}^{K} \left(\frac{\nabla f'(s_{k-1} + x_{k-1})}{L} - \frac{\nabla f(x_{k-1})}{L} - \Delta s_{k-1} \right)^\top ((s_k + x_k) - (s^* + x^*)) \\
&+ \frac{L}{2} \sum_{k=1}^{K} \|(s_{k-1} + x_{k-1}) - (s^* + x^*)\|^2 - \frac{L}{2} \|(s_k + x_k) - (s^* + x^*)\|^2, \\
= & L \sum_{k=1}^{K} \left(\frac{\nabla f'(s_{k-1} + x_{k-1})}{L} - \frac{\nabla f(x_{k-1})}{L} - \Delta s_{k-1} \right)^\top (x_k - x^* + s_k - s^*) \\
&+ \frac{L}{2} (\|x_0 - x^* + s_0 - s^*\|^2 - \|x_K - x^* + s_K - s^*\|^2), \\
= & L \sum_{k=1}^{K} \left(\frac{\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})}{2L} - \mathrm{diag}(\mathbf{J}_1 s'_{k-1}) \nabla f'(x_{k-1} + s_{k-1}) - \mathbf{J}_2 s'_{k-1} \right)^\top (x_k - x^* + s_k - s^*) \\
&+ \frac{L}{2} (\|x_0 - x^* + s_0 - s^*\|^2 - \|x_K - x^* + s_K - s^*\|^2), \\
= & L \sum_{k=1}^{K} \left(\frac{\nabla f'(s_{k-1} + x_{k-1})}{L} + x_k - x_{k-1} + s_k - s_{k-1} \right)^\top (x_k - x^* + s_k - s^*) \\
&+ \frac{L}{2} (\|x_0 - x^* + s_0 - s^*\|^2 - \|x_K - x^* + s_K - s^*\|^2).
\end{aligned}
\tag{33}
$$

Since we have demonstrated that there may be no convergence guarantees in each iteration, we cannot directly split $F'(x_K + s_K) - F'(x^* + s^*)$ from the left-hand side of the above inequalities, we denote that $\min_{k=1,\ldots,K} F'(x_k + s_k) - F'(x^* + s^*)$ is minimal over $F'(x_j + s_j) - F'(x^* + s^*), j = 1, \ldots, K$, which is a scenario that leads to the best convergence rate upper bound. Without loss of generality, we can always find a k that minimizes all $F'(x_k + s_k) - F'(x^* + s^*), k \in [1, K]$.

After rearrangement, we have the following two equivalent expressions:

$$
\begin{aligned}
&\min_{k=1,\ldots,K} F'(x_k + s_k) - F'(x^* + s^*) \\
\leq & \frac{L}{2} \|x_0 - x^* + s_0 - s^*\|^2 - \frac{L}{2} \|x_K - x^* + s_K - s^*\|^2 \\
&+ \frac{1}{K} \sum_{k=1}^{K} \nabla f'(x_{k-1} + s_{k-1})^\top (x_k + s_k - x^* - s^*) + \frac{L}{K} \sum_{k=1}^{K} (x_k + s_k - x_{k-1} - s_{k-1})^\top (x_k + s_k - x^* - s^*), \\
= & \frac{L}{2} \|x_0 - x^* + s_0 - s^*\|^2 - \frac{L}{2} \|x_K - x^* + s_K - s^*\|^2 \\
&+ \frac{L}{K} \sum_{k=1}^{K} (x_k + s_k - x^* - s^*)^\top (x_k + s_k - (x_{k-1} + s_{k-1} - \frac{\nabla f'(x_{k-1} + s_{k-1})}{L})).
\end{aligned}
\tag{34}
$$

$$
\begin{aligned}
&\min_{k=1,\dots,K} F'(x_k+s_k) - F'(x^*+s^*) \\
\leq &\frac{L}{2}\|x_0 - x^* + s_0 - s^*\|^2 - \frac{L}{2}\|x_K - x^* + s_K - s^*\|^2 \\
&+ \frac{1}{K}\sum_{k=1}^{K} \nabla f'(x_{k-1}+s_{k-1})^\top (x_k + s_k - x^* - s^*) + \frac{L}{K}\sum_{k=1}^{K}(x_k + s_k - x_{k-1} - s_{k-1})^\top (x_k + s_k - x^* - s^*), \\
= &\frac{L}{2}\|x_0 + s_0 - x^* - s^*\|^2 - \frac{L}{2}\|x_K + s_K - x^* - s^*\|^2 \\
&+ \frac{1}{K}\sum_{k=1}^{K} \nabla f'(x_{k-1}+s_{k-1})^\top (x_k + s_k - x^* - s^*) + \frac{L}{K}\sum_{k=1}^{K}\left(-\frac{\nabla f(x_{k-1})}{L} - \Delta s'_{k-1}\right)^\top (x_k + s_k - x^* - s^*).
\end{aligned}
\tag{35}
$$

□

The above results imply that there is no convergence guarantee in OOD scenarios.

If not OOD, which means both variable and objective are from the InD scenario, i.e., $s := 0$, $f' = f$, the convergence rate upper bound is as follows, which is precisely that of gradient-descent:

$$
\begin{aligned}
&\frac{L}{2}\|x_0 - x^* + s_0 - s^*\|^2 - \frac{L}{2}\|x_K - x^* + s_K - s^*\|^2 \\
&+ \frac{1}{K}\sum_{k=1}^{K} \nabla f'(x_{k-1}+s_{k-1})^\top (x_k + s_k - x^* - s^*) + \frac{L}{K}\sum_{k=1}^{K}(x_k + s_k - x_{k-1} - s_{k-1})^\top (x_k + s_k - x^* - s^*) \\
= &\frac{L}{2}\|x_0 - x^*\|^2 - \frac{L}{2}\|x_K - x^*\|^2 + \frac{L}{K}\sum_{k=1}^{K}\frac{\nabla f(x_{k-1})}{L}^\top (x_k - x^*) + \frac{L}{K}\sum_{k=1}^{K}\left(-\frac{\nabla f(x_{k-1})}{L}\right)^\top (x_k - x^*), \\
= &\frac{L}{2}\|x_0 - x^*\|^2 - \frac{L}{2}\|x_K - x^*\|^2.
\end{aligned}
\tag{36}
$$

Note that $s := 0$ cannot lead to the convergence rate of gradient descent since the third term in equation 34 is non-zero and cannot be eliminated. Such an **upper** bound is trivially upper bounded by:

$$
\begin{aligned}
&F'(x_k+s_k) - F'(x^*+s^*) \\
\leq &\frac{L}{2}\|x_0 - x^* + s_0 - s^*\|^2 - \frac{L}{2}\|x_K - x^* + s_K - s^*\|^2 + \frac{1}{K}\sum_{k=1}^{K} \nabla f'(x_{k-1}+s_{k-1})^\top (x_k + s_k - x^* - s^*) \\
&+ \frac{L}{K}\sum_{k=1}^{K}(x_k + s_k - x_{k-1} - s_{k-1})^\top (x_k + s_k - x^* - s^*), \\
\leq &\frac{L}{2}\|x_0 - x^* + s_0 - s^*\|^2 - \frac{L}{2}\|x_K - x^* + s_K - s^*\|^2 + \frac{1}{K}\sum_{k=1}^{K} \|\nabla f'(x_{k-1}+s_{k-1})\|\|x_k + s_k - x^* - s^*\| \\
&+ \frac{L}{K}\sum_{k=1}^{K}\|x_k + s_k - x_{k-1} - s_{k-1}\|\|x_k + s_k - x^* - s^*\|.
\end{aligned}
\tag{37}
$$

Such an **upper** bound is trivially lower bounded as follows:

$$
\begin{aligned}
&\frac{L}{2}\|x_0 - x^* + s_0 - s^*\|^2 - \frac{L}{2}\|x_K - x^* + s_K - s^*\|^2 + \frac{1}{K}\sum_{k=1}^{K}\nabla f'(x_{k-1}+s_{k-1})^\top (x_k + s_k - x^* - s^*) \\
&+ \frac{L}{K}\sum_{k=1}^{K}(x_k + s_k - x_{k-1} - s_{k-1})^\top (x_k + s_k - x^* - s^*) \\
\geq &\frac{L}{2}\|x_0 - x^* + s_0 - s^*\|^2 - \frac{L}{2}\|x_K - x^* + s_K - s^*\|^2 - \frac{1}{K}\sum_{k=1}^{K}\|\nabla f'(x_{k-1}+s_{k-1})\|\|x_k + s_k - x^* - s^*\| \\
&- \frac{L}{K}\sum_{k=1}^{K}\|x_k + s_k - x_{k-1} - s_{k-1}\|\|x_k + s_k - x^* - s^*\|.
\end{aligned}
\tag{38}
$$

8.7. Proof of Corollary 3

Proof. This proof derives the upper bound with respect to the magnitude of virtual input feature $\|s'\|$ of the L2O model, where s' is proposed in Sec. 3 to formulate the difference between input features of the OOD and InD scenarios.

Substituting $\Delta s'_{k-1}$ yields:

$$
\begin{aligned}
&F'(x_k + s_k) - F'(x^* + s^*) \\
\leq &\frac{1}{K}\sum_{k=1}^{K}\nabla f'(x_{k-1}+s_{k-1})^\top (x_k + s_k - x^* - s^*) + \frac{L}{2}\|x_0 + s_0 - x^* - s^*\|^2 - \frac{L}{2}\|x_K + s_K - x^* - s^*\|^2 \\
&+ \frac{L}{K}\sum_{k=1}^{K}\left(-\frac{\nabla f(x_{k-1})}{L} - \Delta s'_{k-1}\right)^\top (x_k + s_k - x^* - s^*) \\
= &\frac{1}{K}\sum_{k=1}^{K}\nabla f'(x_{k-1}+s_{k-1})^\top (x_k + s_k - x^* - s^*) + \frac{L}{2}\|x_0 + s_0 - x^* - s^*\|^2 - \frac{L}{2}\|x_K + s_K - x^* - s^*\|^2 \\
&- \frac{L}{K}\sum_{k=1}^{K}\left(\frac{\nabla f(x_{k-1})}{L} + \frac{\nabla f'(x_{k-1}+s_{k-1}) - \nabla f(x_{k-1})}{2L} + \mathrm{diag}(\mathbf{J}_1 s'_{k-1})\nabla f'(x_{k-1}+s_{k-1}) + \mathbf{J}_2 s'_{k-1}\right)^\top \\
&\qquad (x_k + s_k - x^* - s^*).
\end{aligned}
\tag{39}
$$

By Cauchy-Schwarz inequality and Triangle inequality, we have:

$$
\begin{aligned}
&F'(x_k+s_k)-F'(x^*+s^*)\\
\leq&\frac{L}{2}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2}\|x_K+s_K-x^*-s^*\|^2\\
&+\frac{1}{2K}\sum_{k=1}^{K}\left(\nabla f'(x_{k-1}+s_{k-1})-\nabla f(x_{k-1})\right)^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}\left(-\operatorname{diag}(\mathbf{J}_1 s'_{k-1})\nabla f'(x_{k-1}+s_{k-1})-\mathbf{J}_2 s'_{k-1}\right)^\top(x_k+s_k-x^*-s^*)\\
\leq&\frac{L}{2}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2}\|x_K+s_K-x^*-s^*\|^2\\
&+\frac{1}{2K}\sum_{k=1}^{K}\left(\nabla f'(x_{k-1}+s_{k-1})-\nabla f(x_{k-1})\right)^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}(\left\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})\nabla f'(x_{k-1}+s_{k-1})\right\|+\|\mathbf{J}_2 s'_{k-1}\|)\|x_k+s_k-x^*-s^*\|\\
\leq&\frac{L}{2}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2}\|x_K+s_K-x^*-s^*\|^2\\
&+\frac{1}{2K}\sum_{k=1}^{K}\left(\nabla f'(x_{k-1}+s_{k-1})-\nabla f(x_{k-1})\right)^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}(C_1\sqrt{n}\|s'_{k-1}\|\|\nabla f'(x_{k-1}+s_{k-1})\|+C_2\sqrt{n}\|s'_{k-1}\|)\|x_k+s_k-x^*-s^*\|\\
=&\frac{L}{2}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2}\|x_K+s_K-x^*-s^*\|^2\\
&+\frac{1}{2K}\sum_{k=1}^{K}\left(\nabla f'(x_{k-1}+s_{k-1})-\nabla f(x_{k-1})\right)^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}C_1\sqrt{n}\|\nabla f'(x_{k-1}+s_{k-1})\|\|s'_{k-1}\|+\frac{L}{K}\sum_{k=1}^{K}C_2\sqrt{n}\|x_k+s_k-x^*-s^*\|\|s'_{k-1}\|.
\end{aligned}
\tag{40}
$$

□

8.8. Proof of Theorem 3

This proof demonstrates that if the FP and GC conditions hold, the L2O model follows the structure defined in Theorem 3 and yields a unique solution at each iteration. We construct the proof by following the workflow proposed in [14].

First, we define our gradient-only input feature construction. Then, we apply the lemma proposed in [14] to construct a few candidate parameter matrices. We propose a new formulation to achieve the two sufficient conditions, GC and FP. For the non-smooth part r in the objective, we apply the proximal gradient method [10] as in [14] to solve a solution.

As the summation of two convex functions, $F(x)$ is convex on x. We have $\mathbf{0}\in\partial F(x)$ that $\mathbf{0}\in\nabla f(x^*)+\partial r(x)$. We choose g_{x^*} as $\nabla f(x^*)$. When $k\to\infty$, by making the following denotation:

$$\hat{d}_k=d_k(\nabla f(x^*),-\nabla f(x^*),\mathbf{0}),$$

where $\nabla f(x^*)$ denotes gradient of optimal solution x^*. In the above definition, $\mathbf{0}$ means the results of historical modeling operator u reaching zero when $k\to\infty$. Our following demonstrations will also derive the formulation of u to ensure such a condition.

With $\hat{d}_k$, we can rewrite the L2O update formula of k-th iteration in equation 8 as:

$$x_k=x_{k-1}-d_k(\nabla f(x_{k-1}),g_k,v_{k-1})+d_k(\nabla f(x^*),-\nabla f(x^*),\nabla f(x^*),-\nabla f(x^*))-\hat{d}_k.$$

$g_k \in \partial r(x_k)$ represents an implicit subgradient vector at desired x_k, which yields the application of the proximal gradient algorithm [14].

We assume that there are following bounded parameter matrices:

$$\mathbf{J}_{j,k} \in \mathbb{R}^{n\times n}, \|\mathbf{J}_{j,k}\| \leq C\sqrt{n}, \quad \forall j = 1, 2, 3,$$

where C is the upper bound on the Jacobian matrix of d's function space. Without loss of generality, such an assumption is a general setting by setting a bounded activation function in machine learning [14].

Based on Lemma 1 in Section A.1. of [14], we can represent d_k with the above bounded parameter matrices as follows:

$$\begin{aligned}
x_k &= x_{k-1} - \mathbf{J}_{1,k}(\nabla f(x_{k-1}) - \nabla f(x^*)) - \mathbf{J}_{2,k}(g_k + \nabla f(x^*)) - \mathbf{J}_{3,k}(v_{k-1} - \mathbf{0}) - \hat{d}_k, \\
&= x_{k-1} - \mathbf{J}_{1,k}(\nabla f(x_{k-1}) - \nabla f(x^*)) - \mathbf{J}_{2,k}(g_k + \nabla f(x^*)) - \mathbf{J}_{2,k}(\nabla f(x_{k-1}) - \nabla f(x^*)) + \mathbf{J}_{2,k}(\nabla f(x_{k-1}) - \nabla f(x^*)) \\
&\quad - \mathbf{J}_{3,k}v_{k-1} - \hat{d}_k, \\
&= x_{k-1} - \mathbf{J}_{2,k}\nabla f(x_{k-1}) - \mathbf{J}_{2,k}g_k - \mathbf{J}_{3,k}v_{k-1} \\
&\quad - (\mathbf{J}_{1,k} - \mathbf{J}_{2,k})(\nabla f(x_{k-1}) - \nabla f(x^*)) - \hat{d}_k.
\end{aligned}$$

In the second and third steps, we unify the parameters for smooth part and non-smooth part by the $-(\mathbf{J}_{1,k} - \mathbf{J}_{2,k})(\nabla f(x_k) - \nabla f(x^*))$ term. When $k \to \infty$, the above equality becomes:

$$\begin{aligned}
x_k =& x_{k-1} - \mathbf{J}_{2,k}\nabla f(x^*) - \mathbf{J}_{2,k}(-\nabla f(x^*)) - \mathbf{J}_{3,k}\mathbf{0} - (\mathbf{J}_{1,k} - \mathbf{J}_{2,k})(\nabla f(x^*) - \nabla f(x^*)) - \mathbf{0}, \\
=& x_{k-1},
\end{aligned}$$

where we define $\lim_{k\to\infty} v_k = \mathbf{0}$. At each iteration, given a group of parameter $\mathbf{J}$ and b, the solution x_k is uniquely constructed. Based on the method proposed in [14], we construct such parameters by learning. We define the learnable parameters as follows:

$$\begin{aligned}
\mathbf{R}_k :=& \mathbf{J}_{2,k}, \\
\mathbf{Q}_k :=& \mathbf{J}_{3,k}, \\
b_{1,k} :=& (\mathbf{J}_{1,k} - \mathbf{J}_{2,k})(\nabla f(x_{k-1}) - \nabla f(x^*)) + \hat{d}_k.
\end{aligned}$$

Thus, the update of solution is given by:

$$x_k = x_{k-1} - \mathbf{R}_k\nabla f(x_{k-1}) - \mathbf{R}_k g_k - \mathbf{Q}_k v_{k-1} - b_{1,k}. \tag{41}$$

As demonstrated in [14], all terms in $b_{1,k}$ reach zero as the iteration reaches ∞. We note that all defined parameter matrices are bounded by Lemma 1 in [14]. From Triangle and Cauchy Schwarz inequalities, $b_{1,k}$ is also bounded by:

$$\|b_{1,k}\| \leq 2\sqrt{n}C\|\nabla f(x_k) - \nabla f(x^*)\| + \|\hat{d}_k\|.$$

Here, we eliminate the requirement on an extra parameter matrix to control the boundness of $b_{1,k}$ in [14]. Moreover, we note that $b_{1,k}$ can be arbitrarily defined, which means it may be either non-negative or non-positive. This observation implies that both negative and positive implementations are available. In our implementation, we following [14] and use non-negative $b_{1,k}$.

Then, we derive the update formulation for v_k. Following [14], we set the length of historical information $T = 2$. We define the following operator to generate the historical feature vector v_k:

$$v_k = u_k(\nabla f(x_{k-1}) + g_{k-1}, v_{k-1}),$$

where we use explicit subgradient vector g_{k-1}. As defined in Sec. 5, we can recover subgradient vector g_k after solving x_k. Based on the L2O model defined in equation 41, assume $\mathbf{R}_k \succ 0$, we have the following equation to get the summation of gradient and subgradient:

$$\nabla f(x_{k-1}) + g_{k-1} = \mathbf{R}_k^{-1}(x_{k-1} - x_k - \mathbf{Q}_k v_{k-1} - b_{1,k}).$$

Moreover, we take a recurrent definition of the operator u, which takes the output of the last iteration v_{k-1} as the second input.

Suppose $\mathbf{0} := u_k(\mathbf{0}, \mathbf{0})$, which means when $k \to \infty$, the inputs of u are all the gradient (and subgradient) at optimal solution and the output of u converge to the gradient (and subgradient) at optimal solution as well. Suppose there are following bounded parameter matrices:

$$\mathbf{J}_{j,k} \in \mathbb{R}^{n\times n}, \|\mathbf{J}_{j,k}\| \leq \sqrt{n}C, \quad \forall j = 5, 6.$$

Denote $G_{k-1} = \nabla f(x_{k-1}) + g_{k-1}$, we have:

$$\begin{aligned}
v_k =& u_k(G_{k-1}, v_{k-1}) - u(\mathbf{0}, \mathbf{0}) + \mathbf{0},\\
=& \mathbf{J}_{5,k}(G_{k-1} - \mathbf{0}) + \mathbf{J}_{6,k}(v_{k-1} - \mathbf{0}) + \mathbf{0},\\
=& \mathbf{J}_{5,k}G_{k-1} + \mathbf{J}_{6,k}v_{k-1} + (\mathbf{I} - \mathbf{J}_{5,k} - \mathbf{J}_{6,k})\mathbf{0},\\
=& \mathbf{J}_{5,k}G_{k-1} + (\mathbf{I} - \mathbf{J}_{5,k} - \mathbf{J}_{6,k})G_{k-1} + \mathbf{J}_{6,k}v_{k-1} - (\mathbf{I} - \mathbf{J}_{5,k} - \mathbf{J}_{6,k})(G_{k-1} - \mathbf{0}),\\
=& (\mathbf{I} - \mathbf{J}_{6,k})G_{k-1} + \mathbf{J}_{6,k}v_{k-1} - (\mathbf{I} - \mathbf{J}_{5,k} - \mathbf{J}_{6,k})(G_{k-1} - \mathbf{0}).
\end{aligned}$$

Here, we construct a reaching zero term $G_{k-1} - \mathbf{0}$ w.r.t. G_{k-1}, which imply that G_{k-1} may not be exactly equal to $\mathbf{0}$. We define the learnable parameters as follows:

$$\begin{aligned}
\mathbf{B}_k :=& \mathbf{J}_{6,k},\\
b_{2,k} :=& (\mathbf{I} - \mathbf{J}_{5,k} - \mathbf{J}_{6,k})(G_{k-1} - \mathbf{0}).
\end{aligned}$$

Assume $v_0 := 0$, at k-th iteration, the historical information v_k is given by the following equation:

$$v_k = (\mathbf{I} - \mathbf{B}_k)G_{k-1} + \mathbf{B}_k v_{k-1} - b_{2,k}. \tag{42}$$

Motivated by the momentum scheme in FISTA [5], we set $\mathbf{B}_k$ to be negative semi-definite. Thus, v_k illustrates the momentum of the gradient at x_{k-1}. It is worth noting that the above formulation is more like the classical momentum method, different from the Nesterov gradient method in FISTA [5] and Math-L2O [14].

