DEBIASING A FIRST-ORDER HEURISTIC FOR APPROXIMATE BI-LEVEL OPTIMIZATION: SUPPLEMENTARY MATERIALS

A Related work

Unbiased gradient estimation. Stochastic gradient descent (SGD) (Bottou et al., 2018) is an essential component of large-scale machine learning. Unbiased gradient estimation, as a part of SGD, guarantees convergence to a stationary point of the optimization objective. For this reason, many algorithms were proposed to perform unbiased gradient estimation in various applications, e.g. REINFORCE (Williams, 1992) and its low-variance modifications (Tucker et al., 2017; Gu et al., 2016) with applications in reinforcement learning and evolution strategies (Wierstra et al., 2014). The variational autoencoder (Kingma & Welling, 2014) and variational dropout (Kingma et al., 2015) are based on a reparametrization trick for unbiased back-propagation through continuous or, involving a relaxation (Jang et al., 2016; Gal et al., 2017), discrete random variables.

Theory of meta-learning. Our proof technique fits into the realm of theoretical understanding for meta-learning, which has been explored in (Fallah et al., 2019; Ji et al., 2020) for nonconvex functions (see also (Ablin et al., 2020) for convergence analysis in certain bi-level optimization setups), as well as (Finn et al., 2019; Balcan et al., 2019) for convex functions and their extensions, such as online convex optimization (Hazan, 2019). While (Fallah et al., 2019) provides a brief counterexample for which $(r = 1)$-step FOM does not converge, we establish a rigorous non-convergence counterexample proof for FOM with any number of steps r when using *stochastic* gradient descent. Our proof is based on arguments using expectations and probabilities, providing new insights into stochastic optimization during meta-learning. Furthermore, while (Ji et al., 2020) touches on the *zero-order* case found in (Song et al., 2020), which is mainly focused on reinforcement learning, our work studies the case where exact gradients are available, which is suited for supervised learning.

B Synthetic Experiment Details

For the synthetic experiments shown in Figures 1a, 1b, and 1c, we set the following parameters from Theorem 1 and proof of Theorem 2:

$$r = 10, \quad \alpha = 0.1, \quad q = 0.1 \text{ (UFOM)}, \quad \forall k \in \mathbb{N} : \gamma_k = \frac{10}{k},$$
$$a_1 = 0.5, \quad a_2 = 1.5, \quad b_1 = 0, \quad b_2 = 17.39, \quad A = 12.59$$

(b_2 and A values are obtained by setting $D = 0.06$ in the Theorem 2 proof). We do 5 simulations for FOM and UFOM, where we sample θ_0 from a uniform distribution on a segment $[-10, 30]$.

To demonstrate a wider range of q^* values, for Figures 1d, 1e, and 1f, we opt for a slightly different set of parameters:

$$r = 10, \quad \alpha = 0.1, \quad \forall k \in \mathbb{N} : \gamma_k = \frac{10}{k}, \quad a_1 = 0.5, \quad a_2 = 1.5, \quad b_2 = 10, \quad A = 10.$$

To approximate $\mathbb{V}^2, \mathbb{D}^2$ on Figure 1d, we find a maximal value of the corresponding expectation (computed precisely for two tasks with equal probability) on a grid of 10000 θ values on $[-50, 50]$.

To output the "experiment" curve on Figure 1e, for each value of α, we search for q on a grid of 20 elements between 0.02 and 0.4. For each q on a grid we simulate 10000 SGD loops (100 iterations each) from a starting point drawn uniformly on $[-50, 50]$. Then we compute the average of the curves corresponding to $|\frac{\partial}{\partial\theta}\mathcal{M}^{(p)}|$ for these 10000 simulations. Given the best q for each α, we choose the one which achieves the minimal average value of $|\frac{\partial}{\partial\theta}\mathcal{M}^{(p)}|$ in the fastest time, computed for $q = 0.02$.

For Figure 1f, we report mean and standard error over 1000 curves starting from a point drawn uniformly on $[-50, 50]$.

C Data Hypercleaning Details

Validation loss and ϕ_0 do not depend on θ, and thus to use UFOM, we include the last inner optimization step into the definition of $\mathcal{L}^{out}$. This implies that $\mathcal{L}^{out}$ depends on θ. We partition the original MNIST train set into sets of size

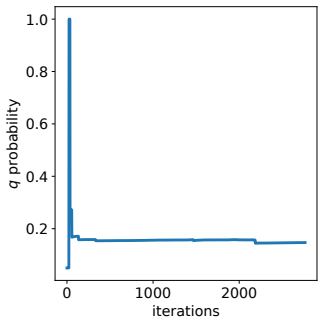

Figure 4: Adaptive q probabilities during training for the hypercleaning setup.

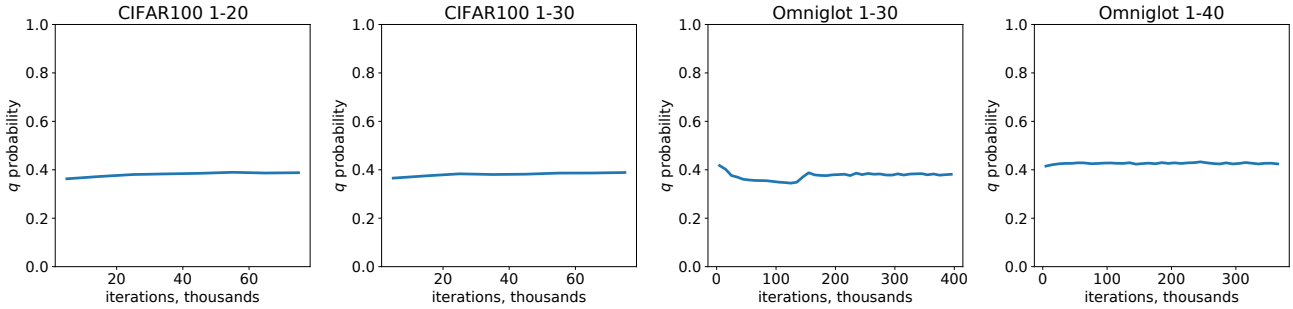

Figure 5: Adaptive q probabilities during training for the few-shot learning setup.

5000 for the hyperclearning task's train and validation sets. We use the MNIST test set for testing. We corrupt half of training examples by drawing labels uniformly from $0, \ldots, 9$. For the classifier, we use a 2-layer feedforward network with dimensions $784 \to 256 \to 10$, with ReLU nonlinearities. We use Adam (Kingma & Ba, 2014) as an outer-loop optimizer with a learning rate of 0.1. We run all methods for a number of function calls equivalent to 500 outer-loop iterations for the memory-efficient exact gradient.

Figure 4 demonstrates adaptively chosen q during optimization using Adaptive UFOM. We observe that q stabilizes soon in the beginning of optimization and doesn't change much during training. This could mean that statistics $\overline{\mathbb{D}^2}, \overline{\mathbb{V}^2}$ are roughly the same along the whole optimization trajectory.

D Few-Shot Learning Details

All results are reported in a transductive setting (Nichol et al., 2018). In all setups for Reptile, we reuse the code from (Nichol et al., 2018). For the exact ABLO and Adaptive UFOM, we clip each entry of the gradient to be in $[-0.1, 0.1]$. We use the following hyperparameters for the two datasets:

- **Omniglot**. We run all methods for the number of function calls equivalent to $\tau = 200000$ outer iterations of the memory-efficient exact ABLO. For exact ABLO/FOM/Adaptive UFOM, we set $\forall k : \gamma_k = 0.1$, meta-batch size of 5, $\alpha = 0.005$. In all setups for Reptile, we set hyperparameter values to be equal to the ones found in the 1-shot 20-way case (Nichol et al., 2018). This is because Reptile underperforms if its hyperparameters are set to the values used for exact ABLO/FOM/Adaptive UFOM. We take train and test splits as in (Finn et al., 2017; Nichol et al., 2018).

- **CIFAR100**. We use the same hyperparameters as in Omniglot setup, but run all methods for the number of function calls equivalent to $\tau = 40000$ outer iterations of the memory-efficient exact ABLO. For a train-test split, we combine CIFAR100's train and test sets and randomly split classes into 80 train and 20 test classes.

Figure 5 demonstrates adaptively chosen q during optimization using Adaptive UFOM. Again, we observe that q stabilizes and doesn't change much during training, most probably meaning that statistics $\overline{\mathbb{D}^2}, \overline{\mathbb{V}^2}$ are roughly the same along the optimization trajectory.

E Proofs

In this section, we provide proofs for Theorems 1 and 2 from the main body of the paper.

E.1 Theorem 1

We start by formulating and proving three helpful lemmas. In proofs we use the fact that, as a direct consequence of Assumption 1, for all $\theta \in \mathbb{R}^s, \phi \in \mathbb{R}^p, \mathcal{T} \in \Omega_{\mathcal{T}}$

$$\max(\|\frac{\partial^2}{\partial\theta\partial\phi}\mathcal{L}^{in}(\theta,\phi,\mathcal{T})\|_2, \|\frac{\partial^2}{\partial\phi^2}\mathcal{L}^{in}(\theta,\phi,\mathcal{T})\|_2) \leq L_2.$$

Lemma 1. *Let $p, r, s \in \mathbb{N}$, $\{\alpha_j > 0\}_{j=1}^{\infty}$ be any sequence, $q \in (0,1]$, $p(\mathcal{T})$ be a distribution on a nonempty set $\Omega_{\mathcal{T}}$, $\xi \sim \text{Bernoulli}(q)$ be independent of $p(\mathcal{T})$, $V : \mathbb{R}^s \times \Omega_{\mathcal{T}} \to \mathbb{R}^p, \mathcal{L}^{in}, \mathcal{L}^{out} : \mathbb{R}^s \times \mathbb{R}^p \times \Omega_{\mathcal{T}} \to \mathbb{R}$ be functions satisfying Assumption 1, and let $U^{(r)} : \mathbb{R}^s \times \Omega_{\mathcal{T}} \to \mathbb{R}^p$ be defined according to (2-3), $\mathcal{M}^{(r)} : \mathbb{R}^s \to \mathbb{R}$ be defined according to (4) and satisfy Assumption 2. Define $\mathcal{G}_{FO} : \mathbb{R}^s \times \Omega_{\mathcal{T}} \to \mathbb{R}^s$ and $\mathcal{G} : \mathbb{R}^s \times \Omega_{\mathcal{T}} \times \{0,1\} \to \mathbb{R}^s$ as*

$$\mathcal{G}_{FO}(\theta,\mathcal{T}) = \frac{\partial}{\partial\theta}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T}) + (\frac{\partial}{\partial\theta}V(\theta,\mathcal{T}))^{\top}\frac{\partial}{\partial\phi}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T}), \quad \phi_r = U^{(r)}(\theta,\mathcal{T}),$$
$$\mathcal{G}(\theta,\mathcal{T},x) = \mathcal{G}_{FO}(\theta,\mathcal{T}) + \frac{x}{q}(\nabla_\theta\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T}) - \mathcal{G}_{FO}(\theta,\mathcal{T})).$$

