# Supplementary Material

The supplementary material is organized as follows. Appendix A offers more algorithm illustration for the 4 topics we discuss in Sec. 3. In Appendix B we discuss our choice for EPG formulation and the truncated setting in GMRL. In Appendix C we briefly summarise biased Hessian estimation issue in MAML-RL mentioned in Section 4.2, In Appendix D we illustrate how realistic are Assumption 4.1-4.3 of Section 4. Appendices E to G contain the proofs for the results presented in the paper. In Appendix H we provide statements and proofs for some auxiliary lemmas which are instrumental for the main results. For convenience of the reader, before each proof we also restate the corresponding theorem. Finally, in Appendix I we present additional experiments results.

## A More topics on GMRL

### A.1 Few-shot Reinforcement Learning

One important research field in Meta Reinforcement Learning is few-shot Reinforcement Learning. The main objective of this research field is to enable Reinforcement Learning agent with fast adaptation ability. Instead of thousands of interactions in traditional Reinforcement Learning algorithms, agent in few-shot setting is only allowed to interact with the new environment for a few trajectories. One of the most classical gradient based algorithms in this field is **Model Agnostic Meta Learning (MAML-RL)**. [10] aims at learning neural network's initial parameters for fast adaptation on new environments. It assumes distribution $\rho(\mathcal{T})$ over RL environment $\mathcal{T}$ and tries to optimise $\boldsymbol{\theta}$ which leads to high-performing updated policy $\boldsymbol{\theta}'$. The objective equation for one-step MAML-RL can be shown as follows:

$$\begin{aligned} J(\boldsymbol{\theta}) &= \mathbb{E}_{\mathcal{T}\sim\rho(\mathcal{T})}\left[\mathbb{E}_{\boldsymbol{\tau}'\sim P_{\mathcal{T}}(\boldsymbol{\tau}'|\boldsymbol{\theta}')}\left[R\left(\boldsymbol{\tau}'\right)\right]\right] \\ \text{with} \quad \boldsymbol{\theta}' &= \boldsymbol{\theta} + \alpha\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\tau}\sim P_{\mathcal{T}}(\boldsymbol{\tau}|\boldsymbol{\theta})}[R(\boldsymbol{\tau})] \end{aligned} \tag{11}$$

where in practice we use the limited trajectories sampled from the new environment to estimate $\nabla_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\tau}\sim P_{\mathcal{T}}(\boldsymbol{\tau}|\boldsymbol{\theta})}[R(\boldsymbol{\tau})]$. During training, by estimating meta policy gradient $\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta})$, MAML can conduct meta update on the initial policy parameters.

In the scope of Eq. (2), MAML-RL optimizes over meta initial parameters to maximize the return of one-step adapted policy: $\boldsymbol{\theta}' = \boldsymbol{\theta} + \alpha\nabla_{\boldsymbol{\theta}}J^{\text{In}}(\boldsymbol{\theta})$. In MAML-RL, $J^{\text{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}')$ degenerates to $J^{\text{Out}}(\boldsymbol{\theta}')$ and $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ represent the same initial parameters. The meta-gradient can be derived with the following equation:

$$\nabla_{\boldsymbol{\theta}}J(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\boldsymbol{\theta}'\nabla_{\boldsymbol{\theta}'}J^{\text{Out}}\left(\boldsymbol{\theta}'\right), \nabla_{\boldsymbol{\theta}}\boldsymbol{\theta}' = I + \alpha\nabla^2_{\boldsymbol{\theta}}J^{\text{In}}(\boldsymbol{\theta}) \tag{12}$$

### A.2 Meta-gradient in Opponent Shaping

Opponent shaping [11, 19, 22] is a powerful tool in multi-agent learning process for different purposes. For instance, Foerster et al. [11] and Letcher et al. [22] have shown that putting other-players learning dynamic into self-learning process can bring in cooperation behaviors, which may help to reach better social welfare compared with purely independent learning. Meta-gradient estimation is needed when ego-agent takes derivatives of other-agent policy gradient step. **Learning with Opponent-Learning Awareness (LOLA)** [11] proposed a new learning objective by including an additional term accounting for the impact of ego policy to the anticipated opponent gradient update. Specifically, in the two-player setting, with agent 1 policy $\boldsymbol{\phi}$ and agent 2 policy $\boldsymbol{\theta}$, the traditional independent learning (IL) and 1-step LOLA algorithm can result in different updates for agent 1:

$$\begin{aligned} \boldsymbol{\phi}'_{\text{IL}} &= \boldsymbol{\phi} + \beta\nabla_{\boldsymbol{\phi}}J^{\text{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}) \\ \boldsymbol{\phi}'_{\text{LOLA}} &= \boldsymbol{\phi} + \beta\nabla_{\boldsymbol{\phi}}J^{\text{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}') \\ \text{where } \boldsymbol{\theta}' &= \boldsymbol{\theta} + \alpha\nabla_{\boldsymbol{\theta}}J^{\text{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}) \end{aligned} \tag{13}$$