At $k-1$-th iteration, v_{k-1} is yielded by:

$$v_{k-1} = (\mathbf{I} - \mathbf{B}_{k-1})G_{k-2} + \mathbf{B}_{k-1}v_{k-2} - b_{2,k-1}.$$

Substituting v_{k-1} into equation 41 yields the following complete formulation to generate x_k at k-th iteration:

$$\begin{aligned}
x_k &= x_{k-1} - \mathbf{R}_k\nabla f(x_{k-1}) - \mathbf{R}_k g_k - \mathbf{Q}_k v_{k-1} - b_{1,k}\\
&= x_{k-1} - \mathbf{R}_k\nabla f(x_{k-1}) - \mathbf{Q}_k((\mathbf{I} - \mathbf{B}_{k-1})G_{k-2} + \mathbf{B}_{k-1}v_{k-2} - b_{2,k-1}) - b_{1,k} - \mathbf{R}_k g_k.
\end{aligned} \tag{43}$$

We follow the method proposed in [14] to derive a unique solution of x_k based on the first-order derivative condition of non-smooth convex optimization. For non-smooth convex objective $r(x)$, $0 \in \partial r(x)$ is a sufficient and necessary condition for its optimality. We rewrite equation 43 as:

$$x_k + \mathbf{R}_k g_k = x_{k-1} - \mathbf{R}_k\nabla f(x_{k-1}) - \mathbf{Q}_k((\mathbf{I} - \mathbf{B}_{k-1})G_{k-2} + \mathbf{B}_{k-1}v_{k-2} - b_{2,k-1}) - b_{1,k}.$$

Since $g_k \in \partial r(x_k)$, we have:

$$x_{k-1} - \mathbf{R}_k\nabla f(x_{k-1}) - \mathbf{Q}_k((\mathbf{I} - \mathbf{B}_{k-1})G_{k-2} + \mathbf{B}_{k-1}v_{k-2} - b_{2,k-1}) - b_{1,k} \in x_k + \mathbf{R}_k\partial r(x_k).$$

After rearrangement, we have:

$$0 \in \mathbf{R}_k\partial r(x_k) + x_k - (x_{k-1} - \mathbf{R}_k\nabla f(x_{k-1}) - \mathbf{Q}_k((\mathbf{I} - \mathbf{B}_{k-1})G_{k-2} + \mathbf{B}_{k-1}v_{k-2} - b_{2,k-1}) - b_{1,k})\,.$$

Given $\mathbf{R}_k$ as a symmetric positive definite matrix, we have:

$$0 \in \partial r(x_k) + \mathbf{R}_k^{-1}\left(x_k - (x_{k-1} - \mathbf{R}_k\nabla f(x_{k-1}) - \mathbf{Q}_k((\mathbf{I} - \mathbf{B}_{k-1})G_{k-2} + \mathbf{B}_{k-1}v_{k-2} - b_{2,k-1}) - b_{1,k})\right), \tag{44}$$

where $x_{k-1} - \mathbf{R}_k\nabla f(x_{k-1}) - \mathbf{Q}_k((\mathbf{I} - \mathbf{B}_{k-1})G_{k-2} + \mathbf{B}_{k-1}v_{k-2} - b_{2,k-1}) - b_{1,k}$ are exactly calculated, we denote it as $\bar{x}$.

Then, based on first-order condition, x_k can be uniquely solved by following the proximal operator:

$$x_k = \arg\min_x r(x) + (1/2)(x - \bar{x})^\top \mathbf{R}_k^{-1}(x - \bar{x}),$$

where taking x as the variable, $r(x) + (1/2)(x - \bar{x})^\top \mathbf{R}_k^{-1}(x - \bar{x})$ is the mathematical integration of right-hand side of equation 44.

In the experiments, we set $\mathbf{R}$, $\mathbf{Q}$, and $\mathbf{B}$ to be positive definite matrices by Sigmoid activation functions.

9. Composite Case Results

This section introduces several more theoretical findings on the composite case where the smooth and non-smooth parts in objective P are non-degenerated. Similar to the results in the smooth case of main pages, we derive several theorems and corollaries on per iteration and multi-iteration convergence of the L2O model. We follow the proofs of the vanilla proximal point method and proximal gradient algorithm (PGA) in [10] and [28] to derive our demonstrations for theorems and corollaries, where we use a gradient map to represent the $\arg\min$ operation for non-smooth optimization in PGA.

9.1. Preliminary

As in equation P, the objective of composite case is as below:

$$\min_x f(x) + r(x),$$

where $f(x) \in \mathcal{F}_L$ is a L-smooth and convex function and $r(x) \in \mathcal{F}$ is a proper, convex but probably non-smooth function. Notably, in the composite case, $f(x)$ and $r(x)$ are non-degenerated.

Based on the definition of L-smoothness on $f(x)$, $\frac{L'}{2}x^\top x - f(x)$ is convex [32]. Thus, for any points $y, x \in \mathbb{R}^n$, the convexity of f yields the following upper bound of $f(y)$:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2. \tag{45}$$

We take the definition of the k-th iteration update formulation given by the L2O model in [14] as below:

$$x_k = x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\left(\nabla f(x_{k-1}) + g_k\right) - \mathbf{N}_2(z_{k-1}), \tag{46}$$

where we set $\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) \succeq 0$ as a symmetric positive definition matrix and $g_k \in \partial r(x)$ as an implicit subgradient value at x_k [14]. We have the following reformulation of the above equation:

$$\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)^{-1}\left(x_k - \left(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1}) - \mathbf{N}_2(z_{k-1})\right)\right) + g_k = \mathbf{0}.$$

Since $g_k \in \partial r(x)$, we can represent the above equation with the following relationship:

$$\mathbf{0} \in \partial r(x_k) + \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)^{-1}\left(x_k - \left(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1}) - \mathbf{N}_2(z_{k-1})\right)\right).$$

Due to the first-order condition for convex optimization, as in [14], appling the proximal gradient method in [18], we can use the following proximal operator to solve a x_k:

$$\begin{aligned}
&\operatorname{prox}_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}\left(x_k - \left(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1}) - \mathbf{N}_2(z_{k-1})\right)\right) \\
=&\arg\min_{x_k} r(x_k) + \frac{1}{2}\left\|x_k - \left(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1}) - \mathbf{N}_2(z_{k-1})\right)\right\|^2_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}},
\end{aligned} \tag{47}$$

where the norm $\|\cdot\|_{\mathbf{P}_k^{-1}}$ is defined as $\|x\|_{\mathbf{P}_k^{-1}} := \sqrt{x^\top \mathbf{P}_k^{-1} x}$ [14]. From our definition in Sec. 2.1, the non-smooth function r is solvable. The $\arg\min$ operation in the above operator will explicitly generate an optimal solution.

The unknown implicit process in $\arg\min$ makes it hard to explicitly analyze the update from x_{k-1} to x_k. However, there are precisely two parts in the above operator, i.e., the $\arg\min$ operation on non-smooth function r and gradient descent operation $\left(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1}) - \mathbf{N}_2(z_{k-1})\right)$ on the smooth function f. We can regard the non-smooth function r update as an implicit subgradient descent and combine the two parts into one proximal gradient descent with both smooth and non-smooth gradients [28]. As in [28], the proximal gradient is named the gradient map.

Different from [28], we define the gradient map for L2O model in this work, denote as $G_{\mathbf{N}_1(z)}(x_{k-1})$. To represent the update from x_{k-1} to the x_k given by the L2O model defined in equation 46, we define the it as following operations:

$$\begin{aligned}
&G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) \\
:=&\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)^{-1}\left(x_{k-1} - \operatorname{prox}_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}\left(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1}) - \mathbf{N}_2(z_{k-1})\right) - \mathbf{N}_2(z_{k-1})\right).
\end{aligned}$$

Then, $G_{\mathbf{N}_1(z)}(x_{k-1})$ yields the following update formulation from x_{k-1} to x_k, which is similar to the L2O model in the smooth case.

$$\begin{aligned} x_k &= \mathrm{prox}_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}\Big(x_{k-1} - \mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1}) - \mathbf{N}_2(z_{k-1})\Big), \\ &= x_{k-1} - \mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \mathbf{N}_2(z_{k-1}). \end{aligned} \tag{48}$$

Substitute the above x_k's representation with gradient map into the L-smoothness inequation in equation 45, we have the following upper bound of $f(x_k)$ from L-smoothness:

$$\begin{aligned} f(x_k) &\leq f(x_{k-1}) + \nabla f(x_{k-1})^\top (x_k - x_{k-1}) + \frac{L}{2}\|x_k - x_{k-1}\|^2, \\ &\leq f(x_{k-1}) - \nabla f(x_{k-1})^\top (\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1})), \\ &\quad + \frac{L}{2}\left\|\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1})\right\|^2. \end{aligned} \tag{49}$$

Moreover, we would like to construct the representation of $\nabla f(x_{k-1}) + g_k$ by the gradient map. From the gradient map definition in equation 48 and the L2O model definition in equation 46, we directly have the following equality of $G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})$:

$$G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) = \nabla f(x_{k-1}) + g_k, \tag{50}$$

where $g_k \in \partial r(x_k)$ is the virtual subgradient of the non-smooth part r of objective. Thus, we have $G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \nabla f(x_{k-1}) \in \partial r(x_k)$, along with the definition of the convexity of r, for any $x, t \in \mathbb{R}^n$, we have the following inequality between $r(t)$ and $r(x)$:

$$r(t) \geq r(x_k) + \left(G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \nabla f(x_{k-1})\right)^\top (t - x_k).$$

After rearrangement, we have the following upper bound of $r(x_k)$ with any arbitrary $t \in \mathbb{R}^n$:

$$r(x_k) \leq r(t) - \left(G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \nabla f(x_{k-1})\right)^\top (t - x_k). \tag{51}$$

Finally, we present the following lemma to construct a general relationship between the objectives of any arbitrary two points:

Lemma 2. *$\forall x_k, t \in \mathbf{R}^n$:*

$$\begin{aligned} &F(x_k) \\ \leq &F(t) + G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (x_{k-1} - t) \\ &+ \frac{L}{2}\left(\left\|\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) - \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\|^2\right. \\ &\left. - \left\|\frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\|^2\right). \end{aligned}$$

The proof is as follows.

Proof. First, based on the definition that $\forall x \in \mathbb{R}^n, F(x) = f(x) + r(x)$, adding $r(x_k)$ into inequality 49 yields a full representation of objective F:

$$\begin{aligned} &F(x_k) \\ \leq &f(x_{k-1}) - \nabla f(x_{k-1})^\top \left(\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1})\right) \\ &+ \frac{L}{2}\left\|\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1})\right\|^2 + r(x_k). \end{aligned}$$

Since f is convex and differentiable, $\forall x_{k-1}, t \in \mathbb{R}^n$, we have $f(x_{k-1}) \leq f(t) - \nabla f(x_{k-1})^\top (t - x_{k-1})$, adding it into the above inequation yields:

$$
\begin{aligned}
&F(x_k) \\
\leq & f(t) - \nabla f(x_{k-1})^\top (t - x_{k-1}) - \nabla f(x_{k-1})^\top \big(\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) \big) \\
& + \frac{L}{2} \left\| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) \right\|^2 + r(x_k).
\end{aligned}
$$

Moreover, adding the upper bound of $r(x_k)$ in inequation 51 yields:

$$
\begin{aligned}
&F(x_k) \\
\leq & f(t) - \nabla f(x_{k-1})^\top (t - x_{k-1}) - \nabla f(x_{k-1})^\top \big(\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) \big) \\
& + \frac{L}{2} \left\| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) \right\|^2 \\
& + r(t) - \left(G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \nabla f(x_{k-1}) \right)^\top \Big(t - \big(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \mathbf{N}_2(z_{k-1}) \big) \Big).
\end{aligned}
$$

Then, we make the following rearrangement on the right-hand side of the above inequation:

$$
\begin{aligned}
&F(x_k) \\
\leq & f(t) - \nabla f(x_{k-1})^\top (t - x_{k-1}) - \nabla f(x_{k-1})^\top \big(\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) \big) \\
& + \frac{L}{2} \left\| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) \right\|^2 \\
& + r(t) - \left(G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \nabla f(x_{k-1}) \right)^\top \Big(t - \big(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \mathbf{N}_2(z_{k-1}) \big) \Big), \\
= & f(t) + r(t) + \frac{L}{2} \left\| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) \right\|^2 \\
& - G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (t - (x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \mathbf{N}_2(z_{k-1}))), \\
= & F(t) + G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (x_{k-1} - t) \\
& + \frac{L}{2} \left\| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) \right\|^2 \\
& - G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top \big(\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) \big),
\end{aligned}
$$

where we put $f(t)$ and $r(t)$ together and combine a similar term to achieve the simplification in the second step. In the third step, we combine $f(t) + r(t)$ as $F(t)$ based on the objective definition.

Finally, making up a perfect square between the last two terms finishes the proof:

$$
\begin{aligned}
&F(x_k)\\
=&F(t)+G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top(x_{k-1}-t)\\
&+\frac{L}{2}\left\|\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right)G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})+\mathbf{N}_2(z_{k-1})\right\|^2\\
&-G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top\left(\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right)G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})+\mathbf{N}_2(z_{k-1})\right),\\
=&F(t)+G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top(x_{k-1}-t)\\
&+\frac{L}{2}\Big(\|\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right)G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})+\mathbf{N}_2(z_{k-1})\|^2\\
&\quad-\frac{2}{L}G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top\left(\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right)G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})+\mathbf{N}_2(z_{k-1})\right)\Big)\\
=&F(t)+G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top(x_{k-1}-t)\\
&+\frac{L}{2}\left(\left\|\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right)G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})+\mathbf{N}_2(z_{k-1})-\frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\|^2\right.\\
&\quad\left.-\left\|\frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\|^2\right).
\end{aligned}
$$

□

We are ready to derive convergence analysis based on Lemma 2. We will iteratively apply Lemma 2 to construct the difference in objective between one and last iteration and between one iteration and the optimum.

9.2. InD Convergence Upper Bound

Similar to Lemma 1 for the smooth case, for the composite case, we propose the following lemma for per iteration convergence gain to ensure the L2O model is robust in the InD scenario.

Lemma 3. *For $\forall z_{k-1}\in\mathcal{Z}_P,\forall x_{k-1}\in\mathcal{S}_P$, if $\mathbf{N}_1(z_{k-1})$ and $\mathbf{N}_2(z_{k-1})$ are bounded by following compact sets:*

$$
\begin{aligned}
&\mathbf{N}_1(z_{k-1})\in\left[\mathbf{0},\frac{2}{L}\mathbf{1}\right],\\
&\left\|\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right)\left(\nabla f(x_{k-1})+g_k\right)+\mathbf{N}_2(z_{k-1})-\frac{\nabla f(x_{k-1})+g_k}{L}\right\|\le\left\|\frac{\nabla f(x_{k-1})+g_k}{L}\right\|,\forall\mathbf{N}_1(z_{k-1})\in\left[\mathbf{0},\frac{2}{L}\mathbf{1}\right],
\end{aligned}
$$

where $g_k\in\partial r(x_k)$, for any x_k generated by L2O model in equation 47, we have the following homogeneous derease on objective:

$$F(x_k)-F(x_{k-1})\le 0.$$

Proof. Based on Lemma 2, set $t:=x_{k-1}$, we have the following inequation between two objectives:

$$
\begin{aligned}
&F(x_k)-F(x_{k-1})\\
\le&\frac{L}{2}\left(\left\|\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right)G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})+\mathbf{N}_2(z_{k-1})-\frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\|^2\right.\\
&\quad\left.-\left\|\frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\|^2\right).
\end{aligned}
\tag{52}
$$

To ensure $F(x_{k-1})\le F(x_{k-1})$, we should keep right-hand side non-positive. Thus, we have the following inequality:

$$
\left\|\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right)G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})+\mathbf{N}_2(z_{k-1})-\frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\|\le\frac{\left\|G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})\right\|}{L}.
\tag{53}
$$

Similarly, we first freeze $\mathbf{N}_2(z_{k-1})$ and discuss $\mathbf{N}_1(z_{k-1})$-only terms, which yields:

$$\left\|\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\| \leq \frac{\left\|G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})\right\|}{L},$$

$$\left\|\left(L\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) - \mathbf{I}\right) \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\| \leq \frac{\left\|G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})\right\|}{L}.$$

Solve the above inequation, we have the following upper bound of $\mathbf{N}_1(z_{k-1})$:

$$\mathbf{N}_1(z_{k-1}) \in \left[\mathbf{0}, \frac{2}{L}\mathbf{1}\right].$$

Furthermore, each choice of $\mathbf{N}_1(z_{k-1})$ yields a range of $\mathbf{N}_2(z_{k-1})$. For example, $\mathbf{N}_1(z_{k-1}) := 0$ yields the following inequality for $\mathbf{N}_1(z_{k-1})$:

$$\left\|\mathbf{N}_2(z_{k-1}) - \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\| \leq \frac{\left\|G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})\right\|}{L}.$$

Solve the above inequation, $\mathbf{N}_2(z_{k-1})$ is bounded as follows:

$$\mathbf{N}_2(z_{k-1}) \in \left[\mathbf{0}, \frac{2}{L}|G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})|\right].$$

For example, $\mathbf{N}_1(z_{k-1}) := \frac{2}{L}\mathbf{1}$ yields:

$$\mathbf{N}_2(z_{k-1}) \in \left[-\frac{2}{L}|G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})|, \mathbf{0}\right].$$

Replacing the $G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})$ in inequation 53 with $G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) = \nabla f(x_{k-1}) + g_k$ in equation 50 yields:

$$\left\|\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\left(\nabla f(x_{k-1}) + g_k\right) + \mathbf{N}_2(z_{k-1}) - \frac{\nabla f(x_{k-1}) + g_k}{L}\right\| \leq \left\|\frac{\nabla f(x_{k-1}) + g_k}{L}\right\|.$$

□

Similar to Corollary 1 for the smooth case, for the composite case, we propose the following corollary to achieve the best robust L2O model with the largest per iteration convergence gain.

Corollary 4. *For any $z_{k-1} \in \mathcal{Z}_P$, we let:*

$$\mathbf{N}_1(z_{k-1}) := \frac{1}{L}\mathbf{1}, \mathbf{N}_2(z_{k-1}) := \mathbf{0},$$

the Math-L2O model in equation 6 is exactly gradient descent update with convergence rate:

$$F(x_K) - F(x^*) \leq \frac{L}{2K}\|x_0 - x^*\|^2.$$

Proof. In the last term of inequality 52, the best convergence gain yields:

$$\left\|\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) - \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\| := 0. \tag{54}$$

$\mathbf{N}_1(z_{k-1}) = \frac{1}{L}\mathbf{1}, \mathbf{N}_2(z_{k-1}) = \mathbf{0}$ is a feasible solution.

Given $\mathbf{N}_1(z_{k-1}) = \frac{1}{L}\mathbf{1}, \mathbf{N}_2(z_{k-1}) = \mathbf{0}$, the update formula in equation 48 is:

$$x_k = x_{k-1} - \frac{1}{L} G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}). \tag{55}$$

Based on Lemma 2, set $t := x^*$, we have the following inequality between the objective at k-th iteration and the optimum:

$$\begin{aligned}
&F(x_k) - F(x^*) \\
\leq & G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (x_{k-1} - x^*) \\
&+ \frac{L}{2}\left(\left\| \mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) - \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right. \\
&\qquad \left. - \left\| \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right), \\
= & G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (x_{k-1} - x^*) - \frac{L}{2} \left\| \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2, \\
= & \frac{L}{2}\left(\frac{2}{L} G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (x_{k-1} - x^*) - \left\| \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right), \\
= & \frac{L}{2}\left(\|x_{k-1} - x^*\|^2 - \left\| x_{k-1} - x^* - \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right), \\
= & \frac{L}{2}(\|x_{k-1} - x^*\|^2 - \|x_k - x^*\|^2),
\end{aligned}$$

where in the second step, we apply the equality in equation 54 and remove the degenerated terms. In the 4th step, we make up a perfect square. In the 5th step, we apply the L2O model's update formula in equation 55.

Sum over K iterations yields:

$$F(x_K) - F(x^*) \leq \frac{L}{2K}(\|x_0 - x^*\|^2 - \|x_K - x^*\|^2) \leq \frac{L}{2K}\|x_0 - x^*\|^2. \tag{56}$$

□

OOD Definitions

We first derive some preliminary formulations before the convergence analysis for OOD scenarios.

We make the following assumptions identical to those in the smooth case. Suppose $z, \tilde{z}, z' \in \mathcal{Z}$ are input feature vectors of the L2O model. There exists a vector $\alpha \in [0, 1]$ that $z' := \alpha z + (1 - \alpha)\tilde{z}, z' \in \mathcal{Z}$. Denote virtual Jacobian matrix of $\mathbf{N}_1(z')$ and $\mathbf{N}_2(z')$ at point z' as $\mathbf{J}_1$ and $\mathbf{J}_2$.

Since $\mathbf{N}_1(z)$ and $\mathbf{N}_2(z)$ are smooth, due to the Mean Value Theorem, we have the following equalities:

$$\mathbf{N}_1(z) = \mathbf{N}_1(\tilde{z}) + \mathbf{J}_1(z - \tilde{z}), \quad \mathbf{N}_2(z) = \mathbf{N}_2(\tilde{z}) + \mathbf{J}_2(z - \tilde{z}).$$

As demonstrated in the preliminary of the smooth case, we have $\|\mathbf{J}_1\| \leq \sqrt{n}C_1$, and $\|\mathbf{J}_2\| \leq \sqrt{n}C_2$.

