	Deterministic	Stochastic
Warm-start	$\alpha_x \frac{2\widetilde{M}M}{m^{val}} \cdot \frac{((1+\alpha_y L)^K+\alpha_x \widetilde{L})^T-1}{(1+\alpha_y L)^K+\alpha_x \widetilde{L}-1}$	$\inf_{0\le t'\le T}\left[\frac{2\alpha_x \widetilde{M}}{m^{val}} \frac{\left((1+\alpha_y L)^K+\alpha_x \widetilde{L}'\right)^{T-t'}-1}{(1+\alpha_y L)^K+\alpha_x \widetilde{L}'-1} + \frac{t'}{m^{val}} s(\ell)\right]$
Cold-start	$\left(1+\sqrt{\frac{L+\mu}{L-\mu}}\right) \frac{2\widetilde{M}M}{\widetilde{L}m^{val}} \cdot \left[(1+\alpha_x \widetilde{L})^T - 1\right]$	$\inf_{0\le t'\le T}\left[\frac{2\widetilde{M}}{m^{val}\widetilde{L}'}\left[(1+\alpha_x \widetilde{L}')^{t-t'}-1\right] + \frac{t'}{m^{val}} s(\ell)\right]$

Table 1. Uniformly stability constant β.

	$\widetilde{M}$	$\widetilde{L}$
ITD	$M\left(1+\frac{L}{\mu}\left(1-(1-\alpha_y\mu)^K\right)\right)$	$\mathcal{O}\left((1-(1-\alpha_y\mu)^K\right)$
AID	$M\left(1+\frac{L}{\mu}\left(1-(1-\alpha_y\mu)^D\right)\right)$	$\mathcal{O}\left((1-(1-\alpha_y\mu)^D\right)$

Table 2. Lipschitz continuity and smoothness properties.

1. Outlines

- The main results is provided in 2.
- The Proof of Theoretical Results
  - The results of Lipschitz continuous and smooth in the hypergradient estimation $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}(\boldsymbol{x}))$ of ITD-based method is provided in 4.2, 4.3 and 4.4.
  - The generalization bound of deterministic HPT algorithm is provided in 4.5.
  - The uniformly stability constant β of deterministic ITD with cold-start is provided in 4.6.
  - The uniformly stability constant β of deterministic ITD with cold-start (random initialization) is provided in 4.6.
  - The uniformly stability constant β of deterministic ITD with warm-start is provided in 4.7.
  - The results of Lipschitz continuous and smooth in the hypergradient estimation $\widehat{\nabla}_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}(\boldsymbol{x}))$ of AID-based method is provided in 4.8, 4.9 and 4.10.
  - The uniformly stability constant β of deterministic AID with cold-start is provided in 4.11.
  - The uniformly stability constant β of deterministic AID with warm-start is provided in 4.12.
  - The generalization bound of stochastic HPT algorithm is provided in 4.13.
  - The uniformly stability constant β of stochastic HPT algorithm with cold-start is provided in 4.14.
  - The uniformly stability constant β of stochastic HPT algorithm with warm-start is provided in 4.15.
- The experiments with neural networks is provided in 5.
- The discussion of the boundedness assumption of the loss function is provided in 6.
- The discussion on the inapplicability of warm-start strategy in meta-learning is provided in 7.

2. Main results

The main Results are presented in Table 1 and Table 2.

The results indicate that whether for AID/ITD, or their stochastic settings, cold-start achieves better generalization than warm-start (since β grows more slowly with T), which can be attributed to the tighter coupling of warm-start with the inner dynamics.

For AID/ITD methods, the key factor is the continuity of the estimated hypergradient—that is, the terms $\widetilde{M}$ and $\widetilde{L}$ (for stochastic setting, $\widetilde{L}' = (1 - 1/m^{val})\widetilde{L}$). Table 2 provides their specific forms (D is the size of terms in the Neumann series).

We find that regardless of whether AID or ITD is used, the specific form of the uniformly stability constant β is not affected by the method itself; rather, these methods influence the Lipschitz continuity and smoothness properties of the hypergradient $\nabla_x f(x, \hat{y}(x))$, which in turn are related to β. In contrast, the strategies (cold/warm-start, stochastic/deterministic) directly affect the particular form of β. In short, different hypergradient estimation methods (AID/ITD) impact β through their

Algorithm 1 ITD-based Bilevel Optimization

1: **Input:** The total number of outer iterations T, the total number of inner iterations K, learning rate $\alpha_{\boldsymbol{x}}$ and $\alpha_{\boldsymbol{y}}$.
2: **Initialize:** $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$.
3: **for** $t = 0$ **to** $T-1$ **do**
4:   # *Step 1: Obtain* $\hat{\boldsymbol{y}}$
5:   Set $\boldsymbol{y}_t^0$:

$$\boldsymbol{y}_t^0 = \begin{cases} \text{Warm-start:}\boldsymbol{y}_{t-1}^K \ \text{ if } t > 0 \text{ and } \boldsymbol{y}_0 \ \text{ otherwise} \\ \text{Cold-start:}\boldsymbol{y}_0 \end{cases}$$

6:   **for** $k = 0$ **to** $K-1$ **do**
7:     $\boldsymbol{y}_t^{k+1}(\boldsymbol{x}_t) = \boldsymbol{y}_t^k - \alpha_{\boldsymbol{y}} \nabla_{\boldsymbol{y}} g(\boldsymbol{x}_t, \boldsymbol{y}_t^k)$
8:   **end for**
9:   Set $\hat{\boldsymbol{y}}(\boldsymbol{x}_t) = \boldsymbol{y}_t^K(\boldsymbol{x}_t)$.
10:   # *Step 2: Update* $\boldsymbol{x}$
11:   $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t))|_{\boldsymbol{x}=\boldsymbol{x}_t}$
12: **end for**

effects on the continuity and smoothness properties of the hypergradient, while different algorithm strategies (cold/warm start, stochastic/deterministic) directly alter the expression of β, ultimately influencing generalization.

3. The Theoretical Results in the Original paper

In this section, we aim to analyze and compare cold-start and warm-start strategies for bilevel optimization from the view of generalization. Before introducing this result, we give some notations and assumptions, which have been widely adopted in current works (Ghadimi & Wang, 2018; Ji et al., 2021). We use $\|\cdot\|$ to denote the l^2-norm, and present two sets of assumptions on objective functions of the outer and inner problems in the form of Eq. (**??**).

Assumption 3.1. Let $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{y})$ denote all parameters. Functions f and g satisfy

a) $f(\boldsymbol{w})$ is M-Lipschitz, i.e., for any $\boldsymbol{w}, \boldsymbol{w}'$,

$$|f(\boldsymbol{w}) - f(\boldsymbol{w}')| \leq M \|\boldsymbol{w} - \boldsymbol{w}'\| .$$

b) $\nabla f(\boldsymbol{w})$ and $\nabla g(\boldsymbol{w})$ are L-Lipschitz, i.e., for any $\boldsymbol{w}, \boldsymbol{w}'$,

$$|\nabla f(\boldsymbol{w}) - \nabla f(\boldsymbol{w}')| \leq L \|\boldsymbol{w} - \boldsymbol{w}'\| ,$$
$$|\nabla g(\boldsymbol{w}) - \nabla g(\boldsymbol{w}')| \leq L \|\boldsymbol{w} - \boldsymbol{w}'\| .$$

c) g are μ-strong-convex w.r.t. $\boldsymbol{y}$, i.e., $\mu I \preceq \nabla^2_{\boldsymbol{y}\boldsymbol{y}} g$.

The following assumption imposes the Lipschitz conditions on such high-order derivatives, as also made in Ghadimi & Wang (2018) and Ji et al. (2021).

Assumption 3.2. Suppose the derivatives $\nabla^2_{12} g(\boldsymbol{w})$ and $\nabla^2_{22} g(\boldsymbol{w})$ are τ- and ρ-Lipschitz, i.e., for any $\boldsymbol{w}, \boldsymbol{w}'$,

a) $\left\|\nabla^2_{12} g(\boldsymbol{w}) - \nabla^2_{12} g(\boldsymbol{w}')\right\| \leq \tau \|\boldsymbol{w} - \boldsymbol{w}'\|$.

b) $\left\|\nabla^2_{22} g(\boldsymbol{w}) - \nabla^2_{22} g(\boldsymbol{w}')\right\| \leq \rho \|\boldsymbol{w} - \boldsymbol{w}'\|$.

We firstly characterize the joint Lipschitz continuity of hypergradient. For warm-start strategy, we could get the following Lemma 3.3.

Lemma 3.3. *Suppose Assumptions 3.1-3.2 hold. Let $\alpha_{\boldsymbol{y}} \leq \frac{1}{L}$, then warm-start strategy for ITD-based algorithm, we have*

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t))\| \leq L_{\hat{f}} \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|\right) ,$$

where

$$\hat{\boldsymbol{y}}(\boldsymbol{x}_t) = \boldsymbol{y}_{t_K}(\boldsymbol{x}_t) = \boldsymbol{y}_{t_0} - \alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1} \nabla_2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i}(\boldsymbol{x}_t)),$$
$$\hat{\boldsymbol{y}}'(\boldsymbol{x}'_t) = \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t) = \boldsymbol{y}'_{t_0} - \alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1} \nabla_2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_i}(\boldsymbol{x}'_t)), \tag{1}$$

and

$$L_{\hat{f}} := \frac{M(\tau\mu + L\rho) + L\mu(L+\mu)}{\mu^2} \left(\frac{\alpha_{\boldsymbol{y}} L}{\sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}} + 1 \right). \tag{2}$$

For cold-start strategy, we can get the following Lemma.

Lemma 3.4. *Suppose Assumptions 3.1-3.2 hold. Let $\alpha_{\boldsymbol{y}} \leq \frac{1}{L}$, then cold-start strategy for ITD-based algorithm, we have*

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t))\| \leq L_{\hat{f}} \|\boldsymbol{x}_t - \boldsymbol{x}'_t\|,$$

where $L_{\hat{f}}$ is defined in Eq. (2).

Lemma 3.5. *Suppose Assumptions 3.1-3.2 hold, then for ITD-based Algorithm 1, we have $\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t))\| \leq \widetilde{M}$, where $\widetilde{M} = \mathcal{O}(1)$ is defined in Eq. (11).*

Next, we can derive a high probability bound for ITD-based warm-start strategy. Fristly, we adopt the definition of uniform stability on validation data, as introduced in Bao et al. (2021), as an analytical tool.

Definition 3.6. A HPT algorithm $\mathbf{A}_{hpt}$ is β-uniformly stable on validation in expectation if for all validation datasets $S^{val}, S'^{val} \in Z^m$ such that S^{val}, S'^{val} differ in at most one sample, we have

$$\forall S^{tr} \in Z^n, \forall z \in Z, \ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z) - \ell(\mathbf{A}_{hpt}(S^{tr}, S'^{val}), z) \leq \beta.$$

If a HPT algorithm is β-uniformly stable on validation, then we have the following generalization bound.