Where $\beta$ refers to the outer learning rate and $J^{\text{In/Out}}$ refers to the value function for agent 2 and agent 1 respectively. For meta-agent 1 with parameters $\boldsymbol{\phi}$, it will optimise its return over one-step-lookahead opponent parameters $\boldsymbol{\theta}'$. Thus the meta-gradient of meta-agent corresponds exactly to Eq. (2) with $\nabla_{\boldsymbol{\phi}}\boldsymbol{\theta}' = \alpha\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}}J^{\text{In}}(\boldsymbol{\phi}, \boldsymbol{\theta})$. Note that this one-step-lookahead is a just virtual update considered in the optimisation of agent 1. Agent 2 can also choose this LOLA update by conducting one-step-lookahead over agent 1.

### A.3 Single-lifetime Meta-gradient RL

In this setting, the main objective is to self-tune the meta parameters ($\gamma$ in [39]) or meta models (intrinsic model in [42]) along with the underlying normal RL updates. It is called online because it only involves one single RL life-time. This research field is also related with online hyperparameter optimisation in supervised learning such as [2, 14]. Xu et al. [39] proposed meta-gradient reinforcement learning (MGRL) to tune the discount factor $\gamma$ and bootstrapping parameter $\lambda$ in an online manner. It tries to differentiate through one RL inner update to optimize the meta-parameters and maximise one-step policy return.

$$\begin{aligned} &\max_{\boldsymbol{\eta}} V^{\pi_{\boldsymbol{\theta}'}}, \text{where } \boldsymbol{\theta}' = \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\eta}), \text{and} \\ \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\tau}, \boldsymbol{\theta}, \boldsymbol{\eta}) &= \left(g_{\boldsymbol{\eta}}(\boldsymbol{\tau}) - v_{\boldsymbol{\theta}}(S)\right) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(A \mid S) \\ &\quad + \left(g_{\boldsymbol{\eta}}(\boldsymbol{\tau}) - v_{\boldsymbol{\theta}}(S)\right) \nabla_{\boldsymbol{\theta}} v_{\boldsymbol{\theta}}(S) \\ &\quad + \nabla_{\boldsymbol{\theta}} H\left(\pi_{\boldsymbol{\theta}}(\cdot \mid S)\right) \end{aligned} \tag{14}$$

where $\eta$ refers to $(\gamma, \lambda)$, $\boldsymbol{\tau}$ refers to trajectories, $g_{\boldsymbol{\eta}}$, $v_{\boldsymbol{\theta}}$, H represent GAE estimation, value function and entropy respectively. Eq. (14) combines actor loss, critic loss and entropy loss, which are commonly used in typical Actor-Critic [25] algorithms. Specifically, the meta parameters $(\gamma, \lambda)$ corresponds to $\boldsymbol{\phi}$ in Eq. (2) . After the policy parameters $\boldsymbol{\theta}$ take one policy gradient update to become $\boldsymbol{\theta}'(\boldsymbol{\theta}' = \boldsymbol{\theta} + \alpha \nabla_{\boldsymbol{\theta}} J^{\text{In}}(\boldsymbol{\theta}, \boldsymbol{\phi}))$, we can calculate the meta-gradient by backpropogating from $J^{\text{Out}}$ to meta parameters. In MGRL, $J^{\text{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}')$ degenerates to $J^{\text{Out}}(\boldsymbol{\theta}')$. The meta-gradient can be shown as:

$$\nabla_{\boldsymbol{\phi}} J(\boldsymbol{\phi}) = \nabla_{\boldsymbol{\phi}} \boldsymbol{\theta}' \nabla_{\boldsymbol{\theta}'} J^{\text{Out}}\left(\boldsymbol{\theta}'\right), \nabla_{\boldsymbol{\phi}} \boldsymbol{\theta}' = \alpha \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}} J^{\text{In}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \tag{15}$$

Here for simplicity we omit the critic and entropy loss. Usually work in this research field only conduct one-step inner-loop update before taking meta update. Some recent works such as [34, 4] have also shown that multi-step online meta-gradient can achieve better performance.

