A APPENDIX

A.1 RESTATEMENT OF ASSUMPTION 3

Assumption 5 (Restatement of Assumption 3)**.** *We assume Lipschitz properties for all functions* $l_i(\theta)(i = 1, 2, 3)$ *as follows:*

a) $l_i(\theta)$ *is* M*-Lipschitz, i.e., for any* θ_1 *and* θ_2, $\|l_i(\theta_1) - l_i(\theta_2)\| \leq M\|\theta_1 - \theta_2\|(i = 1, 2, 3)$.

b) $\nabla l_i(\theta)$ *is* L*-Lipschitz, i.e., for any* θ_1 *and* θ_2, $\|\nabla l_i(\theta_1) - \nabla l_i(\theta_2)\| \leq L\|\theta_1 - \theta_2\|(i = 1, 2, 3)$.

c) $\nabla^2 l_i(\theta)$ *is* ρ*-Lipschitz, i.e., for any* θ_1 *and* θ_2, $\|\nabla^2 l_i(\theta_1) - \nabla^2 l_i(\theta_2)\| \leq \rho\|\theta_1 - \theta_2\|(i = 1, 2, 3)$.

d) $m(z, \phi)$ *is* M_{m1}*-Lipschitz w.r.t.* z *and* M_{m2}*-Lipschitz w.r.t.* ϕ*, i.e.,*

$$\begin{aligned} \|m(z_1, \phi) - m(z_2, \phi)\| &\leq M_{m1}\|z_1 - z_2\| \quad \textit{for any } z_1 \textit{ and } z_2, \\ \|m(z, \phi_1) - m(z, \phi_2)\| &\leq M_{m2}\|\phi_1 - \phi_2\| \quad \textit{for any } \phi_1 \textit{ and } \phi_2. \end{aligned}$$

e) $\nabla^2_\phi \theta_T(\phi)$ *is* ρ_θ*-Lipschitz, i.e., for any* ϕ_1 *and* ϕ_2, $\|\nabla^2_\phi \theta_T(\phi_1) - \nabla^2_\phi \theta_T(\phi_2)\| \leq \rho_\theta\|\phi_1 - \phi_2\|$.

The above Assumptions (a)(b)(c) also hold for stochastic $\hat{l}_i(\theta)$, $\nabla\hat{l}_i(\theta)$ *and* $\nabla^2_\theta\hat{l}_i(\theta)(i = 1, 2, 3)$.

A.2 PROOF OF SUPPORTING LEMMAS (LEMMA 12 CORRESPONDS TO PROPOSITION 1)

Lemma 1. *Based on update procedure of* $\theta_t(\theta)$*, we obtain*

$$\nabla_\phi \theta_T(\phi) = \sum_{i=0}^{T-1} \left(\prod_{j=i+1}^{T-1} \left(I + \nabla_1 m(\nabla_\theta \hat{l}(\theta_{T+i-j}), \phi)\nabla^2_\theta \hat{l}(\theta_{T+i-j}) \right) \nabla_2 m(\nabla_\theta \hat{l}(\theta_i), \phi) \right).$$

Proof. The $\theta_t(\phi)$ update process is shown below:

$$\theta_{t+1}(\phi) = \theta_t(\phi) + m(z_t, \phi).$$

If we only consider $z_t(\theta_t; \zeta_t) = \nabla_\theta l(\theta_t(\phi); \zeta_t) = \nabla_\theta \hat{l}(\theta_t)$, then we obtain

$$\begin{aligned} \nabla_\phi \theta_{t+1}(\phi) &= \nabla_\phi \theta_t(\phi) + \nabla_\phi m(\nabla_\theta \hat{l}(\theta_t), \phi) \\ &= \nabla_\phi \theta_t(\phi) + \nabla_1 m(\nabla_\theta \hat{l}(\theta_t), \phi)\nabla^2_\theta \hat{l}(\theta_t)\nabla_\phi \theta_t(\phi) + \nabla_2 m(\nabla_\theta \hat{l}(\theta_t), \phi) \\ &= (I + \nabla_1 m(\nabla_\theta \hat{l}(\theta_t), \phi)\nabla^2_\theta \hat{l}(\theta_t))\nabla_\phi \theta_t(\phi) + \nabla_2 m(\nabla_\theta \hat{l}(\theta_t), \phi). \end{aligned}$$

If we iterate the above equation from $t = 0$ to T, then we obtain

$$\begin{aligned} \nabla_\phi \theta_T(\phi) = &\sum_{i=0}^{T-1} \left(\prod_{j=i+1}^{T-1} \left(I + \nabla_1 m(\nabla_\theta \hat{l}(\theta_{T+i-j}), \phi)\nabla^2_\theta \hat{l}(\theta_{T+i-j}) \right) \nabla_2 m(\nabla_\theta \hat{l}(\theta_i), \phi) \right) \\ &+ \prod_{i=1}^{T} \left(I + \nabla_1 m(\nabla_\theta \hat{l}(\theta_{T-i}), \phi)\nabla^2_\theta \hat{l}(\theta_{T-i}) \right) \nabla_\phi \theta_0, \end{aligned}$$

We assume θ_0 is randomly sampled and independent from ϕ, then we obtain

$$\nabla_\phi \theta_T(\phi) = \sum_{i=0}^{T-1} \left(\prod_{j=i+1}^{T-1} \left(I + \nabla_1 m(\nabla_\theta \hat{l}(\theta_{T+i-j}), \phi)\nabla^2_\theta \hat{l}(\theta_{T+i-j}) \right) \nabla_2 m(\nabla_\theta \hat{l}(\theta_i), \phi) \right).$$

□

Lemma 2. *If we assume that* $\theta_0(\phi_1) = \theta_0(\phi_2)$*, based on Assumption 3, then we obtain*

$$\|\theta_T(\phi_1) - \theta_T(\phi_2)\| \leq \left(((M_{m1}L + 1)^{T-1} - 1)\frac{M_{m2}}{M_{m1}L} \right) \|\phi_1 - \phi_2\| = M_{\theta T}\|\phi_1 - \phi_2\|. \quad (9)$$

Proof. Based on the iterate procedure of $\theta_T(\phi)$, we obtain

$$
\begin{aligned}
&\|\theta_T(\phi_1) - \theta_T(\phi_2)\| \\
\overset{(i)}{=}&\left\| \sum_{t=1}^{T-1} (m(\nabla_\theta \hat{l}(\theta_t(\phi_1)), \phi_1) - m(\nabla_\theta \hat{l}(\theta_t(\phi_2)), \phi_2)) \right\| \\
=&\Big\| \sum_{t=1}^{T-1} (m(\nabla_\theta \hat{l}(\theta_t(\phi_1)), \phi_1) - m(\nabla_\theta \hat{l}(\theta_t(\phi_1)), \phi_2) + m(\nabla_\theta \hat{l}(\theta_t(\phi_1)), \phi_2) \\
&- m(\nabla_\theta \hat{l}(\theta_t(\phi_2)), \phi_2)) \Big\| \\
\leq&\left\| \sum_{t=1}^{T-1} (m(\nabla_\theta \hat{l}(\theta_t(\phi_1)), \phi_1) - m(\nabla_\theta \hat{l}(\theta_t(\phi_1)), \phi_2)) \right\| \\
&+ \left\| \sum_{t=1}^{T-1} m(\nabla_\theta \hat{l}(\theta_t(\phi_1)), \phi_2) - m(\nabla_\theta \hat{l}(\theta_t(\phi_2)), \phi_2)) \right\| \\
\overset{(ii)}{\leq}&\left\| \sum_{t=1}^{T-1} M_{m2} \|\phi_1 - \phi_2\| \right\| + \left\| \sum_{t=1}^{T-1} M_{m1} \|\nabla_\theta \hat{l}(\theta_t(\phi_1)) - \nabla_\theta \hat{l}(\theta_t(\phi_2))\| \right\| \\
\leq&(T-1) M_{m2} \|\phi_1 - \phi_2\| + M_{m1} \sum_{t=1}^{T-1} \|\nabla_\theta \hat{l}(\theta_t(\phi_1)) - \nabla_\theta \hat{l}(\theta_t(\phi_2))\| \\
\overset{(iii)}{\leq}&(T-1) M_{m2} \|\phi_1 - \phi_2\| + M_{m1} L \sum_{t=1}^{T-1} \|\theta_t(\phi_1) - \theta_t(\phi_2)\|,
\end{aligned}
$$