Theorem 3.7. *(Generalization bound of a uniformly stable algorithm). For the given samples $S^{tr} \sim (\mathcal{D}^{tr})^{m^{tr}}$, $S^{val} \sim (\mathcal{D}^{val})^{m^{val}}$ and S^{tr} and S^{val} are independent. Suppose a deterministic HPT algorithm $\mathbf{A}_{hpt}$ is β-uniformly stable on validation and the loss function ℓ is bounded by $s(\ell) \geq 0$, then for all $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), \mathcal{D}^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), S^{val}) \leq \beta + \sqrt{\frac{(2\beta m^{val} + s(\ell))^2 \ln \delta^{-1}}{2m^{val}}}.$$

Then, we can derive the specific form of β for ITD-based cold-start strategy.

Theorem 3.8. *Suppose that Assumptions 3.1-3.2 hold. Let $\alpha_{\boldsymbol{y}} \leq \frac{1}{L}$, then ITD-based cold-start strategy with T-step gradient descent is β-uniformly stable on validation with*

$$\beta = \left(1 + \sqrt{\frac{L+\mu}{L-\mu}}\right) \frac{2\widetilde{M}M}{L_{\hat{f}} m^{val}} \cdot \left[(1 + \alpha_{\boldsymbol{x}} L_{\hat{f}})^T - 1\right], \tag{3}$$

where $L_{\hat{f}}$ is defined in Eq. (2) and $\widetilde{M}$ is defined in Eq. (11).

As a comparison, we give the specific form of β for ITD-based warm-start strategy in Theorem 3.9.

Theorem 3.9. *Suppose that Assumptions 3.1-3.2 hold. Let $\alpha_{\boldsymbol{y}} \leq \frac{1}{L}$, then ITD-based warm-start strategy with T-step gradient descent is β-uniformly stable on validation with*

$$\beta = \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}M}{m^{val}} \cdot \frac{((1 + \alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} L_{\hat{f}})^T - 1}{(1 + \alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} L_{\hat{f}} - 1}, \tag{4}$$

where $L_{\hat{f}}$ is defined in Eq. (2) and $\widetilde{M}$ is defined in Eq. (11).

4. Proofs of Main Theoretical Results

4.1. Useful Lemmas

Lemma 4.1. *(Ji et al., 2021) Suppose Assumptions 3.1-3.2 hold, Let $\alpha \le \frac{1}{L}$. Then for Algorithm 1, we have*

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \boldsymbol{y}^*(\boldsymbol{x}_t))\| \le \left(\frac{L(L+\mu)(1-\alpha\mu)^{\frac{K}{2}}}{\mu} + \frac{2M\,(\tau\mu + L\rho)}{\mu^2}(1-\alpha\mu)^{\frac{K-1}{2}}\right)\Delta + \frac{LM(1-\alpha\mu)^K}{\mu}.$$

and using $\|\boldsymbol{y}_{t_0} - \boldsymbol{y}^*(\boldsymbol{x}_t)\| \le \Delta$.

Lemma 4.2. *Suppose Assumption 3.1 hold, then for Algorithm 1, we have $f(\boldsymbol{x}, \boldsymbol{y}^*(\boldsymbol{x}))$ as a function of $\boldsymbol{x}$ is M_{f^*}-Lipschitz, where $M_{f^*} = M(1+\frac{L}{\mu})$.*

Proof. Firstly, for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^m$, we have

$$\begin{aligned}
|f(\boldsymbol{x}_1, \boldsymbol{y}^*(\boldsymbol{x}_1)) - f(\boldsymbol{x}_2, \boldsymbol{y}^*(\boldsymbol{x}_2))| \le & |f(\boldsymbol{x}_1, \boldsymbol{y}^*(\boldsymbol{x}_1)) - f(\boldsymbol{x}_2, \boldsymbol{y}^*(\boldsymbol{x}_1))| + |f(\boldsymbol{x}_2, \boldsymbol{y}^*(\boldsymbol{x}_1)) - f(\boldsymbol{x}_2, \boldsymbol{y}^*(\boldsymbol{x}_2))| \\
\le & M\left(\|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \|\boldsymbol{y}^*(\boldsymbol{x}_1) - \boldsymbol{y}^*(\boldsymbol{x}_2)\|\right) \\
\le & M(1+\frac{L}{\mu})\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|,
\end{aligned}$$

where the last inequality follows from b) of Lemma 2.2 in Ghadimi & Wang (2018). □

4.2. Proof of Lemma 3.3

Proof. Firstly, combined with the fact that $\nabla_2 g(\boldsymbol{x}, \boldsymbol{y})$ is differentiable w.r.t. $\boldsymbol{x}$, indicates that the inner output $\hat{\boldsymbol{y}}$ is differentiable w.r.t. $\boldsymbol{x}$. Then, based on the chain rule, we have

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) = \nabla_1 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) + \nabla\hat{\boldsymbol{y}}(\boldsymbol{x}_t)\nabla_2 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}). \tag{5}$$

Based on the updates that $\hat{\boldsymbol{y}}(\boldsymbol{x}_t) = \boldsymbol{y}_{t_0} - \alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1} \nabla_2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i}(\boldsymbol{x}_t))$, we have

$$\nabla\hat{\boldsymbol{y}}(\boldsymbol{x}_t) = -\alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1}\left[\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i}) \times \prod_{j=i+1}^{K-1}(I - \alpha_{\boldsymbol{y}}\nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_j}))\right].$$

Then, we have

$$\|\nabla\hat{\boldsymbol{y}}(\boldsymbol{x}_t)\| \le \alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1}\left[\|\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i})\| \cdot \prod_{j=i+1}^{K-1}(1-\alpha_{\boldsymbol{y}}\mu)\right] \le \alpha_{\boldsymbol{y}} L \sum_{i=0}^{K-1}(1-\alpha_{\boldsymbol{y}}\mu)^{K-1-i} = \frac{L}{\mu}(1-(1-\alpha_{\boldsymbol{y}}\mu)^K).$$

Similarly, $\|\nabla\hat{\boldsymbol{y}}'(\boldsymbol{x}_t')\| \le \frac{L}{\mu}(1-(1-\alpha_{\boldsymbol{y}}\mu)^K)$. Next, using eq. (20) and the triangle inequality, we have

$$\begin{aligned}
& \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'))\| \\
\le & \|\nabla_1 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_1 f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'))\| + \|\nabla\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla\hat{\boldsymbol{y}}'(\boldsymbol{x}_t')\|\|\nabla_2 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}})\| + \|\nabla\hat{\boldsymbol{y}}'(\boldsymbol{x}_t')\|\|\nabla_2 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}) - \nabla_2 f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}')\| \\
\le & L(\|\boldsymbol{x}_t - \boldsymbol{x}_t'\| + \|\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \hat{\boldsymbol{y}}'(\boldsymbol{x}_t')\|) + M\|\nabla\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla\hat{\boldsymbol{y}}'(\boldsymbol{x}_t')\| + \frac{L^2}{\mu}(1-(1-\alpha_{\boldsymbol{y}}\mu)^K) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}_t'\| + \|\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \hat{\boldsymbol{y}}'(\boldsymbol{x}_t)\|).
\end{aligned}$$

For $\nabla\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla\hat{\boldsymbol{y}}'(\boldsymbol{x}_t')$, we have

$$\begin{aligned}
& \nabla\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla\hat{\boldsymbol{y}}'(\boldsymbol{x}_t') = \nabla\boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \nabla\boldsymbol{y}_{t_K}'(\boldsymbol{x}_t') \\
= & \nabla\boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla\boldsymbol{y}_{t_{K-1}}'(\boldsymbol{x}_t') - \alpha_{\boldsymbol{y}}\left(\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) - \nabla_{12}^2 g(\boldsymbol{x}_t', \boldsymbol{y}_{t_{K-1}}')\right) - \alpha_{\boldsymbol{y}}\left(\nabla\boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla\boldsymbol{y}_{t_{K-1}}'(\boldsymbol{x}_t')\right)\nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) \\
& + \alpha_{\boldsymbol{y}}\nabla\boldsymbol{y}_{t_{K-1}}'(\boldsymbol{x}_t')\left(\nabla_{22}^2 g(\boldsymbol{x}_t', \boldsymbol{y}_{t_{K-1}}') - \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}})\right).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) - \alpha_{\boldsymbol{y}} \left(\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\right) \nabla^2_{22} g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}})\| \\
\leq &\|I - \alpha_{\boldsymbol{y}} \nabla^2_{22} g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}})\| \|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\|,
\end{aligned}$$

and

$$\begin{aligned}
&\| - \alpha_{\boldsymbol{y}} \left(\nabla^2_{12} g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) - \nabla^2_{12} g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}})\right) + \alpha_{\boldsymbol{y}} \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) \left(\nabla^2_{22} g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}}) - \nabla^2_{22} g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}})\right) \| \\
\leq &\alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu}(1 - (1 - \alpha_{\boldsymbol{y}}\mu)^{K-1})\right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\|\right).
\end{aligned}$$

Then, we have

$$\begin{aligned}
\|\nabla \boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t)\| \leq &(1 - \alpha_{\boldsymbol{y}}\mu) \|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\| \\
&+ \alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu}(1 - (1 - \alpha_{\boldsymbol{y}}\mu)^{K-1})\right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\|\right) \\
\leq &(1 - \alpha_{\boldsymbol{y}}\mu) \|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\| + \alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu}\right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\|\right).
\end{aligned}$$

As for $\|\boldsymbol{x}_t - \boldsymbol{x}'_t\|$ and $\|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\|$, according to the conclusion of Lemma3 in Bao et al. (2021), we have

$$\|\boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t)\| \leq L^G_1 \frac{(L^G_2)^K - 1}{L^G_2 - 1} \|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + (L^G_2)^K \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\| \leq \frac{L^G_1}{1 - L^G_2} \|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + L^G_2 \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|, \tag{6}$$

where $L^G_2 := \sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}$, and L^G_1 means the function $G(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y} - \alpha_{\boldsymbol{y}} \nabla_2 g(\boldsymbol{x}, \boldsymbol{y})$ is L^G_1-Lipschitz w.r.t $\boldsymbol{x}$. So we let $L^G_1 = \alpha_{\boldsymbol{y}} L$ cause

$$\|G(\boldsymbol{x}, \boldsymbol{y}) - G(\boldsymbol{x}', \boldsymbol{y})\| = \alpha_{\boldsymbol{y}} \|\nabla_2 g(\boldsymbol{x}, \boldsymbol{y}) - \nabla_2 g(\boldsymbol{x}', \boldsymbol{y})\| \leq \alpha_{\boldsymbol{y}} L \|\boldsymbol{x} - \boldsymbol{x}'\|.$$

Then we have

$$\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \hat{\boldsymbol{y}}'(\boldsymbol{x}_t)\| \leq \left(\frac{\alpha_{\boldsymbol{y}} L}{1 - \sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}} + 1\right) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|),$$

$$\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\| \leq \left(\frac{\alpha_{\boldsymbol{y}} L}{1 - \sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}} + 1\right) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|).$$