Then

$$\mathbb{D}^2 \leq \mathbb{D}^2_{bound}, \tag{25}$$

$$\mathbb{V}^2 \leq \mathbb{V}^2_{bound}, \quad \mathbb{V}_{bound} = L_1 + M_1L_1\prod_{j=1}^{r}(1+\alpha_jL_2) + L_1L_2\sum_{j=1}^{r}\alpha_j\prod_{j'=j}^{r}(1+\alpha_{j'}L_2), \tag{26}$$

where $\mathbb{D}, \mathbb{V}, \mathbb{D}_{bound}$ are defined in (14), (16), (15) respectively. Further, for all $\theta \in \mathbb{R}^s$

$$\mathbb{E}_{\xi,p(\mathcal{T})}\left[\mathcal{G}(\theta,\mathcal{T},\xi)\right] = \frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta), \tag{27}$$

$$\mathbb{E}_{\xi,p(\mathcal{T})}\left[\|\mathcal{G}(\theta,\mathcal{T},\xi)\|_2^2\right] \leq (\frac{1}{q}-1)\mathbb{D}^2 + \mathbb{V}^2. \tag{28}$$

Proof. First, we show (25). Let $\phi_0, \ldots, \phi_r$ be inner-GD rollouts (2-3) corresponding to θ and $\mathcal{T}$. Observe that by Assumption 1

$$\|\frac{\partial}{\partial\phi}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T})\|_2 = \|\nabla_{\phi_r}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T})\|_2 \leq L_1$$

and according to (8-9) for each $1 \leq j \leq r$

$$\begin{aligned}
\|\nabla_{\phi_{j-1}}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T})\|_2 &= \|\nabla_{\phi_j}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T}) - \alpha_j\left(\frac{\partial^2}{\partial\phi^2}\mathcal{L}^{in}(\theta,\phi_{j-1},\mathcal{T})\right)^{\top}\nabla_{\phi_j}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T})\|_2 \\
&= \|\nabla_{\phi_j}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T})\|_2 + \alpha_j\|\frac{\partial^2}{\partial\phi^2}\mathcal{L}^{in}(\theta,\phi_{j-1},\mathcal{T})\|_2\|\nabla_{\phi_j}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T})\|_2 \\
&\leq (1+\alpha_jL_2)\|\nabla_{\phi_j}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T})\|_2 \leq \cdots \leq \|\nabla_{\phi_r}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T})\|_2\prod_{j'=j}^{r}(1+\alpha_{j'}L_2) \\
&\leq L_1\prod_{j'=j}^{r}(1+\alpha_{j'}L_2).
\end{aligned} \tag{29}$$

In addition, we deduce that

$$\|\nabla_{\phi_0}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T}) - \frac{\partial}{\partial\phi}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T})\|_2 = \|\nabla_{\phi_1}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T}) - \alpha_1\left(\frac{\partial^2}{\partial\phi^2}\mathcal{L}^{in}(\theta,\phi_0,\mathcal{T})\right)^{\top}\nabla_{\phi_1}\mathcal{L}^{out}(\theta,\phi_r,\mathcal{T})$$

$$
\begin{aligned}
&- \frac{\partial}{\partial \phi}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\|_2 \\
&\leq \|\nabla_{\phi_1}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}) - \frac{\partial}{\partial \phi}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\|_2 + \alpha_1 \|\frac{\partial^2}{\partial \phi^2}\mathcal{L}^{in}(\theta, \phi_0, \mathcal{T})\|_2 \|\nabla_{\phi_1}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\|_2 \\
&\leq \|\nabla_{\phi_1}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}) - \frac{\partial}{\partial \phi}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\|_2 + \alpha_1 L_1 L_2 \prod_{j'=1}^{r}(1 + \alpha_{j'} L_2) \leq \dots \\
&\leq \|\nabla_{\phi_r}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}) - \frac{\partial}{\partial \phi}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\|_2 + L_1 L_2 \sum_{j=1}^{r}\alpha_j \prod_{j'=j}^{r}(1 + \alpha_{j'} L_2) = L_1 L_2 \sum_{j=1}^{r}\alpha_j \prod_{j'=j}^{r}(1 + \alpha_{j'} L_2).
\end{aligned}
$$

Then:

$$
\begin{aligned}
&\|\mathcal{G}_{FO}(\theta, \mathcal{T}) - \nabla_\theta \mathcal{L}^{out}(\theta, U^{(r)}(\theta, \mathcal{T}), \mathcal{T})\|_2 = \|\left(\frac{\partial}{\partial \theta}V(\theta, \mathcal{T})\right)^\top (\nabla_{\phi_0}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}) - \frac{\partial}{\partial \phi}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})) \\
&- \sum_{j=1}^{r}\alpha_j \left(\frac{\partial^2}{\partial \theta \partial \phi}\mathcal{L}^{in}(\theta, \phi_{j-1}, \mathcal{T})\right)^\top \nabla_{\phi_j}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\|_2 \\
&\leq \|\frac{\partial}{\partial \theta}V(\theta, \mathcal{T})\|_2 \|\nabla_{\phi_0}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}) - \frac{\partial}{\partial \phi}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\|_2 + \sum_{j=1}^{r}\alpha_j \|\frac{\partial^2}{\partial \theta \partial \phi}\mathcal{L}^{in}(\theta, \phi_{j-1}, \mathcal{T})\|_2 \\
&\cdot \|\nabla_{\phi_j}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\|_2 \\
&\leq M_1 L_1 L_2 \sum_{j=1}^{r}\alpha_j \prod_{j'=j}^{r}(1 + \alpha_{j'} L_2) + \sum_{j=1}^{r}\alpha_j L_2 L_1 \prod_{j'=j}^{r}(1 + \alpha_{j'} L_2) = \mathbb{D}_{bound}.
\end{aligned}
$$

Hence, for each $\theta \in \mathbb{R}^s$ $\mathbb{E}_{p(\mathcal{T})}\left[\|\mathcal{G}_{FO}(\theta, \mathcal{T}) - \nabla_\theta \mathcal{L}^{out}(\theta, U^{(r)}(\theta, \mathcal{T}), \mathcal{T})\|_2^2\right]$ is well-defined and bounded by $\mathbb{D}_{bound}^2$. Therefore, $\mathbb{D}^2$ is well-defined and bounded by $\mathbb{D}_{bound}^2$.

Next, we show (26). From (29) it follows that

$$
\begin{aligned}
&\|\nabla_\theta \mathcal{L}^{out}(\theta, U^{(r)}(\theta, \mathcal{T}), \mathcal{T})\|_2 = \|\frac{\partial}{\partial \theta}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}) + \left(\frac{\partial}{\partial \theta}V(\theta, \mathcal{T})\right)^\top \nabla_{\phi_0}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}) \\
&- \sum_{j=1}^{r}\alpha_j \left(\frac{\partial^2}{\partial \theta \partial \phi}\mathcal{L}^{in}(\theta, \phi_{j-1}, \mathcal{T})\right)^\top \nabla_{\phi_j}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\|_2 \\
&\leq \|\frac{\partial}{\partial \theta}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\|_2 + \|\frac{\partial}{\partial \theta}V(\theta, \mathcal{T})\|_2 \|\nabla_{\phi_0}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\|_2 + \sum_{j=1}^{r}\alpha_j \|\frac{\partial^2}{\partial \theta \partial \phi}\mathcal{L}^{in}(\theta, \phi_{j-1}, \mathcal{T})\|_2 \\
&\cdot \|\nabla_{\phi_j}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\|_2 \\
&\leq L_1 + M_1 L_1 \prod_{j=1}^{r}(1 + \alpha_j L_2) + \sum_{j=1}^{r}\alpha_j L_2 L_1 \prod_{j'=j}^{r}(1 + \alpha_{j'} L_2) = \mathbb{V}_{bound}.
\end{aligned}
$$

Hence, for each $\theta \in \mathbb{R}^s$ $\mathbb{E}_{p(\mathcal{T})}\left[\|\nabla_\theta \mathcal{L}^{out}(\theta, U^{(r)}(\theta, \mathcal{T}), \mathcal{T})\|_2^2\right]$ is well-defined and bounded by $\mathbb{V}_{bound}^2$ for all $\theta \in \mathbb{R}^s$. Consequently, $\mathbb{V}^2$ is well-defined and is also bounded by $\mathbb{V}_{bound}^2$.

(27) is satisfied by observing that

$$
\begin{aligned}
\mathbb{E}_{\xi, p(\mathcal{T})}\left[\mathcal{G}(\theta, \mathcal{T}, \xi)\right] &= \mathbb{E}_{p(\mathcal{T})}\left[\mathbb{E}_\xi\left[\mathcal{G}(\theta, \mathcal{T}, \xi)\right]\right] \\
&= \mathbb{E}_{p(\mathcal{T})}\left[\mathcal{G}_{FO}(\theta, \mathcal{T}) + \frac{q}{q}(\nabla_\theta \mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}) - \mathcal{G}_{FO}(\theta, \mathcal{T}))\right] \\
&= \mathbb{E}_{p(\mathcal{T})}\left[\nabla_\theta \mathcal{L}^{out}(\theta, \phi_r, \mathcal{T})\right] = \mathbb{E}_{p(\mathcal{T})}\left[\nabla_\theta \mathcal{L}^{out}(\theta, U^{(r)}(\theta, \mathcal{T}), \mathcal{T})\right]
\end{aligned}
$$

$$= \nabla_\theta \mathbb{E}_{p(\mathcal{T})}\left[\mathcal{L}^{out}(\theta, U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\right] = \frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta).$$

To show (28), we fix $\mathcal{T}\in\Omega_\mathcal{T}$. Let $\phi_0,\dots,\phi_r$ be inner-GD rollouts (2-3) corresponding to θ and $\mathcal{T}$.