where (i) follows from Equation (1), (ii) and (iii) from Assumption 3. If we further iterate it from $t = 0$ to T, we obtain

$$
\|\theta_T(\phi_1) - \theta_T(\phi_2)\| \leq \left(((M_{m1} L + 1)^{T-1} - 1) \frac{M_{m2}}{M_{m1} L} \right) \|\phi_1 - \phi_2\| = M_{\theta T} \|\phi_1 - \phi_2\|.
$$

□

Lemma 3. *If we define $A_i(\phi) = \nabla_2 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_1)$, based on Assumption 3 and Lemma 2, we obtain*

$$
\|A_i(\phi_1) - A_i(\phi_2)\| \leq M_{Ai} \|\phi_1 - \phi_2\|,
$$

where $M_{Ai} = L_{m2} + L_{m1} L M_{\theta i}$.

Proof. Based on the definition of $A_i(\phi)$, we have

$$
\begin{aligned}
&\|A_i(\phi_1) - A_i(\phi_2)\| \\
=&\|\nabla_2 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_1) - \nabla_2 m(\nabla_\theta \hat{l}(\theta_i(\phi_2)), \phi_2)\| \\
=&\|\nabla_2 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_1) - \nabla_2 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_2) \\
&+ \nabla_2 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_2) - \nabla_2 m(\nabla_\theta \hat{l}(\theta_i(\phi_2)), \phi_2)\| \\
\overset{(i)}{\leq}&L_{m2} \|\phi_1 - \phi_2\| + L_{m1} \|\nabla_\theta \hat{l}(\theta_i(\phi_1)) - \nabla_\theta \hat{l}(\theta_i(\phi_2))\| \\
\leq&L_{m2} \|\phi_1 - \phi_2\| + L_{m1} L \|\theta_i(\phi_1) - \theta_i(\phi_2)\| \\
\overset{(ii)}{\leq}&(L_{m2} + L_{m1} L M_{\theta i}) \|\phi_1 - \phi_2\| = M_{Ai} \|\phi_1 - \phi_2\|,
\end{aligned}
$$

where (i) follows from Assumption 3, (ii) follows from Lemma 2. □

Lemma 4. *We first define $B_i(\phi) = \nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi)), \phi) \nabla_\theta^2 \hat{l}(\theta_i(\phi))$. Based on the Lemma 2 and Assumption 3, we obtain*

$$
\|B_i(\phi_1) - B_i(\phi_2)\| \leq M_{Bi} \|\phi_1 - \phi_2\|,
$$

where $M_{Bi} = M_{m1} \rho M_{\theta i} + L L_{m2} + L^2 L_{m1} M_{\theta i}$.

Proof. Based on the definition of $B_i(\phi)$, we have

$$
\begin{aligned}
&\|B_i(\phi_1) - B_i(\phi_2)\| \\
&\quad = \|\nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_1)\nabla_\theta^2 \hat{l}(\theta_i(\phi_1)) - \nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi_2)), \phi_2)\nabla_\theta^2 \hat{l}(\theta_i(\phi_2))\| \\
&\quad = \|\nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_1)\nabla_\theta^2 \hat{l}(\theta_i(\phi_1)) - \nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_1)\nabla_\theta^2 \hat{l}(\theta_i(\phi_2)) \\
&\qquad + \nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_1)\nabla_\theta^2 \hat{l}(\theta_i(\phi_2)) - \nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi_2)), \phi_2)\nabla_\theta^2 \hat{l}(\theta_i(\phi_2))\| \\
&\quad \overset{(i)}{\leq} M_{m1}\|\nabla_\theta^2 \hat{l}(\theta_i(\phi_1)) - \nabla_\theta^2 \hat{l}(\theta_i(\phi_2))\| + L\|\nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_1) - \nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi_2)), \phi_2)\| \\
&\quad \overset{(ii)}{\leq} M_{m1}\rho M_{\theta i}\|\phi_1 - \phi_2\| + L\|\nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_1) - \nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_2) \\
&\qquad + \nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_2) - \nabla_1 m(\nabla_\theta \hat{l}(\theta_i(\phi_2)), \phi_2)\| \\
&\quad \overset{(iii)}{\leq} M_{m1}\rho M_{\theta i}\|\phi_1 - \phi_2\| + LL_{m2}\|\phi_1 - \phi_2\| + LL_{m1}\|\nabla_\theta \hat{l}(\theta_i(\phi_1)) - \nabla_\theta \hat{l}(\theta_i(\phi_2))\| \\
&\quad \leq (M_{m1}\rho M_{\theta i} + LL_{m2} + L^2 L_{m1} M_{\theta i})\|\phi_1 - \phi_2\| = M_{Bi}\|\phi_1 - \phi_2\|,
\end{aligned}
$$

where (i) and (iii) follows from Assumption 3, (ii) follows from Lemma 2. □

Lemma 5. *Based on Assumption 3 and Lemmas 1, 3 and 4, then we obtain*

$$\|\nabla_\phi \theta_T(\phi_1) - \nabla_\phi \theta_T(\phi_2)\| \leq L_{\theta T}\|\phi_1 - \phi_2\|, \tag{10}$$

where $L_{\theta T} = \sum_{i=0}^{T-1}(1+M_{m1}L)^{T-i-1}M_{Ai} + \sum_{i=0}^{T-1} M_{m2}(1+M_{m1}L)^{T-i-2}\sum_{j=i+1}^{T-1} M_{B(T+i-j)}$.

Proof. Based on the definition of $\nabla_\phi \theta_T(\phi)$ in Lemma 1, we obtain