Then, we have

$$\begin{aligned}
&\|\nabla \boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t)\| \\
\leq &(1 - \alpha_{\boldsymbol{y}}\mu) \|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\| + \alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu}\right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\|\right) \\
\leq &\left(\frac{\tau}{\mu} + \frac{L\rho}{\mu^2}\right) \left(\frac{\alpha_{\boldsymbol{y}} L}{1 - \sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}} + 1\right) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|)
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t))\| &\leq \frac{M(\tau\mu + L\rho) + L\mu(L + \mu)}{\mu^2} \left(\frac{\alpha_{\boldsymbol{y}} L}{\sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}} + 1\right) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|) \\
&\approx \mathcal{O}(1) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|). \qquad (7)
\end{aligned}$$

Then, the proof is completed. □

4.3. Proof of Lemma 3.4

Proof. Firstly, combined with the fact that $\nabla_2 g(\boldsymbol{x}, \boldsymbol{y})$ is differentiable w.r.t. $\boldsymbol{x}$, indicates that the inner output $\hat{\boldsymbol{y}}$ is differentiable w.r.t. $\boldsymbol{x}$. Then, based on the chain rule, we have

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) = \nabla_1 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) + \nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) \nabla_2 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}). \tag{8}$$

Based on the updates that $\hat{\boldsymbol{y}}(\boldsymbol{x}_t) = \boldsymbol{y}_{t_0} - \alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1} \nabla_2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i}(\boldsymbol{x}_t))$, we have

$$\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) = -\alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1} \left[\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i}) \times \prod_{j=i+1}^{K-1} (I - \alpha_{\boldsymbol{y}} \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_j})) \right].$$

Then, we have

$$\|\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t)\| \leq \alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1} \left[\|\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i})\| \cdot \prod_{j=i+1}^{K-1} (1 - \alpha_{\boldsymbol{y}} \mu) \right] \leq \alpha_{\boldsymbol{y}} L \sum_{i=0}^{K-1} (1 - \alpha_{\boldsymbol{y}} \mu)^{K-1-i} = \frac{L}{\mu} (1 - (1 - \alpha_{\boldsymbol{y}} \mu)^K).$$

Similarly, $\|\nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\| \leq \frac{L}{\mu}(1 - (1 - \alpha_{\boldsymbol{y}}\mu)^K)$. Next, using eq. (8) and the triangle inequality, we have

$$\begin{aligned}
&\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t))\| \\
\leq &\|\nabla_1 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_1 f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t))\| + \|\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\| \|\nabla_2 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}})\| + \|\nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\| \|\nabla_2 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}) - \nabla_2 f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}')\| \\
\leq &L(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\|) + M \|\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\| + \frac{L^2}{\mu}(1 - (1 - \alpha_{\boldsymbol{y}}\mu)^K) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \hat{\boldsymbol{y}}'(\boldsymbol{x}_t)\|).
\end{aligned}$$

For $\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)$, we have

$$\begin{aligned}
&\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t) = \nabla \boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t) \\
= &\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) - \alpha_{\boldsymbol{y}} \left(\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) - \nabla_{12}^2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}}) \right) - \alpha_{\boldsymbol{y}} \left(\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) \right) \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) \\
&+ \alpha_{\boldsymbol{y}} \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) \left(\nabla_{22}^2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}}) - \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) \right).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) - \alpha_{\boldsymbol{y}} \left(\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) \right) \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}})\| \\
\leq &\|I - \alpha_{\boldsymbol{y}} \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}})\| \|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\|,
\end{aligned}$$

and

$$\begin{aligned}
&\| - \alpha_{\boldsymbol{y}} \left(\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) - \nabla_{12}^2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}}) \right) + \alpha_{\boldsymbol{y}} \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) \left(\nabla_{22}^2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}}) - \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) \right) \| \\
\leq &\alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu} (1 - (1 - \alpha_{\boldsymbol{y}}\mu)^{K-1}) \right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\| \right).
\end{aligned}$$

Then, we have

$$\begin{aligned}
\|\nabla \boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t)\| \leq &(1 - \alpha_{\boldsymbol{y}}\mu) \|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\| \\
&+ \alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu} (1 - (1 - \alpha_{\boldsymbol{y}}\mu)^{K-1}) \right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\| \right) \\
\leq &(1 - \alpha_{\boldsymbol{y}}\mu) \|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\| + \alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu} \right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\| \right)
\end{aligned}$$

As for $\|\boldsymbol{x}_t - \boldsymbol{x}'_t\|$ and $\|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\|$, according to the conclusion of Lemma3 in Bao et al. (2021), we have

$$\|\boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t)\| \leq L_1^G \frac{(L_2^G)^K - 1}{L_2^G - 1} \|\boldsymbol{x}_t - \boldsymbol{x}'_t\| \leq \frac{L_1^G}{1 - L_2^G} \|\boldsymbol{x}_t - \boldsymbol{x}'_t\|, \tag{9}$$

where $L_2^G := \sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}$, and L_1^G means the function $G(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y} - \alpha_{\boldsymbol{y}} \nabla_2 g(\boldsymbol{x}, \boldsymbol{y})$ is L_1^G-Lipschitz w.r.t $\boldsymbol{x}$. So we let $L_1^G = \alpha_{\boldsymbol{y}} L$ cause

$$\|G(\boldsymbol{x}, \boldsymbol{y}) - G(\boldsymbol{x}', \boldsymbol{y})\| = \alpha_{\boldsymbol{y}} \|\nabla_2 g(\boldsymbol{x}, \boldsymbol{y}) - \nabla_2 g(\boldsymbol{x}', \boldsymbol{y})\| \le \alpha_{\boldsymbol{y}} L \|\boldsymbol{x} - \boldsymbol{x}'\|.$$

Then, we have

$$\begin{aligned}
&\|\nabla \boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t)\| \\
\le &(1 - \alpha_{\boldsymbol{y}}\mu)\|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\| + \alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu}\right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\|\right) \\
\le &(1 - \alpha_{\boldsymbol{y}}\mu)\|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\| + \alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu}\right) \left(\frac{\alpha_{\boldsymbol{y}} L}{\sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}} + 1\right) \cdot \|\boldsymbol{x}_t - \boldsymbol{x}'_t\| \\
\le &\left(\frac{\tau}{\mu} + \frac{L\rho}{\mu^2}\right) \left(\frac{\alpha_{\boldsymbol{y}} L}{\sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}} + 1\right) \cdot \|\boldsymbol{x}_t - \boldsymbol{x}'_t\|.
\end{aligned}$$

Thus, we have

$$\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t))\| \le \frac{M(\tau\mu + L\rho) + L\mu(L+\mu)}{\mu^2} \left(\frac{\alpha_{\boldsymbol{y}} L}{\sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}} + 1\right) \cdot \|\boldsymbol{x}_t - \boldsymbol{x}'_t\| \approx \mathcal{O}(1)\|\boldsymbol{x}_t - \boldsymbol{x}'_t\|. \tag{10}$$

Then, the proof is completed. □

4.4. Proof of Lemma 3.5

Proof. According to Lemma 4.1, we have

$$\begin{aligned}
\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t))\| &\le \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \boldsymbol{y}^*(\boldsymbol{x}_t))\| + \Big(\frac{L(L+\mu)(1-\alpha\mu)^{\frac{K}{2}}}{\mu} + \frac{2M\,(\tau\mu + L\rho)}{\mu^2}(1-\alpha\mu)^{\frac{K-1}{2}}\Big)\Delta + \frac{LM(1-\alpha\mu)^K}{\mu} \\
&\le M(1 + \frac{2L}{\mu}) + \Big(\frac{L\mu(L+\mu) + 2M\,(\tau\mu + L\rho)}{\mu^2}\Big)\Delta + \frac{LM}{\mu} := \widetilde{M} \approx \mathcal{O}(1).
\end{aligned} \tag{11}$$

Combined with Lemma 4.2, this proof is completed. □

4.5. Proof of Theorem 3.7

Proof. Let $\Phi(S^{tr}, S^{val}) = \ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), \mathcal{D}^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), S^{val})$. Suppose $S^{val}, S^{'val}$ differ in at most one point, then

$$\begin{aligned}
&|\Phi(S^{tr}, S^{val}) - \Phi(S^{tr}, S^{'val})| \\
\le &|\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), \mathcal{D}^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{'val}), \mathcal{D}^{val})| + |\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), S^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{'val}), S^{'val})|.
\end{aligned}$$

For the first term,

$$|\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), \mathcal{D}^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{'val}), \mathcal{D}^{val})| = |\mathbb{E}_{z\sim\mathcal{D}^{val}} \left[\ell(\mathbf{A}(S^{tr}, S^{val}), z) - \ell(\mathbf{A}(S^{tr}, S^{'val}), z)\right]| \le \beta.$$

For the second term,

$$\begin{aligned}
|\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), S^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{'val}), S^{'val})| \le &\frac{1}{m^{val}} \sum_{i=1}^{m^{val}} |\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z_i^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{'val}), z_i^{'val})| \\
\le &\frac{s(\ell)}{m^{val}} + \frac{m^{val} - 1}{m^{val}}\beta.
\end{aligned}$$

As a result,

$$|\Phi(S^{tr}, S^{val}) - \Phi(S^{tr}, S^{'val})| \leq \frac{s(\ell)}{m^{val}} + 2\beta.$$

According to McDiarmid's inequality, we have for all $\epsilon \in \mathbb{R}^+$,

$$P_{S^{val}\sim(\mathcal{D}^{val})^{m^{val}}}(\Phi(S^{tr}, S^{val}) - \mathbb{E}_{S^{val}\sim(\mathcal{D}^{val})^{m^{val}}}\left[\Phi(S^{tr}, S^{val})\right] \geq \epsilon) \leq \exp(-2\frac{m^{val}\epsilon^2}{(s(\ell)+2m^{val}\beta)^2}).$$

Besides, we have

$$\begin{aligned}
&\mathbb{E}_{S^{val}\sim(\mathcal{D}^{val})^{m^{val}}}\left[\Phi(S^{tr}, S^{val})\right] = \mathbb{E}_{S^{val}\sim(\mathcal{D}^{val})^{m^{val}}}\left[\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), \mathcal{D}^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), S^{val})\right] \\
=&\mathbb{E}_{S^{val}\sim(\mathcal{D}^{val})^{m^{val}}, z\sim\mathcal{D}^{val}}\left[\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z_1^{val})\right] \\
=&\mathbb{E}_{S^{val}\sim(\mathcal{D}^{val})^{m^{val}}, z\sim\mathcal{D}^{val}}\left[\ell(\mathbf{A}_{hpt}(S^{tr}, z, z_2^{val}, \cdots, z_{m^{val}}^{val}), z_1^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z_1^{val})\right] \leq \beta.
\end{aligned}$$