Given a virtual variable $s \in \mathbb{R}^n$ to represent the OOD shifting on variable x, define the virtual feature (difference of L2O model's input feature between OOD and InD scenarios) as $s' = [s^\top, (\nabla f'(x + s) - \nabla f(x))^\top, (g' - g)^\top]^\top$, where $g' \in \partial r'(x + s), g \in \partial r(x)$ are subgradient instances of OOD and InD scenarios respectively. We have the following equations to formulate the L2O model's behaviors in OOD and InD scenarios:

$$\begin{aligned}
\mathbf{N}_1(z + s') &= \mathbf{N}_1(z) + \mathbf{J}_1(z + s' - z) = \mathbf{N}_1(z) + \mathbf{J}_1 s' \\
\mathbf{N}_2(z + s') &= \mathbf{N}_2(z) + \mathbf{J}_2(z + s' - z) = \mathbf{N}_2(z) + \mathbf{J}_2 s'.
\end{aligned} \tag{57}$$

Based on Lemma 2, $\forall x_k \in \mathcal{S}_p, s_k \in \mathbb{R}^n$, OOD yields the following inequality between any two values of objective:

$$\begin{aligned}
&F'(x_k + s_k) \\
\leq &F'(t) + G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1})^\top (x_{k-1} + s_{k-1} - t) \\
&+ \frac{L}{2}\left\| \mathrm{diag}(\mathbf{N}_1(z_{k-1} + s'_{k-1}))G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1}) + \mathbf{N}_2(z_{k-1} + s'_{k-1}) \right. \\
&\left. - \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1})}{L} \right\|^2 - \frac{L}{2}\left\| \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1})}{L} \right\|^2.
\end{aligned} \tag{58}$$

For gradient mapping, OOD yields:

$$\begin{aligned}
&-\mathrm{diag}(\mathbf{N}_1(z_{k-1} + s'_{k-1}))^{-1}\left(x_k + s_k - \left(x_{k-1} + s_{k-1} - \mathrm{diag}(\mathbf{N}_1(z_{k-1} + s'_{k-1}))\nabla f'(x_{k-1} + s_{k-1}) - \mathbf{N}_2(z_{k-1} + s'_{k-1})\right)\right) \\
=&-\mathrm{diag}(\mathbf{N}_1(z_{k-1} + s'_{k-1}))^{-1} \\
&\Big(x_{k-1} + s_{k-1} - \mathrm{diag}(\mathbf{N}_1(z_{k-1} + s'_{k-1}))G'_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1}) - \mathbf{N}_2(z_{k-1} + s'_{k-1}) \\
&\quad - \left(x_{k-1} + s_{k-1} - \mathrm{diag}(\mathbf{N}_1(z_{k-1} + s'_{k-1}))\nabla f'(x_{k-1} + s_{k-1}) - \mathbf{N}_2(z_{k-1} + s'_{k-1})\right)\Big), \\
=&G'_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1}) - \nabla f'(x_{k-1} + s_{k-1}),
\end{aligned}$$

where we use $G'_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}$ to represent the gradient map in the OOD scenario.

Moreover, similar to equation 50, we have the following formulation of using gradient map to represent gradient and subgradient in the OOD scenario:

$$G'_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1}) = \nabla f'(x_{k-1} + s_{k-1}) + g'_k, \tag{59}$$

where $g'_k \in \partial r'(x_k + s_k)$.

Similar to Assumption 1 for the smooth case, we derive the following assumption to ensure the most robust L2O model for the composite case.

Assumption 2. *After training,* $\forall x_{k-1} \in \mathcal{S}_P, \forall z_{k-1} \in \mathcal{Z}_P$, $\mathbf{N}_1(z_{k-1}) := \frac{1}{L}\mathbf{1}$ *and* $\mathbf{N}_2(z_{k-1}) := \mathbf{0}$.

9.3. OOD Per-Iteration Convergence Gain

Based on the Lemma 3 and Corollary 4, Assumption 2 leads to an L2O model with best robustness on all InD instances. In the following theorem, we quantify the diminution in convergence rate instigated by the virtual feature s' defined in Sec. 3.

Theorem 4. *Under Assumption 2, there exists virtual Jacobian matrices* $\mathbf{J}_{1,k-1}, \mathbf{J}_{2,k-1}, k = 1, 2, \ldots, K$ *that OOD's convergence improvement of one iteration is upper bounded by following inequality:*

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &-\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + \frac{L}{2}\|\mathrm{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\|^2,
\end{aligned}$$

where $g'_k \in \partial r'(x_k + s_k)$.

Proof. From equation 57, we have the following reformulations of $\mathbf{N}_1(z_{k-1} + s'_{k-1})$ and $\mathbf{N}_2(z_{k-1} + s'_{k-1})$:

$$\begin{aligned}
\mathbf{N}_1(z_{k-1} + s'_{k-1}) &= \mathbf{N}_1(z_{k-1}) + \mathbf{J}_1 s'_{k-1}, \\
\mathbf{N}_2(z_{k-1} + s'_{k-1}) &= \mathbf{N}_2(z_{k-1}) + \mathbf{J}_2 s'_{k-1}.
\end{aligned}$$

Substituting the definitions of $\mathbf{N}_1(z_{k-1})$ and $\mathbf{N}_2(z_{k-1})$ in Assumption 2 yields:

$$\begin{aligned}
\mathbf{N}_1(z_{k-1} + s'_{k-1}) &= \frac{1}{L}\mathbf{1} + \mathbf{J}_1 s'_{k-1} \\
\mathbf{N}_2(z_{k-1} + s'_{k-1}) &= \mathbf{J}_2 s'_{k-1}.
\end{aligned} \tag{60}$$

We then apply construct inequality between objectives of two adjacent iterations. Substituting $t := x_{k-1} + s_{k-1}$ into inequality 58 yields:

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &\frac{L}{2}\left\| \operatorname{diag}(\mathbf{N}_1(z_{k-1} + s'_{k-1}))G'_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1}) + \mathbf{N}_2(z_{k-1} + s'_{k-1}) \right. \\
&\left. - \frac{G'_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1})}{L} \right\|^2 - \frac{L}{2}\left\| \frac{G'_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1})}{L} \right\|^2.
\end{aligned}$$

Substituting equation 60 into above inequality yields:

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &\frac{L}{2}\| \operatorname{diag}(\mathbf{J}_1 s'_{k-1})G'_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1}) + \mathbf{J}_2 s'_{k-1}\|^2 - \frac{L}{2}\left\| \frac{G'_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1})}{L} \right\|^2.
\end{aligned}$$

Based on equation 60, we recover gradient and subgradient from gradient map:

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &-\frac{L}{2}\left\| \frac{\nabla f'(x_{k-1} + s_{k-1}) + g'_k}{L} \right\|^2 + \frac{L}{2}\| \operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\|^2, \\
= &-\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + \frac{L}{2}\| \operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\|^2.
\end{aligned}$$

□

Moreover, we derive the upper bound of per iteration convergence gain in the following Corollary 5.

Corollary 5. *Under Assumption 2, the convergence improvement for one iteration of the OOD scenario can be upper bounded w.r.t.* $\|s'_{k-1}\|$ *by:*

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &-\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + (Ln^2C_1^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + Ln^2C_2^2)\|s'_{k-1}\|^2,
\end{aligned}$$

where $g'_k \in \partial r'(x_k + s_k)$.

Proof. Based on Triangle and Cauchy-Schwarz inequalities, we have:

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &-\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + \frac{L}{2}\| \operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\|^2, \\
\leq &-\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + L\| \operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k)\|^2 + L\|\mathbf{J}_2 s'_{k-1}\|^2, \\
\leq &-\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + L\|\mathbf{J}_1 s'_{k-1}\|^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + L\|\mathbf{J}_2 s'_{k-1}\|^2, \\
\leq &-\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + Ln^2C_1^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2\|s'_{k-1}\|^2 + Ln^2C_2^2\|s'_{k-1}\|^2, \\
= &-\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + (Ln^2C_1^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + Ln^2C_2^2)\|s'_{k-1}\|^2,
\end{aligned}$$

where $g'_k \in \partial r'(x_k + s_k)$. □

9.4. OOD Multi-Iteration Convergence Rate

Theorem 5. *Under Assumption 2, OOD's convergence rate of K iterations is upper bounded by:*

$$\begin{aligned}
&\min_{k=1,\dots,K} F'(x_k+s_k)-F'(x^*+s^*)\\
\leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2\\
&+\frac{L}{K}\sum_{k=1}^{K}\left(x_k+s_k-x_{k-1}-s_{k-1}+\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k}{L}\right)^\top(x_k+s_k-x^*-s^*).
\end{aligned}$$

Proof. We construct the relationship between k-th iteration's and optimal objectives by substituting $t:=x^*+s^*$ into inequality 58 yields:

$$\begin{aligned}
&F'(x_k+s_k)-F'(x^*+s^*)\\
\leq&G'_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1}+s_{k-1})^\top(x_{k-1}+s_{k-1}-x^*-s^*)\\
&+\frac{L}{2}\left\|\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))G'_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1}+s_{k-1})+\mathbf{N}_2(z_{k-1}+s'_{k-1})\right.\\
&\left.-\frac{G'_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1}+s_{k-1})}{L}\right\|^2-\frac{L}{2}\left\|\frac{G'_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1}+s_{k-1})}{L}\right\|^2.
\end{aligned}$$

We eliminate InD terms by substituting equation 60 into above inequality yields:

$$\begin{aligned}
&F'(x_k+s_k)-F'(x_{k-1}+s_{k-1})\\
\leq&G'_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1}+s_{k-1})^\top(x_{k-1}+s_{k-1}-x^*-s^*)\\
&+\frac{L}{2}\left\|\mathrm{diag}(\mathbf{J}_1s'_{k-1})G'_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1}+s_{k-1})+\mathbf{J}_2s'_{k-1}\right\|^2-\frac{L}{2}\left\|\frac{G'_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1}+s_{k-1})}{L}\right\|^2.
\end{aligned}$$

Then, we recover the gradient and subgradient from the gradient map by substituting equation 59 into the above inequality yields:

$$\begin{aligned}
&F'(x_k+s_k)-F'(x_{k-1}+s_{k-1})\\
\leq&(\nabla f'(x_{k-1}+s_{k-1})+g'_k)^\top(x_{k-1}+s_{k-1}-x^*-s^*)-\frac{L}{2}\left\|\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k}{L}\right\|^2\\
&+\frac{L}{2}\left\|\mathrm{diag}(\mathbf{J}_1s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)+\mathbf{J}_2s'_{k-1}\right\|^2,\\
=&\frac{L}{2}\left(2\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k}{L}^\top(x_{k-1}+s_{k-1}-x^*-s^*)-\left\|\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k}{L}\right\|^2\right)\\
&+\frac{L}{2}\|\mathrm{diag}(\mathbf{J}_1s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)+\mathbf{J}_2s'_{k-1}\|^2.
\end{aligned}\tag{61}$$

By the definition of $G'_{\mathrm{diag}(\mathbf{N}_1(z_k))^{-1}}(x_{k-1}+s_{k-1})$ in equation 59, we can represent x_k+s_k by the following equation:

$$\begin{aligned}
x_k+s_k&=\mathrm{prox}_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1}+s_{k-1}-\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))\nabla f(x_{k-1}+s_{k-1})-\mathbf{N}_2(z_{k-1}+s'_{k-1})),\\
&=x_{k-1}+s_{k-1}-\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))G'_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1}+s_{k-1})-\mathbf{N}_2(z_{k-1}+s'_{k-1}),\\
&=x_{k-1}+s_{k-1}-\mathrm{diag}(\frac{1}{L}\mathbf{1}+\mathbf{J}_1s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)-\mathbf{J}_2s'_{k-1},\\
&=x_{k-1}+s_{k-1}-\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k}{L}-\mathrm{diag}(\mathbf{J}_1s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)-\mathbf{J}_2s'_{k-1},
\end{aligned}\tag{62}$$

where $g'_k \in \partial r'(x_k + s_k)$ is a subgradient vector.

Similarly, we aim to make up a perfect square in equation 61 with the above formulation of the update given by the L2O model in the OOD scenario. The demonstration is as follows.

First, in order to apply equation 62, we would like to make up several terms of equation 62:

$$
\begin{aligned}
& F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq & \frac{L}{2}(2\frac{\nabla f'(x_{k-1} + s_{k-1}) + g'_k}{L}^\top (x_{k-1} + s_{k-1} - x^* - s^*) - \|\frac{\nabla f'(x_{k-1} + s_{k-1}) + g'_k}{L}\|^2) \\
& + \frac{L}{2}\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\|^2, \\
= & \frac{L}{2}\Big(2\big(\operatorname{diag}(\frac{1}{L}\mathbf{1} + \mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) - \operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k)\big)^\top (x_{k-1} + s_{k-1} - x^* - s^*) \\
& - \|\operatorname{diag}(\frac{1}{L}\mathbf{1} + \mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) - \operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k)\|^2\Big) \\
& + \frac{L}{2}\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\|^2, \\
= & \frac{L}{2}\Big(2\big(\operatorname{diag}(\frac{1}{L}\mathbf{1} + \mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1} \\
& \quad - \operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) - \mathbf{J}_2 s'_{k-1}\big)^\top (x_{k-1} + s_{k-1} - x^* - s^*) \\
& \quad - \|\operatorname{diag}(\frac{1}{L}\mathbf{1} + \mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1} - \operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) - \mathbf{J}_2 s'_{k-1}\|^2\Big) \\
& + \frac{L}{2}\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\|^2,
\end{aligned}
$$

where in the first step, we makeup the $\operatorname{diag}(\frac{1}{L}\mathbf{1} + \mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k)$ of equation 62. In the second step, we makeup the $\mathbf{J}_2 s'_{k-1}$ of equation 62.

Then, we expand the quadratic term in the second line of above inequation's right-hand side and merge similar terms:

$$
\begin{aligned}
& F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq & \frac{L}{2}\Big(2\big(\operatorname{diag}(\frac{1}{L}\mathbf{1} + \mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\big)^\top (x_{k-1} + s_{k-1} - x^* - s^*) \\
& \quad - \|\operatorname{diag}(\frac{1}{L}\mathbf{1} + \mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\|^2\Big) \\
& - 2\frac{L}{2}\big(\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\big)^\top (x_{k-1} + s_{k-1} - x^* - s^*) \\
& + 2\frac{L}{2}\big(\operatorname{diag}(\frac{1}{L}\mathbf{1} + \mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\big)^\top \big(\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\big) \\
& - \frac{L}{2}\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\|^2 + \frac{L}{2}\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\|^2.
\end{aligned}
$$

Finially, we are able to make up a perfect square on the first two lines of the right-hand side:

$$
\begin{aligned}
& F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq & \frac{L}{2}(\|x_{k-1} + s_{k-1} - x^* - s^*\|^2 - \|x_{k-1} + s_{k-1} - x^* - s^* - \operatorname{diag}(\frac{1}{L}\mathbf{1} + \mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) - \mathbf{J}_2 s'_{k-1}\|^2) \\
& - 2\frac{L}{2}\big(\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\big)^\top (x_{k-1} + s_{k-1} - x^* - s^*) \\
& + 2\frac{L}{2}\big(\operatorname{diag}(\frac{1}{L}\mathbf{1} + \mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\big)^\top \big(\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1}\big).
\end{aligned}
$$

Moreover, we can apply the update formula in equation 62 to simplify the above inequation as follows:

$$\begin{aligned}
&F'(x_k+s_k)-F'(x_{k-1}+s_{k-1})\\
\leq&\frac{L}{2}(\|x_{k-1}+s_{k-1}-x^*-s^*\|^2-\|x_k+s_k-x^*-s^*\|^2)\\
&-L\big(\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)+\mathbf{J}_2 s'_{k-1}\big)^\top\\
&\quad\Big(x_{k-1}+s_{k-1}-x^*-s^*-\operatorname{diag}(\frac{1}{L}\mathbf{1}+\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)-\mathbf{J}_2 s'_{k-1}\Big),
\end{aligned}\tag{63}$$

where we replace the update on $x_{k-1}+s_{k-1}$ with x_k+s_k in the first line.

Similarly, we propose to maintain the InD update formula on x_{k-1} as $x_k = x_{k-1} - \frac{\nabla f(x_{k-1}+s_{k-1})+g_k}{L}$, which further yields:

$$\begin{aligned}
x_k+s_k=&x_{k-1}+s_{k-1}-\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k}{L}-\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)-\mathbf{J}_2 s'_{k-1},\\
=&x_{k-1}-\frac{\nabla f(x_{k-1}+s_{k-1})+g_k}{L}\\
&+s_{k-1}-\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k-\nabla f(x_{k-1}+s_{k-1})-g_k}{L}-\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)-\mathbf{J}_2 s'_{k-1},
\end{aligned}$$

where, we construct s_k by following equation:

$$s_k = s_{k-1}-\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k-\nabla f(x_{k-1}+s_{k-1})-g_k}{L}-\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)-\mathbf{J}_2 s'_{k-1}.$$

Substituting above equation into right-hand side of inequality 63 yields:

$$\begin{aligned}
&F'(x_k+s_k)-F'(x_{k-1}+s_{k-1})\\
\leq&\frac{L}{2}(\|x_{k-1}+s_{k-1}-x^*-s^*\|^2-\|x_k+s_k-x^*-s^*\|^2)\\
&-L\big(\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)+\mathbf{J}_2 s'_{k-1}\big)^\top\\
&\quad\Big(x_{k-1}+s_{k-1}-x^*-s^*-\operatorname{diag}(\frac{1}{L}\mathbf{1}+\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)-\mathbf{J}_2 s'_{k-1}\Big),\\
=&\frac{L}{2}(\|x_{k-1}+s_{k-1}-x^*-s^*\|^2-\|x_k+s_k-x^*-s^*\|^2)\\
&+L\left(x_k+s_k-x_{k-1}-s_{k-1}+\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k}{L}\right)^\top\left(x_{k-1}-\frac{\nabla f(x_{k-1}+s_{k-1})+g_k}{L}-x^*\right.\\
&\left.+s_{k-1}-\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k-\nabla f(x_{k-1}+s_{k-1})-g_k}{L}-\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)-\mathbf{J}_2 s'_{k-1}-s^*\right),\\
=&\frac{L}{2}(\|x_{k-1}+s_{k-1}-x^*-s^*\|^2-\|x_k+s_k-x^*-s^*\|^2)\\
&+L(x_k+s_k-x_{k-1}-s_{k-1}+\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k}{L})^\top(x_k+s_k-x^*-s^*).
\end{aligned}\tag{64}$$

Over K-iterations, we have:

$$\begin{aligned}
&\min_{k=1,\dots,K}F'(x_k+s_k)-F'(x^*+s^*)\\
\leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2\\
&+\frac{L}{K}\sum_{k=1}^{K}(x_k+s_k-x_{k-1}-s_{k-1}+\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k}{L})^\top(x_k+s_k-x^*-s^*).
\end{aligned}$$

□

Moreover, we derive the upper bound of multi-iteration convergence rate in the following Corollary 6.

Corollary 6. *Under Assumption 2, L2O model d's (equation 47) convergence rate is upper bounded by w.r.t.* $\|s'_{k-1}\|$ *by:*

$$\begin{aligned}
&\min_{k=1,\ldots,K} F'(x_k+s_k)-F'(x^*+s^*)\\
\leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2-\frac{1}{K}\sum_{k=1}^{K}(\nabla f'(x_{k-1}+s_{k-1}))^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}(\sqrt{n}C_1\|\nabla f'(x_{k-1}+s_{k-1})\|+\sqrt{n}C_2)\|x_k+s_k-x^*-s^*\|\|s'_{k-1}\|.
\end{aligned}$$

Proof. First, we rewrite the convergence rate upper bound as follows:

$$\begin{aligned}
&\min_{k=1,\ldots,K} F'(x_k+s_k)-F'(x^*+s^*)\\
\leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2+\frac{L}{K}\sum_{k=1}^{K}(x_k+s_k-x_{k-1}-s_{k-1})^\top(x_k+s_k-x^*-s^*),\\
=&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2-\frac{1}{K}\sum_{k=1}^{K}(\nabla f'(x_{k-1}+s_{k-1}))^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}(-\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)-\mathbf{J}_2 s'_{k-1})^\top(x_k+s_k-x^*-s^*).
\end{aligned}$$

Next, we derive its upper bound w.r.t. $\|s'_{k-1}\|$. Cauchy-Schwarz inequality and Triangle inequality yield:

$$\begin{aligned}
&\min_{k=1,\ldots,K} F'(x_k+s_k)-F'(x^*+s^*)\\
\leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2-\frac{1}{K}\sum_{k=1}^{K}(\nabla f'(x_{k-1}+s_{k-1}))^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}(-\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)-\mathbf{J}_2 s'_{k-1})^\top(x_k+s_k-x^*-s^*),\\
\leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2-\frac{1}{K}\sum_{k=1}^{K}(\nabla f'(x_{k-1}+s_{k-1}))^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}(\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)\|+\|\mathbf{J}_2 s'_{k-1}\|)\|x_k+s_k-x^*-s^*\|,\\
\leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2-\frac{1}{K}\sum_{k=1}^{K}(\nabla f'(x_{k-1}+s_{k-1}))^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}(\sqrt{n}C_1\|s'_{k-1}\|\|\nabla f'(x_{k-1}+s_{k-1})\|+\sqrt{n}C_2\|s'_{k-1}\|)\|x_k+s_k-x^*-s^*\|,\\
=&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2-\frac{1}{K}\sum_{k=1}^{K}(\nabla f'(x_{k-1}+s_{k-1}))^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}\left(\sqrt{n}C_1\|\nabla f'(x_{k-1}+s_{k-1})\|+\sqrt{n}C_2\right)\|x_k+s_k-x^*-s^*\|\|s'_{k-1}\|.
\end{aligned}$$

□

10. Non-Smooth Case Results

For the non-smooth case, we set the smooth part in the objective of problem P to be zero $f(x) := 0$, and the objective becomes:

$$\min_x r(x), \tag{P}$$

where $x \in \mathcal{S}_P$ and $r \in \mathcal{F}_P$. Based on the definition, $r(x)$ is proper and convex, where the "proper" means $r(x)$ is trivially solvable for any x.

In the OOD scenario, the optimization problem becomes:

$$\min_{x'} r'(x'), \tag{O}$$

where $x' \in \mathcal{S}_O$ and $r \in \mathcal{F}_O$.

Based on the definition, $r'(x)$ is still proper and convex. We can directly get the solution from:

$$x'^* = \arg\min_{x'} r'(x').$$

Thus, constructing a L2O model is unnecessary for the smooth case. We eliminate the demonstrations for this case.