$$
\begin{aligned}
&\|\nabla_\phi \theta_T(\phi_1) - \nabla_\phi \theta_T(\phi_2)\| \\
&\overset{(i)}{\leq} \sum_{i=0}^{T-1}\Bigg\| \prod_{j=i+1}^{T-1}\Big(I + \nabla_1 m(\nabla_\theta \hat{l}(\theta_{T+i-j}(\phi_1)), \phi_1)\nabla_\theta^2 \hat{l}(\theta_{T+i-j}(\phi_1))\Big)\nabla_2 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)), \phi_1) \\
&\quad - \prod_{j=i+1}^{T-1}\Big(I + \nabla_1 m(\nabla_\theta \hat{l}(\theta_{T+i-j}(\phi_2)), \phi_2)\nabla_\theta^2 \hat{l}(\theta_{T+i-j}(\phi_2))\Big)\nabla_2 m(\nabla_\theta \hat{l}(\theta_i(\phi_2)), \phi_2)\Bigg\| \\
&\overset{(ii)}{=} \sum_{i=0}^{T-1}\Bigg\| \prod_{j=i+1}^{T-1}\Big(I + B_{T+i-j}(\phi_1)\Big)A_i(\phi_1) - \prod_{j=i+1}^{T-1}\Big(I + B_{T+i-j}(\phi_2)\Big)A_i(\phi_2)\Bigg\| \\
&\quad = \sum_{i=0}^{T-1}\Bigg\| \prod_{j=i+1}^{T-1}\Big(I + B_{T+i-j}(\phi_1)\Big)A_i(\phi_1) - \prod_{j=i+1}^{T-1}\Big(I + B_{T+i-j}(\phi_1)\Big)A_i(\phi_2) \\
&\qquad + \prod_{j=i+1}^{T-1}\Big(I + B_{T+i-j}(\phi_1)\Big)A_i(\phi_2) - \prod_{j=i+1}^{T-1}\Big(I + B_{T+i-j}(\phi_2)\Big)A_i(\phi_2)\Bigg\| \\
&\overset{(iii)}{\leq} \sum_{i=0}^{T-1}\Bigg((1+M_{m1}L)^{T-i-1}\|A_i(\phi_1) - A_i(\phi_2)\| + M_{m2}\Bigg\|\prod_{j=i+1}^{T-1}\big(I + B_{T+i-j}(\phi_1)\big) \\
&\qquad - \prod_{j=i+1}^{T-1}\big(I + B_{T+i-j}(\phi_2)\big)\Bigg\|\Bigg) \\
&\quad \leq \sum_{i=0}^{T-1}\Bigg((1+M_{m1}L)^{T-i-1}\|A_i(\phi_1) - A_i(\phi_2)\| + M_{m2}(1+M_{m1}L)^{T-i-2} \\
&\qquad \sum_{j=i+1}^{T-1}\|B_{T+i-j}(\phi_1) - B_{T+i-j}(\phi_2)\|\Bigg) \\
&\overset{(iv)}{\leq} \sum_{i=0}^{T-1}\Bigg((1+M_{m1}L)^{T-i-1}M_{Ai}\|\phi_1 - \phi_2\| + M_{m2}(1+M_{m1}L)^{T-i-2}
\end{aligned}
$$

$$\sum_{j=i+1}^{T-1} M_{B(T+i-j)}\|\phi_1-\phi_2\|\Big)$$
$$=\Big(\sum_{i=0}^{T-1}(1+M_{m1}L)^{T-i-1}M_{Ai}+\sum_{i=0}^{T-1}M_{m2}(1+M_{m1}L)^{T-i-2}\sum_{j=i+1}^{T-1}M_{B(T+i-j)}\Big)\|\phi_1-\phi_2\|$$
$$=L_{\theta T}\|\phi_1-\phi_2\|,$$

where (i) is based on Lemma 1, (ii) is based on the fact that $A_i(\phi_1)=\nabla_2 m(\nabla_\theta \hat{l}(\theta_i(\phi_1)),\phi_1)$, $B_{T+i-j}(\phi_1)=\nabla_1 m(\nabla_\theta \hat{l}(\theta_{T+i-j}(\phi_1)),\phi_1)\nabla^2_\theta \hat{l}(\theta_{T+i-j}(\phi_1))$, (iii) follows from Assumption 3 and (iv) follows from Lemma 3 and 4. □

Lemma 6. *Based on Lemmas 2, 5 and Assumption 3, we obtain*

$$\|\nabla_\phi \hat{g}_T(\phi_1)-\nabla_\phi \hat{g}_T(\phi_2)\|\le L_{g_T}\|\phi_1-\phi_2\|,$$

where $L_{g_T}=ML_{\theta T}+LM^2_{\theta T}$, $L_{\theta T}$ is defined in Lemma 5, $M_{\theta T}$ is defined in Lemma 2.

Proof. We assume all functions share the same starting point θ_0, then we have

$$\begin{aligned}
&\|\nabla_\phi \hat{g}_T(\phi_1)-\nabla_\phi \hat{g}_T(\phi_2)\|\\
&\quad=\|\nabla_\phi \hat{l}(\theta_T(\phi_1))-\nabla_\phi \hat{l}(\theta_T(\phi_2))\|\\
&\quad=\|\nabla_\theta l(\theta_T(\phi_1))\nabla_\phi\theta_T(\phi_1)]-\nabla_\theta l(\theta_T(\phi_2))\nabla_\phi\theta_T(\phi_2)\|\\
&\quad\le\|\nabla_\theta l(\theta_T(\phi_1))\|\|\nabla_\phi\theta_T(\phi_1)-\nabla_\phi\theta_T(\phi_2)\|\\
&\qquad+\|\nabla_\theta l(\theta_T(\phi_1))-\nabla_\theta l(\theta_T(\phi_2))\|\|\nabla_\phi\theta_T(\phi_2)\|\\
&\quad\overset{(i)}{\le}M\|\nabla_\phi\theta_T(\phi_1)-\nabla_\phi\theta_T(\phi_2)\|+L\|\theta_T(\phi_1)-\theta_T(\phi_2)\|\|\nabla_\phi\theta_T(\phi_2)\|\\
&\quad\overset{(ii)}{\le}ML_{\theta T}\|\phi_1-\phi_2\|+LM^2_{\theta T}\|\phi_1-\phi_2\|=(ML_{\theta T}+LM^2_{\theta T})\|\phi_1-\phi_2\|=L_{g_T}\|\phi_1-\phi_2\|,
\end{aligned}$$

where (i) from Assumption 3, (ii) from Lemma 2 and 5. □

Lemma 7. *Based on the Lemma 2, 5 and Assumption 3, we obtain*

$$\|\nabla^2_\phi \hat{g}_T(\phi_1)-\nabla^2_\phi \hat{g}_T(\phi_2)\|\le\rho_{g_T}\|\phi_1-\phi_2\|, \tag{11}$$

where $\rho_{g_T}=3LM_{\theta T}L_{\theta T}+M\rho_\theta+M^3_{\theta T}\rho$.

Proof. We first compute the Lipschitz condition of $\nabla_\phi\nabla_\theta \hat{l}(\theta_T(\phi))$ as follows

$$\begin{aligned}
&\|\nabla_\phi\nabla_\theta \hat{l}(\theta_T(\phi_1))-\nabla_\phi\nabla_\theta \hat{l}(\theta_T(\phi_2))\|\\
&\quad=\|[\nabla_\phi\theta_T(\phi_1)]^T\nabla^2_\theta \hat{l}(\theta_T(\phi_1))-[\nabla_\phi\theta_T(\phi_2)]^T\nabla^2_\theta \hat{l}(\theta_T(\phi_2))\|\\
&\quad\le\|[\nabla_\phi\theta_T(\phi_1)]^T\|\|\nabla^2_\theta \hat{l}(\theta_T(\phi_1))-\nabla^2_\theta \hat{l}(\theta_T(\phi_2))\|\\
&\qquad+\|[\nabla_\phi\theta_T(\phi_1)]^T-[\nabla_\phi\theta_T(\phi_2)]^T\|\|\nabla^2_\theta \hat{l}(\theta_T(\phi_2))\|\\
&\quad\overset{(i)}{\le}M^2_{\theta T}\rho\|\phi_1-\phi_2\|+L_{\theta T}L\|\phi_1-\phi_2\|\\
&\quad=(M^2_{\theta T}\rho+L_{\theta T}L)\|\phi_1-\phi_2\|,
\end{aligned}$$

where (i) follows from Lemma 2, 5 and Assumption 3. Then, based on the definition of $\nabla^2_\phi \hat{g}_T(\phi)$, we have