Next:

$$\begin{aligned}\|\mathcal{G}(\theta,\mathcal{T},\xi)\|_2 &\le \|\mathcal{G}(\theta,\mathcal{T},\xi) - \nabla_\theta\mathcal{L}^{out}(\theta,U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\|_2 + \|\nabla_\theta\mathcal{L}^{out}(\theta,U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\|_2\\ &= (1-\frac{x}{q})\|\mathcal{G}_{FO}(\theta,\mathcal{T}) - \nabla_\theta\mathcal{L}^{out}(\theta,U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\|_2 + \|\nabla_\theta\mathcal{L}^{out}(\theta,U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\|_2.\end{aligned}$$

Take square and then expectation:

$$\begin{aligned}&\mathbb{E}_{\xi,p(\mathcal{T})}\left[\|\mathcal{G}(\theta,\mathcal{T},\xi)\|_2^2\right] = \mathbb{E}_{p(\mathcal{T})}\mathbb{E}_\xi\left[\|\mathcal{G}(\theta,\mathcal{T},\xi)\|_2^2\right]\\
&\le \mathbb{E}_{p(\mathcal{T})}\mathbb{E}_\xi\Bigg[(1-2\frac{\xi}{q}+\frac{\xi^2}{q^2})\|\mathcal{G}_{FO}(\theta,\mathcal{T}) - \nabla_\theta\mathcal{L}^{out}(\theta,U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\|_2^2\\
&+ 2(1-\frac{\xi}{q})\|\mathcal{G}_{FO}(\theta,\mathcal{T}) - \nabla_\theta\mathcal{L}^{out}(\theta,U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\|_2\|\nabla_\theta\mathcal{L}^{out}(\theta,U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\|_2\\
&+ \|\nabla_\theta\mathcal{L}^{out}(\theta,U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\|_2^2\Bigg]\\
&= \mathbb{E}_{p(\mathcal{T})}\Bigg[(\frac{1}{q}-1)\|\mathcal{G}_{FO}(\theta,\mathcal{T}) - \nabla_\theta\mathcal{L}^{out}(\theta,U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\|_2^2\\
&+ 2\cdot 0\cdot\|\mathcal{G}_{FO}(\theta,\mathcal{T}) - \nabla_\theta\mathcal{L}^{out}(\theta,U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\|_2\|\nabla_\theta\mathcal{L}^{out}(\theta,U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\|_2\\
&+ \|\nabla_\theta\mathcal{L}^{out}(\theta,U^{(r)}(\theta,\mathcal{T}),\mathcal{T})\|_2^2\Bigg]\\
&\le (\frac{1}{q}-1)\mathbb{D}^2 + \mathbb{V}^2.\end{aligned}$$

□

Lemma 2. *Let $p,r,s\in\mathbb{N}$, $\{\alpha_j>0\}_{j=1}^\infty$ be any sequence, $p(\mathcal{T})$ be a distribution on a nonempty set $\Omega_\mathcal{T}$, $V:\mathbb{R}^s\times\Omega_\mathcal{T}\to\mathbb{R}^p$, $\mathcal{L}^{in},\mathcal{L}^{out}:\mathbb{R}^s\times\mathbb{R}^p\times\Omega_\mathcal{T}\to\mathbb{R}$ be functions satisfying Assumption 1, and let $U^{(r)}:\mathbb{R}^s\times\Omega_\mathcal{T}\to\mathbb{R}^p$ be defined according to (2-3), $\mathcal{M}^{(r)}:\mathbb{R}^s\to\mathbb{R}$ be defined according to (4) and satisfy Assumption 2. Then for all $\theta',\theta''\in\mathbb{R}^s$ it holds that*

$$\|\frac{\partial}{\partial\theta}\mathcal{M}(\theta') - \frac{\partial}{\partial\theta}\mathcal{M}(\theta'')\|_2 \le \mathcal{C}\|\theta'-\theta''\|_2,$$

where

$$\mathcal{C} = L_2 + L_2\mathcal{A}_r + \sum_{j=1}^r\alpha_j\left(L_2\mathcal{B}_j + L_3(1+\mathcal{A}_{j-1})L_1\prod_{j'=j+1}^r(1+\alpha_{j'}L_2)\right) + M_1\mathcal{B}_0 + M_2L_1\prod_{j=1}^r(1+\alpha_jL_2), \tag{30}$$

$$\mathcal{B}_j = \left(L_2(1+\mathcal{A}_r)(1+\alpha_jL_2) + L_1L_3\sum_{j'=j+1}^r\alpha_{j'}(1+\mathcal{A}_{j'-1})\right)\prod_{j'=j+1}^r(1+\alpha_{j'}L_2), \tag{31}$$

$$\mathcal{A}_j = \left(M_1\prod_{j'=1}^j(1+\alpha_{j'}L_2) + L_2\sum_{j'=1}^j\alpha_{j'}\prod_{j''=j'+1}^j(1+\alpha_{j''}L_2)\right). \tag{32}$$

Proof. Fix $\mathcal{T}\in\Omega_\mathcal{T}$. Let $\phi_0' = V(\theta',\mathcal{T}),\dots,\phi_r'$ and $\phi_0'' = V(\theta'',\mathcal{T}),\dots,\phi_r''$ be inner-GD rollouts (2-3) for θ' and θ'' respectively. For each $1\le j\le r$ inequalities applies:

$$\begin{aligned}\|\phi_j'-\phi_j''\|_2 &= \|\phi_{j-1}'-\phi_{j-1}'' - \alpha_j(\frac{\partial}{\partial\phi}\mathcal{L}^{in}(\theta',\phi_{j-1}',\mathcal{T}) - \frac{\partial}{\partial\phi}\mathcal{L}^{in}(\theta'',\phi_{j-1}'',\mathcal{T}))\|_2\\
&\le \|\phi_{j-1}'-\phi_{j-1}''\|_2 + \alpha_j\|\frac{\partial}{\partial\phi}\mathcal{L}^{in}(\theta',\phi_{j-1}',\mathcal{T}) - \frac{\partial}{\partial\phi}\mathcal{L}^{in}(\theta'',\phi_{j-1}'',\mathcal{T})\|_2\end{aligned}$$

$$\leq \|\phi'_{j-1} - \phi''_{j-1}\|_2 + \alpha_j L_2 \|\phi'_{j-1} - \phi''_{j-1}\|_2 + \alpha_j L_2 \|\theta' - \theta''\|_2$$
$$= (1 + \alpha_j L_2)\|\phi'_{j-1} - \phi''_{j-1}\|_2 + \alpha_j L_2 \|\theta' - \theta''\|_2.$$

Therefore, for each $0 \leq j \leq r$

$$\|\phi'_j - \phi''_j\|_2 \leq \|\phi'_0 - \phi''_0\|_2 \prod_{j'=1}^{j} (1 + \alpha_{j'} L_2) + L_2 \|\theta' - \theta''\|_2 \sum_{j'=1}^{j} \alpha_{j'} \prod_{j''=j'+1}^{j} (1 + \alpha_{j''} L_2)$$
$$= \mathcal{A}_j \cdot \|\theta' - \theta''\|_2. \tag{33}$$

Therefore,

$$\|\nabla_{\phi'_r} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T})) - \nabla_{\phi''_r} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T})\|_2 = \|\frac{\partial}{\partial \phi} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T})) - \frac{\partial}{\partial \phi} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T})\|_2$$
$$\leq L_2 \|\phi'_r - \phi''_r\|_2 + L_2 \|\theta' - \theta''\|_2 \leq L_2 (1 + \mathcal{A}_r) \|\theta' - \theta''\|_2.$$

For each $1 \leq j \leq r$ the following chain of inequalities applies as a result of (8-9):

$$\|\nabla_{\phi'_{j-1}} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T}) - \nabla_{\phi''_{j-1}} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T})\|_2 = \|\nabla_{\phi'_j} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T}) - \nabla_{\phi''_j} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T})$$
$$- \alpha_j \Big(\Big(\frac{\partial^2}{\partial \phi^2} \mathcal{L}^{in}(\theta', \phi'_{j-1}, \mathcal{T}) \Big)^\top \nabla_{\phi'_j} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T}) - \Big(\frac{\partial^2}{\partial \phi^2} \mathcal{L}^{in}(\phi''_{j-1}, \mathcal{T}) \Big)^\top \nabla_{\phi''_j} \mathcal{L}^{out}(\phi''_r, \mathcal{T})) \|_2$$
$$= \|\nabla_{\phi'_j} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T}) - \nabla_{\phi''_j} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T}) - \alpha_j \Big(\frac{\partial^2}{\partial \phi^2} \mathcal{L}^{in}(\theta', \phi'_{j-1}, \mathcal{T}) \Big)^\top (\nabla_{\phi'_j} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T})$$
$$- \nabla_{\phi''_j} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T})) - \alpha_j \Big(\frac{\partial^2}{\partial \phi^2} \mathcal{L}^{in}(\theta', \phi'_{j-1}, \mathcal{T}) - \frac{\partial^2}{\partial \phi^2} \mathcal{L}^{in}(\theta'', \phi''_{j-1}, \mathcal{T}) \Big)^\top \nabla_{\phi''_j} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T}) \|_2$$
$$\leq \|\nabla_{\phi'_j} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T}) - \nabla_{\phi''_j} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T})\|_2 + \alpha_j \|\frac{\partial^2}{\partial \phi^2} \mathcal{L}^{in}(\theta', \phi'_{j-1}, \mathcal{T})\|_2 \|\nabla_{\phi'_j} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T})$$
$$- \nabla_{\phi''_j} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T})\|_2 + \alpha_j \|\frac{\partial^2}{\partial \phi^2} \mathcal{L}^{in}(\theta', \phi'_{j-1}, \mathcal{T}) - \frac{\partial^2}{\partial \phi^2} \mathcal{L}^{in}(\theta'', \phi''_{j-1}, \mathcal{T})\|_2 \cdot \|\nabla_{\phi''_j} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T})\|_2$$
$$\leq (1 + \alpha_j L_2) \|\nabla_{\phi'_j} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T}) - \nabla_{\phi''_j} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T})\|_2 + \alpha_j L_3 L_1 (\|\phi'_{j-1} - \phi''_{j-1}\|_2$$
$$+ \|\theta' - \theta''\|_2) \prod_{j'=j+1}^{r} (1 + \alpha_{j'} L_2)$$
$$\leq (1 + \alpha_j L_2) \|\nabla_{\phi'_j} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T}) - \nabla_{\phi''_j} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T})\|_2 + \alpha_j L_1 L_3 (1 + \mathcal{A}_{j-1}) \|\theta' - \theta''\|_2 \prod_{j'=j+1}^{r} (1 + \alpha_{j'} L_2)$$
$$\leq \dots$$
$$\leq \|\nabla_{\phi'_r} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T}) - \nabla_{\phi''_r} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T})\|_2 \prod_{j'=j}^{r} (1 + \alpha_{j'} L_2)$$
$$+ L_1 L_3 \|\theta' - \theta''\|_2 \sum_{j'=j}^{r} \alpha_{j'} (1 + \mathcal{A}_{j'-1}) \prod_{j''=j'}^{r} (1 + \alpha_{j''} L_2) \prod_{j''=j+1}^{j'-1} (1 + \alpha_{j''} L_2) \leq \mathcal{B}_{j-1} \|\theta' - \theta''\|_2,$$

where we use (29). Using (6-7), we deduce that

$$\|\nabla_{\theta'} \mathcal{L}^{out}(\theta', U^{(r)}(\theta', \mathcal{T}), \mathcal{T}) - \nabla_{\theta''} \mathcal{L}^{out}(\theta'', U^{(r)}(\theta'', \mathcal{T}), \mathcal{T})\|_2 = \|\frac{\partial}{\partial \theta} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T}) - \frac{\partial}{\partial \theta} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T})$$
$$- \sum_{j=1}^{r} \alpha_j \Big(\Big(\frac{\partial^2}{\partial \theta \partial \phi} \mathcal{L}^{in}(\theta', \phi'_{j-1}, \mathcal{T}) \Big)^\top \nabla_{\phi'_j} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T}) - \Big(\frac{\partial^2}{\partial \theta \partial \phi} \mathcal{L}^{in}(\theta'', \phi''_{j-1}, \mathcal{T}) \Big)^\top \nabla_{\phi''_j} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T}) \Big)$$
$$+ \Big(\frac{\partial}{\partial \theta} V(\theta', \mathcal{T}) \Big)^\top \nabla_{\phi'_0} \mathcal{L}^{out}(\theta', \phi'_r, \mathcal{T}) - \Big(\frac{\partial}{\partial \theta} V(\theta'', \mathcal{T}) \Big)^\top \nabla_{\phi''_0} \mathcal{L}^{out}(\theta'', \phi''_r, \mathcal{T}) \|_2$$