11. Longer Horizon Case Results

In the smooth and composite cases, we have demonstrated convergence analysis per iteration and multi-iteration convergence analysis for L2O. Modern algorithms utilize historical information to accelerate convergence, such as Nesterov momentum in FISTA algorithm [5] and long short-term memory in LSTM-based unrolling algorithms [14, 15]. This case establishes convergence analysis with historical modeling in L2O. We take the SOTA Math-L2O framework [14] to define the historical modeling part, a general and problem-independent approach that ensures our proposed theorems are also general.

We first establish that the input features of neural networks should be consistent with the definition of L2O model d. Suppose there exist a $d \in \mathcal{D}_C(\mathbb{R}^{m\times n}) \to \mathbb{R}^n$. Due to Lemma 1 in [14], for any $x_1, y_1, x_2, y_2, \ldots, x_m, y_m \in \mathbb{R}^n$, there exists matrices $\mathbf{J}_1, \mathbf{J}_2, \ldots, \mathbf{J}_m$ that:

$$d(x_1, x_2, \ldots, x_m) = d(x_1^*, x_2^*, \ldots, x_m^*) + \sum_{j=1}^{m} \mathbf{J}_j (x_j - x_j^*),$$

where $\mathbf{J}_j$ is j-th block of d's Jacobian matrix at a interior point between $[x_1^\top, x_2^\top, \ldots, x_m^\top]^\top$ and $[x_1^{*\top}, x_2^{*\top}, \ldots, x_m^{*\top}]^\top$.

The L2O is constructing such $\mathbf{J}$s by learning [14]. Denote a NN as $\mathbf{N}_j$ and its input feature vector as s. We propose the following lemma to formulate s to at least include all variables of d.

Lemma 4. *For any feature vector s such that $\mathbf{J}_j = \mathbf{N}_j(s), j = 1, 2, \ldots, m$, s should follows:*

$$\{x_1^\top, x_2^\top, \ldots, x_m^\top\} \subseteq s.$$

Proof. We prove the above lemma by contradiction. Suppose not, which means there exists a s' that $\exists x_i \notin s'$. Then for $x_j, j = 1, 2, \ldots, m$, we have $\mathbf{J}_j = \mathbf{N}_j(s')$ by the definition. First, suppose $j \neq i$. Since d is arbitrary, x_j is not guaranteed linear with x_i in d. Hence, x_i should be one of the input features of $\mathbf{N}_j$. Moreover, suppose $j = i$. d is not guaranteed to always be less than the first order on x_i. Hence, x_i should be one of the input features of $\mathbf{N}_j$.

The above scenarios cause contradictions with the assumption that $\exists x_i \notin s'$, leading to the lemma's conclusion. □

11.1. Preliminary

Similar to the composite case in Sec. 9. We make the following preliminary constructions. The objective is as follows:

$$\min_x f(x) + r(x),$$

where $f(x) \in \mathcal{F}_L$ is a L-smooth and convex function and $r(x) \in \mathcal{F}$ is a proper and convex function.

The definition of L-smoothness yields following upper bound of $f(y)$ for $\forall x, y \in \mathbb{R}^n$:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2. \tag{65}$$

For k-th iteration, we use $y_{k-1} \in \mathbb{R}^n$ to represent the historical information and use z_{k-1} to represent the input feature vector for the L2O model. The L2O model is defined as follows:

$$x_k = x_{k-1} - d(z_{k-1}),$$

where z_{k-1} is defined as $z_{k-1} := [x_{k-1}^\top, \nabla f(x_{k-1})^\top, g_k^\top, y_{k-1}^\top]$, $g_k \in \partial r(x_k)$ [14]. Without loss of generality, y_{k-1} represents the result of any historical modeling methods. For example, we can use neural network models to achieve momentum-like modeling [14].

Utilizing T to denote the number of iterations in historical modeling, following [14], we set $T := 1$. Inductively, we define y_{k-1} as:

$$y_{k-1} = \left(\mathbf{I} - \operatorname{diag}\left(\mathbf{N}_4([v_{k-1}^\top, v_{k-2}^\top]^\top)\right)\right)v_{k-1} + \mathbf{N}_4([v_{k-1}^\top, v_{k-2}^\top]^\top)v_{k-2} + \mathbf{N}_5([v_{k-1}^\top, v_{k-2}^\top]^\top),$$

where $\mathbf{N}_4 \in \mathcal{D}_{C_4}(\mathbb{R}^{2n})$ and $\mathbf{N}_5 \in \mathcal{D}_{C_5}(\mathbb{R}^{2n})$ are two neural network operators of the L2O model. $v \in \mathbb{R}^n$ denotes an input vector. Without loss of generality, such a definition covers any historical modeling methods with any feature selection. For example, v can be a variable x [14], a gradient $\nabla f(x)$ or a subgradient $g \in \partial r(x)$.

We are ready to demonstrate convergence analysis for longer-horizon cases. We focus on two kinds of historical feature selections: the horizon of variable x's sequence and [14] the horizon of gradient $\nabla f(x)$'s (and subgradient $\partial r(x)$'s) sequence(s).

Variable Method

The variable method is from [14], where variable features are utilized to model the historical information. First, the neural network models' input vector z_{k-1} is defined by:

$$z_{k-1} = [x_{k-1}^\top, \nabla f(x_{k-1})^\top, g_k^\top, y_{k-1}^\top]^\top, \tag{66}$$

where $g_k \in \partial r(x_k)$ is a subgradient vector. And y_{k-1} denotes the feature from historical modeling. Then, the update given by the L2O model d is defined as:

$$x_k = x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1}) - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)g_k - \operatorname{diag}(\mathbf{N}_3(z_{k-1}))(y_{k-1} - x_{k-1}) - \mathbf{N}_2(z_{k-1}).$$

In this case, we use variable x to construct the input vector v for historical model $\mathbf{N}_4$ and denote $u_{k-1} := [x_{k-1}^\top, x_{k-2}^\top]^\top$. Based on Lemma 1 in Section A.1. of [14], we define historical modeling result y_{k-1} as a linear-like combination of x_{k-1} and x_{k-2}:

$$y_{k-1} = (\mathbf{I} - \operatorname{diag}(\mathbf{N}_4(u_{k-1})))x_{k-1} + \operatorname{diag}(\mathbf{N}_4(u_{k-1}))x_{k-2},$$

where we eliminate the reaching zero bias term in [14].

Based on [14], we define the L2O model as:

$$\begin{aligned} x_k = x_{k-1} &- \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\left(\nabla f(x_{k-1}) + g_k\right) - \mathbf{N}_2(z_{k-1}) \\ &- \operatorname{diag}(\mathbf{N}_3(z_{k-1}))\operatorname{diag}(\mathbf{N}_4(u_{k-1}))(-x_{k-1} + x_{k-2}), \end{aligned}$$

where we set $\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) \succ 0$ and g_k is an implicit subgradient value at x_k. Moreover, we omit all bias terms since they are demonstrated to vanish along iteration [14].

Assume $\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)$ is a symmetric positive definite, similar to those in the composite case, we have the following necessary and sufficient conditions from the definition of the convex function r:

$$\begin{aligned} &\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)^{-1}\Big(x_k - \big(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1}) - \mathbf{N}_2(z_{k-1}) \\ &\quad - \operatorname{diag}(\mathbf{N}_3(z_{k-1}))\operatorname{diag}(\mathbf{N}_4(u_{k-1}))(-x_{k-1} + x_{k-2})\big)\Big) + g_k = 0, \\ &0 \in \partial r(x_k) + \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)^{-1}\Big(x_k - \big(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1}) - \mathbf{N}_2(z_{k-1}) \\ &\quad - \operatorname{diag}(\mathbf{N}_3(z_{k-1}))\operatorname{diag}(\mathbf{N}_4(u_{k-1}))(-x_{k-1} + x_{k-2})\big)\Big), \end{aligned} \tag{67}$$

which yields the following proximal operator:

$$\begin{aligned}
&\operatorname{prox}_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}\left(x_{k-1}-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1})-\mathbf{N}_2(z_{k-1})-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))\operatorname{diag}(\mathbf{N}_4(u_{k-1}))(-x_{k-1}+x_{k-2})\right)\\
=&\arg\min_{x_k} r(x_k)+\frac{1}{2}\Big\|x_k-\big(x_{k-1}-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1})-\mathbf{N}_2(z_{k-1})\\
&\qquad-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))\operatorname{diag}(\mathbf{N}_4(u_{k-1}))(-x_{k-1}+x_{k-2})\big)\Big\|^2_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}.
\end{aligned}\tag{68}$$

Gradient (and Subgradient) Method

This method utilizes gradient-related features to achieve historical modeling. First, the neural network models' input vector z_{k-1} is defined by:

$$z_{k-1}=[\nabla f(x_{k-1})^\top, g_k^\top, y_{k-1}^\top]^\top,\tag{69}$$

where $g_k\in\partial r(x_k)$. And y_{k-1} denotes the feature from historical modeling. Compared with variable method in [14], we remove variable x_{k-1} from equation 66. Thus, compared with the variable method, y_{k-1} represents a different feature based on the historical modeling method without the variable.

Similarly, in this case, we use gradient and subgradient to construct the input vector v for historical model $\mathbf{N}_4$ and denote $w_{k-1}:=[(\nabla f(x_{k-1})+g_{k-1})^\top,(\nabla f(x_{k-2})+g_{k-2})^\top]^\top$. For any $x\in\mathbb{R}^n$, we denote the lower bound and the upper bound of $\partial r(x)$ as $\partial r(x)_{\text{lb}}$ and $\partial r(x)_{\text{ub}}$ respectively. Based on Lemma 1 in Section A.1. of [14], using explicit subgradient, we define y_{k-1} by:

$$\begin{aligned}
y_{k-1}=&(\mathbf{I}-\operatorname{diag}(\mathbf{N}_4(w_{k-1}))(\nabla f(x_{k-1})+g_{k-1}))+\operatorname{diag}(\mathbf{N}_4(w_{k-1}))(\nabla f(x_{k-2})+g_{k-2}),\\
g_{k-1}=&(\mathbf{I}-\operatorname{diag}(\mathbf{N}_5(r_{k-1})))\partial r(x_{k-1})_{\text{lb}}+\operatorname{diag}(\mathbf{N}_5(r_{k-1}))\partial r(x_{k-1})_{\text{ub}}\\
g_{k-2}=&(\mathbf{I}-\operatorname{diag}(\mathbf{N}_5(r_{k-2})))\partial r(x_{k-2})_{\text{lb}}+\operatorname{diag}(\mathbf{N}_5(r_{k-2}))\partial r(x_{k-2})_{\text{ub}},\\
x_k=&x_{k-1}-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1})-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)g_k-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))(y_{k-1}-(\nabla f(x_{k-1})+g_{k-1}))\\
&-\mathbf{N}_2(z_{k-1}),
\end{aligned}$$

where $g_k\in\partial r(x_k)$, $g_{k-1}\in\partial r(x_{k-1})$, and $g_{k-2}\in\partial r(x_{k-2})$, $r_{k-1}:=[\partial r(x_{k-1})_{\text{lb}}^\top,\partial r(x_{k-1})_{\text{ub}}^\top]^\top$. In the second and third equations, we apply two extra neural network models, denoted as $\mathbf{N}_5$ and $\mathbf{N}_6$ to learn subgradient vectors.

Based on our proposed L2O model in equation 43, the L2O model of gradient method is given by:

$$\begin{aligned}
x_k=&x_{k-1}-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\left(\nabla f(x_{k-1})+g_k\right)-\mathbf{N}_2(z_{k-1})\\
&-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))\operatorname{diag}(\mathbf{N}_4(w_{k-1}))\Big(-\big(\nabla f(x_{k-1})+\big((\mathbf{I}-\operatorname{diag}(\mathbf{N}_5(r_{k-1})))\partial r(x_{k-1})_{\text{lb}}\\
&+\operatorname{diag}(\mathbf{N}_5(r_{k-1}))\partial r(x_{k-1})_{\text{ub}}\big)\big)+\nabla f(x_{k-2})+\big((\mathbf{I}-\operatorname{diag}(\mathbf{N}_5(r_{k-2})))\partial r(x_{k-2})_{\text{lb}}+\operatorname{diag}(\mathbf{N}_5(r_{k-2}))\partial r(x_{k-2})_{\text{ub}}\big)\Big),
\end{aligned}$$

where we set $\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\succ 0$ and g_k is an implicit subgradient value at x_k. Notably, in this definition, without loss of generality, we make a simpification by setting $\mathbf{Q}=\mathbf{H}$ and $\mathbf{B}=\mathbf{C}$ in equation 43, which are defined as $\operatorname{diag}(\mathbf{N}_3(z_{k-1}))$ and $\operatorname{diag}(\mathbf{N}_4(w_{k-1}))$ respectively. Moreover, we take explicit subgradient longer horizon modeling of two subgradient values of iteation $k-1$ and $k-2$ by $(\mathbf{I}-\operatorname{diag}(\mathbf{N}_5(r_{k-1})))\partial r(x_{k-1})_{\text{lb}}+\operatorname{diag}(\mathbf{N}_5(r_{k-1}))$ and $(\mathbf{I}-\operatorname{diag}(\mathbf{N}_5(r_{k-2})))\partial r(x_{k-2})_{\text{lb}}+\operatorname{diag}(\mathbf{N}_5(r_{k-2}))$ respectively. We also omit all bias terms since they are demonstrated to vanish along iteration in Sec. 8.8.

Assume $\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)$ is symmetric positive definite, the necessary and sufficient conditions from convexity definition are:

$$\begin{aligned}
&\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)^{-1}\Big(x_k-\big(x_{k-1}-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1})-\mathbf{N}_2(z_{k-1})\\
&\qquad-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))\operatorname{diag}(\mathbf{N}_4(w_{k-1}))(-(\nabla f(x_{k-1})+g_{k-1})+\nabla f(x_{k-2})+g_{k-2})\big)\Big)+g_k=0,\\
&0\in\partial r(x_k)+\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)^{-1}\Big(x_k-\big(x_{k-1}-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1})-\mathbf{N}_2(z_{k-1})\\
&\qquad-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))\operatorname{diag}(\mathbf{N}_4(w_{k-1}))(-(\nabla f(x_{k-1})+g_{k-1})+\nabla f(x_{k-2})+g_{k-2})\big)\Big),
\end{aligned}\tag{70}$$

which yields the following proximal operator:

$$\begin{aligned}
&\operatorname{prox}_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}\left(x_{k-1}-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1})-\mathbf{N}_2(z_{k-1})\right.\\
&\qquad\left.-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))\operatorname{diag}(\mathbf{N}_4(w_{k-1}))(-(\nabla f(x_{k-1})+g_{k-1})+\nabla f(x_{k-2})+g_{k-2})\right)\\
=&\arg\min_{x_k} r(x_k)+\frac{1}{2}\|x_k-(x_{k-1}-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1})-\mathbf{N}_2(z_{k-1})\\
&\qquad-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))\operatorname{diag}(\mathbf{N}_4(w_{k-1}))(-(\nabla f(x_{k-1})+g_{k-1})+\nabla f(x_{k-2})+g_{k-2}))\|^2_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}.
\end{aligned}\tag{71}$$

Gradient Map

As introduced in the composite case, we still apply the gradient map method to facilitate convergence analysis. We note that both cases share a similar definition of gradient map. Utilizing a denotation P_{k-1} to represent the historical modeling results in both cases, we can represent the update of the L2O model in both cases with the following equation:

$$x_k=x_{k-1}-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\left(\nabla f(x_{k-1})+g_k\right)-\mathbf{N}_2(z_{k-1})-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1},\tag{72}$$

where, in the variable method, P_{k-1} is conducted by:

$$P_{k-1}:=\operatorname{diag}(\mathbf{N}_4(u_{k-1}))(-x_{k-1}+x_{k-2}).\tag{73}$$

In the gradient method, P_{k-1} is conducted by:

$$\begin{aligned}
P_{k-1}:=\operatorname{diag}(\mathbf{N}_4(w_{k-1}))\Big(&-\left(\nabla f(x_{k-1})+(\mathbf{I}-\operatorname{diag}(\mathbf{N}_5(r_{k-1})))\partial r(x_{k-1})_{\text{lb}}+\operatorname{diag}(\mathbf{N}_5(r_{k-1}))\partial r(x_{k-1})_{\text{ub}}\right)\\
&+\nabla f(x_{k-2})+(\mathbf{I}-\operatorname{diag}(\mathbf{N}_5(r_{k-2})))\partial r(x_{k-2})_{\text{lb}}+\operatorname{diag}(\mathbf{N}_5(r_{k-2}))\partial r(x_{k-2})_{\text{ub}}\Big).
\end{aligned}\tag{74}$$

Then, we define a gradient map $G_{\mathbf{N}_1(z)}(x_{k-1})$ that:

$$\begin{aligned}
&G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})\\
=&\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)^{-1}\Big(x_{k-1}-\operatorname{prox}_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}\left(x_{k-1}-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1})-\mathbf{N}_2(z_{k-1})-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1}\right)\\
&\qquad-\mathbf{N}_2(z_{k-1})-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1}\Big).
\end{aligned}$$

And we can represent x_k with $G_{\mathbf{N}_1(z)}(x_{k-1})$ by:

$$\begin{aligned}
x_k=&\operatorname{prox}_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}\Big(x_{k-1}-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)\nabla f(x_{k-1})-\mathbf{N}_2(z_{k-1})\Big)-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1},\\
=&x_{k-1}-\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})-\mathbf{N}_2(z_{k-1})-\operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1}.
\end{aligned}\tag{75}$$

Substitute the above x_k's representation into equation 65, we have the following upper bound of $f(x_k)$:

$$\begin{aligned}
f(x_k)\leq&f(x_{k-1})+\nabla f(x_{k-1})^\top(x_k-x_{k-1})+\frac{L}{2}\|x_k-x_{k-1}\|^2,\\
\leq&f(x_{k-1})-\nabla f(x_{k-1})^\top(\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})+\mathbf{N}_2(z_{k-1})+\operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1})\\
&+\frac{L}{2}\|\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right)G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})+\mathbf{N}_2(z_{k-1})+\operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1}\|^2.
\end{aligned}\tag{76}$$

Similar to equation 50 in the composite case, we still have the following representation of gradient and subgradient:

$$G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})=\nabla f(x_{k-1})+g_k,\tag{77}$$

where $g_k\in\partial r(x_k)$.

Similar to Lemma 2 in the composite case, the general relationship between the objectives of any arbitrary two points in the longer horizon case is as follows:

Lemma 5.

$$
\begin{aligned}
&F(x_k)\\
\leq& F(t) + G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (x_{k-1} - t)\\
&+ \frac{L}{2}\left(\left\| \mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} - \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right.\\
&\qquad \left. - \left\| \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right).
\end{aligned}
$$

The above inequation differs from that of Lemma 2 on the right-hand side. An extra term on historical modeling result P_{k-1} exists. There are two different modeling methods to construct P_{k-1}, i.e. variable method in equation 73 and gradient method in equation 74.

Proof. Workflow of the proof is identical to that of Lemma 2 with a additional but stable term $\mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1}$.

First, we make objective F and apply an upper bound from the convexity definition and gradient map.

$$
\begin{aligned}
&F(x_k)\\
\leq& f(x_{k-1}) - \nabla f(x_{k-1})^\top \left(\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \right)\\
&+ \frac{L}{2} \| \mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \|^2 + r(x_k),\\
\leq& f(t) - \nabla f(x_{k-1})^\top (t - x_{k-1})\\
&- \nabla f(x_{k-1})^\top (\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1})\\
&+ \frac{L}{2} \| \mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \|^2 + r(x_k),\\
\leq& f(t) - \nabla f(x_{k-1})^\top (t - x_{k-1})\\
&- \nabla f(x_{k-1})^\top (\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1})\\
&+ \frac{L}{2} \| \mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \|^2\\
&+ r(t) - \left(G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \nabla f(x_{k-1}) \right)^\top\\
&\qquad \left(t - \left(x_{k-1} - \mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \mathbf{N}_2(z_{k-1}) + \mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \right) \right).
\end{aligned}
$$

In first step, we add $r(x_k)$ to inequality 76. In the second step, we substitute the first-order condition of convex f on x_{k-1}. In the third step, we substitute the gradient map representation of the first-order condition of convex r on x_{k-1}.

Then, we make up the first perfect square:

$$
\begin{aligned}
&F(x_k)\\
\leq& f(t) - \nabla f(x_{k-1})^\top (t - x_{k-1})\\
&- \nabla f(x_{k-1})^\top \big(\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \big)\\
&+ \frac{L}{2} \| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \|^2\\
&+ r(t) - \Big(G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \nabla f(x_{k-1}) \Big)^\top\\
&\qquad \Big(t - \big(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \mathbf{N}_2(z_{k-1}) - \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \big) \Big),\\
=& f(t) + r(t) + \frac{L}{2} \| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1} + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \|^2\\
&- G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top\\
&\qquad \Big(t - \big(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \mathbf{N}_2(z_{k-1}) - \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \big) \Big),\\
=& F(t) + G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (x_{k-1} - t)\\
&+ \frac{L}{2} \| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \|^2\\
&- G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1}).
\end{aligned}
$$

Second perfect square:

$$
\begin{aligned}
&F(x_k)\\
\leq& F(t) + G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (x_{k-1} - t)\\
&+ \frac{L}{2} \Big(\| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \|^2\\
&- \frac{2}{L} G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top \big(\operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \big) \Big),\\
=& F(t) + G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (x_{k-1} - t)\\
&+ \frac{L}{2} \left(\left\| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} - \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right.\\
&\qquad \left. - \left\| \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right).
\end{aligned}
\tag{78}
$$

□

Similarly, we can derive convergence analysis by iteratively applying Lemma 5.

11.2. InD Convergence Upper Bound

Similar to Lemma 3 in the composite case, we use the following lemma to ensure an InD robust L2O model in the longer horizon case.