### A.4 Multi-lifetime Meta-gradient RL

Existing work like [26, 43, 40, 9] are trying to learn some fundamental/generalizable meta module across different environments such as a neural RL algorithm in [26](LPG). An important feature of multi-lifetime Meta-gradient RL is that it inherently needs multi-step inner-loop to account for the effect of fundamental meta module over the RL process. The objective of LPG is to learn a neural network based RL algorithm, by which a RL agent can be properly trained. The mathmatical formulation can be shown as follows:

$$J(\boldsymbol{\phi}) = \mathbb{E}_{\mathcal{T} \sim \rho(\mathcal{T})}\left[\mathbb{E}_{\boldsymbol{\tau}^K \sim P_{\mathcal{T}}\left(\boldsymbol{\tau}^K \mid \theta^K\right)}\left[R\left(\boldsymbol{\tau}^K\right)\right]\right], \text{with} \tag{16}$$

$$\theta^i = \theta^{i-1} + \alpha \nabla_{\theta^{i-1}} \mathbb{E}_{\boldsymbol{\tau} \sim P_{\mathcal{T}}(\boldsymbol{\tau} \mid \theta^{i-1})}\left[f_{\boldsymbol{\phi}}(\boldsymbol{\tau})\right] \tag{17}$$

where $f_{\boldsymbol{\phi}}(\boldsymbol{\tau})$ is the output of meta-network $\boldsymbol{\phi}$ for conducting inner-loop neural policy gradient and $k$ can be large to show the long-range impact brought by neural RL algorithm. We omit the kl inner loss used in [26] for simplicity. In the scope of Eq. (1), $J^{\text{In/Out}}$ refers to the value function, $\boldsymbol{\theta}$ represents the RL agent policy parameters and $\boldsymbol{\phi}$ is the meta-parameter of neural RL algorithm. Most of works are under a multi-task/environment (or a distribution over environment) and multi-lifetime setting. [40] is a special case in these work because it is also under the online setting. We believe the main reason is that the training iterations/sample complexity in [40] is real large (1e9) and makes it become a special case of 'multi-lifetime' setting.

## B Discussion of expected policy gradient (EPG) formulation and truncated setting

We discuss 4 research topics in Section 3: few-shot RL(MAML-RL), opponent shaping(LOLA-DiCE), online meta gradient RL(MGRL) and meta gradient based inverse design(LPG). And we need to discuss how this multi-step EPG inner-loop formulation differs in these topics. Though they all need meta policy gradient estimation, the differences between setting and final objective require us to discuss them separately.

**Different setting**: MAML-RL and most inverse design algorithms are under multi-lifetime setting which can renew an environment and restart the RL training from the very beginning. Work in online

meta gradient RL/LOLA only happen in a single lifetime RL process. There only exists one RL training process.

**Different objectives**: For MAML-RL, the main objective is to maximise the return of few-step adapted policy. Thus the objective corresponds exactly to few-step inner-loop formulation. However, for topics beyond few-shot RL, in most case they need to measure the influence of meta module over RL final (after thousands of steps) performance.

There are two important issues in this EPG formulation. The first one is that it assumes an expected policy gradient inner-loop update. And the second one is because we only consider few-step inner-loop update so they are under a truncated estimation setting which might bring in bias. Recently, one work [36] argues that: (1) the general unbiased meta gradient for MAML-RL ([10]) and Online Meta Gradient ([39],[42]) should be the K-sample inner-loop meta gradient shown in E-MAML [1] rather than the expected policy gradient inner-loop meta gradient used in many recent work [23, 29, 33]. (2) The gradient estimator in online meta gradient utilise truncated optimization and the unbiased meta gradient should be the one in untruncated setting.

Overall we agree that: (1) The K-sample inner-loop meta gradient estimator is unbiased for MAML-RL problem when sampled policy gradient are used. (2) To learn an schedule (rather than a global meta module) of meta-parameter/meta-module for MGRL or to learn some fundamental concepts in inverse-design, the gradient estimator in untruncated setting is unbiased. However, we argue that (1) For MAML-related problem, the variance of sampling correction term in K-sample inner-loop meta gradient estimator is large because it needs to sum up all $k$ terms and that is why [36] proposes to use one coefficient to control. The EPG can achieve lower variance estimation and perform better empirically [29] (2) For meta gradient based inverse design with multi-lifetime, the few-step meta gradient estimation under truncated setting is biased.