$$
\begin{aligned}
&\leq \|\frac{\partial}{\partial\theta}\mathcal{L}^{out}(\theta',\phi'_r,\mathcal{T}) - \frac{\partial}{\partial\theta}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T})\|_2 \\
&+ \sum_{j=1}^{r}\alpha_j\|\left(\frac{\partial^2}{\partial\theta\partial\phi}\mathcal{L}^{in}(\theta',\phi'_{j-1},\mathcal{T})\right)^\top\nabla_{\phi'_j}\mathcal{L}^{out}(\theta',\phi'_r,\mathcal{T}) - \left(\frac{\partial^2}{\partial\theta\partial\phi}\mathcal{L}^{in}(\theta'',\phi''_{j-1},\mathcal{T})\right)^\top\nabla_{\phi''_j}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T})\|_2 \\
&+ \|\left(\frac{\partial}{\partial\theta}V(\theta',\mathcal{T})\right)^\top\nabla_{\phi'_0}\mathcal{L}^{out}(\theta',\phi'_r,\mathcal{T}) - \left(\frac{\partial}{\partial\theta}V(\theta'',\mathcal{T})\right)^\top\nabla_{\phi''_0}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T})\|_2 \\
&\leq \|\frac{\partial}{\partial\theta}\mathcal{L}^{out}(\theta',\phi'_r,\mathcal{T}) - \frac{\partial}{\partial\theta}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T})\|_2 \\
&+ \sum_{j=1}^{r}\alpha_j\|\left(\frac{\partial^2}{\partial\theta\partial\phi}\mathcal{L}^{in}(\theta',\phi'_{j-1},\mathcal{T})\right)^\top\nabla_{\phi'_j}\mathcal{L}^{out}(\theta',\phi'_r,\mathcal{T}) - \left(\frac{\partial^2}{\partial\theta\partial\phi}\mathcal{L}^{in}(\theta',\phi'_{j-1},\mathcal{T})\right)^\top\nabla_{\phi''_j}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T}) \\
&+ \left(\frac{\partial^2}{\partial\theta\partial\phi}\mathcal{L}^{in}(\theta',\phi'_{j-1},\mathcal{T})\right)^\top\nabla_{\phi''_j}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T}) - \left(\frac{\partial^2}{\partial\theta\partial\phi}\mathcal{L}^{in}(\theta'',\phi''_{j-1},\mathcal{T})\right)^\top\nabla_{\phi''_j}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T})\|_2 \\
&+ \|\left(\frac{\partial}{\partial\theta}V(\theta',\mathcal{T})\right)^\top\nabla_{\phi'_0}\mathcal{L}^{out}(\theta',\phi'_r,\mathcal{T}) - \left(\frac{\partial}{\partial\theta}V(\theta',\mathcal{T})\right)^\top\nabla_{\phi''_0}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T}) \\
&+ \left(\frac{\partial}{\partial\theta}V(\theta',\mathcal{T})\right)^\top\nabla_{\phi''_0}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T}) - \left(\frac{\partial}{\partial\theta}V(\theta'',\mathcal{T})\right)^\top\nabla_{\phi''_0}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T})\|_2 \\
&\leq \|\frac{\partial}{\partial\theta}\mathcal{L}^{out}(\theta',\phi'_r,\mathcal{T}) - \frac{\partial}{\partial\theta}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T})\|_2 \\
&+ \sum_{j=1}^{r}\alpha_j\|\frac{\partial^2}{\partial\theta\partial\phi}\mathcal{L}^{in}(\theta',\phi'_{j-1},\mathcal{T})\|_2\|\nabla_{\phi'_j}\mathcal{L}^{out}(\theta',\phi'_r,\mathcal{T}) - \nabla_{\phi''_j}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T})\|_2 \\
&+ \sum_{j=1}^{r}\alpha_j\|\frac{\partial^2}{\partial\theta\partial\phi}\mathcal{L}^{in}(\theta',\phi'_{j-1},\mathcal{T}) - \frac{\partial^2}{\partial\theta\partial\phi}\mathcal{L}^{in}(\theta'',\phi''_{j-1},\mathcal{T})\|_2\|\nabla_{\phi'_j}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T})\|_2 \\
&+ \|\frac{\partial}{\partial\theta}V(\theta',\mathcal{T})\|_2\|\nabla_{\phi'_0}\mathcal{L}^{out}(\theta',\phi'_r,\mathcal{T}) - \nabla_{\phi''_0}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T})\|_2 \\
&+ \|\frac{\partial}{\partial\theta}V(\theta',\mathcal{T}) - \frac{\partial}{\partial\theta}V(\theta'',\mathcal{T})\|_2\|\nabla_{\phi'_0}\mathcal{L}^{out}(\theta'',\phi''_r,\mathcal{T})\|_2 \\
&\leq L_2\|\theta'-\theta''\|_2 + L_2\|\phi'_r-\phi''_r\|_2 + \sum_{j=1}^{r}\alpha_j\Bigg(L_2\mathcal{B}_j\|\theta'-\theta''\|_2 + L_3(\|\theta'-\theta''\|_2 \\
&+ \|\phi'_{j-1}-\phi''_{j-1}\|_2)L_1\prod_{j'=j+1}^{r}(1+\alpha_{j'}L_2)\|_2\Bigg) + M_1\mathcal{B}_0\|\theta'-\theta''\|_2 + M_2\|\theta'-\theta''\|_2L_1\prod_{j=1}^{r}(1+\alpha_jL_2) \\
&\leq L_2\|\theta'-\theta''\|_2 + L_2\mathcal{A}_r\|\theta'-\theta''\|_2 + \sum_{j=1}^{r}\alpha_j\Bigg(L_2\mathcal{B}_j\|\theta'-\theta''\|_2 + L_3(\|\theta'-\theta''\|_2 \\
&+ \mathcal{A}_{j-1}\|\theta'-\theta''\|_2)L_1\prod_{j'=j+1}^{r}(1+\alpha_{j'}L_2)\|_2\Bigg) + M_1\mathcal{B}_0\|\theta'-\theta''\|_2 + M_2\|\theta'-\theta''\|_2L_1\prod_{j=1}^{r}(1+\alpha_jL_2) \leq \mathcal{C}\|\theta'-\theta''\|_2,
\end{aligned}
$$

where we use (29) and (33). Finally, by taking expectation with respect to $\mathcal{T}\sim p(\mathcal{T})$ and applying Jensen inequality we get

$$
\begin{aligned}
\|\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta') - \frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta'')\|_2^2 &\leq \mathbb{E}_{p(\mathcal{T})}\left[\|\nabla_{\theta'}\mathcal{L}^{out}(\theta',U^{(r)}(\theta',\mathcal{T}),\mathcal{T}) - \nabla_{\theta''}\mathcal{L}^{out}(\theta'',U^{(r)}(\theta'',\mathcal{T}),\mathcal{T})\|_2^2\right] \\
&\leq \mathcal{C}^2\|\theta'-\theta''\|_2^2,
\end{aligned}
$$

which is equivalent to the statement of Lemma. □

Lemma 3. *Let $p,r,s\in\mathbb{N}$, $\{\alpha_j>0\}_{j=1}^{\infty}$ and $\{\gamma_k>0\}_{k=1}^{\infty}$ be any sequences, $q\in(0,1],\theta_0\in\mathbb{R}^p$, $p(\mathcal{T})$ be a distribution on a nonempty set $\Omega_{\mathcal{T}}$, $V:\mathbb{R}^s\times\Omega_{\mathcal{T}}\to\mathbb{R}^p$, $\mathcal{L}^{in},\mathcal{L}^{out}:\mathbb{R}^s\times\mathbb{R}^p\times\Omega_{\mathcal{T}}\to\mathbb{R}$ be functions satisfying Assumption 1, and let $U^{(r)}:\mathbb{R}^s\times\Omega_{\mathcal{T}}\to\mathbb{R}^p$ be defined according to (2-3), $\mathcal{M}^{(r)}:\mathbb{R}^p\to\mathbb{R}$ be defined according to (4) and satisfy Assumption 2.*

Define $\mathcal{G}_{FO} : \mathbb{R}^s \times \Omega_{\mathcal{T}} \to \mathbb{R}^s$ *and* $\mathcal{G} : \mathbb{R}^s \times \Omega_{\mathcal{T}} \times \{0,1\} \to \mathbb{R}^s$ *as*

$$\mathcal{G}_{FO}(\theta, \mathcal{T}) = \frac{\partial}{\partial \theta}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}) + (\frac{\partial}{\partial \theta}V(\theta, \mathcal{T}))^\top \frac{\partial}{\partial \phi}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}), \quad \phi_r = U^{(r)}(\theta, \mathcal{T}),$$
$$\mathcal{G}(\theta, \mathcal{T}, x) = \mathcal{G}_{FO}(\theta, \mathcal{T}) + \frac{x}{q}(\nabla_\theta \mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}) - \mathcal{G}_{FO}(\theta, \mathcal{T})).$$

Let $\{\mathcal{T}_k\}_{k=1}^\infty, \{\xi_k\}_{k=1}^\infty$ *be sequences of i.i.d. samples from* $p(\mathcal{T})$ *and* $\text{Bernoulli}(q)$ *respectively, such that* σ*-algebras populated by both sequences are independent. Let* $\{\theta_k \in \mathbb{R}^s\}_{k=0}^\infty$ *be a sequence where for all* $k \in \mathbb{N}$ $\theta_k = \theta_{k-1} - \gamma_k \mathcal{G}(\theta_{k-1}, \mathcal{T}_k, \xi_k)$*. Then for each* $k \in \mathbb{N}$

$$\sum_{u=1}^k \gamma_u \mathbb{E}\left[\|\frac{\partial}{\partial \theta}\mathcal{M}^{(r)}(\theta_{u-1})\|_2^2\right] \leq \mathcal{M}^{(r)}(\theta_0) - \mathcal{M}_*^{(r)} + \mathcal{C}\left((\frac{1}{q} - 1)\mathbb{D}^2 + \mathbb{V}^2\right)\sum_{u=1}^k \gamma_u^2 \tag{34}$$

where $\mathcal{C}$ *is defined in (30-32),* $\mathbb{D}, \mathbb{V}$ *are defined in (14), (16) respectively.*

Proof. Let $\mathcal{F}_u$ denote a σ-algebra populated by $\{\mathcal{T}_\kappa, \xi_\kappa\}_{\kappa<u}$. Using Lemma 2, we apply Inequality 4.3 from (Bottou et al., 2018) to obtain that for all $\theta', \theta'' \in \mathbb{R}^s$

$$\mathcal{M}^{(r)}(\theta') \leq \mathcal{M}^{(r)}(\theta'') + \left(\frac{\partial}{\partial \theta}\mathcal{M}^{(r)}(\theta'')\right)^\top (\theta' - \theta'') + \frac{1}{2}\mathcal{C}\|\theta' - \theta''\|_2^2.$$