$$\begin{aligned}
&\|\nabla^2_\phi \hat{g}_T(\phi_1)-\nabla^2_\phi \hat{g}_T(\phi_2)\|\\
&\quad=\|\nabla^2_\phi \hat{l}(\theta_T(\phi_1))-\nabla^2_\phi \hat{l}(\theta_T(\phi_2))\|\\
&\quad=\|\nabla^2_\phi\theta_T(\phi_1)\nabla_\theta \hat{l}(\theta_T(\phi_1))+[\nabla^2_\phi\theta^i_T(\phi_1)]^T\nabla_\phi\nabla_\theta \hat{l}(\theta_T(\phi_1))-\nabla^2_\phi\theta_T(\phi_2)\nabla_\theta \hat{l}(\theta_T(\phi_2))\\
&\qquad-[\nabla^2_\phi\theta_T(\phi_2)]^T\nabla_\phi\nabla_\theta l_i(\theta_T(\phi_2))\|\\
&\quad\le\|\nabla^2_\phi\theta_T(\phi_1)\nabla_\theta \hat{l}(\theta_T(\phi_1))-\nabla^2_\phi\theta_T(\phi_2)\nabla_\theta \hat{l}(\theta_T(\phi_2))\|
\end{aligned}$$

$$
\begin{aligned}
&+\|[\nabla_\phi\theta_T(\phi_1)]^T\nabla_\phi\nabla_\theta\hat{l}(\theta_T(\phi_1))-[\nabla_\phi\theta_T(\phi_2)]^T\nabla_\phi\nabla_\theta\hat{l}(\theta_T(\phi_2))\| \\
\leq&\|\nabla^2_\phi\theta_T(\phi_1)\|\|\nabla_\theta\hat{l}(\theta_T(\phi_1))-\nabla_\theta\hat{l}(\theta_T(\phi_2))\| \\
&+\|\nabla^2_\phi\theta_T(\phi_1)-\nabla^2_\phi\theta_T(\phi_2)\|\|\nabla_\theta\hat{l}(\theta_T(\phi_2))\| \\
&+\|[\nabla_\phi\theta_T(\phi_1)]^T\|\|\nabla_\phi\nabla_\theta\hat{l}(\theta_T(\phi_1))-\nabla_\phi\nabla_\theta\hat{l}(\theta_T(\phi_2))\| \\
&+\|[\nabla_\phi\theta_T(\phi_1)]^T-[\nabla_\phi\theta_T(\phi_2)]^T\|\|\nabla_\phi\nabla_\theta\hat{l}(\theta_T(\phi_2))\| \\
\overset{(i)}{\leq}&LL_{\theta T}\|\theta_T(\phi_1)-\theta_T(\phi_2)\|+M\|\nabla^2_\phi\theta_T(\phi_1)-\nabla^2_\phi\theta_T(\phi_2)\| \\
&+M_{\theta T}\|\nabla_\phi\nabla_\theta\hat{l}(\theta_T(\phi_1))-\nabla_\phi\nabla_\theta\hat{l}(\theta_T(\phi_2))\|+L_{\theta T}\|\phi_1-\phi_2\|M_{\theta T}L \\
\leq&LM_{\theta T}L_{\theta T}\|\phi_1-\phi_2\|+M\rho_\theta\|\phi_1-\phi_2\| \\
&+(M^2_{\theta T}\rho+L_{\theta T}L)M_{\theta T}\|\phi_1-\phi_2\|+M_{\theta T}LL_{\theta T}\|\phi_1-\phi_2\| \\
=&(3LM_{\theta T}L_{\theta T}+M\rho_\theta+M^3_{\theta T}\rho)\|\phi_1-\phi_2\|=\rho_{g_T}\|\phi_1-\phi_2\|,
\end{aligned}
$$

where (i) follows from Lemma 2 and 5.

□

Lemma 8. *If we assume $\theta^1_0(\phi)=\theta^2_0(\phi)$, based on Assumption 3 and 4, we obtain*

$$\|\theta^1_T(\phi)-\theta^2_T(\phi)\|\leq\sigma_{\theta T},$$

where T is the iteration number and $\sigma_{\theta T}=(1+M_{m1}L)^T\frac{\Delta_{12}}{L}-\frac{\Delta_{12}}{L}$.

Proof. Based on the iterative process of $\theta_t(\phi)$, we obtain

$$
\begin{aligned}
&\|\theta^1_T(\phi)-\theta^2_T(\phi)\| \\
\overset{(i)}{\leq}&\|\theta^1_{T-1}(\phi)+m(\nabla_\theta\hat{l}_1(\theta^1_{T-1}),\phi)-\theta^2_{T-1}(\phi)-m(\nabla_\theta\hat{l}_2(\theta^2_{T-1}),\phi)\| \\
\overset{(ii)}{\leq}&\|\theta^1_{T-1}(\phi)-\theta^2_{T-1}(\phi)\|+M_{m1}\|\nabla_\theta\hat{l}_1(\theta^1_{T-1})-\nabla_\theta\hat{l}_2(\theta^2_{T-1})\| \\
\leq&\|\theta^1_{T-1}(\phi)-\theta^2_{T-1}(\phi)\|+M_{m1}\|\nabla_\theta\hat{l}_1(\theta^1_{T-1})-\nabla_\theta\hat{l}_2(\theta^1_{T-1})\| \\
&+M_{m1}\|\nabla_\theta\hat{l}_2(\theta^1_{T-1})-\nabla_\theta\hat{l}_2(\theta^2_{T-1})\| \\
\overset{(iii)}{\leq}&(1+M_{m1}L)\|\theta^1_{T-1}(\phi)-\theta^2_{T-1}(\phi)\|+M_{m1}\Delta_{12},
\end{aligned}
$$

where (i) follows from Equation (1), (ii) follows from Assumption 3, (iii) follows from Assumption 4. If we iterate above inequalities from $t=0$ to $T-1$, then we obtain:

$$\|\theta^1_T(\phi)-\theta^2_T(\phi)\|\leq(1+M_{m1}L)^T\frac{\Delta_{12}}{L}-\frac{\Delta_{12}}{L}=\sigma_{\theta T}.$$

□

Lemma 9. *Based on Assumptions 3 and 4, Lemma 8, we have following inequality:*

$$\|C^1_i-C^2_i\|\leq\Delta_{Ci},$$

where $C^j_i=\nabla_2 m(\nabla_\theta\hat{l}_j(\theta_i),\phi)$ $(i=0:T,j\in\{1,2\})$ and $\Delta_{Ci}=L_{m1}(1+M_{m1}L)^i\Delta_{12}$.