Lemma 6. *For $\forall z_{k-1} \in \mathcal{Z}_P, \forall x_{k-1} \in \mathcal{S}_P$, if $\mathbf{N}_1(z_{k-1}), \mathbf{N}_2(z_{k-1}), \mathbf{N}_3(z_{k-1})$ are bounded by following compact sets:*

$$
\begin{aligned}
&\mathbf{N}_1(z_{k-1}) \in \left[\mathbf{0}, \frac{2}{L}\mathbf{1} \right],\\
&\left\| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) \left(\nabla f(x_{k-1}) + g_k \right) + \mathbf{N}_2(z_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} - \frac{\nabla f(x_{k-1}) + g_k}{L} \right\| \leq \left\| \frac{\nabla f(x_{k-1}) + g_k}{L} \right\|,\\
&\forall \mathbf{N}_1(z_{k-1}) \in \left[\mathbf{0}, \frac{2}{L}\mathbf{1} \right],
\end{aligned}
$$

where $g_k \in \partial r(x_k)$.

Then, for any x_k *generated by L2O model in equation 68, we have the following homogeneous derease on objective:*

$$F(x_k) - F(x_{k-1}) \leq 0.$$

Proof. The proof is similar that of Lemma 3 in the composite case with an extra $\operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1}$ term.

We first freeze operator $\mathbf{N}_2$ and $\mathbf{N}_3$ and derive the bound for $\mathbf{N}_1$. Then, each given $\mathbf{N}_1$ yields a bound for $\mathbf{N}_2$ and $\mathbf{N}_3$.

Based on the Lemma 5, substituting $t := x_{k-1}$ yields the following upper bound of the objective decrease:

$$\begin{aligned}
&F(x_k) - F(x_{k-1}) \\
\leq &\frac{L}{2}\left(\left\| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} - \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right. \\
&\left. - \left\| \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right).
\end{aligned} \tag{79}$$

To ensure $F(x_k) \leq F(x_{k-1})$, we should keep right-hand side non-positive, which yields:

$$\begin{aligned}
&\frac{L}{2}\left(\left\| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} - \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right. \\
&\left. - \left\| \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right) \leq 0.
\end{aligned}$$

After rearrangement, we have the following inequality:

$$\begin{aligned}
&\left\| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} - \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\| \\
&\leq \frac{\left\| G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) \right\|}{L}.
\end{aligned} \tag{80}$$

Similarly, we first freeze $\mathbf{N}_2(z_{k-1})$, $\operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1}$ and discuss $\mathbf{N}_1(z_{k-1})$-only terms, which yields:

$$\begin{aligned}
&\left\| \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\| \leq \frac{\left\| G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) \right\|}{L}, \\
&\left\| (L \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) - \mathbf{I}) \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\| \leq \frac{\left\| G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) \right\|}{L}.
\end{aligned}$$

Solving the inequation, we have the following boundness for $\mathbf{N}_1(z_{k-1})$:

$$\mathbf{N}_1(z_{k-1}) \in \left[\mathbf{0}, \frac{2}{L}\mathbf{1} \right].$$

Similarly, each choice of $\mathbf{N}_1(z_{k-1})$ yields a pair of bounds for $\mathbf{N}_2(z_{k-1})$ and $\operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1}$. For example, $\mathbf{N}_1(z_{k-1}) := 0$ yields:

$$\left\| \mathbf{N}_2(z_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} - \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\| \leq \frac{\left\| G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) \right\|}{L}.$$

Inductively, freezing $\operatorname{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1}$ yields:

$$\left\| \mathbf{N}_2(z_{k-1}) - \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\| \leq \frac{\left\| G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) \right\|}{L}.$$

Solving the above inequation, we have the following boundness on $\mathbf{N}_2(z_{k-1})$:

$$\mathbf{N}_2(z_{k-1}) \in \left[\mathbf{0}, \frac{2}{L}|G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})|\right].$$

Then, if $\mathbf{N}_2(z_{k-1}) = 0$, we can construct following inequality for $\mathbf{N}_3(z_{k-1})$:

$$\left\|\mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} - \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\| \leq \frac{\left\|G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})\right\|}{L},$$

which yields:

$$\mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \in \left[\mathbf{0}, \frac{2}{L}|G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})|\right].$$

If $\mathbf{N}_2(z_{k-1}) = \frac{1}{L}G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})$, the inequality is:

$$\|\mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1}\| \leq \frac{\left\|G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})\right\|}{L},$$

which yields:

$$\mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} \in \left[-\frac{1}{L}|G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})|, \frac{1}{L}|G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})|\right].$$

Recovering the $G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})$ in inequation 80 with $G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) = \nabla f(x_{k-1}) + g_k$ in equation 77 yields:

$$\|\,\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right)\left(\nabla f(x_{k-1}) + g_k\right) + \mathbf{N}_2(z_{k-1}) + \mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} - \frac{\nabla f(x_{k-1}) + g_k}{L}\| \leq \|\frac{\nabla f(x_{k-1}) + g_k}{L}\|,$$

where $g_k \in \partial r(x_k)$. □

Similar to Corollary 4 in the composite case, we present the following corollary to ensure a robust L2O model in the InD scenario.

Corollary 7. *For any $z_{k-1} \in \mathcal{Z}_P$, we let:*

$$\mathbf{N}_1(z_{k-1}) := \frac{1}{L}\mathbf{1}, \mathbf{N}_2(z_{k-1}) := \mathbf{0}, \mathbf{N}_3(z_{k-1}) := \mathbf{0}, P_{k-1} := \mathbf{0},$$

the Math-L2O model in equation 72 is exactly gradient descent update with convergence rate:

$$F(x_K) - F(x^*) \leq \frac{L}{2K}\|x_0 - x^*\|^2.$$

Proof. In the last term of inequation 79, the best convergence gain yields:

$$\left\|\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right)G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} - \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L}\right\| := 0.$$

$\mathbf{N}_1(z_{k-1}) = \frac{1}{L}\mathbf{1}, \mathbf{N}_2(z_{k-1}) = \mathbf{0}, \mathbf{N}_3(z_{k-1}) = \mathbf{0}, P_{k-1} = \mathbf{0}$ is a feasible solution.

Given $\mathbf{N}_1(z_{k-1}) = \frac{1}{L}\mathbf{1}, \mathbf{N}_2(z_{k-1}) = \mathbf{0}, \mathbf{N}_3(z_{k-1}) = \mathbf{0}, P_{k-1} = \mathbf{0}$, the update formula in equation 75 is:

$$x_k = x_{k-1} - \frac{1}{L}G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}). \tag{81}$$

Based on Lemma 5, when $t := x^*$, we have the following inequality to evaluate per iteration convergence gain:

$$
\begin{aligned}
&F(x_k) - F(x^*) \\
\leq & G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (x_{k-1} - x^*) \\
&+ \frac{L}{2}\left(\left\| \mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) + \mathbf{N}_2(z_{k-1}) + \mathrm{diag}(\mathbf{N}_3(z_{k-1}))P_{k-1} - \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right. \\
&\qquad \left. - \left\| \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right), \\
= & G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (x_{k-1} - x^*) - \frac{L}{2}\left\| \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2, \\
= & \frac{L}{2}\left(\frac{2}{L} G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})^\top (x_{k-1} - x^*) - \left\| \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right), \\
= & \frac{L}{2}\left(\|x_{k-1} - x^*\|^2 - \left\| x_{k-1} - x^* - \frac{G_{\mathrm{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})}{L} \right\|^2 \right), \\
= & \frac{L}{2}(\|x_{k-1} - x^*\|^2 - \|x_{k-1} - x^*\|^2).
\end{aligned}
$$

Sum over K iterations, we have the following InD multi-iteration convergence rate:

$$
F(x_K) - F(x^*) \leq \frac{L}{2K}(\|x_0 - x^*\|^2 - \|x_K - x^*\|^2) \leq \frac{L}{2K}\|x_0 - x^*\|^2. \tag{82}
$$

□

Based on Corollary 7, we further analyze the difference between such a constraint in the variable and gradient methods. The definition in equation 73 and equation 74 yields the following two different conditions for two methods, respectively.

Variable Method:

$$
\mathrm{diag}(\mathbf{N}_3(z_{k-1}))\,\mathrm{diag}(\mathbf{N}_4(u_{k-1}))(-x_{k-1} + x_{k-2}) = 0,
$$

where $u_{k-1} = [x_{k-1}^\top, x_{k-2}^\top]^\top$ is the feature constructed with variable.

Gradient Method:

$$
\mathrm{diag}(\mathbf{N}_3(z_{k-1}))\,\mathrm{diag}(\mathbf{N}_4(w_{k-1}))\,(-(\nabla f(x_{k-1}) + g_{k-1}) + \nabla f(x_{k-2}) + g_{k-2}) = 0,
$$

where $w_{k-1} = [(\nabla f(x_{k-1}) + g_{k-1})^\top, (\nabla f(x_{k-2}) + g_{k-2})^\top]^\top$ is the feature from gradient and subgradient. Moreover, there are two extra neural network operators, $\mathbf{N}_5$ and $\mathbf{N}_6$, to construct the subgradient vectors g_{k-1} and g_{k-2}, respectively:

$$
\begin{aligned}
r_{k-1} &= [\partial r(x_{k-1})_{\mathrm{lb}}^\top, \partial r(x_{k-1})_{\mathrm{ub}}^\top]^\top, \\
r_{k-2} &= [\partial r(x_{k-2})_{\mathrm{lb}}^\top, \partial r(x_{k-2})_{\mathrm{ub}}^\top]^\top, \\
g_{k-1} &= (\mathbf{I} - \mathrm{diag}(\mathbf{N}_5(r_{k-1})))\partial r(x_{k-1})_{\mathrm{lb}} + \mathrm{diag}(\mathbf{N}_5(r_{k-1}))\partial r(x_{k-1})_{\mathrm{ub}}, \\
g_{k-2} &= (\mathbf{I} - \mathrm{diag}(\mathbf{N}_5(r_{k-2})))\partial r(x_{k-2})_{\mathrm{lb}} + \mathrm{diag}(\mathbf{N}_5(r_{k-2}))\partial r(x_{k-2})_{\mathrm{ub}}.
\end{aligned}
$$

We denote $\mathcal{U}_P$ and $\mathcal{W}_P$ as feature spaces upon InD variable space $\mathcal{S}_P$, which are similarly defined as $\mathcal{Z}_P$. Inductively, we define that $P_{k-1} := 0$ is given by $\mathbf{N}_4(u_{k-1}) := 0$ and $\mathbf{N}_4(w_{k-1}) := 0$ in the variable and gradient methods respectively, $\forall u_{k-1} \in \mathcal{U}_P$ and $\forall w_{k-1} \in \mathcal{W}_P$.

We assume both the variable method and gradient methods when modeling longer horizon modeling achieve robustness after training, as in the following assumption:

Assumption 3. *After training,* $\forall x_{k-1} \in \mathcal{S}_P, \forall z_{k-1} \in \mathcal{Z}_P, \forall u_{k-1} \in \mathcal{U}_P, \forall w_{k-1} \in \mathcal{W}_P$, $\mathbf{N}_1(z_{k-1}) := \frac{1}{L}\mathbf{1}$, $\mathbf{N}_2(z_{k-1}) := \mathbf{0}$, $\mathbf{N}_3(z_{k-1}) := \mathbf{0}$, $\mathbf{N}_4(u_{k-1}) := \mathbf{0}$, *and* $\mathbf{N}_4(w_{k-1}) := \mathbf{0}$.

OOD Definitions

Similar to the composite case, we first derive some preliminary formulations for OOD scenarios before the demonstrations. The following definitions follow the same workflow for those in the composite case.

Suppose $z, \tilde{z}, z' \in \mathcal{Z}$, there exists a $\alpha \in [0,1]$ that $z' := \alpha z + (1-\alpha)\tilde{z}, z' \in \mathcal{Z}$. Denote the virtual Jacobian matrix of $\mathbf{N}_1(z')$, $\mathbf{N}_2(z')$, $\mathbf{N}_3(z')$ at point z' as $\mathbf{J}_1$, $\mathbf{J}_2$, $\mathbf{J}_3$, respectively, and $\|\mathbf{J}_1\| \leq \sqrt{n}C_1$, $\|\mathbf{J}_2\| \leq \sqrt{n}C_2$, and $\|\mathbf{J}_3\| \leq \sqrt{n}C_3$.

Since $\mathbf{N}_1(z)$, $\mathbf{N}_2(z)$, $\mathbf{N}_3(z)$ are smooth, due to the Mean Value Theorem, we have the following equalities:

$$\mathbf{N}_1(z) = \mathbf{N}_1(\tilde{z}) + \mathbf{J}_1(z-\tilde{z}), \quad \mathbf{N}_2(z) = \mathbf{N}_2(\tilde{z}) + \mathbf{J}_2(z-\tilde{z}), \quad \mathbf{N}_3(z) = \mathbf{N}_3(\tilde{z}) + \mathbf{J}_3(z-\tilde{z}).$$

Given an OOD virtual variable $s \in \mathbb{R}^n$ (difference in variables between OOD and InD scenarios), we denote the virtual feature (difference in L2O model's input features between OOD and InD scenarios) as s'. For $z+s'$, based on Assumption 3, we have the following representations of the L2O model's outputs in the OOD scenario by those in the InD scenario:

$$\begin{aligned}
\mathbf{N}_1(z+s') &= \mathbf{N}_1(z) + \mathbf{J}_1(z+s'-z) = \mathbf{N}_1(z) + \mathbf{J}_1 s', \\
\mathbf{N}_2(z+s') &= \mathbf{N}_2(z) + \mathbf{J}_2(z+s'-z) = \mathbf{N}_2(z) + \mathbf{J}_2 s', \\
\mathbf{N}_3(z+s') &= \mathbf{N}_3(z) + \mathbf{J}_3(z+s'-z) = \mathbf{N}_3(z) + \mathbf{J}_3 s'.
\end{aligned} \tag{83}$$

Further, for historical modeling operator $\mathbf{N}_4$, we have the following two definitions for variable and gradient methods since different modeling methods have different input feature selections.

Variable Method

We take a similar construction to represent OOD output with InD output for $\mathbf{N}_4$. Suppose $u, \tilde{u}, u' \in \mathcal{U}$ where $\mathcal{U}$ denotes variable space of operator $\mathbf{N}_4$ for the gradient method. There exists a $\alpha \in [0,1]$ that $u' := \alpha u + (1-\alpha)\tilde{u}, u' \in \mathcal{U}$. Denote the virtual Jacobian matrix of $\mathbf{N}_4(u')$ at point u' as $\mathbf{J}_4$, $\|\mathbf{J}_4\| \leq \sqrt{n}C_4$, $\mathbf{N}_4(u')$ follows:

$$\mathbf{N}_4(u) = \mathbf{N}_4(\tilde{u}) + \mathbf{J}_4(u-\tilde{u}).$$

Given two virtual variables $s_1, s_2 \in \mathbb{R}^n$ (variable difference), the difference of neural network $\mathbf{N}_4$ between OOD and InD scenarios u' is defined as:

$$u' = [s_1^\top, s_2^\top]^\top. \tag{84}$$

For $u+u'$, based on Assumption 3, $\mathbf{N}_4(u) = 0$, we have the following representation of OOD output with InD output:

$$\mathbf{N}_4(u+u') = \mathbf{N}_4(u) + \mathbf{J}_4(u+u'-u) = \mathbf{N}_4(u) + \mathbf{J}_4 u' = \mathbf{J}_4 u'. \tag{85}$$

The OOD historical modeling result y'_{k-1} of the variable method is given by:

$$\begin{aligned}
y'_{k-1} &= -\mathbf{N}_4(u'_{k-1})x'_{k-1} + \mathbf{N}_4(u'_{k-1})x'_{k-2}, \\
&= -\mathbf{N}_4(u'_{k-1})(x_{k-1}+s_{k-1}) + \mathbf{N}_4(u'_{k-1})(x_{k-2}+s_{k-2}), \\
&= -\operatorname{diag}(\mathbf{J}_4 u')(x_{k-1}+s_{k-1}) + \operatorname{diag}(\mathbf{J}_4 u')(x_{k-2}+s_{k-2}).
\end{aligned}$$

The InD historical modeling result y_{k-1} of the variable method is given by:

$$y_{k-1} = -\mathbf{N}_4(u_{k-1})x_{k-1} + \mathbf{N}_4(u_{k-1})x_{k-2} = 0,$$

Their difference between OOD and InD scenarios is given by:

$$\begin{aligned}
y'_{k-1} - y_{k-1} &= -\operatorname{diag}(\mathbf{J}_4 u')(x_{k-1}+s_{k-1}) + \operatorname{diag}(\mathbf{J}_4 u')(x_{k-2}+s_{k-2}) - 0 \\
&= -\operatorname{diag}(\mathbf{J}_4 u')(x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2}).
\end{aligned}$$

Based on the above definitions, at k-th iteration, virtual feature s' (difference of features between OOD and InD scenarios) of the variable method is defined by:

$$\begin{aligned}
s'_{k-1} &= \left[s_{k-1}^\top, (\nabla f'(x_{k-1}+s_{k-1}) - \nabla f(x))^\top, (g'_k - g_k)^\top, (y'_{k-1} - y_{k-1})^\top\right]^\top, \\
&= \left[s_{k-1}^\top, (\nabla f'(x_{k-1}+s_{k-1}) - \nabla f(x))^\top, (g'_k - g_k)^\top, \left(-\operatorname{diag}(\mathbf{J}_4 u')(x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2})\right)^\top\right]^\top,
\end{aligned} \tag{86}$$

where $g'_k \in \partial r'(x_k + s_k)$ and $g_k \in \partial r(x_k)$.

Gradient Method

Similarly, suppose $w, \tilde{w}, w' \in \mathcal{W}$, where $\mathcal{W}$ denotes variable space of operator $\mathbf{N}_4$ for the gradient method. There exists a $\alpha \in [0,1]$ that $w' := \alpha w + (1-\alpha)\tilde{w}, w' \in \mathcal{W}$. Denote virtual Jacobian matrix of $\mathbf{N}_4(w')$ at point w' as $\mathbf{J}_4$, $\|\mathbf{J}_4\| \leq \sqrt{n}C_4$, $\mathbf{N}_4(w')$ follows:

$$\mathbf{N}_4(w) = \mathbf{N}_4(\tilde{w}) + \mathbf{J}_4(w - \tilde{w}). \tag{87}$$

Given two virtual variable $s_1, s_2 \in \mathbb{R}^n$ (difference of variables between OOD and InD scenarios), the difference of neural network $\mathbf{N}_4$ between OOD and InD scenarios w' is defined as:

$$w' = [(\nabla f(x_1 + s_1) + g_1')^\top - (\nabla f(x_1) + g_1)^\top, (\nabla f(x_2 + s_2) + g_2')^\top - (\nabla f(x_2) + g_2)^\top]^\top. \tag{88}$$

For $w + w'$, based on Assumption 3, $\mathbf{N}_4(w) = 0$,, we have the following equalities:

$$\mathbf{N}_4(w + w') = \mathbf{N}_4(w) + \mathbf{J}_4(w + w' - W) = \mathbf{J}_4 w'. \tag{89}$$

We eliminate the definition for $\mathbf{N}_5$ since we have defined it to be a diagonal matrix whose diagonal entries $\in [0,1]$.
The OOD historical modeling result y_{k-1}' of the gradient method is given by:

$$\begin{aligned}
y_{k-1}' &= -\mathbf{N}_4(w_{k-1}')(\nabla f'(x_{k-1}') + g_{k-1}') + \mathbf{N}_4(w_{k-1}')(\nabla f'(x_{k-2}') + g_{k-2}') \\
&= -\mathbf{N}_4(w_{k-1}')(\nabla f'(x_{k-1} + s_{k-1}) + g_{k-1}') + \mathbf{N}_4(w_{k-1}')(\nabla f'(x_{k-2} + s_{k-2}) + g_{k-2}') \\
&= -\operatorname{diag}(\mathbf{J}_4 w')(\nabla f'(x_{k-1} + s_{k-1}) + g_{k-1}') + \operatorname{diag}(\mathbf{J}_4 w')(\nabla f'(x_{k-2} + s_{k-2}) + g_{k-2}')
\end{aligned}$$

The InD historical modeling result y_{k-1} of the variable method is given by:

$$y_{k-1} = -\mathbf{N}_4(w_{k-1})(\nabla f(x_{k-1}) + g_{k-1}) + \mathbf{N}_4(w_{k-1})(\nabla f'(x_{k-2}) + g_{k-2}') = 0.$$

Their difference between OOD and InD scenarios is given by:

$$\begin{aligned}
y_{k-1}' - y_{k-1} &= -\operatorname{diag}(\mathbf{J}_4 w')(\nabla f'(x_{k-1} + s_{k-1}) + g_{k-1}') + \operatorname{diag}(\mathbf{J}_4 w')(\nabla f'(x_{k-2} + s_{k-2}) + g_{k-2}') - 0 \\
&= -\operatorname{diag}(\mathbf{J}_4 w')\big(\nabla f'(x_{k-1} + s_{k-1}) - \nabla f'(x_{k-2} + s_{k-2}) + g_{k-1}' - g_{k-2}'\big).
\end{aligned}$$

Based on above definitions, at k-th iteration, virtual feature s' (difference between OOD and InD scenarios) of the gradient method is defined by:

$$\begin{aligned}
& s_{k-1}' \\
=& \left[(\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x))^\top, (g_k' - g_k)^\top, (y_{k-1}' - y_{k-1})^\top\right]^\top \\
=& \Bigg[(\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x))^\top, (g_k' - g_k)^\top, \\
& \Big(-\operatorname{diag}(\mathbf{J}_4 w')\big(\nabla f'(x_{k-1} + s_{k-1}) - \nabla f'(x_{k-2} + s_{k-2}) + g_{k-1}' - g_{k-2}'\big)\Big)^\top\Bigg]^\top,
\end{aligned} \tag{90}$$

where $g_k' \in \partial r'(x_k + s_k)$ and $g_k \in \partial r(x_k)$.