Thereby, we have for all $\epsilon \in \mathbb{R}^+$,

$$P_{S^{val}\sim(\mathcal{D}^{val})^{m^{val}}}(\Phi(S^{tr}, S^{val}) - \beta \geq \epsilon) \leq \exp(-2\frac{m^{val}\epsilon^2}{(s(\ell)+2m^{val}\beta)^2}).$$

Notice the above inequality holds for all S^{tr}, we further have $\epsilon \in \mathbb{R}^+$,

$$P_{S^{tr}\sim(D^{tr})^{m^{tr}}, S^{val}\sim(\mathcal{D}^{val})^{m^{val}}}(\Phi(S^{tr}, S^{val}) - \beta \geq \epsilon) \leq \exp(-2\frac{m^{val}\epsilon^2}{(s(\ell)+2m^{val}\beta)^2}).$$

Equivalently, we have $\forall\delta \in (0, 1)$,

$$P_{S^{tr}\sim(D^{tr})^{m^{tr}}, S^{val}\sim(\mathcal{D}^{val})^{m^{val}}}\left(\Phi(S^{tr}, S^{val}) \leq \beta + \sqrt{\frac{(2\beta m^{val} + s(\ell))^2 \ln\delta^{-1}}{2m^{val}}}\right) \geq 1 - \delta.$$

Then, the proof is completed. $\square$

4.6. Proof of Theorem 3.8

Proof. We use the following equation to denote the updating rule in the outer level,

$$\Upsilon(\boldsymbol{x}_t, S^{val}) = \boldsymbol{x}_t - \alpha_{\boldsymbol{x}}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}),$$

where $\hat{\boldsymbol{y}}(\boldsymbol{x}_t) = \boldsymbol{y}_{t_0} - \alpha_{\boldsymbol{y}}\sum_{i=0}^{K-1}\nabla_2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i}; S^{tr})$.

We suppose S^{val} and $S^{'val}$ are different in at most one sample point, and let $\{\boldsymbol{x}_t\}_{t\geq 0}$ and $\{\boldsymbol{x}'_t\}_{t\geq 0}$ be the trace by gradient descent with S^{val} and $S^{'val}$ respectively. Let $\delta_t = \|\boldsymbol{x}_t - \boldsymbol{x}'_t\|$, and then

$$\begin{aligned}
\delta_{t+1} &= \|\Upsilon(\boldsymbol{x}_t, S^{val}) - \Upsilon(\boldsymbol{x}'_t, S^{'val})\| \\
&= \|\boldsymbol{x}_t - \alpha_{\boldsymbol{x}}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \boldsymbol{x}'_t + \alpha_{\boldsymbol{x}}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{'val})\| \\
&\leq \|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \alpha_{\boldsymbol{x}}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{'val})\|,
\end{aligned}$$

where $\hat{\boldsymbol{y}'}(\boldsymbol{x}'_t) = \boldsymbol{y}'_{t_0} - \alpha_{\boldsymbol{y}}\sum_{i=0}^{K-1}\nabla_2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_i}; S^{tr})$. $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val})$ and $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{'val})$ denote $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}(\boldsymbol{x}); S^{val})|_{\boldsymbol{x}=\boldsymbol{x}_t}$ and $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}'(\boldsymbol{x}); S^{'val})|_{\boldsymbol{x}=\boldsymbol{x}'_t}$ respectively.

Next, we have

$$\begin{aligned}
&\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{'val})\| \\
\leq&\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{val})\| + \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{'val})\| \\
\leq&\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{val})\| + \frac{2\widetilde{M}}{m^{val}} \\
\leq&L_{\hat{f}}\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \frac{2\widetilde{M}}{m^{val}},
\end{aligned}$$

Therefore, we have

$$\delta_{t+1} \leq \delta_t + \alpha_{\boldsymbol{x}} L_{\hat{f}} \delta_t + \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}}{m^{val}} = (1 + \alpha_{\boldsymbol{x}} L_{\hat{f}})\delta_t + \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}}{m^{val}}, \tag{12}$$

Then we have $\delta_t \leq \frac{2\widetilde{M}}{m^{val}} \cdot \frac{(1+\alpha_{\boldsymbol{x}} L_{\hat{f}})^t - 1}{L_{\hat{f}}}$. Thus, we have

$$\begin{aligned} &|\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{'val}), z)| := |f(\boldsymbol{x}_T, \boldsymbol{y}_T(\boldsymbol{x}_{T-1}); z) - f(\boldsymbol{x}'_T, \boldsymbol{y}'_T(\boldsymbol{x}'_{T-1}); z)| \\ \leq & M \left(1 + \sqrt{\frac{L+\mu}{L-\mu}}\right) \delta_T \leq \left(1 + \sqrt{\frac{L+\mu}{L-\mu}}\right) \frac{2\widetilde{M}M}{L_{\hat{f}} m^{val}} \cdot \left[(1 + \alpha_{\boldsymbol{x}} L_{\hat{f}})^T - 1\right]. \end{aligned}$$

where utilizes Eq. (9). Then, the proof is completed. □

Corollary 4.3. *Under the same assumptions with Theorem 3.8, and for the cold-start with random initialization, we assume* $\forall\, \boldsymbol{y}_1, \boldsymbol{y}_2 \sim \mathcal{D}_{\boldsymbol{y}}$, *we have* $\|\boldsymbol{y}_1 - \boldsymbol{y}_2\| \leq a$. *Thus we have the uniformly stable constant is*

$$\beta = \left(1 + \sqrt{\frac{L+\mu}{L-\mu}}\right) \cdot \left(\frac{2\widetilde{M}M}{L_{\hat{f}} m^{val}} + a\right) \cdot \left[(1 + \alpha_{\boldsymbol{x}} L_{\hat{f}})^T - 1\right].$$

4.7. Proof of Theorem 3.9

Proof. We use the following equation to denote the updating rule in the outer level,

$$\Upsilon(\boldsymbol{x}_t, S^{val}) = \boldsymbol{x}_t - \alpha_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}),$$

where $\hat{\boldsymbol{y}}(\boldsymbol{x}_t) = \boldsymbol{y}_{t_0} - \alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1} \nabla_2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i}; S^{tr})$.

We suppose S^{val} and $S^{'val}$ are different in at most one sample point, and let $\{\boldsymbol{x}_t\}_{t\geq 0}$ and $\{\boldsymbol{x}'_t\}_{t\geq 0}$ be the trace by gradient descent with S^{val} and $S^{'val}$ respectively. Let $\delta_t = \|\boldsymbol{x}_t - \boldsymbol{x}'_t\|$, and then

$$\begin{aligned} \delta_{t+1} &= \|\Upsilon(\boldsymbol{x}_t, S^{val}) - \Upsilon(\boldsymbol{x}'_t, S^{'val})\| \\ &= \|\boldsymbol{x}_t - \alpha_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \boldsymbol{x}'_t + \alpha_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{'val})\| \\ &\leq \|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \alpha_{\boldsymbol{x}} \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{'val})\|, \end{aligned}$$

where $\hat{\boldsymbol{y}}'(\boldsymbol{x}'_t) = \boldsymbol{y}'_{t_0} - \alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1} \nabla_2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_i}; S^{tr})$. $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val})$ and $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{'val})$ denote $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}(\boldsymbol{x}); S^{val})|_{\boldsymbol{x}=\boldsymbol{x}_t}$ and $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}'(\boldsymbol{x}); S^{'val})|_{\boldsymbol{x}=\boldsymbol{x}'_t}$ respectively.

Next, we have

$$\begin{aligned} &\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{'val})\| \\ \leq & \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{val})\| + \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{'val})\| \\ \leq & \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t); S^{val})\| + \frac{2\widetilde{M}}{m^{val}} \\ \leq & L_{\hat{f}} \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|\right) + \frac{2\widetilde{M}}{m^{val}}, \end{aligned}$$

and we rewrite $\|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|$ as ζ_{t_0}. Therefore, we have

$$\delta_{t+1} \leq \delta_t + \alpha_{\boldsymbol{x}} L_{\hat{f}} \delta_t + \alpha_{\boldsymbol{x}} L_{\hat{f}} \zeta_{t_0} + \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}}{m^{val}} = (1 + \alpha_{\boldsymbol{x}} L_{\hat{f}})\delta_t + \alpha_{\boldsymbol{x}} L_{\hat{f}} \zeta_{t_0} + \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}}{m^{val}}, \tag{13}$$

and

$$\zeta_{t_K} \leq \zeta_{t_{K-1}} + \alpha_{\boldsymbol{y}} L \zeta_{t_{K-1}} + \alpha_{\boldsymbol{y}} L \delta_t = (1 + \alpha_{\boldsymbol{y}} L)^K \zeta_{t_0} + \left((1 + \alpha_{\boldsymbol{y}} L)^K - 1\right) \delta_t. \tag{14}$$

To obtain the upper bounds for δ_t and ζ_t, directly substituting Eq. (27) into Eq. (26) or vice versa makes it challenging to derive their respective upper bounds. Therefore, we further analyze the upper bound of $\delta_{t+1} + \zeta_{t_K}$. By Combining Eqs. (26) and (27), we obtain:

$$\begin{aligned}\delta_{t+1} + \zeta_{t_K} = \delta_{t+1} + \zeta_{(t+1)_0} &\leq (1+\alpha_{\boldsymbol{x}} L_{\hat{f}})\delta_t + \alpha_{\boldsymbol{x}} L_{\hat{f}} \zeta_{t_0} + \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}}{m^{val}} + (1+\alpha_{\boldsymbol{y}} L)^K \zeta_{t_0} + \left((1+\alpha_{\boldsymbol{y}} L)^K - 1\right)\delta_t \\ &= \left((1+\alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} L_{\hat{f}}\right) \cdot (\delta_t + \zeta_{t_0}) + \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}}{m^{val}}. \end{aligned} \tag{15}$$

Then we have $\delta_t + \zeta_{t_0} \leq \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}}{m^{val}} \cdot \frac{((1+\alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} L_{\hat{f}})^t - 1}{(1+\alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} L_{\hat{f}} - 1}$. Thus, we have

$$\begin{aligned}&|\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{'val}), z)| := |f(\boldsymbol{x}_T, \boldsymbol{y}_{T-1_K}(\boldsymbol{x}_{T-1}); z) - f(\boldsymbol{x}'_T, \boldsymbol{y}'_{T-1_K}(\boldsymbol{x}'_{T-1}); z)| \\ \leq & M(\delta_T + \zeta_{(T-1)_K}) = M(\delta_T + \zeta_{T_0}) \leq \alpha_{\boldsymbol{x}} \frac{2\widetilde{M} M}{m^{val}} \cdot \frac{((1+\alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} L_{\hat{f}})^T - 1}{(1+\alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} L_{\hat{f}} - 1},\end{aligned}$$

Then, the proof is completed. □

4.8. The Lipschitz Smooth Constant of AID with Warm-Start

Proof. Firstly, combined with the fact that $\nabla_2 g(\boldsymbol{x}, \boldsymbol{y})$ is differentiable w.r.t. $\boldsymbol{x}$, indicates that the inner output $\hat{\boldsymbol{y}}$ is differentiable w.r.t. $\boldsymbol{x}$. Then, based on the chain rule, we have