Proof. Based on the definition of C^j_i, we obtain

$$
\begin{aligned}
\|C^1_i-C^2_i\|=&\|\nabla_2 m(\nabla_\theta\hat{l}_1(\theta^1_i),\phi)-\nabla_2 m(\nabla_\theta\hat{l}_2(\theta^2_i),\phi)\| \\
\leq&L_{m1}\|\nabla_\theta\hat{l}_1(\theta^1_i)-\nabla_\theta\hat{l}_2(\theta^1_i)+\nabla_\theta\hat{l}_2(\theta^1_i)-\nabla_\theta\hat{l}_2(\theta^2_i)\| \\
\overset{(i)}{\leq}&L_{m1}\Delta_{12}+L_{m1}L\|\theta^1_i-\theta^2_i\| \\
\overset{(ii)}{\leq}&L_{m1}(1+M_{m1}L)^i\Delta_{12}=\Delta_{Ci},
\end{aligned}
$$

where (i) follows from Assumption 3 and 4, (ii) follows from Lemma 8. □

Lemma 10. *Then based on Assumptions 3 and 4, Lemma 8, we have following inequality:*

$$\|D_i^1 - D_i^2\| \leq \Delta_{Di},$$

where $D_i^j = \nabla_j m(\nabla_\theta \hat{l}_j(\theta_i^j), \phi)\nabla_\theta^2 \hat{l}_j(\theta_i^j)$ $(i = 0 : T, j \in 1, 2)$, $\Delta_{Di} = M_{m1}(\rho\sigma_{\theta_i} + \widetilde{\Delta}_{12}) + L_{m1}L(1 + M_{m1}L)^i\Delta_{12}$ *and* σ_{θ_i} *is defined in Lemma 8.*

Proof. Based on the definition of D_i^j, we obtain

$$\begin{aligned}
\|D_i^1 - D_i^2\| =& \|\nabla_1 m(\nabla_\theta \hat{l}_1(\theta_i^1), \phi)\nabla_\theta^2 \hat{l}_1(\theta_i^1) - \nabla_1 m(\nabla_\theta \hat{l}_2(\theta_i^2), \phi)\nabla_\theta^2 \hat{l}_2(\theta_i^2)\| \\
=& \|\nabla_1 m(\nabla_\theta \hat{l}_1(\theta_i^1), \phi)\nabla_\theta^2 \hat{l}_1(\theta_i^1) - \nabla_1 m(\nabla_\theta \hat{l}_1(\theta_i^1), \phi)\nabla_\theta^2 \hat{l}_2(\theta_i^2) \\
&+ \nabla_1 m(\nabla_\theta \hat{l}_1(\theta_i^1), \phi)\nabla_\theta^2 \hat{l}_2(\theta_i^2) - \nabla_1 m(\nabla_\theta \hat{l}_2(\theta_i^2), \phi)\nabla_\theta^2 \hat{l}_2(\theta_i^2)\| \\
\leq& \|\nabla_1 m(\nabla_\theta \hat{l}_1(\theta_i^1), \phi)\|\|\nabla_\theta^2 \hat{l}_1(\theta_i^1) - \nabla_\theta^2 \hat{l}_2(\theta_i^2)\| \\
&+ \|\nabla_1 m(\nabla_\theta \hat{l}_1(\theta_i^1), \phi) - \nabla_1 m(\nabla_\theta \hat{l}_2(\theta_i^2), \phi)\|\|\nabla_\theta^2 \hat{l}_2(\theta_i^2)\| \\
\leq& M_{m1}\|\nabla_\theta^2 \hat{l}_1(\theta_i^1) - \nabla_\theta^2 \hat{l}_1(\theta_i^2) + \nabla_\theta^2 \hat{l}_1(\theta_i^2) - \nabla_\theta^2 \hat{l}_2(\theta_i^2)\| \\
&+ L_{m1}L\|\nabla_\theta \hat{l}_1(\theta_i^1) - \nabla_\theta \hat{l}_1(\theta_i^2) + \nabla_\theta \hat{l}_1(\theta_i^2) - \nabla_\theta \hat{l}_2(\theta_i^2)\| \\
\leq& M_{m1}(\rho\|\theta_i^1 - \theta_i^2\| + \widetilde{\Delta}_{12}) + L_{m1}L(L\sigma_{\theta i} + \Delta_{12}) \\
\overset{(i)}{\leq}& M_{m1}(\rho\sigma_{\theta_i} + \widetilde{\Delta}_{12}) + L_{m1}L(1 + M_{m1}L)^i\Delta_{12} = \Delta_{Di},
\end{aligned}$$

where (i) follow from Lemma 8. □

Lemma 11. *Based on Assumptions 3, 4 and Lemma 1, we obtain*

$$\begin{aligned}
&\|\nabla_\phi \theta_T^1(\phi) - \nabla_\phi \theta_T^2(\phi)\| \\
&\leq \sum_{i=0}^{T-1}\left((1 + M_{m1}L)^{T-i-1}\Delta_{Ci} + M_{m2}(1 + M_{m1}L)^{T-i-2}\sum_{j=i+1}^{T-1}\Delta_{Dj}\right),
\end{aligned}$$

where Δ_{Ci} *and* Δ_{Dj} *have been defined in Lemmas 9 and 10.*

Proof. Based on the Lemma 1, we obtain

$$\begin{aligned}
&\|\nabla_\phi \theta_T^1(\phi) - \nabla_\phi \theta_T^2(\phi)\| \\
&\overset{(i)}{\leq} \sum_{i=0}^{T-1}\Bigg\|\prod_{j=i+1}^{T-1}(I + D_{T+i-j}^1)C_i^1 - \prod_{j=i+1}^{T-1}(I + D_{T+i-j}^1)C_i^2 + \prod_{j=i+1}^{T-1}(I + D_{T+i-j}^1)C_i^2 \\
&\quad - \prod_{j=i+1}^{T-1}(I + D_{T+i-j}^2)C_i^2\Bigg\| \\
&\leq \sum_{i=0}^{T-1}\left(\left\|\prod_{j=i+1}^{T-1}(I + D_{T+i-j}^1)\right\|\|C_i^1 - C_i^2\| + \left\|\prod_{j=i+1}^{T-1}(I + D_{T+i-j}^1)\right.\right. \\
&\quad \left.\left. - \prod_{j=i+1}^{T-1}(I + D_{T+i-j}^2)\right\|\|C_i^2\|\right) \\
&\overset{(ii)}{\leq} \sum_{i=0}^{T-1}\Bigg((1 + M_{m1}L)^{T-i-1}\|C_i^1 - C_i^2\| + M_{m2}(1 + M_{m1}L)^{T-i-2} \\
&\quad \sum_{j=i+1}^{T-1}\|D_{T+i-j}^1 - D_{T+i-j}^2\|\Bigg) \\
&\overset{(iii)}{\leq} \sum_{i=0}^{T-1}\left((1 + M_{m1}L)^{T-i-1}\Delta_{Ci} + M_{m2}(1 + M_{m1}L)^{T-i-2}\sum_{j=i+1}^{T-1}\Delta_{Dj}\right),
\end{aligned}$$

where (i) follows from the definitions that $D_i^j = \nabla_j m(\nabla_\theta \hat{l}_j(\theta_i^j), \phi)\nabla_\theta^2 \hat{l}_j(\theta_i^j)$, $C_i^j = \nabla_2 m(\nabla_\theta \hat{l}_j(\theta_i), \phi)$, (ii) follows from Assumption 3 and (iii) follows from Lemma 9 and 10. □

Lemma 12. ***(Correspond to Proposition 1)*** *Based on Assumptions 3 and 4, Lemmas 8 and 11, we obtain*

$$\|\nabla_\phi \hat{g}_T^1(\phi) - \nabla_\phi \hat{g}_T^2(\phi)\| = \mathcal{O}(TQ^{T-1}\widetilde{\Delta}_{12} + Q^{2T-1}\Delta_{12}),$$

where $Q = 1 + M_{m1}L$.

Proof. We first consider Δ_{Ci} and Δ_{Di}, we obtain

$$\begin{aligned}
\Delta_{Ci} =& L_{m1}(1 + M_{m1}L)^i\Delta_{12}) = \mathcal{O}(Q^i\Delta_{12}), && (12)\\
\Delta_{Di} =& \mathcal{O}(M_{m1}(\rho\sigma_{\theta_i} + \widetilde{\Delta}_{12}) + L_{m1}L(1 + M_{m1}L)^i\Delta_{12})\\
\overset{(i)}{=}& \mathcal{O}(Q^i\Delta_{12} + \widetilde{\Delta}_{12} + Q^i\Delta_{12}) = \mathcal{O}(Q^i\Delta_{12} + \widetilde{\Delta}_{12}), && (13)
\end{aligned}$$

where (i) follows because $\sigma_{\theta i} = (1 + M_{m1}L)^i\frac{\Delta_{12}}{L} - \frac{\Delta_{12}}{L} = \mathcal{O}(Q^i\Delta_{12})$.