OOD Update Formulation Similar to that in the composite case, based on Lemma 5, $\forall x_k \in \mathcal{S}_p, s_k \in \mathbb{R}^n$, OOD yields the following inequality between any two values of objective:

$$\begin{aligned}
& F'(x_k + s_k) \\
\leq & F'(t) + G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1} + s_{k-1}'))^{-1}}(x_{k-1} + s_{k-1})^\top (x_{k-1} + s_{k-1} - t) \\
& + \frac{L}{2}\left\| \operatorname{diag}(\mathbf{N}_1(z_{k-1} + s_{k-1}')) G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1} + s_{k-1}'))^{-1}}(x_{k-1} + s_{k-1}) + \mathbf{N}_2(z_{k-1} + s_{k-1}') + \operatorname{diag}(\mathbf{N}_3(z_{k-1} + s_{k-1}')) P_{k-1}' \right. \\
& \left. - \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1} + s_{k-1}'))^{-1}}(x_{k-1} + s_{k-1})}{L} \right\|^2 - \frac{L}{2}\left\| \frac{G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1} + s_{k-1}'))^{-1}}(x_{k-1} + s_{k-1})}{L} \right\|^2.
\end{aligned} \tag{91}$$

Similar to the composite case, we directly get the following formulation for the OOD gradient map:

$$G'_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1}+s_{k-1}) = \nabla f'(x_{k-1}+s_{k-1}) + g'_k, \tag{92}$$

where $g'_k \in \partial r'(x_k + s_k)$.

11.3. OOD Per-Iteration Convergence Gain

Based on the Lemma 6 and Corollary 7, Assumption 3 leads to an L2O model with best robustness on all InD instances.

Based on Assumption 3, in the following theorem, we quantify the diminution in convergence rate instigated by the virtual feature s' defined in Sec. 3.

Theorem 6. *Under Assumption 3, there exists virtual Jacobian matrices $\mathbf{J}_{1,k-1}, \mathbf{J}_{2,k-1}, \mathbf{J}_{3,k-1}, k = 1, 2, \ldots, K$ that OOD's convergence improvement of one iteration is upper bounded by following inequality:*

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &-\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + \frac{L}{2}\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1} + \operatorname{diag}(\mathbf{J}_3 s'_{k-1})P'_{k-1}\|^2,
\end{aligned}$$

where $g'_k \in \partial r'(x_k + s_k)$ and P'_{k-1} represents historical modeling result.

Proof. From equation 83, for operators $\mathbf{N}_1$, $\mathbf{N}_2$, and $\mathbf{N}_3$, we have the following representations of OOD outputs by their InD outputs:

$$\begin{aligned}
\mathbf{N}_1(z_{k-1} + s'_{k-1}) &= \mathbf{N}_1(z_{k-1}) + \mathbf{J}_1 s'_{k-1}, \\
\mathbf{N}_2(z_{k-1} + s'_{k-1}) &= \mathbf{N}_2(z_{k-1}) + \mathbf{J}_2 s'_{k-1}, \\
\mathbf{N}_3(z_{k-1} + s'_{k-1}) &= \mathbf{N}_3(z_{k-1}) + \mathbf{J}_3 s'_{k-1}.
\end{aligned}$$

Substituting the definitions of $\mathbf{N}_1(z_{k-1})$, $\mathbf{N}_2(z_{k-1})$, and $\mathbf{N}_3(z_{k-1})$ in Assumption 3 yields:

$$\begin{aligned}
\mathbf{N}_1(z_{k-1} + s'_{k-1}) &= \frac{1}{L}\mathbf{1} + \mathbf{J}_1 s'_{k-1} \\
\mathbf{N}_2(z_{k-1} + s'_{k-1}) &= \mathbf{J}_2 s'_{k-1} \\
\mathbf{N}_3(z_{k-1} + s'_{k-1}) &= \mathbf{J}_3 s'_{k-1}.
\end{aligned} \tag{93}$$

We substitut $t := x_{k-1} + s_{k-1}$ into inequation 91 to construct the objective difference between two iterations:

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &\frac{L}{2}\left\| \operatorname{diag}(\mathbf{N}_1(z_{k-1} + s'_{k-1}))G'_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1}) + \mathbf{N}_2(z_{k-1} + s'_{k-1}) + \operatorname{diag}(\mathbf{N}_3(z_{k-1} + s'_{k-1}))P'_{k-1} \right.\\
&\left. - \frac{G'_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1})}{L} \right\|^2 - \frac{L}{2}\left\| \frac{G'_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1})}{L} \right\|^2.
\end{aligned}$$

By equation 93, we can represent the OOD outputs and achive the following reformulation:

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &\frac{L}{2}\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})G'_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1}) + \mathbf{J}_2 s'_{k-1} + \operatorname{diag}(\mathbf{J}_3 s'_{k-1})P'_{k-1}\|^2 \\
&- \frac{L}{2}\left\| \frac{G'_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}+s'_{k-1}))^{-1}}(x_{k-1} + s_{k-1})}{L} \right\|^2.
\end{aligned}$$

Then, based on equation 92, we recover gradient and subgradient from the gradient map:

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &-\frac{L}{2}\left\| \frac{\nabla f'(x_{k-1} + s_{k-1}) + g'_k}{L} \right\|^2 + \frac{L}{2}\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1} + \operatorname{diag}(\mathbf{J}_3 s'_{k-1})P'_{k-1}\|^2, \\
= &-\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + \frac{L}{2}\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1} + \operatorname{diag}(\mathbf{J}_3 s'_{k-1})P'_{k-1}\|^2.
\end{aligned}$$

□

For variable and gradient methods, based on Theorem 6, we have the following different theorems of per iteration convergence gain.

Variable Method

From equation 85 and definition in Assumption 3, we have the following representation of $\mathbf{N}_4(u+u')$:

$$\mathbf{N}_4(u+u') = \mathbf{J}_4 u'.$$

For the variable method, Theorem 6 yields the following theorem of the per iteration convergence gain:

Theorem 7. *Under Assumption 3, there exists virtual Jacobian matrices $\mathbf{J}_{1,k-1}, \mathbf{J}_{2,k-1}, \mathbf{J}_{3,k-1}, \mathbf{J}_{4,k-1}, k = 1, 2, \ldots, K$ that OOD's convergence improvement of one iteration is upper bounded by following inequality:*

$$\begin{aligned}
&F'(x_k+s_k) - F'(x_{k-1}+s_{k-1}) \\
\leq &-\frac{\|\nabla f'(x_{k-1}+s_{k-1}) + g'_k\|^2}{2L} + \frac{L}{2}\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1} \\
&+ \operatorname{diag}(\mathbf{J}_3 s'_{k-1})\operatorname{diag}(\mathbf{J}_4 u'_{k-1})(-(x_{k-1}+s_{k-1}) + x_{k-2} + s_{k-2})\|^2,
\end{aligned}$$

where $g'_k \in \partial r'(x_k+s_k)$.

Theorem 7 yields following corollary for its upper bound:

Corollary 8. *Under Assumption 3, the convergence improvement for one iteration of the OOD scenario can be upper bounded w.r.t. s_{k-1} and s_{k-2} by:*

$$\begin{aligned}
&F'(x_k+s_k) - F'(x_{k-1}+s_{k-1}) \\
\leq &-\frac{\|\nabla f'(x_{k-1}+s_{k-1}) + g'_k\|^2}{2L} \\
&+ \Big(LC_1^2\|\nabla f'(x_{k-1}+s_{k-1}) + g'_k\|^2 + LC_2^2 + LC_3^2C_4^2(\|s_{k-1}\|^2 + \|s_{k-2}\|^2)\|x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2}\|^2\Big) \\
&\quad \times \Big(\|s_{k-1}\|^2 + \|\nabla f'(x_{k-1}+s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 \\
&\qquad + C_4^2(\|s_{k-1}\|^2 + \|s_{k-2}\|^2)\|x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2}\|^2\Big),
\end{aligned}$$

where $g'_k \in \partial r'(x_k+s_k)$ and $g_k \in \partial r(x_k)$.

Proof. We iteratively apply Triangle and Cauchy Schwarz inequalities:

$$\begin{aligned}
&F'(x_k+s_k) - F'(x_{k-1}+s_{k-1}) \\
\leq &-\frac{\|\nabla f'(x_{k-1}+s_{k-1}) + g'_k\|^2}{2L} + \frac{L}{2}\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1} \\
&+ \operatorname{diag}(\mathbf{J}_3 s'_{k-1})\operatorname{diag}(\mathbf{J}_4 u'_{k-1})(-(x_{k-1}+s_{k-1}) + x_{k-2} + s_{k-2})\|^2, \\
\leq &-\frac{\|\nabla f'(x_{k-1}+s_{k-1}) + g'_k\|^2}{2L} + L\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1}) + g'_k)\|^2 + L\|\mathbf{J}_2 s'_{k-1}\|^2 \\
&+ L\|\operatorname{diag}(\mathbf{J}_3 s'_{k-1})\operatorname{diag}(\mathbf{J}_4 u'_{k-1})(-(x_{k-1}+s_{k-1}) + x_{k-2} + s_{k-2})\|^2, \\
\leq &-\frac{\|\nabla f'(x_{k-1}+s_{k-1}) + g'_k\|^2}{2L} + L\|\mathbf{J}_1 s'_{k-1}\|^2\|\nabla f'(x_{k-1}+s_{k-1}) + g'_k\|^2 + L\|\mathbf{J}_2 s'_{k-1}\|^2 \\
&+ L\|\mathbf{J}_3 s'_{k-1}\|^2\|\mathbf{J}_4 u'_{k-1}\|^2\| - (x_{k-1}+s_{k-1}) + x_{k-2} + s_{k-2}\|^2, \\
\leq &-\frac{\|\nabla f'(x_{k-1}+s_{k-1}) + g'_k\|^2}{2L} + LC_1^2\|\nabla f'(x_{k-1}+s_{k-1}) + g'_k\|^2\|s'_{k-1}\|^2 + LC_2^2\|s'_{k-1}\|^2 \\
&+ LC_3^2C_4^2\|s'_{k-1}\|^2\|u'_{k-1}\|^2\|x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2}\|^2, \\
= &-\frac{\|\nabla f'(x_{k-1}+s_{k-1}) + g'_k\|^2}{2L} \\
&+ \Big(LC_1^2\|\nabla f'(x_{k-1}+s_{k-1}) + g'_k\|^2 + LC_2^2 + LC_3^2C_4^2\|u'_{k-1}\|^2\|x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2}\|^2\Big)\|s'_{k-1}\|^2,
\end{aligned} \tag{94}$$

where $g'_k \in \partial r'(x_k + s_k)$.

Based on the definition of u'_{k-1} in equation 84, we calculate its vector-norm by:

$$\|u'_{k-1}\|^2 = \|[s_{k-1}, s_{k-2}]\|^2 = \|s_{k-1}\|^2 + \|s_{k-2}\|^2.$$

Substituting it into inequation 94 yields the following upper bound:

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &- \frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} \\
&+ \Big(LC_1^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + LC_2^2 + LC_3^2C_4^2(\|s_{k-1}\|^2 + \|s_{k-2}\|^2)\|x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2}\|^2\Big)\|s'_{k-1}\|^2.
\end{aligned} \tag{95}$$

Moreover, the definition of s'_{k-1} with variable method in equation 86 yields:

$$\begin{aligned}
\|s'_{k-1}\|^2 =&\|s_{k-1}\|^2 + \|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 + \| - \operatorname{diag}(\mathbf{J}_4 u')(x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2})\|^2, \\
\leq&\|s_{k-1}\|^2 + \|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 + \| - \operatorname{diag}(\mathbf{J}_4 u')(x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2})\|^2, \\
\leq&\|s_{k-1}\|^2 + \|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 + \|\mathbf{J}_4 u'\|^2\|x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2}\|^2, \\
=&\|s_{k-1}\|^2 + \|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 + \|\mathbf{J}_4 u'\|^2\|x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2}\|^2, \\
\leq&\|s_{k-1}\|^2 + \|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 + C_4^2\|u'\|^2\|x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2}\|^2, \\
=&\|s_{k-1}\|^2 + \|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 \\
&+ C_4^2\left(\|s_{k-1}\|^2 + \|s_{k-2}\|^2\right)\|x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2}\|^2,
\end{aligned}$$

where $g'_k \in \partial r'(x_k + s_k)$, $g_k \in \partial r(x_k)$, and the third step is based on Cauchy-Schwarz inequality.

Substituting the above vector norm into the inequation 95 yields the final upper bound:

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &- \frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} \\
&+ \Big(LC_1^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + LC_2^2 + LC_3^2C_4^2(\|s_{k-1}\|^2 + \|s_{k-2}\|^2)\|x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2}\|^2\Big) \\
&\times \Big(\|s_{k-1}\|^2 + \|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 \\
&\quad + C_4^2(\|s_{k-1}\|^2 + \|s_{k-2}\|^2)\|x_{k-1} - x_{k-2} + s_{k-1} - s_{k-2}\|^2\Big),
\end{aligned}$$

where $g'_k \in \partial r'(x_k + s_k)$ and $g_k \in \partial r(x_k)$. □

Gradient Method

From equation 89 and definition in Assumption 3, we have the following representation of $\mathbf{N}_4(u + u')$:

$$\mathbf{N}_4(w + w') = \mathbf{J}_4 w'.$$

For the gradient method, Theorem 6 yields the following theorem of the per iteration convergence gain:

Theorem 8. *Under Assumption 3, there exists virtual Jacobian matrices $\mathbf{J}_{1,k-1}, \mathbf{J}_{2,k-1}, \mathbf{J}_{3,k-1}, \mathbf{J}_{4,k-1}, k = 1, 2, \ldots, K$ that OOD's convergence improvement of one iteration is upper bounded by following inequality:*

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq &- \frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + \frac{L}{2}\| \operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1} \\
&+ \operatorname{diag}(\mathbf{J}_3 s'_{k-1}) \operatorname{diag}(\mathbf{J}_4 w'_{k-1})\big(- (\nabla f'(x_{k-1} + s_{k-1}) + g'_{k-1}) + \nabla f'(x_{k-2} + s_{k-2}) + g'_{k-2}\big)\|^2,
\end{aligned}$$

where $g'_k \in \partial r'(x_k + s_k)$ *and* g'_{k-1}, g'_{k-2} *follow:*

$$g'_{k-1} \in [\partial r'(x_{k-1} + s_{k-1})_{lb}, \partial r'(x_{k-1} + s_{k-1})_{ub}],$$
$$g'_{k-2} \in [\partial r'(x_{k-2} + s_{k-2})_{lb}, \partial r'(x_{k-2} + s_{k-2})_{ub}].$$

Theorem 8 yields the following corollary for its upper bound:

Corollary 9. *Under Assumption 3, the convergence improvement for one iteration of the OOD scenario can be upper bounded w.r.t.* s_{k-1} *and* s_{k-2} *by:*

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq& -\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} \\
&+ \Big(Ln^2C_1^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + Ln^2C_2^2 \\
&\quad + Ln^4C_3^2C_4^2\big(L^2(\|s_{k-1}\|^2 + \|s_{k-2}\|^2) + \|g'_{k-1} - g_{k-1}\|^2 + \|g'_{k-2} - g_{k-2}\|^2\big) \\
&\quad\quad \times \big(L^2\|x_{k-2} + s_{k-2} - x_{k-1} - s_{k-1}\|^2 + \|g'_{k-2} - g'_{k-1}\|^2\big)\Big) \\
&\times \Big(\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 \\
&\quad + n^2C_4^2\big(L^2(\|s_{k-1}\|^2 + \|s_{k-2}\|^2) + \|g'_{k-1} - g_{k-1}\|^2 + \|g'_{k-2} - g_{k-2}\|^2\big) \\
&\quad\quad \times \big(L^2\|x_{k-1} + s_{k-1} - x_{k-2} - s_{k-2}\|^2 + \|g'_{k-1} - g'_{k-2}\|^2\big)\Big),
\end{aligned}$$

where $g'_k \in \partial r'(x_k + s_k)$, $g'_{k-1} \in \partial r'(x_{k-1} + s_{k-1})$, *and* $g'_{k-2} \in \partial r'(x_{k-2} + s_{k-2})$.

Proof.

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq& -\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + \frac{L}{2}\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k) + \mathbf{J}_2 s'_{k-1} \\
&+ \operatorname{diag}(\mathbf{J}_3 s'_{k-1})\operatorname{diag}(\mathbf{J}_4 w'_{k-1})\big(-(\nabla f'(x_{k-1} + s_{k-1}) + g'_{k-1}) + \nabla f'(x_{k-2} + s_{k-2}) + g'_{k-2}\big)\|^2, \\
\leq& -\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + L\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1} + s_{k-1}) + g'_k)\|^2 + L\|\mathbf{J}_2 s'_{k-1}\|^2 \\
&+ L\|\operatorname{diag}(\mathbf{J}_3 s'_{k-1})\operatorname{diag}(\mathbf{J}_4 w'_{k-1})\big(-(\nabla f'(x_{k-1} + s_{k-1}) + g'_{k-1}) + \nabla f'(x_{k-2} + s_{k-2}) + g'_{k-2}\big)\|^2, \\
\leq& -\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + L\|\mathbf{J}_1 s'_{k-1}\|^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + L\|\mathbf{J}_2 s'_{k-1}\|^2 \\
&+ L\|\mathbf{J}_3 s'_{k-1}\|^2\|\mathbf{J}_4 w'_{k-1}\|^2\|\nabla f'(x_{k-2} + s_{k-2}) - \nabla f'(x_{k-1} + s_{k-1}) + g'_{k-2} - g'_{k-1}\|^2, \\
\leq& -\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + Ln^2C_1^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2\|s'_{k-1}\|^2 + Ln^2C_2^2\|s'_{k-1}\|^2 \\
&+ Ln^2C_3^2n^2C_4^2\|s'_{k-1}\|^2\|w'_{k-1}\|^2\|\nabla f'(x_{k-2} + s_{k-2}) - \nabla f'(x_{k-1} + s_{k-1}) + g'_{k-2} - g'_{k-1}\|^2, \\
\leq& -\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + Ln^2C_1^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2\|s'_{k-1}\|^2 + Ln^2C_2^2\|s'_{k-1}\|^2 \\
&+ Ln^4C_3^2C_4^2\|s'_{k-1}\|^2\|w'_{k-1}\|^2\big(\|\nabla f'(x_{k-2} + s_{k-2}) - \nabla f'(x_{k-1} + s_{k-1})\|^2 + \|g'_{k-2} - g'_{k-1}\|^2\big), \\
\leq& -\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} + \big(Ln^2C_1^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + Ln^2C_2^2\big)\|s'_{k-1}\|^2 \\
&+ Ln^4C_3^2C_4^2\|w'_{k-1}\|^2\big(L^2\|x_{k-2} + s_{k-2} - x_{k-1} - s_{k-1}\|^2 + \|g'_{k-2} - g'_{k-1}\|^2\big)\|s'_{k-1}\|^2, \\
=& -\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} \\
&+ \Big(Ln^2C_1^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + Ln^2C_2^2 \\
&\quad + Ln^4C_3^2C_4^2\|w'_{k-1}\|^2\big(L^2\|x_{k-2} + s_{k-2} - x_{k-1} - s_{k-1}\|^2 + \|g'_{k-2} - g'_{k-1}\|^2\big)\Big)\|s'_{k-1}\|^2,
\end{aligned} \tag{96}$$

where $g'_k \in \partial r'(x_k + s_k)$, $g'_{k-1} \in \partial r'(x_{k-1} + s_{k-1})$, and $g'_{k-2} \in \partial r'(x_{k-2} + s_{k-2})$. The last step is based on the definition of L-smoothness on f'.

Based on the definition of w'_{k-1} in equation 88, we calculate its vector-norm by:

$$w'_{k-1} = \left[(\nabla f(x_{k-1} + s_{k-1}) + g'_{k-1} - \nabla f(x_{k-1}) - g_{k-1})^\top, (\nabla f(x_{k-2} + s_{k-2}) + g'_{k-2} - \nabla f(x_{k-2}) - g_{k-2})^\top\right]^\top.$$

Thus, we have:

$$\begin{aligned}
&\|w'_{k-1}\|^2 \\
=&\|\nabla f(x_{k-1} + s_{k-1}) + g'_{k-1} - \nabla f(x_{k-1}) - g_{k-1}, \nabla f(x_{k-2} + s_{k-2}) + g'_{k-2} - \nabla f(x_{k-2}) - g_{k-2}\|^2, \\
=&\|\nabla f(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1}) + g'_{k-1} - g_{k-1}, \nabla f(x_{k-2} + s_{k-2}) - \nabla f(x_{k-2}) + g'_{k-2} - g_{k-2}\|^2, \\
=&\|\nabla f(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1}) + g'_{k-1} - g_{k-1}\|^2 + \|\nabla f(x_{k-2} + s_{k-2}) - \nabla f(x_{k-2}) + g'_{k-2} - g_{k-2}\|^2, \\
\leq&\|\nabla f(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_{k-1} - g_{k-1}\|^2 + \|\nabla f(x_{k-2} + s_{k-2}) - \nabla f(x_{k-2})\|^2 + \|g'_{k-2} - g_{k-2}\|^2, \\
\leq&L^2\|x_{k-1} + s_{k-1} - x_{k-1}\|^2 + \|g'_{k-1} - g_{k-1}\|^2 + L^2\|x_{k-2} + s_{k-2} - x_{k-2}\|^2 + \|g'_{k-2} - g_{k-2}\|^2, \\
=&L^2(\|s_{k-1}\|^2 + \|s_{k-2}\|^2) + \|g'_{k-1} - g_{k-1}\|^2 + \|g'_{k-2} - g_{k-2}\|^2.
\end{aligned}$$

In steps 1-3, we rearrange items. The 4th step is based on Triangle inequality. The 4th step is based on Cauchy-Schwarz inequality. The 5th step is based on the definition of L-smoothness on f.

Substituting $\|w'_{k-1}\|^2$'s upper bound into above inequality 96 yields:

$$\begin{aligned}
&F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq& -\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} \\
&+ \Big(Ln^2C_1^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + Ln^2C_2^2 \\
&\quad + Ln^4C_3^2C_4^2\big(L^2(\|s_{k-1}\|^2 + \|s_{k-2}\|^2) + \|g'_{k-1} - g_{k-1}\|^2 + \|g'_{k-2} - g_{k-2}\|^2\big) \\
&\quad\quad \times \big(L^2\|x_{k-2} + s_{k-2} - x_{k-1} - s_{k-1}\|^2 + \|g'_{k-2} - g'_{k-1}\|^2\big)\Big)\|s'_{k-1}\|^2,
\end{aligned}$$

where $g'_k \in \partial r'(x_k + s_k)$, $g_k \in \partial r(x_k)$, $g'_{k-1} \in \partial r'(x_{k-1} + s_{k-1})$, and $g'_{k-2} \in \partial r'(x_{k-2} + s_{k-2})$.