However, in online meta gradient setting (MGRL) or online opponent modelling (LOLA) with single-lifetime, things are completely different thus a direct transform of K-sample inner-loop formulation from MAML to MGRL might not be that straightforward. There exists a large gap between the implementation of online meta gradient algorithm and the final objective (meta-module/hyperparameters schedule) we may wish. First, it's an online setting so the multiple lifetime setting where the algorithm can restart from the very beginning and reiterate the whole process is banned here. This makes the estimation of unbiased meta gradient impossible because the algorithm cannot access to the future dynamic for gradient estimation. The experiments with multi-lifetime training in [36] is in fact out of the scope of online meta gradient setting and are more like meta gradient based inverse design. Second, in implementation of MGRL they only maintain one running $\gamma$ or intrinsic model rather than multiple meta modules as a real schedule needs. Also, recently there exist one work [4] discussing multi-step MGRL and use one fixed meta parameters rather than a schedule for multi-step inner-loop, which may show a different understanding about untruncated gradient. In all, we believe that what online meta algorithm/opponent shaping like MGRL or LOLA optimizes and what the best they can achieve in such online setting are still open questions and remain to be further explored. It is really hard to simply formulate the unbiased meta gradient since the gap between implementation and objective is still not clear.

Thus, in our paper, we still focuses on the previous work (MAML/MGRL and LOLA) objectives with EPG inner-loop setting and use its meta gradient as our target gradient. All bias term we discuss is the bias w.r.t. the expected meta gradient in this EPG inner-loop and truncated setting. That is our work's limitation and we leave more things for future work: (1)The gap between EPG inner-loop meta gradient and K-sample inner-loop meta gradient in MAML-RL related problem. (2) The gap between truncated EPG inner-loop meta gradient and what the best gradient estimation we can get in online meta gradient/opponent shaping. (3) The gap between truncated EPG inner-loop meta gradient and the untruncated gradient in meta gradient based inverse design.

## C Brief summary on biased Hessian estimation in MAML-RL

We will briefly introduce the reasons of biased Hessian estimation with automatic differentiation in one-step MAML-RL. Firstly, we can derive the analytic form of $\boldsymbol{\theta}^1$ and $\nabla_{\boldsymbol{\theta}^0}\boldsymbol{\theta}^1$

$$\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 + \alpha \mathbb{E}_{\boldsymbol{\tau}\sim p(\boldsymbol{\tau};\boldsymbol{\theta}^0)}[\nabla_{\boldsymbol{\theta}^0} \log \pi(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})] \tag{18}$$

$$\nabla_{\theta^0}\theta^1 = I + \mathbb{E}_{\tau\sim p(\tau;\theta^0)}\left[\mathcal{R}(\tau)\left(\nabla^2_{\theta^0}\log\pi_{\theta^0}(\tau) + \nabla_{\theta^0}\log\pi_{\theta^0}(\tau)\nabla_{\theta^0}\log\pi_{\theta^0}(\tau)^\top\right)\right] \tag{19}$$

Typically we need to use trajectory samples $\tau_n$ to estimate the policy gradient, we can get the adapted policy estimate.

$$\hat{\theta}^1 = \theta^0 + \alpha\frac{1}{N}\sum_{\tau_n}\sum_{t=0}^{H-1}\nabla_\theta\log\pi_\theta(a_t^n \mid s_t^n)\left(\sum_{t'=0}^{H}\gamma^t r\left(s_{t'}^n, a_{t'}^n\right)\right) \tag{20}$$

Finally, implementation of MAML-RL derives the gradient estimate by automatic differentation. The corresponding estimation is biased:

$$\begin{aligned}\mathbb{E}[\nabla_{\theta^0}\hat{\theta}^1] &= I + \alpha\mathbb{E}_{\tau\sim p(\tau;\theta^0)}\left[\frac{1}{N}\sum_{\tau_n}\sum_{t=0}^{H-1}\nabla^2_{\theta^0}\log\pi_{\theta^0}(a_t^n \mid s_t^n)\left(\sum_{t'=0}^{H}\gamma^t r(s_{t'}^n, a_{t'}^n)\right)\right] \\ &= I + \alpha\mathbb{E}_{\tau\sim p(\tau;\theta^0)}\left[\mathcal{R}(\tau)\nabla^2_\theta\log\pi_{\theta^0}(\tau)\right] \neq \nabla_{\theta^0}\theta^1\end{aligned} \tag{21}$$