Furthermore, we consider the uniform bound for $\|\nabla_\phi \theta_T^1(\phi) - \nabla_\phi \theta_T^2(\phi)\|$, then we obtain

$$\begin{aligned}
&\|\nabla_\phi \theta_T^1(\phi) - \nabla_\phi \theta_T^2(\phi)\|\\
&\overset{(i)}{=}\mathcal{O}\left(\sum_{i=0}^{T-1}\left(Q^{T-i-2}\left(Q\Delta_{Ci} + \sum_{j=i+1}^{T-1}\Delta_{D(T+i-j)}\right)\right)\right)\\
&\overset{(ii)}{=}\mathcal{O}\left(\sum_{i=0}^{T-1}\left(Q^{T-i-1}Q^i\Delta_{12} + Q^{T-i-2}\sum_{j=i+1}^{T-1}(Q^{T+i-j}\Delta_{12} + \widetilde{\Delta}_{12})\right)\right)\\
&=\mathcal{O}\left(\sum_{i=0}^{T-1}\left(Q^{T-1}\Delta_{12} + (T-i-1)Q^{T-i-2}\widetilde{\Delta}_{12} + (Q^{2T-i-2} - Q^{T-1})\Delta_{12}\right)\right)\\
&=\mathcal{O}\left(\sum_{i=0}^{T-1}((T-i-1)Q^{T-i-2}\widetilde{\Delta}_{12} + Q^{2T-i-2}\Delta_{12})\right)\\
&\overset{(iii)}{=}\mathcal{O}\left(\sum_{j=0}^{T-1}(jQ^{j-1}\widetilde{\Delta}_{12} + Q^{T+j-1}\Delta_{12})\right)\\
&=\mathcal{O}\left(TQ^{T-1}\widetilde{\Delta}_{12} + Q^{2T-1}\Delta_{12}\right),
\end{aligned}$$

where (i) follows from Lemma 11, (ii) follows from Equation (12) and Equation (13), (iii) follows because $j = T - i - 1$. Based on the formulation of $\nabla_\phi \hat{g}_T(\phi)$ in Lemma 6, we have

$$\begin{aligned}
\|\nabla_\phi \hat{g}_T^1(\phi) - \nabla_\phi \hat{g}_T^2(\phi)\| \leq& M\|\nabla_\phi \theta_T^1(\phi) - \nabla_\phi \theta_T^2(\phi)\| + M_{\theta T}Q^T\Delta_{12}\\
\overset{(i)}{=}& \mathcal{O}(TQ^{T-1}\widetilde{\Delta}_{12} + Q^{2T-1}\Delta_{12}),
\end{aligned}$$

where (i) follows because $M_{\theta T}$ defined in Lemma 2 satisfies that $M_{\theta T} = \mathcal{O}(Q^{T-1})$. □

Lemma 13. *Based on the Assumption 3 and Lemma 2, we obtain*

$$\|g_T(\phi_1) - g_T(\phi_2)\| \leq M_{g_T}\|\phi_1 - \phi_2\|,$$

where $M_{g_T} = MM_{\theta T}$ *and* $M_{\theta T}$ *is defined in Lemma 2.*

Proof. Based on the definition of $g_T(\phi)$, we have

$$\begin{aligned}
\|g_T(\phi_1) - g_T(\phi_2)\| =& \|l(\theta_T(\phi_1)) - l(\theta_T(\phi_2))\|\\
\overset{(i)}{\leq}& M\|\theta_T(\phi_1) - \theta_T(\phi_2)\|
\end{aligned}$$

$$\begin{aligned}&\overset{(ii)}{\leq} MM_{\theta T}\|\phi_1-\phi_2\|\\ &=M_{g_T}\|\phi_1-\phi_2\|,\end{aligned}$$

where (i) is based on Assumption 3, (ii) is based on Lemma 2. □

Lemma 14. *Based on the proposition 1 in Fallah et al. (2021), Assumptions 1 and 3, if we set* $\alpha \leq \min\{\frac{1}{2L}, \frac{\mu}{8\rho_{g_T}M_{g_T}}\}$, $\beta_k = \min(\beta, \frac{8}{\mu(k+1)})$ *for* $\beta \leq \frac{8}{\mu}$ *in Algorithm 1, then we have*

$$\mathbb{E}\|\widetilde{\phi}^1_{MK} - \widetilde{\phi}^1_{M*}\|^2 \leq \mathcal{O}(1)\frac{M^2_{g_T}(1+\frac{1}{\beta\mu})}{\mu^3}\left(\frac{L_{g_T}+\rho_{g_T}\alpha M_{g_T}}{K}+\frac{M_{g_T}}{\sqrt{K}}\right),$$

where M_{g_T} *is defined in Lemma 13,* L_{g_T} *is defined in Lemma 6,* ρ_{g_T} *is defiend in Lemma 7.*

Proof. Based on the Proposition 1 in Fallah et al. (2021), we obtain

$$\mathbb{E}\big[\hat{G}^1_T(\widetilde{\phi}^1_{MK}) - \hat{G}^1_T(\widetilde{\phi}^1_{M*})\big] \leq \mathcal{O}(1)\frac{M^2_{g_T}(1+\frac{1}{\beta\mu})}{\mu^2}\left(\frac{L_{g_T}+\rho_{g_T}\alpha M_{g_T}}{K}+\frac{M_{g_T}}{\sqrt{K}}\right),$$

where $\hat{G}_T(\phi)$ is defined in Equation (5). Based on the Assumption 1 and the fact that $\widetilde{\phi}^1_{M*} = \arg\min_\phi \hat{G}^1_T(\phi)$, we have

$$\begin{aligned}\mathbb{E}\|\widetilde{\phi}^1_{MK} - \widetilde{\phi}^1_{M*}\|^2 \leq& \frac{2}{\mu}\mathbb{E}\left(\hat{G}^1_T(\widetilde{\phi}^1_{MK}) - \hat{G}^1_T(\widetilde{\phi}^1_{M*})\right)\\ \leq& \mathcal{O}(1)\frac{M^2_{g_T}(1+\frac{1}{\beta\mu})}{\mu^3}\left(\frac{L_{g_T}+\rho_{g_T}\alpha M_{g_T}}{K}+\frac{M_{g_T}}{\sqrt{K}}\right).\end{aligned}$$

□

Lemma 15. *Based on Assumption 1 and Lemma 13, we have*

$$\|\widetilde{\phi}^1_* - \phi^1_*\| \leq \frac{2\sqrt{2}MM_{\theta T}}{\mu\sqrt{\delta N}},$$

where N *is the sample size.*

Proof. Based on Assumption 1 and Lemma 13, from Theorem 2 in Shalev-Shwartz et al. (2010), with probability at least $1-\delta$, we have

$$g^1_T(\widetilde{\phi}^1_*) - g^1_T(\phi^1_*) \leq \frac{4M^2_{g_T}}{\delta\mu N}.$$