Moreover, the definition of s'_{k-1} with variable method in equation 90 yields:

$$\begin{aligned}
&\|s'_{k-1}\|^2 \\
=&\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 \\
&+ \| - \mathrm{diag}(\mathbf{J}_4 w')\big(\nabla f'(x_{k-1} + s_{k-1}) - \nabla f'(x_{k-2} + s_{k-2}) + g'_{k-1} - g'_{k-2}\big)\|^2, \\
\leq&\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 + \|\mathbf{J}_4 w'\|^2\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f'(x_{k-2} + s_{k-2}) + g'_{k-1} - g'_{k-2}\|^2, \\
\leq&\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 \\
&+ \|\mathbf{J}_4 w'\|^2\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f'(x_{k-2} + s_{k-2})\|^2 + \|\mathbf{J}_4 w'\|^2\|g'_{k-1} - g'_{k-2}\|^2, \\
=&\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 \\
&+ \|\mathbf{J}_4 w'\|^2\big(\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f'(x_{k-2} + s_{k-2})\|^2 + \|g'_{k-1} - g'_{k-2}\|^2\big), \\
\leq&\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 \\
&+ \|\mathbf{J}_4 w'\|^2\big(L^2\|x_{k-1} + s_{k-1} - x_{k-2} - s_{k-2}\|^2 + \|g'_{k-1} - g'_{k-2}\|^2\big), \\
\leq&\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 \\
&+ n^2C_4^2\|w'\|^2\big(L^2\|x_{k-1} + s_{k-1} - x_{k-2} - s_{k-2}\|^2 + \|g'_{k-1} - g'_{k-2}\|^2\big), \\
\leq&\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 \\
&+ n^2C_4^2\big(L^2(\|s_{k-1}\|^2 + \|s_{k-2}\|^2) + \|g'_{k-1} - g_{k-1}\|^2 + \|g'_{k-2} - g_{k-2}\|^2\big) \\
&\quad\quad \times \big(L^2\|x_{k-1} + s_{k-1} - x_{k-2} - s_{k-2}\|^2 + \|g'_{k-1} - g'_{k-2}\|^2\big),
\end{aligned}$$

where $g'_k \in \partial r'(x_k + s_k)$ and $g_k \in \partial r(x_k)$. The second and third steps are based on Cauchy-Schwarz inequality and Triangle inequality. The 5th step is based on the definition of L-smoothness on f'.

Substituting the above formulation into the above inequality yields:

$$
\begin{aligned}
& F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq & -\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} \\
& + \Big(Ln^2 C_1^2 \|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + Ln^2 C_2^2 \\
& \quad + Ln^4 C_3^2 C_4^2 \big(L^2 (\|s_{k-1}\|^2 + \|s_{k-2}\|^2) + \|g'_{k-1} - g_{k-1}\|^2 + \|g'_{k-2} - g_{k-2}\|^2 \big) \\
& \quad\quad \times \big(L^2 \|x_{k-2} + s_{k-2} - x_{k-1} - s_{k-1}\|^2 + \|g'_{k-2} - g'_{k-1}\|^2 \big) \Big) \\
& \times \Big(\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 \\
& \quad + n^2 C_4^2 \big(L^2 (\|s_{k-1}\|^2 + \|s_{k-2}\|^2) + \|g'_{k-1} - g_{k-1}\|^2 + \|g'_{k-2} - g_{k-2}\|^2 \big) \\
& \quad\quad \times \big(L^2 \|x_{k-1} + s_{k-1} - x_{k-2} - s_{k-2}\|^2 + \|g'_{k-1} - g'_{k-2}\|^2 \big) \Big),
\end{aligned}
$$

where $g'_k \in \partial r'(x_k + s_k)$, $g'_{k-1} \in \partial r'(x_{k-1} + s_{k-1})$, and $g'_{k-2} \in \partial r'(x_{k-2} + s_{k-2})$.

□

Comparison between Variable Method and Gradient Method

As introduced in corollaries 8 and 9, we have demonstrated the per iteration convergence gain of the variable and the gradient methods, respectively. We are ready to compare the variable and gradient methods regarding such bounds. We categorically derive the analysis with and without the non-smooth part in the objective. We focus on the case without non-smooth parts based on the assumption that the non-smooth function in the objective is trivially solvable. Such an assumption is achievable in real-world scenarios. For example, in the blurring task for computer vision [5], the non-smooth function is L_1-norm and serves as a regularization term for the smooth objective.

Without Subgradient Case In this case, we remove all subgradient in historical modeling, which yields:

$$
g'_{k-1} := 0, g'_{k-2} := 0, g_{k-1} := 0, g_{k-2} := 0.
$$

Thus, Corollary 9 is simplified into:

$$
\begin{aligned}
& F'(x_k + s_k) - F'(x_{k-1} + s_{k-1}) \\
\leq & -\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L} \\
& + \Big(Ln^2 C_1^2 \|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + Ln^2 C_2^2 \\
& \quad + Ln^4 C_3^2 C_4^{2,g} \big(L^2 (\|s_{k-1}\|^2 + \|s_{k-2}\|^2) \big) \big(L^2 \|x_{k-2} + s_{k-2} - x_{k-1} - s_{k-1}\|^2 \big) \Big) \\
& \times \Big(\|\nabla f'(x_{k-1} + s_{k-1}) - \nabla f(x_{k-1})\|^2 + \|g'_k - g_k\|^2 \\
& \quad + n^2 C_4^{2,g} \big(L^2 (\|s_{k-1}\|^2 + \|s_{k-2}\|^2) \big) \big(L^2 \|x_{k-1} + s_{k-1} - x_{k-2} - s_{k-2}\|^2 \big) \Big),
\end{aligned}
$$

where $g'_k \in \partial r'(x_k + s_k)$ and we use the superscript ^{g} to represent gradient method's C_4.

If we further assume $f'(x_{k-1} + s_{k-1}) := f(x_{k-1} + s_{k-1} + t), t \in \mathbb{R}^n$, which means the OOD on objective is a shifting

on variable, we can further get following upper bound for the gradient method's per iteration convergence gain:

$$
\begin{aligned}
&F'(x_k+s_k)-F'(x_{k-1}+s_{k-1})\\
\leq&-\frac{\|\nabla f'(x_{k-1}+s_{k-1})+g'_k\|^2}{2L}\\
&+\Big(Ln^2C_1^2\|\nabla f'(x_{k-1}+s_{k-1})+g'_k\|^2+n^2C_2^2\\
&\quad+Ln^4C_3^2C_4^{2,g}\big(L^2(\|s_{k-1}\|^2+\|s_{k-2}\|^2)\big)\big(L^2\|x_{k-2}+s_{k-2}-x_{k-1}-s_{k-1}\|^2\big)\Big)\\
&\quad\times\Big(\|\nabla f(x_{k-1}+s_{k-1}+t)-\nabla f(x_{k-1})\|^2+\|g'_k-g_k\|^2\\
&\qquad+n^2C_4^{2,g}\big(L^2(\|s_{k-1}\|^2+\|s_{k-2}\|^2)\big)\big(L^2\|x_{k-1}+s_{k-1}-x_{k-2}-s_{k-2}\|^2\big)\Big),\\
\leq&-\frac{\|\nabla f'(x_{k-1}+s_{k-1})+g'_k\|^2}{2L}\\
&+\Big(Ln^2C_1^2\|\nabla f'(x_{k-1}+s_{k-1})+g'_k\|^2+Ln^2C_2^2\\
&\quad+Ln^4C_3^2C_4^{2,g}\big(L^2(\|s_{k-1}\|^2+\|s_{k-2}\|^2)\big)\big(L^2\|x_{k-2}+s_{k-2}-x_{k-1}-s_{k-1}\|^2\big)\Big)\\
&\quad\times\Big(L^2\|s_{k-1}+t\|^2+\|g'_k-g_k\|^2+n^2C_4^{2,g}\big(L^2(\|s_{k-1}\|^2+\|s_{k-2}\|^2)\big)\big(L^2\|x_{k-1}+s_{k-1}-x_{k-2}-s_{k-2}\|^2\big)\Big),\\
=&-\frac{\|\nabla f'(x_{k-1}+s_{k-1})+g'_k\|^2}{2L}\\
&+\Big(Ln^2C_1^2\|\nabla f'(x_{k-1}+s_{k-1})+g'_k\|^2+Ln^2C_2^2\\
&\quad+Ln^4C_3^2C_4^{2,g}L^4(\|s_{k-1}\|^2+\|s_{k-2}\|^2)(\|x_{k-2}+s_{k-2}-x_{k-1}-s_{k-1}\|^2)\Big)\\
&\quad\times\Big(L^2\|s_{k-1}+t\|^2+\|g'_k-g_k\|^2+n^2C_4^{2,g}L^4(\|s_{k-1}\|^2+\|s_{k-2}\|^2)(\|x_{k-1}+s_{k-1}-x_{k-2}-s_{k-2}\|^2)\Big).
\end{aligned}
\tag{97}
$$

Similarly, we get the bound of the variable method by:

$$
\begin{aligned}
&F'(x_k+s_k)-F'(x_{k-1}+s_{k-1})\\
\leq&-\frac{\|\nabla f'(x_{k-1}+s_{k-1})+g'_k\|^2}{2L}\\
&+\Big(Ln^2C_1^2\|\nabla f'(x_{k-1}+s_{k-1})+g'_k\|^2+Ln^2C_2^2+Ln^4C_3^2C_4^{2,v}(\|s_{k-1}\|^2+\|s_{k-2}\|^2)\|x_{k-1}-x_{k-2}+s_{k-1}-s_{k-2}\|^2\Big)\\
&\quad\times\Big(\|s_{k-1}\|^2+L^2\|s_{k-1}+t\|^2+\|g'_k-g_k\|^2+n^2C_4^{2,v}(\|s_{k-1}\|^2+\|s_{k-2}\|^2)\|x_{k-1}-x_{k-2}+s_{k-1}-s_{k-2}\|^2\Big),
\end{aligned}
\tag{98}
$$

where we use the superscript ^{v} to represent variable method's C_4.

If $L\leq 1$, which means the objective is smooth, the upper bound yielded by the gradient method in inequality 97 is intrinsically smaller than that in inequality 98, which means that gradient-based longer horizon modeling methods are better for functions with small gradients. We note that $L\leq 1$ is general in real-world scenarios. For example, $L\leq 1$ in logistic regression tasks are inherently achieved by average among features.

Otherwise, if $L>1$, to keep an identical boundness in inequalities 97 and 98, we can also achieve a lower convergence bound by shrinking the output of operator $\mathbf{N}_4$ in the gradient method, such as setting $C_4^g=C_4^v/(L^2)$ in 97, which yields:

$$
\begin{aligned}
&F'(x_k+s_k)-F'(x_{k-1}+s_{k-1})\\
\leq&-\frac{\|\nabla f'(x_{k-1}+s_{k-1})+g'_k\|^2}{2L}\\
&+\Big(Ln^2C_1^2\|\nabla f'(x_{k-1}+s_{k-1})+g'_k\|^2+Ln^2C_2^2+Ln^4C_3^2C_4^{2,v}(\|s_{k-1}\|^2+\|s_{k-2}\|^2)(\|x_{k-2}+s_{k-2}-x_{k-1}-s_{k-1}\|^2)\Big)\\
&\quad\times\Big(L^2\|s_{k-1}+t\|^2+\|g'_k-g_k\|^2+n^2C_4^{2,v}(\|s_{k-1}\|^2+\|s_{k-2}\|^2)(\|x_{k-1}+s_{k-1}-x_{k-2}-s_{k-2}\|^2)\Big).
\end{aligned}
\tag{99}
$$

Moreover, the remaining difference between such two upper bounds are $L^2\|s_{k-1}+t\|^2$ in the gradient method and $\|s_{k-1}\|^2+L^2\|s_{k-1}+t\|^2$ in the variable method. Hence, the gradient method yields a smaller convergence gain upper bound.

To sum up, we have the following conclusions:

1) $L\in[0,1]$. The gradient-based longer horizon modeling method is more robust in OOD scenarios.
2) $L\in(1,\infty]$. By setting $C_4^g\leq C_4^v/(L^2)$, the gradient-based longer horizon modeling method is more robust in OOD scenarios.

With Subgradient Case We eliminate this case since we assume $r(x)$ is a proper function that can be trivially solved.

11.4. OOD Multi-Iteration Convergence Rate

Theorem 9. *Under Assumption 3, OOD's convergence rate of K iterations is upper bounded by:*

$$\begin{aligned}&\min_{k=1,\dots,K}F'(x_k+s_k)-F'(x^*+s^*)\\ \leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2\\ &+\frac{L}{K}\sum_{k=1}^{K}(x_k+s_k-x_{k-1}-s_{k-1}+\frac{\nabla f'(x_{k-1}+s_{k-1})+g'_k}{L})^\top(x_k+s_k-x^*-s^*).\end{aligned}$$

Proof. Same as the demonstration for Theorem 5. □

Corollary 10. *Under Assumption 3, L2O model d's (equation 72) convergence rate is upper bounded by w.r.t. $\|s'_{k-1}\|$ by:*

$$\begin{aligned}&\min_{k=1,\dots,K}F'(x_k+s_k)-F'(x^*+s^*)\\ \leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2-\frac{1}{K}\sum_{k=1}^{K}(\nabla f'(x_{k-1}+s_{k-1}))^\top(x_k+s_k-x^*-s^*)\\ &+\frac{L}{K}\sum_{k=1}^{K}\left((\sqrt{n}C_1\|\nabla f'(x_{k-1}+s_{k-1})\|+\sqrt{n}C_2)\|s'_{k-1}\|+\sqrt{n}C_3\|P'_{k-1}\|\right)\|x_k+s_k-x^*-s^*\|.\end{aligned}$$

Proof. First, we rewrite the convergence rate upper bound as the following inequalities:

$$\begin{aligned}&\min_{k=1,\dots,K}F'(x_k+s_k)-F'(x^*+s^*)\\ \leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2+\frac{L}{K}\sum_{k=1}^{K}(x_k+s_k-x_{k-1}-s_{k-1})^\top(x_k+s_k-x^*-s^*)\\ =&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2-\frac{1}{K}\sum_{k=1}^{K}(\nabla f'(x_{k-1}+s_{k-1}))^\top(x_k+s_k-x^*-s^*)\\ &+\frac{L}{K}\sum_{k=1}^{K}(-\operatorname{diag}(\mathbf{J}_1s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)-\mathbf{J}_2s'_{k-1}-\mathbf{J}_3P'_{k-1})^\top(x_k+s_k-x^*-s^*).\end{aligned}$$

Next, we derive its upper bound w.r.t. $\|s'_{k-1}\|$. Cauchy-Schwarz inequality and Triangle inequality yield:

$$
\begin{aligned}
&\min_{k=1,\dots,K} F'(x_k+s_k)-F'(x^*+s^*)\\
\leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2-\frac{1}{K}\sum_{k=1}^{K}(\nabla f'(x_{k-1}+s_{k-1}))^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}(-\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)-\mathbf{J}_2 s'_{k-1}-\mathbf{J}_3 P'_{k-1})^\top(x_k+s_k-x^*-s^*),\\
\leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2-\frac{1}{K}\sum_{k=1}^{K}(\nabla f'(x_{k-1}+s_{k-1}))^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}(\|\operatorname{diag}(\mathbf{J}_1 s'_{k-1})(\nabla f'(x_{k-1}+s_{k-1})+g'_k)\|+\|\mathbf{J}_2 s'_{k-1}\|+\|\mathbf{J}_3 P'_{k-1}\|)\|x_k+s_k-x^*-s^*\|,\\
\leq&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2-\frac{1}{K}\sum_{k=1}^{K}(\nabla f'(x_{k-1}+s_{k-1}))^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}(\sqrt{n}C_1\|s'_{k-1}\|\|\nabla f'(x_{k-1}+s_{k-1})\|+\sqrt{n}C_2\|s'_{k-1}\|+\sqrt{n}C_3\|P'_{k-1}\|)\|x_k+s_k-x^*-s^*\|,\\
=&\frac{L}{2K}\|x_0+s_0-x^*-s^*\|^2-\frac{L}{2K}\|x_K+s_K-x^*-s^*\|^2-\frac{1}{K}\sum_{k=1}^{K}(\nabla f'(x_{k-1}+s_{k-1}))^\top(x_k+s_k-x^*-s^*)\\
&+\frac{L}{K}\sum_{k=1}^{K}\left((\sqrt{n}C_1\|\nabla f'(x_{k-1}+s_{k-1})\|+\sqrt{n}C_2)\|s'_{k-1}\|+\sqrt{n}C_3\|P'_{k-1}\|\right)\|x_k+s_k-x^*-s^*\|.
\end{aligned}
$$

□

12. Details of Experiments

12.1. Implementation Details

Our implementation is conducted with PyTorch based on the open-source code provided by the official implementation of [14] in *https://github.com/xhchrn/MS4L2O*. We follow the settings in [14] to implement our GO-Math-L2O model. We construct a coordinate-wise model where our model takes gradient features according to a variable as an input and generates the update for that coordinate independently on all coordinates.

We implement the learnable parameter matrices $\mathbf{R}$, $\mathbf{Q}$, and $\mathbf{B}$ in Theorem 3 as diagonal matrices. We use neural networks to generate vectors with an identical shape to variable x and use them to conduct the diagonal entries of $\mathbf{R}$, $\mathbf{Q}$, and $\mathbf{B}$. We use different models for $\mathbf{R}$, $\mathbf{Q}$, and $\mathbf{B}$, respectively, both of which take the same inner feature of the former layer. Following L2O-PA in [14], we add an inner linear model between the linear models and the LSTM cell. The complete forward data flow is: the LSTM cell $\rightarrow$ the inner linear model $\rightarrow$ linear models of $\mathbf{R}$, $\mathbf{Q}$, and $\mathbf{B}$.

For LSTM's input features in our GO-Math-L2O, we set it to be smooth gradient $\nabla f(x)$ and two boundaries of non-smooth gradient set, denoted as $\partial r(x)_{\text{lb}}$ and $\partial r(x)_{\text{ub}}$. Thus, the input feature is the concatenation of such three features, $[\nabla f(x)^\top, \partial r(x)_{\text{lb}}^\top, \partial r(x)_{\text{ub}}^\top]^\top$. Moreover, for each block, we normalize input features with the vector norm of the initial point's input feature, i.e., $\|\nabla f(x_0)\|$, $\|\partial r(x_0)_{\text{lb}}\|$, and $\|\partial r(x_0)_{\text{ub}}\|$. At k-th iteration, the input feature is $\left[\frac{\nabla f(x_k)}{\|\nabla f(x_0)\|}^\top, \frac{\partial r(x_k)_{\text{lb}}}{\|\partial r(x_0)_{\text{lb}}\|}^\top, \frac{\partial r(x_k)_{\text{ub}}}{\partial r(x_0)_{\text{ub}}}^\top\right]^\top$.

For L2O-PA [14], the input feature is the concatenation of variable and gradient vectors $[x^\top, \nabla f(x)^\top]^\top$.

For other baseline methods introduced in Sec. 6, we use the implementations provided in the official implementation of [14].

We randomly sample the initial points for all methods. We use deterministic seeds in samplings to ensure reproducibility. Thus, even for non-learning methods, our experimental results are different from those in [14]. However, compared with the origin point set in [14], the random initial point setting is better for robustness evaluation.

12.2. Output Activation

As in Theorem 3, at each iteration, we achieve a symmetric positive definite $\mathbf{R}_k$ by Sigmoid function. The output of the Sigmoid function is in $(0, 1)$. Thus, Frobenius-norm of $\mathbf{R}_k$ is bounded by $\sqrt{n}$, where n is the dimension of variable x. In [14], a larger range is achieved by a direct multiplication with a given constant. We set the constants for $\mathbf{R}_k$, $\mathbf{Q}_k$, and $\mathbf{B}_k$ to be 2, 2, and 1, respectively. Thus, $\|\mathbf{R}_k\| \leq 2\sqrt{n}$, $\|\mathbf{Q}_k\| \leq 2\sqrt{n}$, and $\|\mathbf{B}_k\| \leq \sqrt{n}$.

We follow the activation function setting for LASSO and Logistic regressions in [14]: Sigmoid for LASSO regression and Softplus for Logistic regression. The Sigmoid function is doubled to get $(0, 2)$ range outputs [14]. The activation function is applied on all parameters in Theorem 3.

Following [14], we utilize the objective's smoothness scalar L to shrink the parameters multiplied before gradient, i.e., parameter matrix $\mathbf{R}$ (and $\mathbf{Q}$ for the first two variants in the following section). We set L as the maximal eigenvalue of the Hessian matrix on optimization problem. For *LASSO Regerssion*, L is given by:

$$\|\mathbf{A}\|_2,$$

where $\mathbf{A}$ is the given parameter matrix in objectives defined in Sec. 6.

For *Logistic Regerssion*, L is given by:

$$\left\| \frac{1}{m} \sum_{i=1}^{m} a_i a_i^T h\left(a_i^\top x\right)\left(1 - h\left(a_i^\top x\right)\right) \right\|_2,$$

where h is the sigmoid function and each a_i is a given parameter farture vector defined in Sec. 6. Moreover, since $h\left(a_i^\top x\right)\left(1 - h\left(a_i^\top x\right)\right) \leq 1$, we construct the following upper bound of the above formulation to get a x-unrelated L:

$$\left\| \frac{1}{m} \sum_{i=1}^{m} a_i a_i^T \right\|_2.$$

12.3. Evaluation Metric

Following [14], we use a classical algorithm, named FISTA [5], to generate optimal solutions for both *LASSO Regression* and *Logistic Regression* problems. Based on [14], we run FISTA for 5,000 iterations to ensure the accuracy. Then, all the evaluation solutions are normalized with the optimal solutions by the following equation:

$$\frac{F(x) - F(x^*)}{F(x^*)}.$$

12.4. Gradient Map Ablation

We construct the following three gradient map implementations and select the one with best InD performance. At k-th iteration, the fist one is the standard gradient map (denoted as **STD**) as follows:

$$G_{k-1} = \mathbf{R}_k^{-1}(x_{k-1} - x_k - \mathbf{Q}_k v_{k-1} - b_{1,k}),$$

where G_{k-1} is equivalent to $\nabla f(x_{k-1}) + g_{k-1}$.