The main reason of biased Hessian estimation is that automatic differentiation tools only consider the dependency of $\theta$ in $\nabla_\theta\log\pi_\theta$ while ignoring the dependency in expectation $\mathbb{E}_{\tau\sim p(\tau;\theta^0)}$. In practice, the $\mathbb{E}_{\tau\sim p(\tau;\theta^0)}$ is represented by trajectory sampling so the gradient term $\nabla_\theta\mathbb{E}_{\tau\sim p(\tau;\theta^0)}$ is 0 using automatic differentiation. We need to add additional terms to further derive the gradient $\nabla_\theta\mathbb{E}_{\tau\sim p(\tau;\theta^0)}$ brought by sampling dependency.

# D Limitations on Assumptions

Assumption 4.1-4.3 are standard assumptions used in various theoretical MAML-RL papers [6, 7, 18]. The Lipschitz continuity assumptions in Assumption 4.1 make sure we can work with nonconvex inner and outer objectives. The unbiased first-order gradient estimators assumptions in Assumption 4.2 can highlight our findings on two source of biases, which is also a plausible assumption in GMRL settings. As typically adopted in the analysis for stochastic optimization, we make the bounded-variance assumption in Assumption 4.3. Assumption 4.1-4.3 can be conveniently verified for e.g., inner-loop RL optimization in tabular MDP settings (finite state space and action space) with soft-max parameterisation of the policy, where $\pi_\theta(a \mid s) \propto \exp(\theta(s,a))$ with parameter $\theta = \theta(s,a)$. But in large-scale RL settings like atari games, Assumption 4.1-4.3 will not hold anymore.

# E Proof of Proposition in Section 3

In this section, we provide the proof for Proposition 3.1 in Section 3.

## E.1 Proof of Proposition 3.1

**Proposition 3.1** ($K$-step Meta-Gradient)**.** *The exact meta-gradient to the objective in Eq. (2) can be written as:*

$$\begin{aligned}\nabla_\phi J^K(\phi) &= \nabla_\phi J^{Out}(\phi,\theta^K) + \alpha\nabla_\phi\theta^K\nabla_{\theta^K}J^{Out}(\phi,\theta^K), \\ \nabla_\phi\theta^K &= \sum_{i=0}^{K-1}\nabla_\phi\nabla_{\theta^i}J^{In}(\phi,\theta^i)\prod_{j=i+1}^{K-1}\left(I + \alpha\nabla^2_{\theta^j}J^{In}(\phi,\theta^j)\right).\end{aligned} \tag{3}$$

*Proof.* According to post-update inner parameters $\theta^K = \theta^0 + \alpha\sum_{i=0}^{K-1}\nabla_{\theta^i}J^{\text{In}}(\phi,\theta^i)$ and the fact that $\nabla_{\theta^i}J^{\text{In}}(\phi,\theta^i)$ is differentiable w.r.t. $\phi$, we can treat $\theta^K$ as a differentiable function w.r.t. $\phi$. Based on the chain rule, we can get

$$\nabla_\phi J^K(\phi) = \nabla_\phi J^{\text{Out}}(\phi,\theta^K) + \nabla_\phi\theta^K\nabla_{\theta^K}J^{\text{Out}}(\phi,\theta^K) \tag{22}$$

Based on the iterative updates that $\theta^{i+1} = \theta^i + \alpha\nabla_{\theta^i}J^{In}(\phi,\theta^i)$, for $i = 0,\ldots,K-1$ and similarly treat $\theta^i$ as a differentiable function w.r.t. $\phi$, we have

$$\begin{aligned}\nabla_{\boldsymbol{\phi}}\boldsymbol{\theta}^{i+1} =&\nabla_{\boldsymbol{\phi}}\boldsymbol{\theta}^{i} + \alpha\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i) + \alpha\nabla_{\boldsymbol{\phi}}\boldsymbol{\theta}^{i}\nabla^2_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\\ =&\nabla_{\boldsymbol{\phi}}\boldsymbol{\theta}^{i}\left(I + \alpha\nabla^2_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\right) + \alpha\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\end{aligned} \tag{23}$$