Furthermore, based on Assumption 1 and the fact that $\phi^1_* = \arg\min g^1_T(\phi)$, we obtain

$$\|\widetilde{\phi}^1_* - \phi^1_*\|^2 \leq \frac{2}{\mu}\left(g^1_T(\widetilde{\phi}^1_*) - g^1_T(\phi^1_*)\right) \leq \frac{2}{\mu}\frac{4M^2_{g_T}}{\delta\mu N} = \frac{8M^2_{g_T}}{\delta\mu^2 N}.$$

We take the square root from both side and obtain:

$$\|\widetilde{\phi}^1_* - \phi^1_*\| \leq \frac{2\sqrt{2}M_{g_T}}{\mu\sqrt{\delta N}},$$

with probability at least $1-\delta$. □

A.3 Proof of Theorem 1

Based on our definition of generalization error for the algorithm,

$$\begin{aligned}&g^3_T(\widetilde{\phi}^1_{MK} - \alpha\nabla_\phi\hat{g}^2_T(\widetilde{\phi}^1_{MK})) - g^3_T(\phi^3_*)\\ &\overset{(i)}{\leq} g^3_T(\widetilde{\phi}^1_{MK} - \widetilde{\phi}^1_{M*} + \widetilde{\phi}^1_* + \alpha\nabla_\phi\hat{g}^1_T(\widetilde{\phi}^1_{M*}) - \alpha\nabla_\phi\hat{g}^2_T(\widetilde{\phi}^1_{MK})) - g^3_T(\phi^3_*)\end{aligned}$$

$$
\begin{aligned}
&\overset{(ii)}{\leq} M_{g_T}\|\widetilde{\phi}^1_{MK} - \widetilde{\phi}^1_{M*} + \widetilde{\phi}^1_* - \phi^1_* + \phi^1_* - \phi^3_* + \alpha\nabla_\phi \hat{g}^1_T(\widetilde{\phi}^1_{M*}) - \alpha\nabla_\phi \hat{g}^2_T(\widetilde{\phi}^1_{MK})\| \qquad (14)\\
&\leq M_{g_T}\|\widetilde{\phi}^1_{MK} - \widetilde{\phi}^1_{M*}\| + M_{g_T}\|\widetilde{\phi}^1_* - \phi^1_*\| + M_{g_T}\|\phi^1_* - \phi^3_*\|\\
&\quad + M_{g_T}\alpha\|\nabla_\phi \hat{g}^1_T(\widetilde{\phi}^1_{M*}) - \nabla_\phi \hat{g}^1_T(\widetilde{\phi}^1_{MK})\| + M_{g_T}\alpha\|\nabla_\phi \hat{g}^1_T(\widetilde{\phi}^1_{MK}) - \nabla_\phi \hat{g}^2_T(\widetilde{\phi}^1_{MK})\|\\
&\leq (M_{g_T} + M_{g_T}L_{g_T}\alpha)\|\widetilde{\phi}^1_{MK} - \widetilde{\phi}^1_{M*}\| + M_{g_T}\|\widetilde{\phi}^1_* - \phi^1_*\| + M_{g_T}\|\phi^1_* - \phi^3_*\|\\
&\quad + M_{g_T}\alpha\|\nabla_\phi \hat{g}^1_T(\widetilde{\phi}^1_{MK}) - \nabla_\phi \hat{g}^2_T(\widetilde{\phi}^1_{MK})\|,
\end{aligned}
$$

where (i) follows from Assumption 2, (ii) follows from Lemma 13.

Furthermore, considering Algorithm 1, if we set $\alpha \leq \min\{\frac{1}{2L}, \frac{\mu}{8\rho_{g_T}M_{g_T}}\}$, $\beta_k = \min(\beta, \frac{8}{\mu(k+1)})$ for $\beta \leq \frac{8}{\mu}$, based on Lemma 12, 14, 15, with probability at least $1-\delta$, we obtain

$$
\begin{aligned}
&\mathbb{E}[g^3_T(\widetilde{\phi}^1_{MK} - \alpha\nabla_\phi \hat{g}^2_T(\widetilde{\phi}^1_{MK})) - g^3_T(\phi^3_*)]\\
&\leq (M_{g_T} + M_{g_T}L_{g_T}\alpha)\left\|\mathcal{O}(1)\frac{M_{g_T}}{\mu^2}\sqrt{\frac{L_{g_T} + \rho_{g_T}\alpha M_{g_T}}{\beta K} + \frac{M_{g_T}}{\beta\sqrt{K}}}\right\| + M_{g_T}\left\|\frac{2\sqrt{2}M_{g_T}}{\mu\sqrt{\delta N}}\right\|\\
&\quad + M_{g_T}\delta_{13} + M_{g_T}\alpha\mathcal{O}(TQ^{T-1}\widetilde{\Delta}_{12} + Q^{2T-1}\Delta_{12}),
\end{aligned}
$$

where $\delta_{13} = \|\phi^1_* - \phi^3_*\|$, $Q = (1 + M_{m1}L)$, K is the step number for update, N is the sample size for training.

Then for Lipschitz term M_{g_T} defined in Lemma 13,

$$
M_{g_T} = MM_{\theta T} = \mathcal{O}(Q^{T-1}),
$$

where $M_{\theta T}$ defined in Lemma 2 satisfies $M_{\theta T} = \mathcal{O}(Q^{T-1})$.

For Lipschitz term L_{g_T} defined in Lemma 6, we first compute the order for $L_{\theta T}$ which is defined in Lemma 5, then we obtain

$$
\begin{aligned}
L_{\theta T} =& \mathcal{O}\left(\sum_{i=0}^{T-1} Q^{T-i-1}M_{A_i} + \sum_{i=0}^{T-1} Q^{T-i-2}\sum_{j=i+1}^{T-1} M_{B(T+i-j)}\right)\\
=& \mathcal{O}\left(\sum_{i=0}^{T-1} Q^{T-i-1}Q^{i-1} + \sum_{i=0}^{T-1} Q^{T-i-2}\sum_{j=i+1}^{T-1} Q^{T+i-j-1}\right)\\
\overset{(i)}{=}& \mathcal{O}\left(\sum_{i=0}^{T-1} Q^{T-2} + \sum_{i=0}^{T-1} Q^{T-i-2}Q^{T-1}\right) = \mathcal{O}(TQ^{T-2} + Q^{2T-2}),
\end{aligned}
$$

where (i) follows from Lemmas 3 and 4. Then, we obtain

$$
\begin{aligned}
L_{g_T} =& ML_{\theta T} + LM^2_{\theta T}\\
=& \mathcal{O}(TQ^{T-2} + Q^{2T-2} + Q^{2T-2}) = \mathcal{O}(TQ^{T-2} + Q^{2T-2}).
\end{aligned}
$$

For Lipschitz term ρ_{g_T} defined in Lemma 7, we have

$$
\rho_{g_T} = 3LM_{\theta T}L_{\theta T} + M\rho_\theta + M^3_{\theta T}\rho = \mathcal{O}(TQ^{2T-3} + Q^{2T-3}).
$$

Then, the proof is complete.

