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 $-\nabla f(x)^\top \Big(\operatorname{diag}\big(\mathbf{N}_1(z) - L\mathbf{N}_1(z)^2\big)\Big)\nabla f(x) \leq 0, \forall \nabla f(x)$, we should have:

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

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:

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

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 $-\nabla f(x)^\top \Big(\operatorname{diag}\big(\mathbf{N}_1(z)\big) - L\operatorname{diag}\big(\mathbf{N}_1(z)\big)^2\Big)\nabla f(x)$. 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:

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

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:

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

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

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

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

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

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):

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

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

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

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 $\mathbf{N}_2(z)$:

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

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:

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

$-\frac{\|\nabla f'(x+s)\|^2}{2L}$ 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:

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

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

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

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 $f'(x_k + s_k)$:

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

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}:

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

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:

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

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:

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

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:

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

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 $\min_{k=1,\dots,K} F'(x_k + s_k) - F'(x^* + s^*)$ is minimal over $F'(x_j + s_j) - F'(x^* + s^*), j = 1, \dots, K$, 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:

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

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

□

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:

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

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:

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

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:

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

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

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

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:

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

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 $\big(x_{k-1} - \operatorname{diag}\left(\mathbf{N}_1(z_{k-1})\right) \nabla f(x_{k-1}) - \mathbf{N}_2(z_{k-1})\big)$ 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:

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

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.

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

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:

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

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 $G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1})$:

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

where $g_k \in \partial r(x_k)$ is the virtual subgradient of the non-smooth part r of objective. Thus, we have $G_{\operatorname{diag}(\mathbf{N}_1(z_{k-1}))^{-1}}(x_{k-1}) - \nabla f(x_{k-1}) \in \partial r(x_k)$, 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 $r(x)$:

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

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

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

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

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

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

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:

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

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:

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

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

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

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:

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

□

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:*

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

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:

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

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

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

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

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

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:

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

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

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

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:

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

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 $\mathbf{J}_{1,k-1}, \mathbf{J}_{2,k-1}, k = 1, 2, \dots, K$ 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:

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

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:*

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

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

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

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:

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

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:

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

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

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

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

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

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:

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

Over K-iterations, we have:

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

□

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:*

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

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

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

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

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

□

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 $x_1, y_1, x_2, y_2, \dots, x_m, y_m \in \mathbb{R}^n$, there exists matrices $\mathbf{J}_1, \mathbf{J}_2, \dots, \mathbf{J}_m$ that:

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

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

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 $\mathbf{J}_j = \mathbf{N}_j(s), j = 1, 2, \dots, m$, s should follows:*

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

Proof. We prove the above lemma by contradiction. Suppose not, which means there exists a s' that $\exists x_i \notin s'$. Then for $x_j, j = 1, 2, \dots, m$, 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 $\forall x, y \in \mathbb{R}^n$:

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

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:

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

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:

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

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:

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

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:

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

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:

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

where we set $\mathrm{diag}\left(\mathbf{N}_1(z_{k-1})\right) \succ 0$ 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 $\mathrm{diag}(\mathbf{N}_3(z_{k-1}))$ and $\mathrm{diag}(\mathbf{N}_4(w_{k-1}))$ respectively. Moreover, we take explicit subgradient longer horizon modeling of two subgradient values of iteation $k-1$ and $k-2$ by $(\mathbf{I} - \mathrm{diag}(\mathbf{N}_5(r_{k-1})))\partial r(x_{k-1})_{\text{lb}} + \mathrm{diag}(\mathbf{N}_5(r_{k-1}))$ and $(\mathbf{I} - \mathrm{diag}(\mathbf{N}_5(r_{k-2})))\partial r(x_{k-2})_{\text{lb}} + \mathrm{diag}(\mathbf{N}_5(r_{k-2}))$ respectively. We also omit all bias terms since they are demonstrated to vanish along iteration in Sec. 8.8.