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) = \nabla_1 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) + \nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) \nabla_2 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}). \tag{16}$$

Based on the implicit function theorem and Neumann series to obtain an alternative estimation (Lorraine et al., 2020):, we have

$$\nabla \hat{\boldsymbol{y}}_{\text{AID}}(\boldsymbol{x}_t) = -\alpha_{\boldsymbol{y}} \nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_K}) \sum_{i=0}^{D-1} \left[I - \alpha_{\boldsymbol{y}} \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_K})\right]^i.$$

Then, we have

$$\|\nabla \hat{\boldsymbol{y}}_{\text{AID}}(\boldsymbol{x}_t)\| \leq \alpha_{\boldsymbol{y}} \|\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i})\| \cdot \sum_{i=0}^{D-1} (1-\alpha_{\boldsymbol{y}}\mu)^i \leq \frac{L}{\mu}(1-(1-\alpha_{\boldsymbol{y}}\mu)^D).$$

Similarly, $\|\nabla \hat{\boldsymbol{y}}'_{\text{AID}}(\boldsymbol{x}'_t)\| \leq \frac{L}{\mu}(1-(1-\alpha_{\boldsymbol{y}}\mu)^D)$. Next, using Eq. (20) and the triangle inequality, we have

$$\begin{aligned}&\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t))\| \\ \leq & \|\nabla_1 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_1 f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t))\| + \|\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\| \|\nabla_2 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}})\| + \|\nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\| \|\nabla_2 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}) - \nabla_2 f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}')\| \\ \leq & L(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\|) + M\|\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\| + \frac{L^2}{\mu}(1-(1-\alpha_{\boldsymbol{y}}\mu)^D) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \hat{\boldsymbol{y}}'(\boldsymbol{x}_t)\|).\end{aligned}$$

For $\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)$, we have

$$\begin{aligned}&\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t) = \nabla \boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t) \\ = & \nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) - \alpha_{\boldsymbol{y}} \left(\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) - \nabla_{12}^2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}})\right) - \alpha_{\boldsymbol{y}} \left(\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\right) \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) \\ & + \alpha_{\boldsymbol{y}} \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) \left(\nabla_{22}^2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}}) - \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}})\right).\end{aligned}$$

Then, we have

$$\begin{aligned}&\|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) - \alpha_{\boldsymbol{y}} \left(\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\right) \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}})\| \\ \leq & \|I - \alpha_{\boldsymbol{y}} \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}})\| \|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\|,\end{aligned}$$

and

$$\begin{aligned}
&\| -\alpha_{\boldsymbol{y}} \left(\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) - \nabla_{12}^2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}}) \right) + \alpha_{\boldsymbol{y}} \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) \left(\nabla_{22}^2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}}) - \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) \right) \| \\
\leq &\alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu}(1-(1-\alpha_{\boldsymbol{y}}\mu)^D) \right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\| \right).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\|\nabla \boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t)\| \\
\leq &(1-\alpha_{\boldsymbol{y}}\mu)\|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\| + \alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu}(1-(1-\alpha_{\boldsymbol{y}}\mu)^D) \right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\| \right).
\end{aligned}$$

As for $\|\boldsymbol{x}_t - \boldsymbol{x}'_t\|$ and $\|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\|$, according to the conclusion of Lemma3 in Bao et al. (2021), we have

$$\|\boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t)\| \leq L_1^G \frac{(L_2^G)^K - 1}{L_2^G - 1}\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + (L_2^G)^K \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\| \leq \frac{L_1^G}{1 - L_2^G}\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + L_2^G \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|, \tag{17}$$

where $L_2^G := \sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}$, and L_1^G means the function $G(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{y} - \alpha_{\boldsymbol{y}} \nabla_2 g(\boldsymbol{x}, \boldsymbol{y})$ is L_1^G-Lipschitz w.r.t $\boldsymbol{x}$. So we let $L_1^G = \alpha_{\boldsymbol{y}} L$ cause

$$\|G(\boldsymbol{x}, \boldsymbol{y}) - G(\boldsymbol{x}', \boldsymbol{y})\| = \alpha_{\boldsymbol{y}} \|\nabla_2 g(\boldsymbol{x}, \boldsymbol{y}) - \nabla_2 g(\boldsymbol{x}', \boldsymbol{y})\| \leq \alpha_{\boldsymbol{y}} L \|\boldsymbol{x} - \boldsymbol{x}'\|.$$

Then we have

$$\begin{aligned}
\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \hat{\boldsymbol{y}}'(\boldsymbol{x}_t)\| &\leq \left(\frac{\alpha_{\boldsymbol{y}} L}{1 - \sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}} + 1 \right) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|), \\
\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\| &\leq \left(\frac{\alpha_{\boldsymbol{y}} L}{1 - \sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}} + 1 \right) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\|\nabla \boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t)\| \\
\leq &(1-\alpha_{\boldsymbol{y}}\mu)\|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\| + \alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu}(1-(1-\alpha_{\boldsymbol{y}}\mu)^D) \right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\| \right) \\
\leq &\left(\frac{\tau}{\mu} + \frac{L\rho}{\mu^2}(1-(1-\alpha_{\boldsymbol{y}}\mu)^D) \right) \left(\frac{\alpha_{\boldsymbol{y}} L}{1 - \sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}} + 1 \right) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|)
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}_{\text{AID}}(\boldsymbol{x}_t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'_{\text{AID}}(\boldsymbol{x}'_t))\| \qquad (18) \\
\leq &\left[L + M\frac{\tau}{\mu} + \frac{L(M\rho + L\mu)}{\mu^2} \left(1-(1-\alpha_{\boldsymbol{y}}\mu)^D\right) \right] \cdot \left(\frac{\alpha_{\boldsymbol{y}} L}{\sqrt{1 - 2\alpha_{\boldsymbol{y}} \frac{L\mu}{L+\mu}}} + 1 \right) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|) \\
\approx &\mathcal{O}\left((1-(1-\alpha_{\boldsymbol{y}}\mu)^D\right) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|). \qquad (19)
\end{aligned}$$

Then, the proof is completed. $\square$

4.9. The Lipschitz Smooth Constant of AID with Cold-Start

Proof. Firstly, combined with the fact that $\nabla_2 g(\boldsymbol{x}, \boldsymbol{y})$ is differentiable w.r.t. $\boldsymbol{x}$, indicates that the inner output $\hat{\boldsymbol{y}}$ is differentiable w.r.t. $\boldsymbol{x}$. Then, based on the chain rule, we have

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) = \nabla_1 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) + \nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) \nabla_2 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}). \tag{20}$$

Based on the implicit function theorem and Neumann series to obtain an alternative estimation (Lorraine et al., 2020):, we have

$$\nabla \hat{\boldsymbol{y}}_{\text{AID}}(\boldsymbol{x}_t) = -\alpha_{\boldsymbol{y}} \nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_K}) \sum_{i=0}^{D-1} \left[I - \alpha_{\boldsymbol{y}} \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_K}) \right]^i .$$

Then, we have

$$\|\nabla \hat{\boldsymbol{y}}_{\text{AID}}(\boldsymbol{x}_t)\| \leq \alpha_{\boldsymbol{y}} \|\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i})\| \cdot \sum_{i=0}^{D-1} (1 - \alpha_{\boldsymbol{y}} \mu)^i \leq \frac{L}{\mu} (1 - (1 - \alpha_{\boldsymbol{y}} \mu)^D).$$

Similarly, $\|\nabla \hat{\boldsymbol{y}}'_{\text{AID}}(\boldsymbol{x}'_t)\| \leq \frac{L}{\mu}(1 - (1 - \alpha_{\boldsymbol{y}}\mu)^D)$. Next, using Eq. (20) and the triangle inequality, we have

$$\begin{aligned}
&\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t))\| \\
\leq &\|\nabla_1 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \nabla_1 f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t))\| + \|\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\| \|\nabla_2 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}})\| + \|\nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\| \|\nabla_2 f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}) - \nabla_2 f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}')\| \\
\leq &L(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\|) + M \|\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)\| + \frac{L^2}{\mu}(1 - (1 - \alpha_{\boldsymbol{y}}\mu)^D) \cdot (\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \hat{\boldsymbol{y}}'(\boldsymbol{x}_t)\|).
\end{aligned}$$

For $\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t)$, we have

$$\begin{aligned}
&\nabla \hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \nabla \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t) = \nabla \boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t) \\
= &\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) - \alpha_{\boldsymbol{y}} \left(\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) - \nabla_{12}^2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}}) \right) - \alpha_{\boldsymbol{y}} \left(\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) \right) \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) \\
&+ \alpha_{\boldsymbol{y}} \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) \left(\nabla_{22}^2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}}) - \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) \right).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) - \alpha_{\boldsymbol{y}} \left(\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) \right) \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}})\| \\
\leq &\|I - \alpha_{\boldsymbol{y}} \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}})\| \|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\|,
\end{aligned}$$

and

$$\begin{aligned}
&\| - \alpha_{\boldsymbol{y}} \left(\nabla_{12}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) - \nabla_{12}^2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}}) \right) + \alpha_{\boldsymbol{y}} \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t) \left(\nabla_{22}^2 g(\boldsymbol{x}'_t, \boldsymbol{y}'_{t_{K-1}}) - \nabla_{22}^2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_{K-1}}) \right) \| \\
\leq &\alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu}(1 - (1 - \alpha_{\boldsymbol{y}}\mu)^D) \right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\| \right).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\|\nabla \boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t)\| \\
\leq &(1 - \alpha_{\boldsymbol{y}}\mu) \|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}'_t)\| + \alpha_{\boldsymbol{y}} \left(\tau + \frac{L\rho}{\mu}(1 - (1 - \alpha_{\boldsymbol{y}}\mu)^D) \right) \cdot \left(\|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\| \right).
\end{aligned}$$

As for $\|\boldsymbol{x}_t - \boldsymbol{x}'_t\|$ and $\|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\|$, according to the conclusion of Lemma3 in Bao et al. (2021), we have

$$\|\boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \boldsymbol{y}'_{t_K}(\boldsymbol{x}'_t)\| \leq L_1^G \frac{(L_2^G)^K - 1}{L_2^G - 1} \|\boldsymbol{x}_t - \boldsymbol{x}'_t\| + (L_2^G)^K \|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\| \leq \frac{L_1^G}{1 - L_2^G} \|\boldsymbol{x}_t - \boldsymbol{x}'_t\|, \tag{21}$$

where $L_2^G := \sqrt{1-2\alpha_{\boldsymbol{y}}\frac{L\mu}{L+\mu}}$, and L_1^G means the function $G(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{y} - \alpha_{\boldsymbol{y}}\nabla_2 g(\boldsymbol{x},\boldsymbol{y})$ is L_1^G-Lipschitz w.r.t $\boldsymbol{x}$. So we let $L_1^G = \alpha_{\boldsymbol{y}} L$ cause

$$\|G(\boldsymbol{x},\boldsymbol{y}) - G(\boldsymbol{x}',\boldsymbol{y})\| = \alpha_{\boldsymbol{y}}\|\nabla_2 g(\boldsymbol{x},\boldsymbol{y}) - \nabla_2 g(\boldsymbol{x}',\boldsymbol{y})\| \le \alpha_{\boldsymbol{y}} L\|\boldsymbol{x}-\boldsymbol{x}'\|.$$