A.4 Proof of Remark 2

In terms of M-L2O generalization error, based on the Equation (14) in Appendix A.3, we have

$$
\begin{aligned}
&g^3_T(\widetilde{\phi}^1_{MK} - \alpha\nabla_\phi \hat{g}^2_T(\widetilde{\phi}^1_{MK})) - g^3_T(\phi^3_*)\\
&\leq M_{g_T}\|\widetilde{\phi}^1_{MK} - \widetilde{\phi}^1_{M*} + \widetilde{\phi}^1_* - \phi^1_* + \phi^1_* - \phi^3_* + \alpha\nabla_\phi \hat{g}^1_T(\widetilde{\phi}^1_{M*}) - \alpha\nabla_\phi \hat{g}^2_T(\widetilde{\phi}^1_{MK})\|\\
&\leq M_{g_T}\|\widetilde{\phi}^1_{MK} - \widetilde{\phi}^1_{M*}\| + M_{g_T}\|\widetilde{\phi}^1_* - \phi^1_*\| + M_{g_T}\alpha\|\nabla_\phi \widetilde{g}^2_T(\widetilde{\phi}^1_{MK}) - \nabla_\phi \widetilde{g}^1_T(\widetilde{\phi}^1_{M*})\| + M_{g_T}\delta_{13},
\end{aligned}
$$

where $\delta_{13} = \|\phi_*^1 - \phi_*^3\|$.

In terms of Transfer Learning L2O generalization error with learned initial point $\widetilde{\phi}_K$, we have

$$
\begin{aligned}
&g_T^3(\widetilde{\phi}_K^1 - \alpha\nabla_\phi \hat{g}_T^2(\widetilde{\phi}_K^1)) - g_T^3(\phi_*^3) \\
\leq& M_{g_T}\|\widetilde{\phi}_K^1 - \alpha\nabla_\phi \hat{g}_T^2(\widetilde{\phi}_K^1) - \phi_*^3\| \\
\overset{(i)}{=}& M_{g_T}\|\widetilde{\phi}_K^1 - \alpha\nabla_\phi \hat{g}_T^2(\widetilde{\phi}_K^1) - (\widetilde{\phi}_{M*}^1 - \alpha\nabla_\phi \hat{g}_T^1(\widetilde{\phi}_{M*}^1)) + \widetilde{\phi}_*^1 - \phi_*^3\| \\
\leq& M_{g_T}\|\widetilde{\phi}_K^1 - \widetilde{\phi}_{M*}^1\| + M_{g_T}\|\widetilde{\phi}_*^1 - \phi_*^1\| + M_{g_T}\alpha\|\nabla_\phi \widetilde{g}_T^2(\widetilde{\phi}_K^1) - \nabla_\phi \widetilde{g}_T^1(\widetilde{\phi}_{M*}^1)\| + M_{g_T}\delta_{13},
\end{aligned}
$$

where (i) follows from $\widetilde{\phi}_*^1 = \widetilde{\phi}_{M*}^1 - \alpha\nabla_\phi \hat{g}_T^1(\widetilde{\phi}_{M*}^1)$, $\delta_{13} = \|\phi_*^1 - \phi_*^3\|$. Then, the proof is complete.

A.5 Proof of eq. 8 in Subsection 5.3

We assume that $\widetilde{\phi}_*^3 = \widetilde{\phi}_{M*}^3 - \alpha\nabla_\phi \hat{g}_T^3(\widetilde{\phi}_{M*}^3)$, then we have

$$
\begin{aligned}
&g_T^3(\widetilde{\phi}_{MK}^1 - \alpha\nabla_\phi \hat{g}_T^2(\widetilde{\phi}_{MK}^1)) - g_T^3(\phi_*^3) \\
\leq& M_{g_T}\|\widetilde{\phi}_{MK}^1 - \alpha\nabla_\phi \hat{g}_T^2(\widetilde{\phi}_{MK}^1) - \phi_*^3\| \\
\leq& M_{g_T}\|\widetilde{\phi}_{MK}^1 - \alpha\nabla_\phi \hat{g}_T^2(\widetilde{\phi}_{MK}^1) - \widetilde{\phi}_{M*}^3 + \alpha\nabla_\phi \hat{g}_T^3(\widetilde{\phi}_{M*}^3) + \widetilde{\phi}_*^3 - \phi_*^3\| \\
\leq& M_{g_T}\|\widetilde{\phi}_{MK}^1 - \widetilde{\phi}_{M*}^1\| + M_{g_T}\|\widetilde{\phi}_*^3 - \phi_*^3\| + M_{g_T}\alpha\|\nabla_\phi \hat{g}_T^2(\widetilde{\phi}_{MK}^1) - \nabla_\phi \hat{g}_T^3(\widetilde{\phi}_{M*}^3)\| \\
&+ M_{g_T}\|\widetilde{\phi}_{M*}^1 - \widetilde{\phi}_{M*}^3\|.
\end{aligned}
$$

Then, the proof is complete.

A.6 Additional Experiments

New Optimizees: Rosenbrock We conduct additional experiments with substantially different optimizees, *i.e.* Rosenbrock (Rosenbrock, 1960). In this case, the optimizes are required to minimize a two-dimensional non-convex function taking the following formulation:

$$
f(x, y) = (x - 1)^2 + 100(y - x^2)^2, \tag{15}
$$

which is challenging for algorithms to converge to the global minimum (Tani et al., 2021).

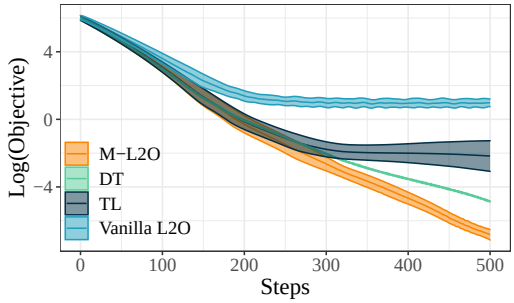
We specify D_{adapt} and D_{test} to be the family of Rosenbrock optimizees with randomly sampled initial points from standard normal distribution. In contrast, the training optimizees are still LASSO with a mixture of uniform distribution from which the coefficient matrices are sampled. The experiments are repeated for 10 times, with all the algorithms receiving identical adaptation and testing samples in each run. Figure A5a shows the curves of the logarithm of the objective values generated by different methods, where our proposed M-L2O outperforms other baselines significantly. At 500-th step, the (mean, standard deviation) of the logarithmic objective values for {Vanilla L2O, TL, DT, M-L2O}are $\{(0.977, 0.225), (-2.170, 1.312), (-4.864, 0.395), (-6.832, 0.445)\}$, which provides numerical supports of the advantage of our methods.

New Evaluation: Interpolation

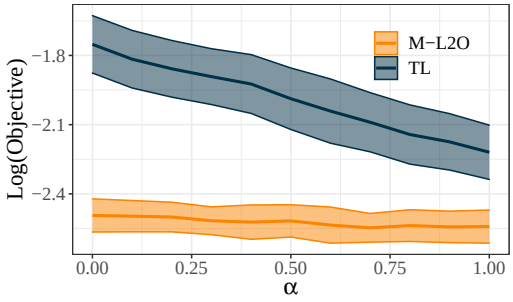
To obtain new optimize weights, we employ a linear interpolation strategy between two adapted optimizers. The first one is optimized on the optimizees that are similar to those used in training, and the second is optimized on the optimizees that are similar to those used in testing. We introduce a factor α to control the interpolation between the two weights, denoted by $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$, respectively, and caluclate the new weights as follows:

$$
\boldsymbol{w} = \alpha\boldsymbol{w}_1 + (1 - \alpha)\boldsymbol{w}_2.
$$

In Figure A5b, we present the mean values of the logarithmic loss, as well as the 95% confidence interval. The results of TL and M-L2O validate our claim that adapting to training-like optimizees tend to yield better performance than adapting to optimizees that more resemble the testing optimizees.



(a) Convergence speeds on Rosenbrock optimizees. We repeat the experiments for 10 times, and present the 95% confidence intervals are shown in the figure.



(b) Convergence speeds on LASSO optimizees, with different interpolation weights α. Both the mean and the 95% confidence intervals are shown in the figure.

Figure A5: Visualization of additional experiment results.