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

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

which yields the following proximal operator:

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

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:

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

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:

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

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.

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

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.

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

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:

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

Second perfect square:

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

□

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:*

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

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:

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

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 $\mathbf{N}_2(z_{k-1})$:

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

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

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

which yields:

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

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

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

which yields:

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

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

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

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:

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

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

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

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:

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

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:

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

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:

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

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:

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

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 $\mathbf{J}_{1,k-1}, \mathbf{J}_{2,k-1}, \mathbf{J}_{3,k-1}, k = 1, 2, \dots, K$ 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:

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

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* $\mathbf{J}_{1,k-1}, \mathbf{J}_{2,k-1}, \mathbf{J}_{3,k-1}, \mathbf{J}_{4,k-1}, k = 1, 2, \dots, K$ *that OOD's convergence improvement of one iteration is upper bounded by following inequality:*

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

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:*

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

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:

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

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:

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

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:

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

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:

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

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:*

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

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}\\
&\quad + \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\\
&\quad + 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\\
&\quad + 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\\
&\quad + 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\\
&\quad + 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\\
&\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)\|s'_{k-1}\|^2,\\
&= -\frac{\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2}{2L}\\
&\quad + \Big(Ln^2C_1^2\|\nabla f'(x_{k-1} + s_{k-1}) + g'_k\|^2 + Ln^2C_2^2\\
&\qquad + 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:

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

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:

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

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:

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

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:

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

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:

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

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

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

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:

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

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:*

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

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:*

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

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

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

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 → the inner linear model → 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.

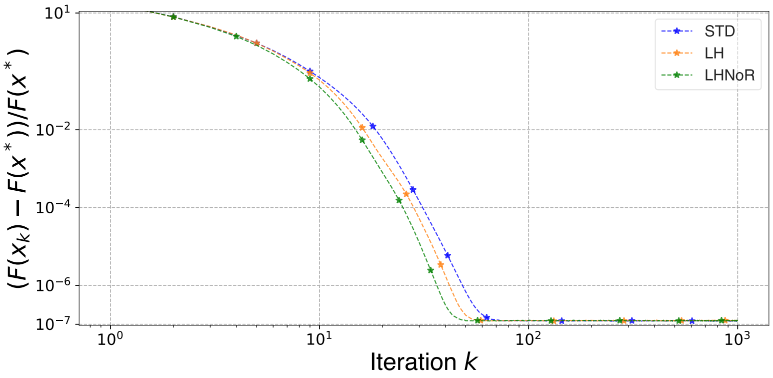

Figure 5. LASSO Regression: Ablation Study on Gradient Map Configurations.

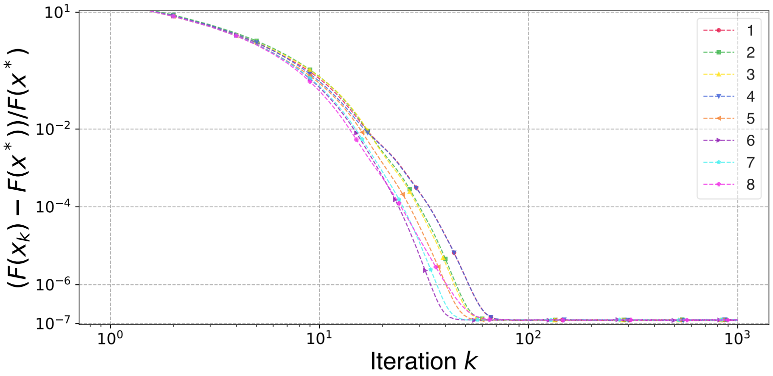

Figure 6. LASSO Regression: Ablation Study on Training Settings, 20/100 BP Frequency.