Then we have

$$\|\boldsymbol{x}_t - \boldsymbol{x}_t'\| + \|\hat{\boldsymbol{y}}(\boldsymbol{x}_t) - \hat{\boldsymbol{y}}'(\boldsymbol{x}_t)\| \le \left(\frac{\alpha_{\boldsymbol{y}} L}{1-\sqrt{1-2\alpha_{\boldsymbol{y}}\frac{L\mu}{L+\mu}}}+1\right)\cdot(\|\boldsymbol{x}_t-\boldsymbol{x}_t'\|),$$

$$\|\boldsymbol{x}_t - \boldsymbol{x}_t'\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\| \le \left(\frac{\alpha_{\boldsymbol{y}} L}{1-\sqrt{1-2\alpha_{\boldsymbol{y}}\frac{L\mu}{L+\mu}}}+1\right)\cdot(\|\boldsymbol{x}_t-\boldsymbol{x}_t'\|).$$

Then, we have

$$\begin{aligned}
&\|\nabla \boldsymbol{y}_{t_K}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_K}(\boldsymbol{x}_t')\| \\
\le&(1-\alpha_{\boldsymbol{y}}\mu)\|\nabla \boldsymbol{y}_{t_{K-1}}(\boldsymbol{x}_t) - \nabla \boldsymbol{y}'_{t_{K-1}}(\boldsymbol{x}_t')\| + \alpha_{\boldsymbol{y}}\left(\tau + \frac{L\rho}{\mu}(1-(1-\alpha_{\boldsymbol{y}}\mu)^D)\right)\cdot\left(\|\boldsymbol{x}_t-\boldsymbol{x}_t'\| + \|\boldsymbol{y}_{t_{K-1}} - \boldsymbol{y}'_{t_{K-1}}\|\right) \\
\le&\left(\frac{\tau}{\mu} + \frac{L\rho}{\mu^2}(1-(1-\alpha_{\boldsymbol{y}}\mu)^D)\right)\left(\frac{\alpha_{\boldsymbol{y}} L}{1-\sqrt{1-2\alpha_{\boldsymbol{y}}\frac{L\mu}{L+\mu}}}+1\right)\cdot\|\boldsymbol{x}_t-\boldsymbol{x}_t'\|
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}_{\text{AID}}(\boldsymbol{x}_t)) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'_{\text{AID}}(\boldsymbol{x}_t'))\| && (22)\\
\le&\left[L + M\frac{\tau}{\mu} + \frac{L(M\rho+L\mu)}{\mu^2}\left(1-(1-\alpha_{\boldsymbol{y}}\mu)^D\right)\right]\cdot\left(\frac{\alpha_{\boldsymbol{y}} L}{\sqrt{1-2\alpha_{\boldsymbol{y}}\frac{L\mu}{L+\mu}}}+1\right)\cdot\|\boldsymbol{x}_t-\boldsymbol{x}_t'\| \\
\approx&\,\mathcal{O}\left((1-(1-\alpha_{\boldsymbol{y}}\mu)^D)\cdot\|\boldsymbol{x}_t-\boldsymbol{x}_t'\|\right. . && (23)
\end{aligned}$$

Then, the proof is completed. □

4.10. The Lipschitz Constant of AID

$$\begin{aligned}
&\|\nabla\hat{F}(\boldsymbol{x})\| = \|\nabla_1 f(\boldsymbol{x},\hat{\boldsymbol{y}}(\boldsymbol{x})) + \nabla\hat{\boldsymbol{y}}_{\text{AID}}(\boldsymbol{x})\nabla_2 f(\boldsymbol{x},\hat{\boldsymbol{y}}(\boldsymbol{x}))\| \\
\le&\|\nabla_1 f(\boldsymbol{x},\hat{\boldsymbol{y}}(\boldsymbol{x}))\| + \|\nabla\hat{\boldsymbol{y}}_{\text{AID}}(\boldsymbol{x})\|\|\nabla_2 f(\boldsymbol{x},\hat{\boldsymbol{y}}(\boldsymbol{x}))\| \\
\le&M + M\cdot\frac{L}{\mu}\left(1-(1-\alpha_{\boldsymbol{y}}\mu)^D\right) = M\left(1+\frac{L}{\mu}\left(1-(1-\alpha_{\boldsymbol{y}}\mu)^D\right)\right) && (24)
\end{aligned}$$

4.11. The Unifomly Stable Constant of AID with Cold-start

Proof. We use the following equation to denote the updating rule in the outer level,

$$\Upsilon(\boldsymbol{x}_t, S^{val}) = \boldsymbol{x}_t - \alpha_{\boldsymbol{x}}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}),$$

where $\hat{\boldsymbol{y}}(\boldsymbol{x}_t) = \boldsymbol{y}_{t_0} - \alpha_{\boldsymbol{y}}\sum_{i=0}^{K-1}\nabla_2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i}; S^{tr})$.

We suppose S^{val} and $S^{'val}$ are different in at most one sample point, and let $\{\boldsymbol{x}_t\}_{t\ge 0}$ and $\{\boldsymbol{x}_t'\}_{t\ge 0}$ be the trace by gradient descent with S^{val} and $S^{'val}$ respectively. Let $\delta_t = \|\boldsymbol{x}_t - \boldsymbol{x}_t'\|$, and then

$$\begin{aligned}
\delta_{t+1} &= \|\Upsilon(\boldsymbol{x}_t, S^{val}) - \Upsilon(\boldsymbol{x}_t', S^{'val})\| \\
&= \|\boldsymbol{x}_t - \alpha_{\boldsymbol{x}}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \boldsymbol{x}_t' + \alpha_{\boldsymbol{x}}\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{'val})\| \\
&\le \|\boldsymbol{x}_t - \boldsymbol{x}_t'\| + \alpha_{\boldsymbol{x}}\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{'val})\|,
\end{aligned}$$

where $\hat{\boldsymbol{y}}'(\boldsymbol{x}_t') = \boldsymbol{y}_{t_0}' - \alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1} \nabla_2 g(\boldsymbol{x}_t', \boldsymbol{y}_{t_i}'; S^{tr})$. $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val})$ and $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{'val})$ denote $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}(\boldsymbol{x}); S^{val})|_{\boldsymbol{x}=\boldsymbol{x}_t}$ and $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}'(\boldsymbol{x}); S^{'val})|_{\boldsymbol{x}=\boldsymbol{x}_t'}$ respectively.

Next, we have

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{'val})\| \\
\leq &\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{val})\| + \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{'val})\| \\
\leq &\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{val})\| + \frac{2\widetilde{M}_{\text{AID}}}{m^{val}} \\
\leq &L_{\hat{f},\text{AID}} \|\boldsymbol{x}_t - \boldsymbol{x}_t'\| + \frac{2\widetilde{M}_{\text{AID}}}{m^{val}},
\end{aligned}
$$

Therefore, we have

$$
\delta_{t+1} \leq \delta_t + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}} \delta_t + \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}_{\text{AID}}}{m^{val}} = (1 + \alpha_{\boldsymbol{x}} L_{\hat{f}}) \delta_t + \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}_{\text{AID}}}{m^{val}}, \tag{25}
$$

Then we have $\delta_t \leq \frac{2\widetilde{M}_{\text{AID}}}{m^{val}} \cdot \frac{(1+\alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}})^t - 1}{L_{\hat{f},\text{AID}}}$. Thus, we have

$$
\begin{aligned}
&|\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{'val}), z)| := |f(\boldsymbol{x}_T, \boldsymbol{y}_T(\boldsymbol{x}_{T-1}); z) - f(\boldsymbol{x}_T', \boldsymbol{y}_T'(\boldsymbol{x}_{T-1}'); z)| \\
\leq &M \left(1 + \sqrt{\frac{L+\mu}{L-\mu}}\right) \delta_T \leq \left(1 + \sqrt{\frac{L+\mu}{L-\mu}}\right) \frac{2\widetilde{M}_{\text{AID}} M}{L_{\hat{f},\text{AID}} m^{val}} \cdot \left[(1 + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}})^T - 1\right] \\
\approx &\mathcal{O}\left(\left(1 - (1 - \alpha_{\boldsymbol{y}} \mu)^D\right)^T\right).
\end{aligned}
$$

where utilizes Eq. (9). Then, the proof is completed. ∎

4.12. The Unifomly Stable Constant of AID with Warm-start

Proof. We use the following equation to denote the updating rule in the outer level,

$$
\Upsilon(\boldsymbol{x}_t, S^{val}) = \boldsymbol{x}_t - \alpha_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}),
$$

where $\hat{\boldsymbol{y}}(\boldsymbol{x}_t) = \boldsymbol{y}_{t_0} - \alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1} \nabla_2 g(\boldsymbol{x}_t, \boldsymbol{y}_{t_i}; S^{tr})$.

We suppose S^{val} and $S^{'val}$ are different in at most one sample point, and let $\{\boldsymbol{x}_t\}_{t\geq 0}$ and $\{\boldsymbol{x}_t'\}_{t\geq 0}$ be the trace by gradient descent with S^{val} and $S^{'val}$ respectively. Let $\delta_t = \|\boldsymbol{x}_t - \boldsymbol{x}_t'\|$, and then

$$
\begin{aligned}
\delta_{t+1} &= \|\Upsilon(\boldsymbol{x}_t, S^{val}) - \Upsilon(\boldsymbol{x}_t', S^{'val})\| \\
&= \|\boldsymbol{x}_t - \alpha_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \boldsymbol{x}_t' + \alpha_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{'val})\| \\
&\leq \|\boldsymbol{x}_t - \boldsymbol{x}_t'\| + \alpha_{\boldsymbol{x}} \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{'val})\|,
\end{aligned}
$$

where $\hat{\boldsymbol{y}}'(\boldsymbol{x}_t') = \boldsymbol{y}_{t_0}' - \alpha_{\boldsymbol{y}} \sum_{i=0}^{K-1} \nabla_2 g(\boldsymbol{x}_t', \boldsymbol{y}_{t_i}'; S^{tr})$. $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val})$ and $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{'val})$ denote $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}(\boldsymbol{x}); S^{val})|_{\boldsymbol{x}=\boldsymbol{x}_t}$ and $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}'(\boldsymbol{x}); S^{'val})|_{\boldsymbol{x}=\boldsymbol{x}_t'}$ respectively.