For any $u \in \mathbb{N}$, by setting $\theta' = \theta_u$, $\theta'' = \theta_{u-1}$ we deduce that

$$\mathcal{M}^{(r)}(\theta_u) \leq \mathcal{M}^{(r)}(\theta_{u-1}) - \gamma_u\left(\frac{\partial}{\partial \theta}\mathcal{M}^{(r)}(\theta_{u-1})\right)^\top \mathcal{G}(\theta_{u-1}, \mathcal{T}_u, \xi_u) + \frac{1}{2}\gamma_u^2 \mathcal{C}\|\mathcal{G}(\theta_{u-1}, \mathcal{T}_u, \xi_u)\|_2^2.$$

Take expectation with respect to $\mathcal{F}_u$:

$$\begin{aligned}\mathbb{E}\left[\mathcal{M}^{(r)}(\theta_u)|\mathcal{F}_u\right] &\leq \mathcal{M}^{(r)}(\theta_{u-1}) - \gamma_u\left(\frac{\partial}{\partial \theta}\mathcal{M}^{(r)}(\theta_{u-1})\right)^\top \mathbb{E}\left[\mathcal{G}(\theta_{u-1}, \mathcal{T}_u, \xi_u)|\mathcal{F}_u\right] \\ &\quad + \frac{1}{2}\gamma_u^2 \mathcal{C}\mathbb{E}\left[\|\mathcal{G}(\theta_{u-1}, \mathcal{T}_u, \xi_u)\|_2^2|\mathcal{F}_u\right] \\ &\leq \mathcal{M}^{(r)}(\theta_{u-1}) - \gamma_u\|\frac{\partial}{\partial \theta}\mathcal{M}^{(r)}(\theta_{u-1})\|_2^2 + \frac{1}{2}\gamma_u^2\mathcal{C}\left((\frac{1}{q} - 1)\mathbb{D}^2 + \mathbb{V}^2\right),\end{aligned}$$

where we use Lemma 1's result. Take the full expectation:

$$\mathbb{E}\left[\mathcal{M}^{(r)}(\theta_u)\right] \leq \mathbb{E}\left[\mathcal{M}^{(r)}(\theta_{u-1})\right] - \gamma_u\mathbb{E}\left[\|\frac{\partial}{\partial \theta}\mathcal{M}^{(r)}(\theta_{u-1})\|_2^2\right] + \frac{1}{2}\gamma_u^2\mathcal{C}\left((\frac{1}{q} - 1)\mathbb{D}^2 + \mathbb{V}^2\right)$$

which is equivalent to

$$\gamma_u\mathbb{E}\left[\|\frac{\partial}{\partial \theta}\mathcal{M}^{(r)}(\theta_{u-1})\|_2^2\right] \leq \mathbb{E}\left[\mathcal{M}^{(r)}(\theta_{u-1})\right] - \mathbb{E}\left[\mathcal{M}^{(r)}(\theta_u)\right] + \frac{1}{2}\gamma_u^2\mathcal{C}\left((\frac{1}{q} - 1)\mathbb{D}^2 + \mathbb{V}^2\right). \tag{35}$$

Sum inequalities (35) for all $1 \leq u \leq k$:

$$\begin{aligned}\sum_{u=1}^k \gamma_u\mathbb{E}\left[\|\frac{\partial}{\partial \theta}\mathcal{M}^{(r)}(\theta_{u-1})\|_2^2\right] &\leq \mathbb{E}\left[\mathcal{M}^{(r)}(\theta_0)\right] - \mathbb{E}\left[\mathcal{M}^{(r)}(\theta_k)\right] + \mathcal{C}\left((\frac{1}{q} - 1)\mathbb{D}^2 + \mathbb{V}^2\right)\sum_{u=1}^k \gamma_u^2 \\ &\leq \mathcal{M}^{(r)}(\theta_0) - \mathcal{M}_*^{(r)} + \mathcal{C}\left((\frac{1}{q} - 1)\mathbb{D}^2 + \mathbb{V}^2\right)\sum_{u=1}^k \gamma_u^2.\end{aligned}$$

□

Theorem 1 proof. Under conditions of the theorem results of Lemma 3 are true.

First, we prove 1. If $\sum_{k=1}^{\infty}\gamma_k^2 < \infty$, then the right-hand side of (34) converges to a finite value when $k \to \infty$. Therefore, the left-hand side also converges to a finite value. Suppose the statement of 1 is false. Then there exists $k_0 \in \mathbb{N}, A > 0$ such that $\forall u \geq k_0 : \mathbb{E}\left[\|\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_{u-1})\|_2^2\right] > A$. But then for all $k \geq k_0$

$$\sum_{u=1}^{k}\gamma_u\mathbb{E}\left[\|\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_{u-1})\|_2^2\right] \geq A\sum_{u=k_0}^{k}\gamma_u \to \infty$$

when $k \to \infty$, which is a contradiction. Therefore, 1 is true.

Next, we prove 2. Observe that

$$\min_{0\leq u<k}\mathbb{E}\left[\|\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_u)\|_2^2\right]\sum_{u=1}^{k}\gamma_u \leq \sum_{u=1}^{k}\gamma_u\mathbb{E}\left[\|\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_{u-1})\|_2^2\right]$$
$$\leq \mathcal{M}^{(r)}(\theta_0) - \mathcal{M}_*^{(r)} + \mathcal{C}\left((\frac{1}{q}-1)\mathbb{D}^2 + \mathbb{V}^2\right)\sum_{u=1}^{k}\gamma_u^2.$$

Divide by $\sum_{u=1}^{k}\gamma_u$:

$$\min_{0\leq u<k}\mathbb{E}\left[\|\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_u)\|_2^2\right] \leq \frac{1}{\sum_{u=1}^{k}\gamma_u}(\mathcal{M}^{(r)}(\theta_0) - \mathcal{M}_*^{(r)}) + \mathcal{C}\left((\frac{1}{q}-1)\mathbb{D}^2 + \mathbb{V}^2\right)\cdot\frac{1}{\sum_{u=1}^{k}\gamma_u}\cdot\sum_{u=1}^{k}\gamma_u^2.$$

2 is satisfied by observing that

$$\sum_{u=1}^{k}\gamma_u = \sum_{u=1}^{k}u^{-0.5} = \Omega(k^{0.5}), \quad \sum_{u=1}^{k}\gamma_u^2 = \sum_{u=1}^{k}u^{-1} = O(\log k) = O(k^{\epsilon})$$

for any $\epsilon > 0$. □

E.2 Theorem 2

Proof. Consider a set $\Omega_{\mathcal{T}}$ consisting of two elements: $\Omega_{\mathcal{T}} = \{\mathcal{T}^{(1)}, \mathcal{T}^{(2)}\}$. Define $p(\mathcal{T})$ so that

$$\mathbb{P}_{p(\mathcal{T})}(\mathcal{T} = \mathcal{T}^{(1)}) = \mathbb{P}_{p(\mathcal{T})}(\mathcal{T} = \mathcal{T}^{(2)}) = \frac{1}{2}.$$

Choose arbitrary numbers $0 < a_1, a_2 < \frac{1}{\alpha}$, $a_1 \neq a_2$ and set $b_1 = 0$. Since $a_1 \neq a_2$, $(1-\alpha a_1)/(1-\alpha a_2) \neq 1$ and, consequently,

$$\left(\frac{1-\alpha a_1}{1-\alpha a_2}\right)^r \neq \left(\frac{1-\alpha a_1}{1-\alpha a_2}\right)^{2r}.$$

Multiply by $\frac{a_1}{a_2} \neq 0$:

$$\frac{a_1}{a_2}\cdot\left(\frac{1-\alpha a_1}{1-\alpha a_2}\right)^r \neq \frac{a_1}{a_2}\cdot\left(\frac{1-\alpha a_1}{1-\alpha a_2}\right)^{2r}. \tag{36}$$

From (36) and since $\frac{a_1}{a_2}(\frac{1-\alpha a_1}{1-\alpha a_2})^r, \frac{a_1}{a_2}(\frac{1-\alpha a_1}{1-\alpha a_2})^{2r} > 0$ it follows that

$$\frac{\frac{a_1}{a_2}(\frac{1-\alpha a_1}{1-\alpha a_2})^{2r}+1}{\frac{a_1}{a_2}(\frac{1-\alpha a_1}{1-\alpha a_2})^{r}+1} - 1 \neq 0.$$

Multiply inequality by $(1-\alpha a_2)^{2r} \neq 0$ and numerator/denominator by $a_2(1-\alpha a_2)^r \neq 0$:

$$\frac{a_1(1-\alpha a_1)^{2r} + a_2(1-\alpha a_2)^{2r}}{a_1(1-\alpha a_1)^r + a_2(1-\alpha a_2)^r}(1-\alpha a_2)^r - (1-\alpha a_2)^{2r} \neq 0.$$

Because of the inequality above, we can define a number b_2 as

$$b_2 = 2\sqrt{2D}\left|\frac{a_1(1-\alpha a_1)^{2r} + a_2(1-\alpha a_2)^{2r}}{a_1(1-\alpha a_1)^r + a_2(1-\alpha a_2)^r}(1-\alpha a_1)^r - (1-\alpha a_2)^{2r}\right|^{-1} > 0 \tag{37}$$

and select arbitrary number A so that

$$A > \left|\frac{b_1}{a_1} - \frac{b_2}{a_2}\right|. \tag{38}$$

Consider two functions $f_i(x)$, $f_i : \mathbb{R} \to \mathbb{R}$, $i \in \{1, 2\}$ defined as follows (denote $z_i = z_i(x) = |x - \frac{b_i}{a_i}|$)

$$f_i(x) = \begin{cases} \frac{1}{2}a_i z_i^2 & \text{if } z_i \le A \\ -\frac{1}{6}a_i(z_i - A)^3 + \frac{1}{2}a_i(z_i - A)^2 + a_i A z_i - \frac{1}{2}a_i A^2 & \text{if } A < z_i \le A + 1 \\ (\frac{1}{2}a_i + a_i A)z_i - \frac{1}{6}a_i - \frac{1}{2}a_i A^2 - \frac{1}{2}a_i A & \text{if } A + 1 < z_i \end{cases}. \tag{39}$$

It is easy to check that for $i \in \{1, 2\}$ $f_i(x)$ is twice differentiable with a global minimum at $\frac{b_i}{a_i}$. The following expressions apply for the first and second derivative:

$$f_i'(x) = \begin{cases} a_i x - b_i & \text{if } z_i \le A \\ \left(-\frac{1}{2}a_i(z_i - A)^2 + a_i z_i\right)\text{sign}(x - \frac{b_i}{a_i}) & \text{if } A < z_i \le A + 1 \\ (\frac{1}{2}a_i + a_i A)\text{sign}(x - \frac{b_i}{a_i}) & \text{if } A + 1 < z_i \end{cases}, \tag{40}$$

$$f_i''(x) = \begin{cases} a_i & \text{if } z_i \le A \\ -a_i z_i + a_i + a_i A & \text{if } A < z_i \le A + 1 \\ 0 & \text{if } A + 1 < z_i \end{cases}. \tag{41}$$

From (40-41) it follows that each f_i has bounded, Lipschitz-continuous gradients and Hessians. Define $V(\theta, \mathcal{T}) \equiv \theta$, $\mathcal{L}^{in}(\theta, \phi, \mathcal{T}_i) \equiv \mathcal{L}^{out}(\theta, \phi, \mathcal{T}_i) \equiv f_i(\phi^{(1)})$ for $i \in \{1, 2\}$, where $\phi^{(1)}$ denotes a first element of ϕ, then Assumption 1 is satisfied. Since $\Omega_{\mathcal{T}}$ is finite, Assumption 2 is also satisfied.