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 $x_1, x_2, \dots, x_T$, 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

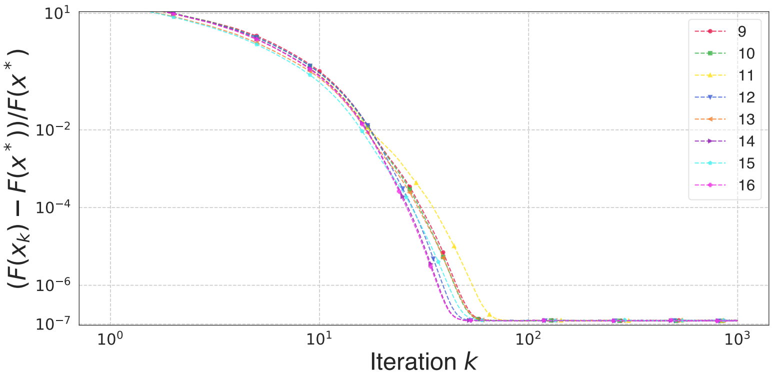

Figure 7. LASSO Regression: Ablation Study on Training Settings, 50/100 BP Frequency.

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”.

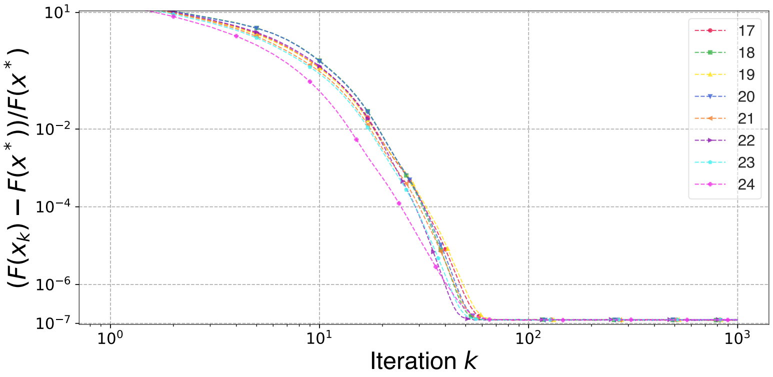

Figure 8. LASSO Regression: Ablation Study on Training Settings, 100/100 BP Frequency.

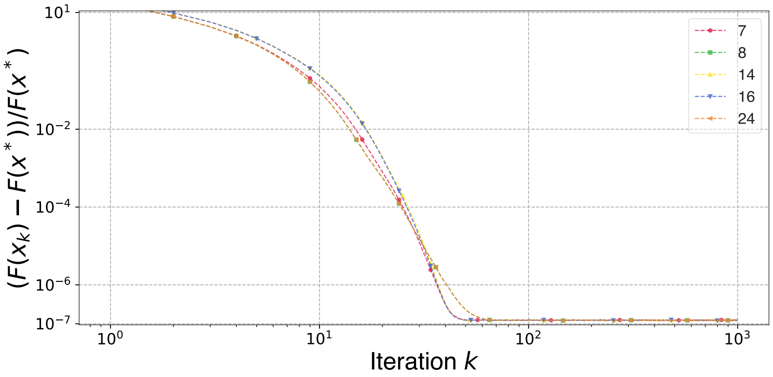

Figure 9. LASSO Regression: Ablation Study on Training Settings, Best.

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,

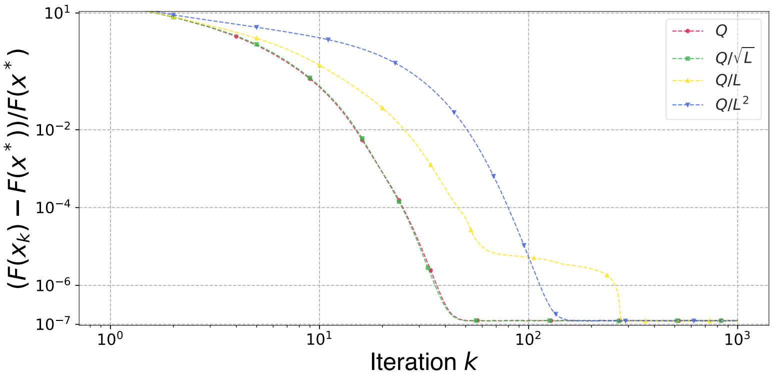

Figure 10. LASSO Regression: Ablation Study on $\mathbf{Q}$ Settings, InD scenario.