Next, we have

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{'val})\| \\
\leq &\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{val})\| + \|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{'val})\| \\
\leq &\|\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t); S^{val}) - \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t', \hat{\boldsymbol{y}}'(\boldsymbol{x}_t'); S^{val})\| + \frac{2\widetilde{M}_{\text{AID}}}{m^{val}} \\
\leq &L_{\hat{f},\text{AID}} \left(\|\boldsymbol{x}_t - \boldsymbol{x}_t'\| + \|\boldsymbol{y}_{t_0} - \boldsymbol{y}_{t_0}'\|\right) + \frac{2\widetilde{M}_{\text{AID}}}{m^{val}},
\end{aligned}
$$

and we rewrite $\|\boldsymbol{y}_{t_0} - \boldsymbol{y}'_{t_0}\|$ as ζ_{t_0}. Therefore, we have

$$\delta_{t+1} \leq \delta_t + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}} \delta_t + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}} \zeta_{t_0} + \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}_{\text{AID}}}{m^{val}} = (1 + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}})\delta_t + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}} \zeta_{t_0} + \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}_{\text{AID}}}{m^{val}}, \tag{26}$$

and

$$\zeta_{t_K} \leq \zeta_{t_{K-1}} + \alpha_{\boldsymbol{y}} L \zeta_{t_{K-1}} + \alpha_{\boldsymbol{y}} L \delta_t = (1 + \alpha_{\boldsymbol{y}} L)^K \zeta_{t_0} + \left((1 + \alpha_{\boldsymbol{y}} L)^K - 1\right) \delta_t. \tag{27}$$

To obtain the upper bounds for δ_t and ζ_t, directly substituting Eq. (27) into Eq. (26) or vice versa makes it challenging to derive their respective upper bounds. Therefore, we further analyze the upper bound of $\delta_{t+1} + \zeta_{t_K}$. By Combining Eqs. (26) and (27), we obtain:

$$\begin{aligned}
\delta_{t+1} + \zeta_{t_K} = \delta_{t+1} + \zeta_{(t+1)_0} &\leq (1 + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}})\delta_t + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}} \zeta_{t_0} + \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}_{\text{AID}}}{m^{val}} + (1 + \alpha_{\boldsymbol{y}} L)^K \zeta_{t_0} + \left((1 + \alpha_{\boldsymbol{y}} L)^K - 1\right) \delta_t \\
&= \left((1 + \alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}}\right) \cdot (\delta_t + \zeta_{t_0}) + \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}_{\text{AID}}}{m^{val}}.
\end{aligned} \tag{28}$$

Then we have $\delta_t + \zeta_{t_0} \leq \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}_{\text{AID}}}{m^{val}} \cdot \frac{((1+\alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}})^t - 1}{(1+\alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}} - 1}$. Thus, we have

$$\begin{aligned}
&|\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z) - \ell(\mathbf{A}_{hpt}(S^{tr}, S'^{val}), z)| := |f(\boldsymbol{x}_T, \boldsymbol{y}_{T-1_K}(\boldsymbol{x}_{T-1}); z) - f(\boldsymbol{x}'_T, \boldsymbol{y}'_{T-1_K}(\boldsymbol{x}'_{T-1}); z)| \\
\leq &M(\delta_T + \zeta_{(T-1)_K}) = M(\delta_T + \zeta_{T_0}) \leq \alpha_{\boldsymbol{x}} \frac{2\widetilde{M}_{\text{AID}} M}{m^{val}} \cdot \frac{((1 + \alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}})^T - 1}{(1 + \alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} L_{\hat{f},\text{AID}} - 1},
\end{aligned}$$

Then, the proof is completed. □

4.13. Generalization Analysis of Stochastic ITD

Theorem 4.4. *Suppose a randomized HPT algorithm $\mathbf{A}_{hpt}$ is $\beta-$uniformly stable on validation in expectation, then*

$$\left|\mathbb{E}_{\mathbf{A}_{hpt}, S^{tr} \sim (\mathcal{D}^{tr})^{m^{tr}}, S^{val} \sim (\mathcal{D}^{val})^{m^{val}}} \left[\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), \mathcal{D}^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), S^{val})\right]\right| \leq \beta. \tag{29}$$

Proof.

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbf{A}_{hpt}, S^{tr}, S^{val}} \left[\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), \mathcal{D}^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), S^{val})\right]\right| \\
=&\left|\mathbb{E}_{\mathbf{A}_{hpt}, S^{tr}, S^{val}, z \sim \mathcal{D}^{val}} \left[\ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z_1^{val})\right]\right| \\
=&\left|\mathbb{E}_{\mathbf{A}_{hpt}, S^{tr}, S^{val}, z \sim \mathcal{D}^{val}} \left[\ell(\mathbf{A}_{hpt}(S^{tr}, z, z_2^{val}, \ldots, , z_m^{val}), z_1^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z_1^{val})\right]\right| \\
\leq&\mathbb{E}_{S^{tr}, S^{val}, z \sim \mathcal{D}^{val}} \left|\mathbb{E}_{\mathbf{A}_{hpt}} \left[\ell(\mathbf{A}_{hpt}(S^{tr}, z, z_2^{val}, \ldots, , z_m^{val}), z_1^{val}) - \ell(\mathbf{A}_{hpt}(S^{tr}, S^{val}), z_1^{val})\right]\right| \leq \beta.
\end{aligned}$$

4.14. The Unifomly Stable Constant of Stochastic Algorithm with Cold-start

Here we prove a stochastic version by considering SGD in the outer level, i.e.,

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \hat{\boldsymbol{y}}(\boldsymbol{x}); z_j^{val}), \tag{30}$$

where j is randomly selected from $\{1, \cdots, m^{val}\}$.

Theorem 4.5. *Under some mild assumptions, Solving Problem (1) with T steps SGD in the outer-level is $\beta-$uniformly stable on validation in expectation with*

$$\beta = \inf_{0 \leq t_0 \leq T} \left[\frac{2\widetilde{M}}{m^{val} \widetilde{L}'} \left[(\alpha_{\boldsymbol{x}} \widetilde{L}' + 1)^{t-t_0} - 1\right] + \frac{t_0}{m^{val}} s(\ell).\right]$$

Proof. Suppose $f(\boldsymbol{x}, \hat{\boldsymbol{y}}(\boldsymbol{x}))$ is $\widetilde{M}-$Lipschitz continuous and $\widetilde{L}-$Lipschitz smooth. Suppose S^{val} and $S^{'val}$ differ in at most one point. Let $\delta_t = \|\boldsymbol{x}_t - \boldsymbol{x}'_t\|$. Suppose $0 \leq t' \leq t$, we have

$$\begin{aligned}\mathbb{E}\left[|f(\boldsymbol{x}_t, \hat{y}(\boldsymbol{x}_t)) - f(\boldsymbol{x}'_t, \hat{y}(\boldsymbol{x}'_t))|\right] &= \mathbb{E}\left[|f(\boldsymbol{x}_t, \hat{y}(\boldsymbol{x}_t)) - f(\boldsymbol{x}'_t, \hat{y}(\boldsymbol{x}'_t))| \cdot 1_{\delta_{t'}=0}\right] + \mathbb{E}\left[|f(\boldsymbol{x}_t, \hat{y}(\boldsymbol{x}_t)) - f(\boldsymbol{x}'_t, \hat{y}(\boldsymbol{x}'_t))| \cdot 1_{\delta_{t'}>0}\right] \\ &\leq \widetilde{M}\mathbb{E}[\delta_t \cdot 1_{\delta_{t'}=0}] + P(\delta_{t'} > 0)s(\ell).\end{aligned}$$

Without loss of generality, we assume S^{val} and $S^{'val}$ at most differ in at the first point. If SGD doesn't selects the first point for the first t' iterations, then $\delta_{t'} = 0$. As a result,

$$P(\delta_{t'} = 0) \geq (1 - \frac{1}{m^{val}})^{t'} \geq 1 - \frac{t'}{m^{val}}.$$

Therefore, $P(\delta_{t'} > 0) \leq \frac{t'}{m^{val}}$ and we have

$$\mathbb{E}\left[|f(\boldsymbol{x}_t, \hat{y}(\boldsymbol{x}_t)) - f(\boldsymbol{x}'_t, \hat{y}(\boldsymbol{x}'_t))|\right] \leq \widetilde{M}\mathbb{E}[\delta_t \cdot 1_{\delta_{t'}=0}] + \frac{t'}{m^{val}}s(\ell). \tag{31}$$

Now we bound $\mathbb{E}[\delta_t \cdot 1_{\delta_{t'}=0}]$. Let $\widetilde{L}' = (1 - 1/m^{val})\widetilde{L}$ and let j be the index selected by SGD at the $t+1$ iteration, then we have

$$\begin{aligned}\mathbb{E}[\delta_{t+1} \cdot 1_{\delta_{t'}=0}] &\leq \mathbb{E}[\delta_{t+1} \cdot 1_{j=1} \cdot 1_{\delta_{t'}=0}] + \mathbb{E}[\delta_{t+1} \cdot 1_{j>1} \cdot 1_{\delta_{t'}=0}] \\ &\leq \frac{1}{m^{val}}(\mathbb{E}[\delta_t \cdot 1_{\delta_{t'}=0}] + 2\alpha_{\boldsymbol{x}}\widetilde{M}) + \frac{m^{val}-1}{m^{val}}(1 + \alpha_{\boldsymbol{x}}\widetilde{L})\mathbb{E}[\delta_t \cdot 1_{\delta_{t'}=0}] \\ &= (1 + \alpha_{\boldsymbol{x}}\widetilde{L}')\mathbb{E}[\delta_t \cdot 1_{\delta_{t'}=0}] + \frac{2\alpha_{\boldsymbol{x}}\widetilde{M}}{m^{val}}.\end{aligned}$$

Thus, we have $\mathbb{E}[\delta_t \cdot 1_{\delta_{t'}=0}] \leq \frac{2\widetilde{M}}{m^{val}\widetilde{L}'}\left[(\alpha_{\boldsymbol{x}}\widetilde{L}' + 1)^{t-t'} - 1\right]$. Then, we have

$$\mathbb{E}\left[|f(\boldsymbol{x}_T, \hat{y}(\boldsymbol{x}_T)) - f(\boldsymbol{x}'_T, \hat{y}(\boldsymbol{x}'_T))|\right] \leq \inf_{0\leq t'\leq T}\left[\frac{2\widetilde{M}}{m^{val}\widetilde{L}'}\left[(\alpha_{\boldsymbol{x}}\widetilde{L}' + 1)^{T-t'} - 1\right] + \frac{t'}{m^{val}}s(\ell)\right].$$

4.15. The Unifomly Stable Constant of Stochastic Algorithm with Warm-Start

Theorem 4.6. *Under some mild assumptions, Solving Problem (1) with T steps SGD in the outer-level is $\beta-$uniformly stable on validation in expectation with*

$$\beta = \inf_{0\leq t'\leq T}\left[\frac{2\alpha_{\boldsymbol{x}}\widetilde{M}}{m^{val}\left[(1+\alpha_{\boldsymbol{y}}L)^K + \alpha_{\boldsymbol{x}}\widetilde{L}' - 1\right]} \cdot \left[\left((1+\alpha_{\boldsymbol{y}}L)^K + \alpha_{\boldsymbol{x}}\widetilde{L}'\right)^{T-t'} - 1\right] + \frac{t'}{m^{val}}s(\ell)\right].$$