Let $I = [\frac{b_2}{a_2} - A, \frac{b_1}{a_1} + A]$. Observe that from (38) it follows that $\frac{b_1}{a_1}, \frac{b_2}{a_2} \in I$ and $I \subseteq [\frac{b_i}{a_i} - A, \frac{b_i}{a_i} + A]$ for $i \in \{1, 2\}$, i.e. I corresponds to a quadratic part of both $f_1(x)$ and $f_2(x)$. If $x \in I$, then for $i \in \{1, 2\}$

$$\begin{aligned} x - \alpha f_i'(x) &= x - \alpha(a_i x - b_i) = (1 - \alpha a_i)x + \alpha b_i \\ &= (1 - \alpha a_i) \cdot x + \alpha a_i \cdot \frac{b_i}{a_i} \in [\min(x, \frac{b_i}{a_i}), \max(x, \frac{b_i}{a_i})] \subseteq I \end{aligned} \tag{42}$$

since $x - \alpha f_i'(x)$ is a convex combination of x and $\frac{b_i}{a_i}$ ($0 < \alpha a_i, 1 - \alpha a_i < 1$). From (42) and the definition of $\mathcal{L}^{in}(\theta, \phi, \mathcal{T}), \mathcal{L}^{out}(\theta, \phi, \mathcal{T})$ it follows that if $\phi_0, \ldots, \phi_r$ is a rollout of inner GD (3) for task $\mathcal{T}^{(i)}$ and $\theta^{(1)} = \phi_0^{(1)} \in I$, then $\phi_1^{(1)}, \ldots, \phi_r^{(1)} \in I$ and, hence,

$$\begin{aligned} \nabla_{\phi_r}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}^{(i)})^{(1)} &= f_i'(\phi_r^{(1)}) = a_i\phi_r^{(1)} - b_i, \\ \forall j \in \{1, \ldots, r\} : \phi_j^{(1)} &= (1 - \alpha a_i)\phi_{j-1}^{(1)} + \alpha b_i. \end{aligned} \tag{43}$$

From (43) we derive that

$$\begin{aligned} \phi_j^{(1)} - \frac{b_i}{a_i} = (1 - \alpha a_i)(\phi_{j-1}^{(1)} - \frac{b_i}{a_i}), \quad \phi_r^{(1)} - \frac{b_i}{a_i} &= (1 - \alpha a_i)^r(\phi_0^{(1)} - \frac{b_i}{a_i}), \\ \phi_r^{(1)} &= (1 - \alpha a_i)^r(\phi_0^{(1)} - \frac{b_i}{a_i}) + \frac{b_i}{a_i}, \\ \nabla_{\phi_r}\mathcal{L}^{out}(\theta, \phi_r, \mathcal{T}^{(i)})^{(1)} = a_i\left((1 - \alpha a_i)^r(\phi_0^{(1)} - \frac{b_i}{a_i}) + \frac{b_i}{a_i}\right) - b_i &= a_i(1 - \alpha a_i)^r(\phi_0^{(1)} - \frac{b_i}{a_i}). \end{aligned} \tag{44}$$

From (5) it follows that there exists a deterministic number $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$

$$\gamma_k < \frac{1}{2} \min_{i \in \{1,2\}} \frac{1}{a_i(1+\alpha a_i)^r}. \tag{45}$$

If (45) holds, then it also holds that

$$\gamma_k < \min_{i \in \{1,2\}} \frac{1}{a_i(1+\alpha a_i)^r}, \quad \gamma_k < \min_{i \in \{1,2\}} \frac{1}{a_i(1-\alpha a_i)^r}. \tag{46}$$

For any $k \geq k_0$ the following cases are possible:

1. Case 1: $\theta_{k-1}^{(1)} \in I$. An identity (44) allows to write that for $i \in \{1,2\}$

$$\mathcal{G}_{FO}(\theta_{k-1}, \mathcal{T}^{(i)})^{(1)} = a_i(1-\alpha a_i)^r(\theta_{k-1}^{(1)} - \frac{b_i}{a_i}). \tag{47}$$

For $i \in \{1,2\}$ let random number $v_i \in 0, 1$ denote an indicator that $\mathcal{T}_k = \mathcal{T}^{(i)}$ ($v_1 + v_2 = 1$). Then from (47) we deduce that

$$\begin{aligned}
\theta_k^{(1)} &= \theta_{k-1}^{(1)} - \gamma_k \sum_{i=1}^{2} v_i a_i (1-\alpha a_i)^r (\theta_{k-1}^{(1)} - \frac{b_i}{a_i}) \\
&= (1 - \gamma_k \sum_{i=1}^{2} v_i a_i (1-\alpha a_i)^r) \cdot \theta_{k-1}^{(1)} + \gamma_k v_1 a_1 (1-\alpha a_1)^r \cdot \frac{b_1}{a_1} \\
&+ \gamma_k v_2 a_2 (1-\alpha a_2)^r \cdot \frac{b_2}{a_2} \\
&\in [\min(\theta_{k-1}^{(1)}, \frac{b_1}{a_1}, \frac{b_2}{a_2}), \max(\theta_{k-1}^{(1)}, \frac{b_1}{a_1}, \frac{b_2}{a_2})] \subseteq I
\end{aligned}$$

since $\theta_k^{(1)}$ is a convex combination of $\theta_{k-1}^{(1)}, \frac{b_1}{a_1}, \frac{b_2}{a_2}$. Indeed, due to (46)

$$0 \leq (1 - \gamma_k \sum_{i=1}^{2} v_i a_i (1-\alpha a_i)^r), \gamma_k v_1 a_1 (1-\alpha a_1)^r, \gamma_k v_2 a_2 (1-\alpha a_2)^r \leq 1$$

and

$$(1 - \gamma_k \sum_{i=1}^{2} v_i a_i (1-\alpha a_i)^r) + \gamma_k v_1 a_1 (1-\alpha a_1)^r + \gamma_k v_2 a_2 (1-\alpha a_2)^r = 1.$$

As a result of this Case we conclude that if $k \geq k_0$ and $\theta_{k-1}^{(1)} \in I$, then for all $k' \geq k$ it also holds that $\theta_{k'}^{(1)} \in I$.

2. Case 2: $\theta_{k-1}^{(1)} > \frac{b_1}{a_1} + A$. From (41) observe that for $i \in \{1,2\}$ and any $x \in \mathbb{R}$ $f_i''(x) \leq a_i$. Hence, f_i''s Lipschitz constant is a_i. Let $\phi_0, \ldots, \phi_r$ and $\overline{\phi}_0, \ldots, \overline{\phi}_r$ be two inner-GD (3) rollouts for task $\mathcal{T}^{(i)}$ and $\phi_0^{(1)} > \overline{\phi}_0^{(1)}$. For $j \in \{1, \ldots, r\}$ suppose that $\phi_{j-1}^{(1)} > \overline{\phi}_{j-1}^{(1)}$. Then

$$\begin{aligned}
\phi_j^{(1)} - \overline{\phi}_j^{(1)} &= \phi_{j-1}^{(1)} - \overline{\phi}_{j-1}^{(1)} - \alpha(f_i'(\phi_{j-1}^{(1)}) - f_i'(\overline{\phi}_{j-1}^{(1)})) \\
&\geq \phi_{j-1}^{(1)} - \overline{\phi}_{j-1}^{(1)} - \alpha|f_i'(\phi_{j-1}^{(1)}) - f_i'(\overline{\phi}_{j-1}^{(1)})| \\
&\geq \phi_{j-1}^{(1)} - \overline{\phi}_{j-1}^{(1)} - \alpha a_i|\phi_{j-1}^{(1)} - \overline{\phi}_{j-1}^{(1)}| \\
&> \phi_{j-1}^{(1)} - \overline{\phi}_{j-1}^{(1)} - |\phi_{j-1}^{(1)} - \overline{\phi}_{j-1}^{(1)}| \\
&= 0
\end{aligned}$$

or $\phi_j^{(1)} > \overline{\phi}_j^{(1)}$ where we use Lipschitz continuity of f_i' and that $\alpha a_i < 1$ by the choice of a_1, a_2. Therefore, since $\phi_0^{(1)} > \overline{\phi}_0^{(1)}$, $\phi_1^{(1)} > \overline{\phi}_1^{(1)}$ and so on, eventually $\phi_r^{(1)} > \overline{\phi}_r^{(1)}$. Observe that $f_i'(x)$ is a strictly monotonously increasing function, therefore $f_i'(\phi_r^{(1)}) > f_i'(\overline{\phi}_r^{(1)})$. To sum up:

$$f_i'(\phi_r^{(1)}) > f_i'(\overline{\phi}_r^{(1)}) \quad \text{when } \phi_0^{(1)} > \overline{\phi}_0^{(1)}. \tag{48}$$

Set $\overline{\phi}_0^{(1)} = \frac{b_i}{a_i}$, then $f'(\overline{\phi}_{j-1}^{(1)}) = 0$ and $\overline{\phi}_1^{(1)} = \overline{\phi}_0^{(1)} - \alpha \cdot 0 = \overline{\phi}_0^{(1)}$ and so on, eventually $\overline{\phi}_r^{(1)} = \frac{b_i}{a_i}$ and $f'(\overline{\phi}_r^{(1)}) = 0$. Therefore, if $\phi_0^{(1)} = \frac{b_1}{a_1} + A > \max(\frac{b_1}{a_1}, \frac{b_2}{a_2})$ then $f_i'(\phi_r^{(1)}) > f_i'(\overline{\phi}_r^{(1)}) = 0$. For $i \in \{1, 2\}$ denote a deterministic value of $f_i'(\phi_r^{(1)})$ by $B_i > 0$. By setting $\phi_0^{(1)} = \theta_{k-1}^{(1)}, \overline{\phi}_0^{(1)} = \frac{b_1}{a_1} + A$ and using (48) we obtain:

$$\mathcal{G}_{FO}(\theta_{k-1}, \mathcal{T}_i)^{(1)} = f_i'(\phi_r^{(1)}) > f_i'(\overline{\phi}_r^{(1)}) = B_i \geq B > 0. \tag{49}$$

where we denote $B = \min(B_1, B_2)$.