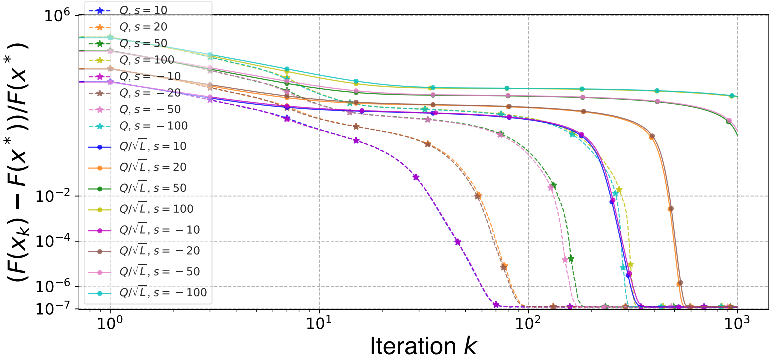

Figure 11. LASSO Regression: Ablation Study on $\mathbf{Q}$ Settings, OOD by Trigger 1.

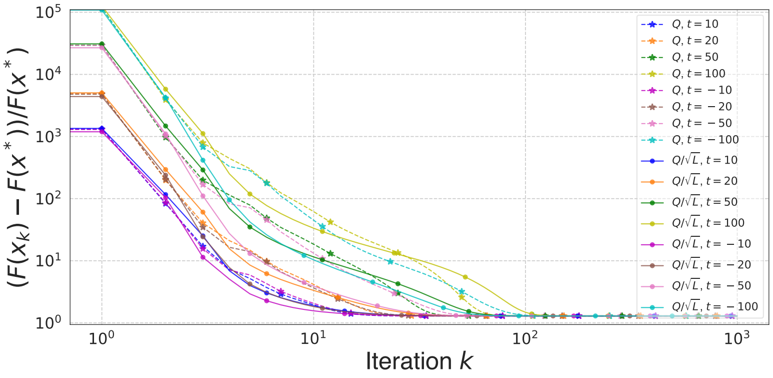

Figure 12. LASSO Regression: Ablation Study on $\mathbf{Q}$ Settings, OOD by Trigger 2.

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

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.

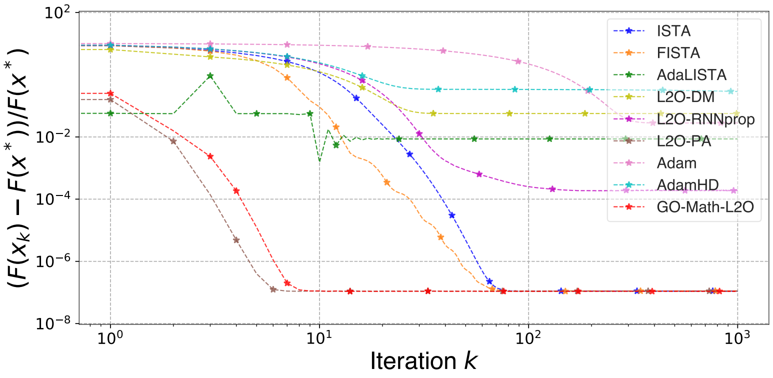

Figure 13. Logistic Regression: InD.