Proof. Suppose $f(\boldsymbol{x}, \hat{\boldsymbol{y}}(\boldsymbol{x}))$ is $\widetilde{M}-$Lipschitz continuous and $\widetilde{L}-$Lipschitz smooth. Suppose S^{val} and $S^{'val}$ differ in at most one point. Let $\delta_t = \|\boldsymbol{x}_t - \boldsymbol{x}'_t\|$. Suppose $0 \leq t' \leq t$, we have

$$\begin{aligned}\mathbb{E}\left[|f(\boldsymbol{x}_t, \hat{y}(\boldsymbol{x}_t)) - f(\boldsymbol{x}'_t, \hat{y}'(\boldsymbol{x}'_t))|\right] &= \mathbb{E}\left[|f(\boldsymbol{x}_t, \hat{y}(\boldsymbol{x}_t)) - f(\boldsymbol{x}'_t, \hat{y}'(\boldsymbol{x}'_t))| \cdot 1_{\delta_{t'}=0}\right] + \mathbb{E}\left[|f(\boldsymbol{x}_t, \hat{y}(\boldsymbol{x}_t)) - f(\boldsymbol{x}'_t, \hat{y}'(\boldsymbol{x}'_t))| \cdot 1_{\delta_{t'}>0}\right] \\ &\leq \widetilde{M}\mathbb{E}[\delta_t \cdot 1_{\delta_{t'}=0}] + P(\delta_{t'} > 0)s(\ell).\end{aligned}$$

Without loss of generality, we assume S^{val} and $S^{'val}$ at most differ in at the first point. If SGD doesn't selects the first point for the first t' iterations, then $\delta_{t'} = 0$. As a result,

$$P(\delta_{t'} = 0) \geq (1 - \frac{1}{m^{val}})^{t'} \geq 1 - \frac{t'}{m^{val}}.$$

Therefore, $P(\delta_{t'} > 0) \leq \frac{t'}{m^{val}}$ and we have

$$\mathbb{E}\left[|f(\boldsymbol{x}_t, \hat{y}(\boldsymbol{x}_t)) - f(\boldsymbol{x}'_t, \hat{y}'(\boldsymbol{x}'_t))|\right] \leq \widetilde{M}\mathbb{E}[(\delta_t + \zeta_{t'_0}) \cdot 1_{\delta_{t'}=0}] + \frac{t'}{m^{val}}s(\ell). \tag{32}$$

K	1	2	4	8	16
Memory (Mb)	4262	6610	11324	20728	39534

Table 3. The Memory cost under various inner-level steps K.

Now we bound $\mathbb{E}[(\delta_t + \zeta_{t_0}) \cdot 1_{\delta_{t'}=0}]$. Before that, we have

$$\begin{aligned}\delta_{t+1} =& \|\boldsymbol{x}_{t+1} - \boldsymbol{x}'_{t+1}\| \\ =& \|\boldsymbol{x}_t - \alpha_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \hat{\boldsymbol{y}}(\boldsymbol{x}_t)) - \boldsymbol{x}'_{t+1} + \alpha_{\boldsymbol{x}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}'_t, \hat{\boldsymbol{y}}'(\boldsymbol{x}'_t))\| \\ \le& \delta_t + \alpha_{\boldsymbol{x}} \widetilde{L}(\delta_t + \zeta_{t_0}) = (1 + \alpha_{\boldsymbol{x}} \widetilde{L})\delta_t + \alpha_{\boldsymbol{x}} \widetilde{L} \zeta_{t_0}.\end{aligned}$$

and $\zeta_{(t+1)_0} \le (1 + \alpha_{\boldsymbol{y}} L)^K \zeta_{t_0} + \left((1 + \alpha_{\boldsymbol{y}} L)^K - 1\right) \delta_t$. Thus, we have

$$\delta_{t+1} + \zeta_{(t+1)_0} \le \left[(1 + \alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} \widetilde{L}\right] (\delta_t + \zeta_{t_0}).$$

Let $\widetilde{L}' = (1 - 1/m^{val})\widetilde{L}$ and let j be the index selected by SGD at the $t+1$ iteration, then we have

$$\begin{aligned}&\mathbb{E}[(\delta_{t+1} + \zeta_{(t+1)_0}) \cdot 1_{\delta_{t'}=0}] \\ \le& \mathbb{E}[(\delta_{t+1} + \zeta_{(t+1)_0}) \cdot 1_{j=1} \cdot 1_{\delta_{t'}=0}] + \mathbb{E}[(\delta_{t+1} + \zeta_{(t+1)_0}) \cdot 1_{j>1} \cdot 1_{\delta_{t'}=0}] \\ \le& \frac{1}{m^{val}} \left[\mathbb{E}[(\delta_t + \zeta_{t_0}) \cdot 1_{\delta_{t'}=0}] \cdot (1 + \alpha_{\boldsymbol{y}} L)^K + 2\alpha_{\boldsymbol{x}} \widetilde{M}\right] + \frac{m^{val} - 1}{m^{val}} \left[\left((1 + \alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} \widetilde{L}\right) \cdot \mathbb{E}[(\delta_t + \zeta_{t_0}) \cdot 1_{\delta_{t'}=0}]\right] \\ =& \left((1 + \alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} \widetilde{L}'\right) \mathbb{E}[(\delta_t + \zeta_{t_0}) \cdot 1_{\delta_{t'}=0}] + \frac{2\alpha_{\boldsymbol{x}} \widetilde{M}}{m^{val}}.\end{aligned}$$

Thus, we have $\mathbb{E}[(\delta_t + \zeta_{t_0}) \cdot 1_{\delta_{t'}=0}] \le \frac{2\alpha_{\boldsymbol{x}} \widetilde{M}}{m^{val}\left[(1+\alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} \widetilde{L}' - 1\right]} \cdot \left[\left((1 + \alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} \widetilde{L}'\right)^{t-t'} - 1\right]$. Then, we have

$$\mathbb{E}\left[|f(\boldsymbol{x}_T, \hat{y}(\boldsymbol{x}_T)) - f(\boldsymbol{x}'_T, \hat{y}(\boldsymbol{x}'_T))|\right] \le \inf_{0 \le t' \le T} \left[\frac{2\alpha_{\boldsymbol{x}} \widetilde{M}}{m^{val}\left[(1 + \alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} \widetilde{L}' - 1\right]} \cdot \left[\left((1 + \alpha_{\boldsymbol{y}} L)^K + \alpha_{\boldsymbol{x}} \widetilde{L}'\right)^{T-t'} - 1\right] + \frac{t'}{m^{val}} s(\ell)\right].$$

5. Experiments with Neural Networks

In this section, we conduct a neural network experiment to further verify the soundness of our theoretical findings and the effectiveness of our method.

We utilize CIFAR-10 and CIFAR-100 as datasets, applying asymmetric label noise at different levels. CMW-Net (Shu et al., 2023) is used as the weighting scheme for the experiment, i.e., the outer-level objective is to optimize the parameters of the weighting network CMW-Net, while the inner-level objective is to optimize the classification network ResNet-18.

Table 4 presents a memory comparison for different steps of inner-level iteration. The results show that when K is large (e.g., $K = 16$), the memory cost is extremely expensive, indicating that cold-start is not suitable for such the complex tasks. On the other hand, when K is small (e.g., $K = 1$), cold-start completely fails, as shown in Table 2. This supports our viewpoint that, in practice, modifying warm-start to improve generalization performance is preferable to directly using cold start.

Table 2 presents the test accuracy under different cases. The results show that when $K = 1$, warm-start achieves good results in most cases. This demonstrates the effectiveness of our second approach (reducing the inner-level steps K), which is consistent with our theoretical findings.

6. Discussion of the Boundedness Assumption of the Loss Function

The bounded assumption is mild and common (e.g., also used in Theorem 3.12 of Hardt et al. (2016) and Section 2 in Shalev-Shwartz et al. (2010)). Indeed, given a machine learning model of a finite number of parameters (e.g. neural networks of finite depth and width used in our experiments), a bounded parameter space, and a bounded input space, the feature space is also bounded. Note that previous work makes a similar assumption (at the bottom of Page 9 in Hardt et al. (2016)) as Assumption 1, as well as and the bottom of Page 3 in Bao et al. (2021).

	CIFAR-10(nr=0.4)	CIFAR-10(nr=0.6)	CIFAR-100(nr=0.4)	CIFAR-100(nr=0.6)
Cold-Start(K=6)	10.00	10.00	0.92	0.86
Warm-Start(K=1)	**91.75**	**90.86**	69.74	**65.41**
Warm-Start(K=2)	91.31	90.82	69.19	64.85
Warm-Start(K=3)	91.42	90.58	68.95	65.09
Warm-Start(K=4)	90.74	90.67	**70.33**	64.69
Warm-Start(K=6)	91.13	90.03	68.93	64.54

Table 4. The test accuracy of CIFAR-10 and CIFAR-100 datasets. The best results are bolded, and the second-best results are underlined. nr: noisy ratio

7. Discussion on the Inapplicability of Warm-Start Strategy in Meta-Learning

Warm-start is not suitable for applications where storing the entire LL solution is costly, such as meta-learning. In fact, meta-learning aims to leverage the "common property" among a set of learning tasks to facilitate the learning process. Therefore, when the number of tasks is large, a common strategy is to solve only a small random subset of tasks in each outer-level iteration. In this case, using warm-start becomes problematic. Specifically, if task i is sampled in iteration t, consistently applying warm starting would require using the solution obtained for the same task i in iteration $t-1$ as the initialization for LL optimization. However, this is not applicable in a meta-learning setup with randomly sampled tasks. Such a discussion can be also found in Grazzi et al. (2023).

References

Bao, F., Wu, G., Li, C., Zhu, J., and Zhang, B. Stability and generalization of bilevel programming in hyperparameter optimization. *Advances in neural information processing systems*, 34:4529–4541, 2021.

Ghadimi, S. and Wang, M. Approximation methods for bilevel programming. *arXiv preprint arXiv:1802.02246*, 2018.

Grazzi, R., Pontil, M., and Salzo, S. Bilevel optimization with a lower-level contraction: Optimal sample complexity without warm-start. *Journal of Machine Learning Research*, 24(167):1–37, 2023.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pp. 1225–1234. PMLR, 2016.

Ji, K., Yang, J., and Liang, Y. Bilevel optimization: Convergence analysis and enhanced design. In *International conference on machine learning*, pp. 4882–4892. PMLR, 2021.

Lorraine, J., Vicol, P., and Duvenaud, D. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pp. 1540–1552. PMLR, 2020.

Shalev-Shwartz, S., Shamir, O., Srebro, N., and Sridharan, K. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 11:2635–2670, 2010.

Shu, J., Yuan, X., Meng, D., and Xu, Z. Cmw-net: Learning a class-aware sample weighting mapping for robust deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11521–11539, 2023.