In addition, set $\phi_0^{(1)} = \theta_{k-1}^{(1)}, \overline{\phi}_0^{(1)} = \frac{b_i}{a_i}$. Then

$$\begin{aligned}
\mathcal{G}_{FO}(\theta_{k-1}, \mathcal{T}^{(i)})^{(1)} &= |\mathcal{G}_{FO}(\theta_{k-1}, \mathcal{T}^{(i)})^{(1)}| = |f_i'(\phi_r^{(1)}) - 0| = |f_i'(\phi_r^{(1)}) - f_i'(\overline{\phi}_r^{(1)})| \\
&\leq a_i |\phi_r^{(1)} - \overline{\phi}_r^{(1)}| = a_i |\phi_{r-1}^{(1)} - \overline{\phi}_{r-1}^{(1)} - \alpha(f_i'(\phi_{r-1}^{(1)}) - f_i'(\overline{\phi}_{r-1}^{(1)}))| \\
&\leq a_i |\phi_{r-1}^{(1)} - \overline{\phi}_{r-1}^{(1)}| + \alpha a_i |f_i'(\phi_{r-1}^{(1)}) - f_i'(\overline{\phi}_{r-1}^{(1)})| \\
&\leq a_i (1 + \alpha a_i) |\phi_{r-1}^{(1)} - \overline{\phi}_{r-1}^{(1)}| \\
&\dots \\
&\leq a_i (1 + \alpha a_i)^r |\phi_0^{(1)} - \overline{\phi}_0^{(1)}| \\
&= a_i (1 + \alpha a_i)^r |\theta_{k-1}^{(1)} - \frac{b_i}{a_i}|.
\end{aligned}$$

Since $\theta_{k-1}^{(1)} > \frac{b_1}{a_1} + A > \max(\frac{b_1}{a_1}, \frac{b_2}{a_2})$, we derive that

$$\begin{aligned}
\mathcal{G}_{FO}(\theta_{k-1}, \mathcal{T}^{(i)})^{(1)} &\leq a_i (1 + \alpha a_i)^r (\theta_{k-1}^{(1)} - \frac{b_i}{a_i}) \leq \frac{1}{\gamma_k} (\theta_{k-1}^{(1)} - \frac{b_i}{a_i}) \\
&\leq \max_{i' \in \{1,2\}} \frac{1}{\gamma_k} (\theta_{k-1}^{(1)} - \frac{b_{i'}}{a_{i'}}) = \frac{1}{\gamma_k} (\theta_{k-1}^{(1)} - \min_{i' \in \{1,2\}} \frac{b_{i'}}{a_{i'}}) \\
&= \frac{1}{\gamma_k} (\theta_{k-1}^{(1)} - \frac{b_1}{a_1})
\end{aligned}$$

where we use (46) and the fact that $\frac{b_1}{a_1} = 0$, $\frac{b_2}{a_2} > 0$. Next, we deduce that

$$\theta_k^{(1)} = \theta_{k-1}^{(1)} - \gamma_k \mathcal{G}_{FO}(\theta_{k-1}, \mathcal{T}_k)^{(1)} \geq \theta_{k-1}^{(1)} - \frac{\gamma_k}{\gamma_k} (\theta_{k-1}^{(1)} - \frac{b_1}{a_1}) = \frac{b_1}{a_1}. \tag{50}$$

On the other hand, from (49)

$$\mathcal{G}_{FO}(\theta_{k-1}, \mathcal{T}_k)^{(1)} > B$$

and

$$\theta_k^{(1)} = \theta_{k-1}^{(1)} - \gamma_k \mathcal{G}_{FO}(\theta_{k-1}, \mathcal{T}_k)^{(1)} < \theta_{k-1}^{(1)} - \gamma_k B. \tag{51}$$

According to (5) there exists a number $k_1 > k$ such that

$$\sum_{k'=k}^{k_1-1} \gamma_{k'} > \frac{1}{B} (\theta_{k-1}^{(1)} - \frac{b_1}{a_1}). \tag{52}$$

In addition, let k_1 be a minimal such number. Suppose that for all $k \leq k' \leq k_1$ $\theta^{(1)}_{k'-1} > \frac{b_1}{a_1} + A$. Then by applying bound (51) for all $k = k'$ we obtain that

$$\theta^{(1)}_{k_1} < \theta^{(1)}_{k_1-1} - \gamma_{k_1} B < \cdots < \theta^{(1)}_{k-1} - B \sum_{k'=k}^{k_1} \gamma_{k'} < \frac{b_1}{a_1}$$

which is a contradiction with the bound (50) applied to $k = k_1$. Therefore, there exists $k \leq k' < k_1$ such that $\theta^{(1)}_{k'} \leq \frac{b_1}{a_1} + A$. Then there exists a number

$$k_2 = \min_{k \leq k' < k_1, \theta^{(1)}_{k'} \leq \frac{b_1}{a_1} + A} k'. \tag{53}$$

Hence, $\theta^{(1)}_{k_2-1} > \frac{b_1}{a_1} + A$ and by applying bound (50) to $k = k_2$ we conclude that $\theta^{(1)}_{k_2} \geq \frac{b_1}{a_1}$. Averall:

$$\theta^{(1)}_{k_2} \in [\frac{b_1}{a_1}, \frac{b_1}{a_1} + A] \subseteq I.$$

As shown in Case 1, for all $k' > k_2$ (including k_1) it also holds that $\theta^{(1)}_{k'} \in I$. To summarize, we have proven that there exists a deterministic number $B > 0$ such that for k_1 defined by (52) $\theta^{(1)}_{k'} \in I$ for all $k' \geq k_1$.

3. Case 3: $\theta^{(1)}_{k-1} < \frac{b_2}{a_2} - A$. Using a symmetric argument as in Case 2 it can be shown that there exists a deterministic number $C > 0$ so that the following holds. According to (5) there exists $k_3 \geq k$ such that

$$\sum_{k'=k}^{k_3-1} \gamma_{k'} > \frac{1}{C}(\frac{b_2}{a_2} - \theta^{(1)}_{k-1}). \tag{54}$$

In addition, let k_3 be a minimal such number. Then $\theta^{(1)}_{k'} \in I$ for all $k' \geq k_3$.

Since $p(\mathcal{T})$ is a discrete distribution, there only exists a finite number of outcomes for a set of random variables $\{\mathcal{T}_k\}_{k<k_0}$. Therefore, there is only a finite set of possible outcomes of the $\theta^{(1)}_{k_0-1}$ random variable. Consequently, there exists a deterministic number $E > 0$ such that $|\theta^{(1)}_{k_0-1}| < E$. According to (5) there exist deterministic numbers $k_4, k_5 \geq k_0$ such that

$$\sum_{k'=k_0}^{k_4-1} \gamma_{k'} > \frac{1}{B}(E - \frac{b_1}{a_1}), \quad \sum_{k'=k_0}^{k_5-1} \gamma_{k'} > \frac{1}{C}(\frac{b_2}{a_2} + E). \tag{55}$$

and let $k_6 = \max(k_4, k_5)$, which is also a deterministic number. Let k_1, k_3 be random numbers from Cases 2, 3 applied to $k = k_0$. Then from (55) and E's definition, it follows that $k_1, k_3 \leq k_6$. In addition, $k_0 \leq k_6$ from k_6's definition. As a result of all cases, we conclude that for any $k \geq k_6$ $\phi^{(1)}_k \in I$.

Denote

$$a^* = \frac{1}{2}(a_1(1 - \alpha a_1)^r + a_2(1 - \alpha a_2)^r), \quad b^* = \frac{1}{2}(b_1(1 - \alpha a_1)^r + b_2(1 - \alpha a_2)^r), \quad x^* = \frac{b^*}{a^*}$$

and consider arbitrary $k > k_6$. Denote $\overline{\mathcal{G}} = \mathcal{G}_{FO}(\theta_{k-1}, \mathcal{T}_k)$ and let $\mathcal{F}_k$ be a σ-algebra populated by $\{\mathcal{T}_\kappa\}, \kappa < k$. From Equation (47) we conclude that

$$\mathbb{E}\left[\overline{\mathcal{G}}^{(1)} | \mathcal{F}_k\right] = a^* \theta^{(1)}_{k-1} - b^*$$

Outer-loop update leads to an expression:

$$\theta^{(1)}_k = \theta^{(1)}_{k-1} - \gamma_k \overline{\mathcal{G}}^{(1)}.$$

Subtract x^*:

$$\theta^{(1)}_k - x^* = \theta^{(1)}_{k-1} - x^* - \gamma_k \overline{\mathcal{G}}^{(1)}.$$

Take a square:

$$(\theta_k^{(1)} - x^*)^2 = (\theta_{k-1}^{(1)} - x^* - \gamma_k \overline{\mathcal{G}}^{(1)})^2 \qquad = (\theta_{k-1}^{(1)} - x^*)^2 - 2\gamma_k(\theta_{k-1}^{(1)} - x^*)\overline{\mathcal{G}}^{(1)} + \gamma_k^2 \overline{\mathcal{G}}^{(1)2}.$$

Take expectation conditioned on $\mathcal{F}_k$:

$$\begin{aligned}
\mathbb{E}\left[(\theta_k^{(1)} - x^*)^2 | \mathcal{F}_k\right] &= (\theta_{k-1}^{(1)} - x^*)^2 - 2\gamma_k(\theta_{k-1}^{(1)} - x^*)\mathbb{E}\left[\overline{\mathcal{G}}^{(1)} | \mathcal{F}_k\right] + \gamma_k^2 \mathbb{E}\left[\left(\overline{\mathcal{G}}^{(1)}\right)^2 | \mathcal{F}_k\right] \\
&= (\theta_{k-1}^{(1)} - x^*)^2 - 2\gamma_k(\theta_{k-1}^{(1)} - x^*)(a^*\theta_{k-1}^{(1)} - b^*) + \gamma_k^2 \mathbb{E}\left[\left(\overline{\mathcal{G}}^{(1)}\right)^2 | \mathcal{F}_k\right] \\
&= (\theta_{k-1}^{(1)} - x^*)^2 - 2\gamma_k a^*(\theta_{k-1}^{(1)} - x^*)(\theta_{k-1}^{(1)} - \frac{b^*}{a^*}) + \gamma_k^2 \mathbb{E}\left[\left(\overline{\mathcal{G}}^{(1)}\right)^2 | \mathcal{F}_k\right] \\
&= (1 - 2\gamma_k a^*)(\theta_{k-1}^{(1)} - x^*)^2 + \gamma_k^2 \mathbb{E}\left[\left(\overline{\mathcal{G}}^{(1)}\right)^2 | \mathcal{F}_k\right].
\end{aligned}$$

For $i \in \{1, 2\}$, $\mathcal{G}_{FO}(\theta_{k-1}, \mathcal{T}^{(i)})^{(1)}$ depends linearly on $\theta_{k-1}^{(1)}$ (47) and, therefore, is bounded on $\theta_{k-1}^{(1)} \in I$. Hence, $\overline{\mathcal{G}}^{(1)2}$ is also bounded by a deterministic number $F > 0$: $\overline{\mathcal{G}}^{(1)2} < F$. Then:

$$\mathbb{E}\left[(\theta_k^{(1)} - x^*)^2 | \mathcal{F}_k\right] \leq (1 - 2\gamma_k a^*)(\theta_{k-1}^{(1)} - x^*)^2 + \gamma_k^2 F.$$