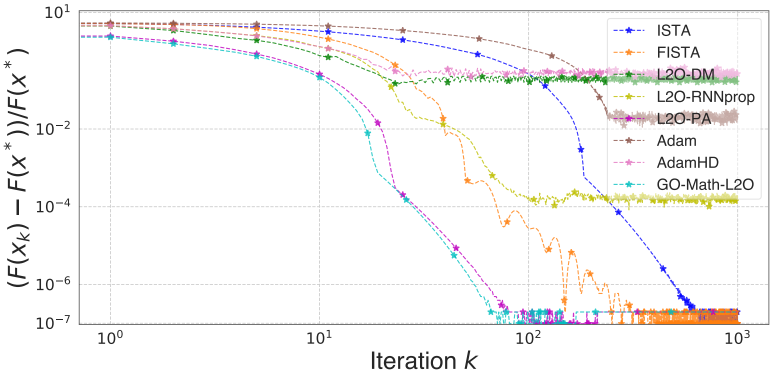

Figure 14. Logistic Regression: Real-World Ionoshpere Dataset.

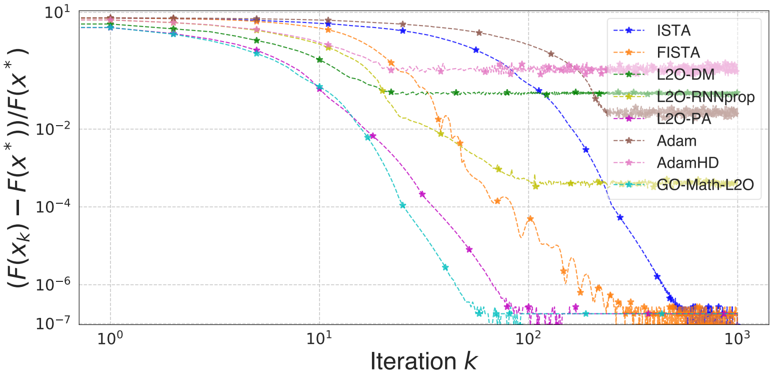

Figure 15. Logistic Regression: Real-World Spambase Dataset.

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 $F'(x) = F(x + t)$,

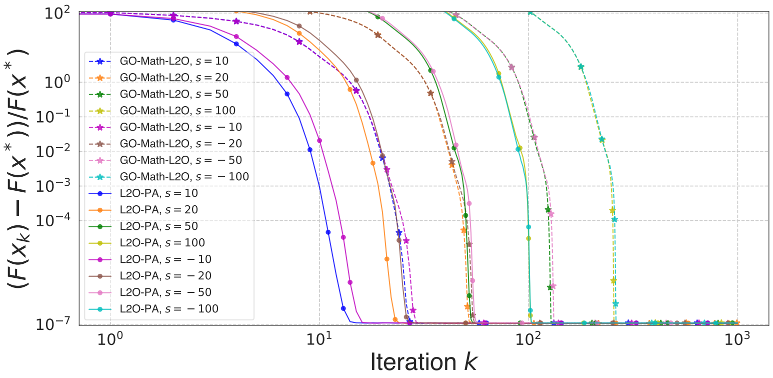

Figure 16. Logistic Regression: OOD by Trigger 1.

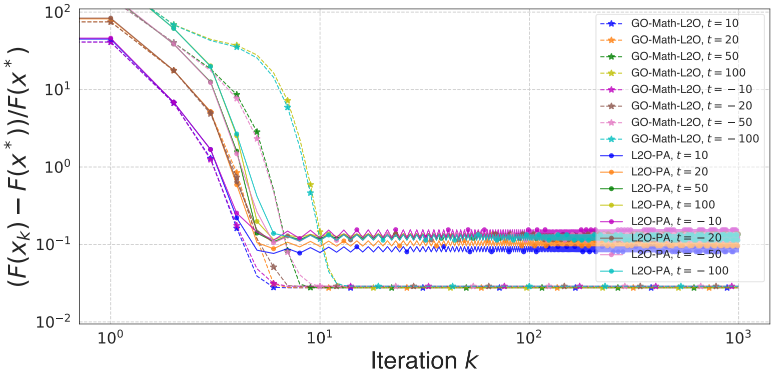

Figure 17. Logistic Regression: OOD by Trigger 2.

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.