Then, we eliminate the minus term for historical information v_{k-1} to let G_{k-1} cover the historical information (denoted as **LH**):

$$G_{k-1} = \mathbf{R}_k^{-1}(x_{k-1} - x_k - b_{1,k}).$$

Moreover, we eliminate $\mathbf{R}_k$ inversion to improve numerical stability (denoted as **LHNoR**):

$$G_{k-1} = x_{k-1} - x_k - b_{1,k}.$$

It is worth noting that such an implementation differs from Math-L2O in [14], where we follow a classical momentum scheme by applying momentum posteriorly. However, Math-L2O use the Nesterov momentum method by adding momentum to the approximation point before the gradient calculation.

The InD results are shown in Figure 5, where **LHNoR** outperforms the other two methods. We use the **LHNoR** version in the following experiments.

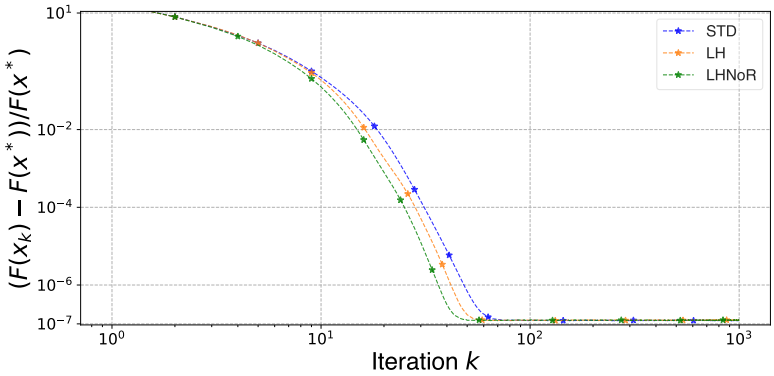

Figure 5. LASSO Regression: Ablation Study on Gradient Map Configurations.

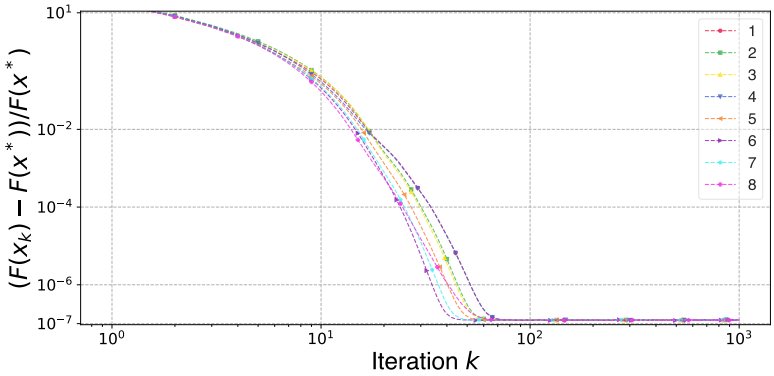

Figure 6. LASSO Regression: Ablation Study on Training Settings, 20/100 BP Frequency.

12.5. Training Configurations

We use Adam [13] as the optimizer to train our model and learning-based baselines. We set the weight decay to zero for all models. For our proposed model, we clamp the norm of the gradient vector to one. However, we do not apply this setting to other baselines since Math-L2O [14] fails to converge at all OOD scenarios. Following [14], we evaluate the in-training model with the evaluation set every 20 iterations.

We evaluate the InD performance of LASSO regression with different training settings in Table 1 to choose the best setting. The "BP Frequency" represents the iterations utilized to achieve one backpropagation and total iterations. For example, 20/100 means backpropagating every 20 iterations in 100 iterations. For training epochs and learning rates, we consider two candidate settings. One epoch and 0.01 is the setting in [14]. We conduct another one with three epochs and a decayed learning rate starting from 0.01. Since our proposed model has more parameters than Math-L2O in [14], we test two different mini-batch settings, 64 and 128, where the 128 mini-batch case has a double training sample that the 64 mini-batch case. Furthermore, we consider alleviating the imbalance problem in a sequence of objective values and design a weighted-sum loss function by the indices of iterations. For a BP length T, given a objective value sequence $x_1, x_2, \ldots, x_T$, the weighted-sum loss is given by:

$$\sum_{i=1}^{T} \frac{i}{\sum_{i=1}^{T} i} F(x_i).$$

In contrast, the mean loss used in [14] is given by:

$$\sum_{i=1}^{T} \frac{1}{T} F(x_i).$$

The results of settings 1 to 8 are shown in Figure 6. The experimental results demonstrate that the best settings for "20/100" BP frequency are settings 7 and 8.

Table 1. Training Settings

Index	BP Frequency	Epochs	Learning Rate	Batch Size	Loss Function
1	20/100	1	0.01	64	Mean
2	20/100	1	0.01	64	Weighted-Sum
3	20/100	1	0.01	128	Mean
4	20/100	1	0.01	128	Weighted-Sum
5	20/100	3	0.01, Decay to 10% Per-Epoch	64	Mean
6	20/100	3	0.01, Decay to 10% Per-Epoch	64	Weighted-Sum
7	20/100	3	0.01, Decay to 10% Per-Epoch	128	Mean
8	20/100	3	0.01, Decay to 10% Per-Epoch	128	Weighted-Sum
9	50/100	1	0.01	64	Mean
10	50/100	1	0.01	64	Weighted-Sum
11	50/100	1	0.01	128	Mean
12	50/100	1	0.01	128	Weighted-Sum
13	50/100	3	0.01, Decay to 10% Per-Epoch	64	Mean
14	50/100	3	0.01, Decay to 10% Per-Epoch	64	Weighted-Sum
15	50/100	3	0.01, Decay to 10% Per-Epoch	128	Mean
16	50/100	3	0.01, Decay to 10% Per-Epoch	128	Weighted-Sum
17	100/100	1	0.01	64	Mean
18	100/100	1	0.01	64	Weighted-Sum
19	100/100	1	0.01	128	Mean
20	100/100	1	0.01	128	Weighted-Sum
21	100/100	3	0.01, Decay to 10% Per-Epoch	64	Mean
22	100/100	3	0.01, Decay to 10% Per-Epoch	64	Weighted-Sum
23	100/100	3	0.01, Decay to 10% Per-Epoch	128	Mean
24	100/100	3	0.01, Decay to 10% Per-Epoch	128	Weighted-Sum

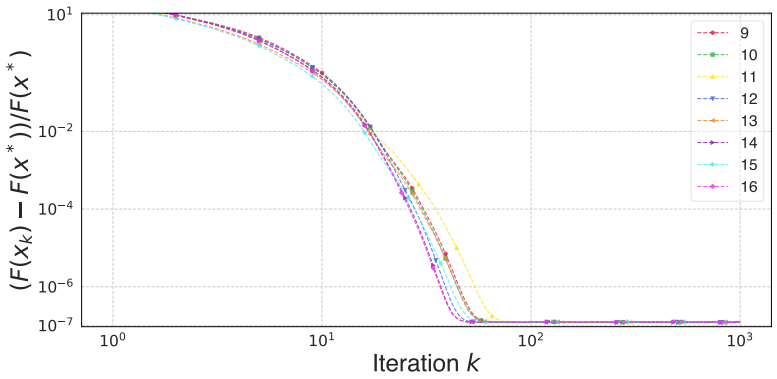

Figure 7. LASSO Regression: Ablation Study on Training Settings, 50/100 BP Frequency.

The results of settings 9 to 16 are shown in Figure 7. The experimental results demonstrate that the best settings for "50/100" BP frequency are settings 14 and 16.

The results of settings 17 to 24 are shown in Figure 8. The experimental results demonstrate that the best setting for "100/100" BP frequency is setting 24.

A further comparison between settings 7, 8, 14, 16, and 24 is illustrated in Figure 9. Based on the result, we conclude that training settings do not dominate the InD performance of our proposed Go-Math-L2O model. We choose setting 7 as our training configuration since we observe that the baseline Math-L2O method fails to converge at all OOD scenarios if we increase the BP frequency to "50/100" or "100/100".

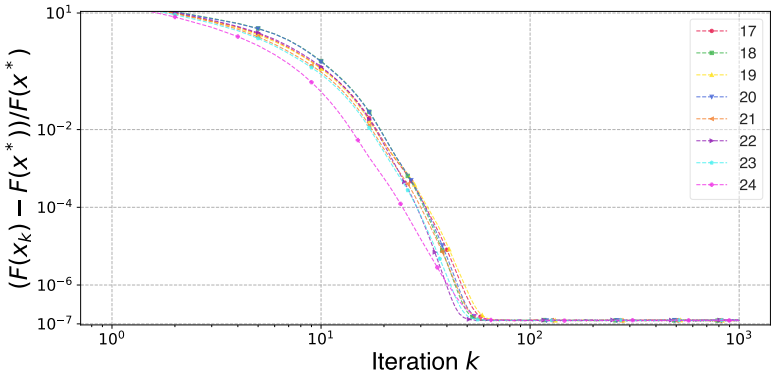

Figure 8. LASSO Regression: Ablation Study on Training Settings, 100/100 BP Frequency.

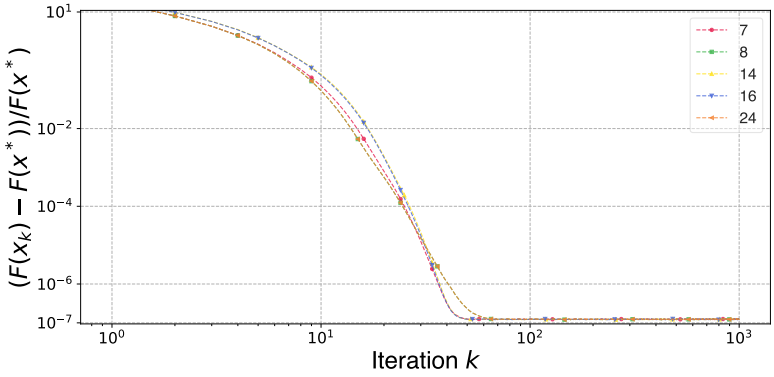

Figure 9. LASSO Regression: Ablation Study on Training Settings, Best.

Table 2. $\mathbf{Q}$-Shrinking Settings

Index	Q Settings
1	$\mathbf{Q}$
2	$\mathbf{Q}/\sqrt{L}$
3	$\mathbf{Q}/L$
4	$\mathbf{Q}/L^2$

12.6. OOD Improvement Configurations

Based on our analysis in Sec. 11.3, by shrinking the range of the parameters in optimizing steep objectives, our proposed model can achieve better robustness than variable-based historical modeling in SOTA Math-L2O [14]. This section evaluates the performance with different extents of parameter shrinking. Specifically, we set the shrinkings on $\mathbf{Q}$ by dividing the smoothness parameter L of smooth objective. Different settings are listed in Table 2.

Since L is calculated individually for each instance by its largest eigenvalue of the Hessian matrix of the objective. Compared with the $\mathbf{Q}$ only version, adding L changes $\mathbf{Q}$'s distribution. Thus, we separately train each setting within Table 2. The InD results of the settings in Table 2 are shown in Figure 10. The illustrated results show that $\mathbf{Q}/L$ and $\mathbf{Q}/L^2$ cause poor InD performance. $\mathbf{Q}/\sqrt{L}$ has a similar InD convergence to $\mathbf{Q}$.

Furthermore, we add an extra experiment to compare their OOD performances, shown in Figures 11 and 12. The results show that the $\mathbf{Q}$ setting outperforms $\mathbf{Q}/\sqrt{L}$ in all initial point OOD scenarios (Figure 11) and achieves better outperformance with larger OOD shiftings. Both methods perform similarly in objective OOD scenarios (Figure 12).

It is worth noting that this result does not violate our theoretical comparison result in Sec. 11.3, where our gradient-only method needs further parameter shrinking strategies to address the deficiency of weaker robustness by larger magnitude in sharp objective cases. Our normalization method on input gradient-only features and our recurrent gradient map setting that eliminates $\mathbf{R}$ inversion have achieved a similar input magnitude to the variable method in [14]. Moreover, in Figure 12,

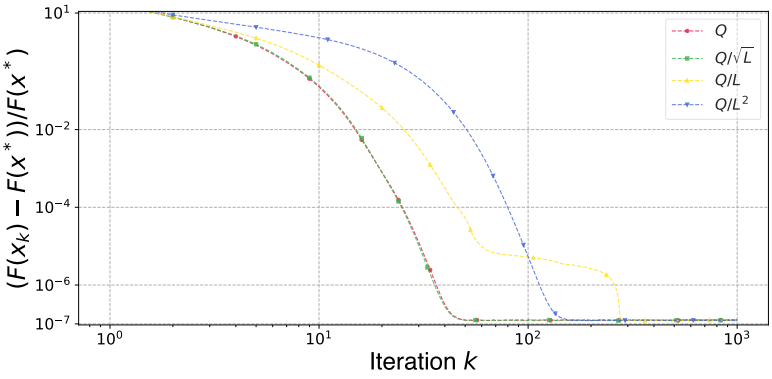

Figure 10. LASSO Regression: Ablation Study on **Q** Settings, InD scenario.

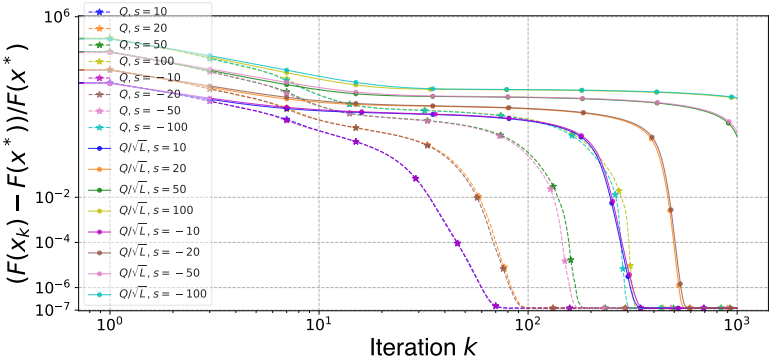

Figure 11. LASSO Regression: Ablation Study on **Q** Settings, OOD by Trigger 1.

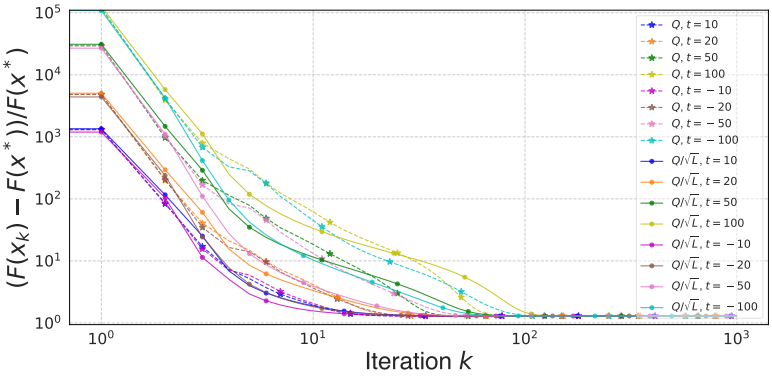

Figure 12. LASSO Regression: Ablation Study on **Q** Settings, OOD by Trigger 2.

$\mathbf{Q}/\sqrt{L}$ setting performs similarly to the **Q**.

12.7. Real-World Evaluation

We further evaluate our model on real-world optimization problems. We follow the methodology proposed in [14] to construct the following real-world datasets:

1) LASSO Regression. 1,000 patches are chosen from the BSDS500 dataset. $\mathbf{A}$ are calculated with K-SVD method and λ is set to be 0.5.
2) Logistic Regression. Ionoshpere dataset contains 4,601 $a_i, b_i \in \mathbb{R}^{34}$ for each sample. Spambase dataset contains 4,601 $a_i, b_i \in \mathbb{R}^{57}$ for each sample.

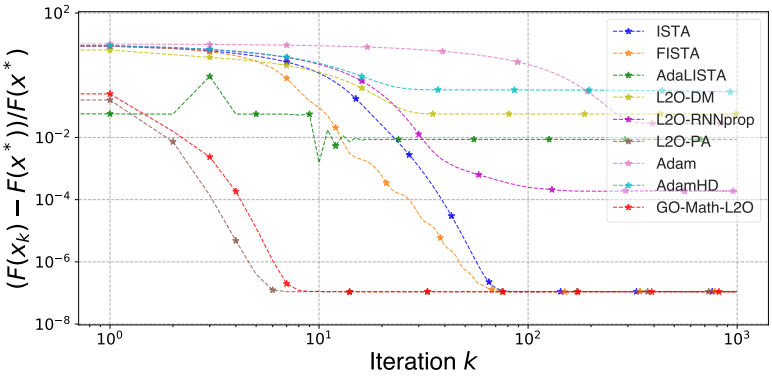

Figure 13. Logistic Regression: InD.

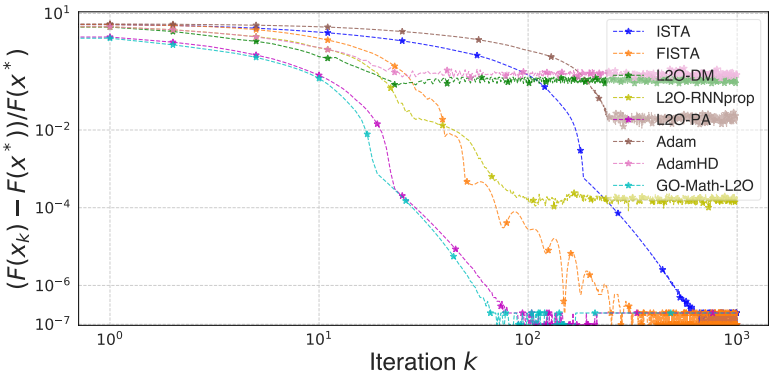

Figure 14. Logistic Regression: Real-World Ionoshpere Dataset.

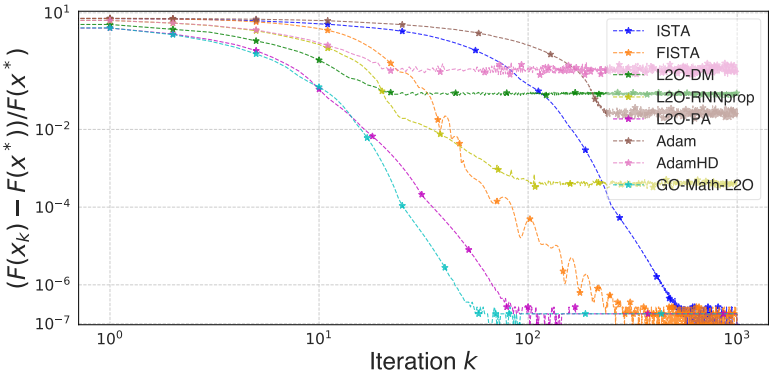

Figure 15. Logistic Regression: Real-World Spambase Dataset.

12.8. Logistic Regression Results

InD comparison is shown in Figure 13. Our proposed Go-Math-L2O performs similarly to L2O-PA and outperforms other baselines.

The OOD comparison on two real-world datasets, Ionoshpere and Spambase, are shown in Figures 14 and 15. Our GO-Math-L2O model outperforms all other baselines.

Figure 16 depicts the OOD scenarios in Logistic regression where the initial point deviates from around zero, i.e., the OOD initial point is $x_0 + s$, where s denotes the extent of the initial point shifting. Under these conditions, our proposed GO-Math-L2O model performs similarly to L2O-PA [14].

Figure 17 presents the results for the OOD scenarios of objective shifting, i.e., the OOD objective is $F'(x) = F(x + t)$,

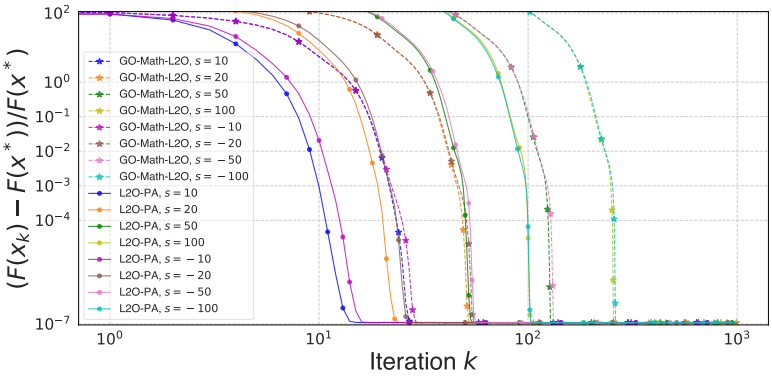

Figure 16. Logistic Regression: OOD by Trigger 1.

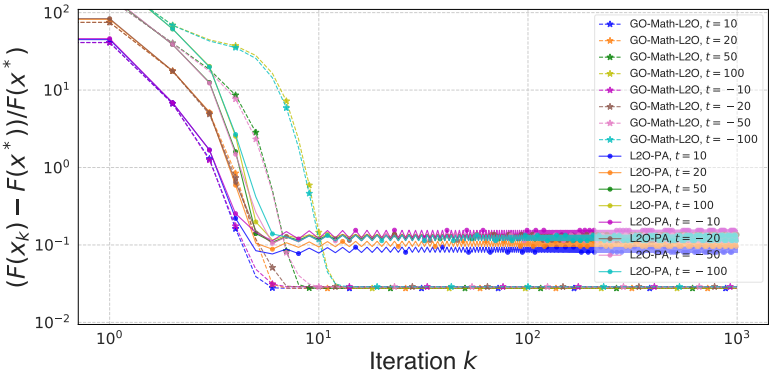

Figure 17. Logistic Regression: OOD by Trigger 2.

where s denotes the extent of initial point shifting. Results in $t = \pm 10, \pm 20$ cases demonstrate that our proposed GO-Math-L2O method converges significantly faster than L2O-PA [14]. For $t = \pm 50, \pm 100$ cases, our model can also converge to better optimums after oscillations. Moreover, the results also show that L2O-PA fails to converge when objective shifts.