Take a full expectation:

$$\mathbb{E}\left[(\theta_k^{(1)} - x^*)^2\right] \leq (1 - 2\gamma_k a^*)\mathbb{E}(\theta_{k-1}^{(1)} - x^*)^2 + \gamma_k^2 F,$$

and denote $y_k = \mathbb{E}\left[(\theta_k^{(1)} - x^*)^2\right]$:

$$y_k \leq (1 - 2\gamma_k a^*) y_{k-1} + \gamma_k^2 F. \tag{56}$$

Now, we prove that $\lim_{k\to\infty} y_k = 0$. Indeed, consider arbitrary $\epsilon > 0$. According to (5) there exists $k_\epsilon > k_6$ such that $\forall k' \geq k_\epsilon : \gamma_{k'} \leq \frac{a^*\epsilon}{F}$. As a result of (56) for every $k \geq k_\epsilon$ it holds

$$y_k \leq (1 - 2\gamma_k a^*) y_{k-1} + \gamma_k^2 F \leq (1 - 2\gamma_k a^*) y_{k-1} + \gamma_k a^* \epsilon.$$

Subtract $\frac{\epsilon}{2}$:

$$y_k - \frac{\epsilon}{2} \leq (1 - 2\gamma_k a^*)(y_{k-1} - \frac{\epsilon}{2}). \tag{57}$$

Observe that by (45), a^*'s definition and since $k \geq k_0$ it holds that $1 - 2\gamma_k a^* > 0$. Therefore and since (57) holds for all $k \geq k_\epsilon$, it can be written that for all $k \geq k_\epsilon$

$$y_k - \frac{\epsilon}{2} \leq \left(\prod_{k'=k_\epsilon}^{k} (1 - 2\gamma_{k'} a^*)\right)(y_{k_\epsilon - 1} - \frac{\epsilon}{2}) \leq \left(\prod_{k'=k_\epsilon}^{k} (1 - 2\gamma_{k'} a^*)\right)|y_{k_\epsilon - 1} - \frac{\epsilon}{2}|.$$

We use inequality $1 - x \leq \exp(-x)$ to deduce that

$$\begin{aligned}
y_k - \frac{\epsilon}{2} &\leq \left(\prod_{k'=k_\epsilon}^{k} (1 - 2\gamma_{k'} a^*)\right)|y_{k_\epsilon - 1} - \frac{\epsilon}{2}| \\
&\leq \exp\left(-2a^* \sum_{k'=k_\epsilon}^{k} \gamma_{k'}\right)|y_{k_\epsilon - 1} - \frac{\epsilon}{2}|.
\end{aligned} \tag{58}$$

If $|y_{k_\epsilon - 1} - \frac{\epsilon}{2}| = 0$, then from (58) it follows that $y_k \leq 0 + \frac{\epsilon}{2} < \epsilon$ for all $k \geq k_\epsilon$. Otherwise, from (5) there exists k'_ϵ such that $\sum_{k'=k_\epsilon}^{k'_\epsilon} \gamma_{k'} > \frac{\log|y_{k_\epsilon - 1} - \frac{\epsilon}{2}| - \log\frac{\epsilon}{2}}{2a^*}$. Then from (58) it follows that for all $k \geq k'_\epsilon$ $y_k - \frac{\epsilon}{2} < \frac{\epsilon}{2}$ or $y_k < \epsilon$. Since $y_k \geq 0$ by definition, we have proven that $\lim_{k\to\infty} y_k = 0$, or

$$\lim_{k\to\infty} \mathbb{E}\left[(\theta_k^{(1)} - x^*)^2\right] = 0. \tag{59}$$

Again, let $k > k_6$. Let $\phi_0 = \theta_k, \ldots, \phi_r$ be a rollout (3) of inner GD for task $\mathcal{T}^{(i)}$. Then according to (6-9)

$$\nabla_{\theta_k}\mathcal{L}^{out}(\theta_k, U^{(r)}(\theta_k, \mathcal{T}^{(i)}), \mathcal{T}^{(i)})^{(1)} = \mathcal{G}_{FO}(\theta_k, \mathcal{T}^{(i)})^{(1)} \prod_{j=0}^{r-1}(1 - \alpha f_i''(\phi_j^{(1)})).$$

From (42) it follows that $f_i''(\phi_j^{(1)}) = a_i$ for $j \in \{0, \ldots, r-1\}$. Moreover, we use (47) to obtain that

$$\nabla_{\theta_k}\mathcal{L}^{out}(\theta_k, U^{(r)}(\theta_k, \mathcal{T}^{(i)}), \mathcal{T}^{(i)})^{(1)} = a_i(1 - \alpha a_i)^{2r}(\theta_k^{(1)} - \frac{b_i}{a_i})$$

and

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_k)^{(1)} &= \mathbb{E}_{p(\mathcal{T})}\left[\nabla_{\theta_k}\mathcal{L}^{out}(\theta_k, U^{(r)}(\theta_k, \mathcal{T}^{(i)}), \mathcal{T}^{(i)})^{(1)}\right] \\ &= \frac{1}{2}\left(a_1(1 - \alpha a_1)^{2r}(\theta_k^{(1)} - \frac{b_1}{a_1}) + a_2(1 - \alpha a_2)^{2r}(\theta_k^{(1)} - \frac{b_2}{a_2})\right) \\ &= \widehat{a}\theta_k^{(1)} - \widehat{b}\end{aligned}$$

where

$$\widehat{a} = \frac{1}{2}(a_1(1 - \alpha a_1)^{2r} + a_2(1 - \alpha a_2)^{2r}), \quad \widehat{b} = \frac{1}{2}(b_1(1 - \alpha a_1)^{2r} + b_2(1 - \alpha a_2)^{2r}).$$

Notice that since $b_1 = 0$ and b_2 is defined by (37), it appears that $|\widehat{a}x^* - \widehat{b}| = \sqrt{2D}$, or $(\widehat{a}x^* - \widehat{b})^2 = 2D$. Multiply (59) by $\widehat{a}$ to obtain that

$$\lim_{k\to\infty}\mathbb{E}\Big[(\widehat{a}\theta_k^{(1)} - \widehat{a}x^*)^2\Big] = 0, \tag{60}$$

$$\lim_{k\to\infty}\mathbb{E}\Big[(\widehat{a}\theta_k^{(1)} - \widehat{b} - (\widehat{a}x^* - \widehat{b}))^2\Big] = 0,$$

$$\lim_{k\to\infty}\mathbb{E}\Big[(\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_k)^{(1)} - (\widehat{a}x^* - \widehat{b}))^2\Big] = 0. \tag{61}$$

For each $k \geq 1$ by Jensen's inequality :

$$\left(\mathbb{E}\left[\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_k)^{(1)}\right] - (\widehat{a}x^* - \widehat{b})\right)^2 \leq \mathbb{E}\left[(\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_k)^{(1)} - (\widehat{a}x^* - \widehat{b}))^2\right].$$

Hence,

$$\lim_{k\to\infty}\left(\mathbb{E}\left[\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_k)^{(1)}\right] - (\widehat{a}x^* - \widehat{b}))\right)^2 = 0, \quad \lim_{k\to\infty}\mathbb{E}\left[\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_k)^{(1)}\right] = (\widehat{a}x^* - \widehat{b})$$

and by expanding (61) we derive that

$$\lim_{k\to\infty}\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_k)^{(1)}\right)^2\right] = 2(\widehat{a}x^* - \widehat{b})\lim_{k\to\infty}\mathbb{E}\left[\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_k)^{(1)}\right] - (\widehat{a}x^* - \widehat{b})^2 = (\widehat{a}x^* - \widehat{b})^2 = 2D.$$

We conclude the proof by observing that

$$\liminf_{k\to\infty}\mathbb{E}\left[\|\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_k)\|_2^2\right] = \lim_{k\to\infty}\mathbb{E}\left[\|\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_k)\|_2^2\right] = \lim_{k\to\infty}\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\mathcal{M}^{(r)}(\theta_k)^{(1)}\right)^2\right] = 2D > D.$$

□

F Optimal Choice of q

The derivative of (19) has the form

$$\begin{aligned} g(q) &= \frac{2}{2\epsilon-1}((1/q-1)\mathbb{D}_{true}^2+\mathbb{V}_{true}^2)^{\frac{1+2\epsilon}{1-2\epsilon}}q^{-2}\mathbb{D}_{true}^2(C_{det}+C_{rnd}q)+\left((1/q-1)\mathbb{D}_{true}^2+\mathbb{V}_{true}^2\right)^{\frac{2}{1-2\epsilon}}C_{rnd} \\ &= q^{-2}((1/q-1)\mathbb{D}_{true}^2+\mathbb{V}_{true}^2)^{\frac{1+2\epsilon}{1-2\epsilon}}\left(\frac{2}{2\epsilon-1}\mathbb{D}_{true}^2(C_{det}+C_{rnd}q)+C_{rnd}(\mathbb{D}_{true}^2+q(\mathbb{V}_{true}^2-\mathbb{D}_{true}^2))q\right) \\ &= q^{-2}((1/q-1)\mathbb{D}_{true}^2+\mathbb{V}_{true}^2)^{\frac{1+2\epsilon}{1-2\epsilon}}\left(C_{rnd}(\mathbb{V}_{true}^2-\mathbb{D}_{true}^2)q^2+\frac{2\epsilon+1}{2\epsilon-1}\mathbb{D}_{true}^2C_{rnd}q+\frac{2}{2\epsilon-1}\mathbb{D}_{true}^2C_{det}\right). \quad (62) \end{aligned}$$

We further deduce that

$$g(1)=(\mathbb{V}_{true}^2)^{\frac{1+2\epsilon}{1-2\epsilon}}\left(C_{rnd}\mathbb{V}_{true}^2+\frac{2}{2\epsilon-1}(C_{det}+C_{rnd})\mathbb{D}_{true}^2\right).$$

Assuming that $\mathbb{V}_{true}^2>0$, we conclude that $g(1)>0$ iff

$$\mathbb{D}_{true}^2<\frac{C_{rnd}}{\frac{2}{1-2\epsilon}(C_{det}+C_{rnd})}\mathbb{V}_{true}^2. \quad (63)$$

$g(q)$ is differentiable on $(0,+\infty)$. Let q^* be the value corresponding to the minimum of (19) on $(0,1]$. q^* exists, since (19) approaches $+\infty$ when $q^*\to 0$. $g(1)>0$ indicates that q^* is not equal to 1, i.e. we obtain a tigher upper bound using UFOM rather than exact gradients. Further, solving $g(q)=0$ reduces to solving a quadratic equation induced by the polynomial inside big brackets of (62):

$$\mathrm{poly}(q)=C_{rnd}(\mathbb{V}_{true}^2-\mathbb{D}_{true}^2)q^2+\frac{2\epsilon+1}{2\epsilon-1}\mathbb{D}_{true}^2C_{rnd}q+\frac{2}{2\epsilon-1}\mathbb{D}_{true}^2C_{det}=0$$

Notice, that if $\epsilon<\frac{1}{2}$ and (63) is satisfied, then $\mathrm{poly}(1)>0$ and $\mathrm{poly}(0)<0$. Hence, from the continuity of $\mathrm{poly}(q)$, it follows that there is an odd number of roots of $\mathrm{poly}(q)$ on $(0,1)$. Since the quadratic equation has at most 2 roots, we conclude that there is a single root on $(0,1)$. Since (19) is differentiable on $(0,1]$, $q=1$ is not a local minimum of (19) on $(0,1]$ and $\lim_{q\to+0}g(q)=+\infty$, we conclude that this single root is a minimum of (19) on $(0,1]$.