Telescoping the above equality over $i$ from 0 to $K-1$, we can get

$$\begin{aligned}\nabla_{\boldsymbol{\phi}}\boldsymbol{\theta}^{K} =&\nabla_{\boldsymbol{\phi}}\boldsymbol{\theta}^{0}\prod_{i=0}^{K-1}\left(I + \alpha\nabla^2_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\right) + \alpha\sum_{i=0}^{K-1}\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\prod_{j=i+1}^{K-1}\left(I + \alpha\nabla^2_{\boldsymbol{\theta}^j}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^j)\right)\\ =&\alpha\sum_{i=0}^{K-1}\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\prod_{j=i+1}^{K-1}\left(I + \alpha\nabla^2_{\boldsymbol{\theta}^j}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^j)\right)\end{aligned} \tag{24}$$

Combining Eq. (22) and Eq. (24) finishes the proof of Proposition 3.1. □

# F Proof of Lemma in Section 4

## F.1 Proof of Lemma 4.4

**Lemma 4.4** (Compositional Bias)**.** *Suppose that Assumption 4.1 and 4.2 hold, let* $\hat{\Delta}_C = \mathbb{E}[\|f(\hat{\boldsymbol{\theta}}^K) - f(\boldsymbol{\theta}^K)\|]$ *be the compositional bias and* $C_0$ *the Lipschitz constant of* $f(\cdot)$*,* $|\boldsymbol{\tau}|$ *denote number of trajectories used to estimate inner-loop gradient in each inner-loop update step,* $\alpha$ *the learning rate, then we have,*

$$\hat{\Delta}_C \leq C_0\mathbb{E}[\|\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^K\|] \leq C_0\left((1+\alpha c_2)^K - 1\right)\frac{\hat{\sigma}_{In}}{c_2\sqrt{|\boldsymbol{\tau}|}}, \tag{6}$$

*where* $\hat{\sigma}_{In} = \max_i\sqrt{\mathbb{V}[\nabla_{\boldsymbol{\theta}^i}\hat{J}^{In}(\boldsymbol{\phi},\boldsymbol{\theta}^i,\boldsymbol{\tau}_0^i)]}$*,* $i \in \{0,..,K-1\}$*.*

*Proof.* In expected policy gradient inner-loop update setting, the iterative updates takes the form

$$\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i + \alpha\nabla_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i),\ i = 0,\ldots,K-1 \tag{25}$$

In Eq. (4), $\boldsymbol{\theta}^{i+1}$ are estimated using samples $\boldsymbol{\tau}_0^{0:i}$, then we have

$$\hat{\boldsymbol{\theta}}^{i+1} = \hat{\boldsymbol{\theta}}^i + \alpha\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_0^i),\ \hat{\boldsymbol{\theta}}^0 = \boldsymbol{\theta}^0,\ i = 0,\ldots,K-1 \tag{26}$$

According to the assumption that non-linear compositional vector-valued $f(\cdot)$ is Lipschitz continuous with constant $C_0$, we can get

$$\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}[\|f(\hat{\boldsymbol{\theta}}^K) - f(\boldsymbol{\theta}^K)\|] \leq C_0\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^K\right\|\right] \tag{27}$$

Based on Eq. (25) and Eq. (26), we can get

$$
\begin{aligned}
&\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^K-\boldsymbol{\theta}^K\right\|\right]\\
=&\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^{K-1}-\boldsymbol{\theta}^{K-1}+\alpha\nabla_{\boldsymbol{\theta}^{K-1}}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^{K-1})-\alpha\nabla_{\hat{\boldsymbol{\theta}}^{K-1}}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^{K-1},\boldsymbol{\tau}_0^{K-1})\right\|\right]\\
\overset{(i)}{\leq}&\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^{K-1}-\boldsymbol{\theta}^{K-1}\right\|\right]+\alpha\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\nabla_{\boldsymbol{\theta}^{K-1}}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^{K-1})-\mathbb{E}_{\tau_0^{K-1}}[\nabla_{\hat{\boldsymbol{\theta}}^{K-1}}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^{K-1},\boldsymbol{\tau}_0^{K-1})]\right\|\right]+\\
&\alpha\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\mathbb{E}_{\tau_0^{K-1}}[\nabla_{\hat{\boldsymbol{\theta}}^{K-1}}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^{K-1},\boldsymbol{\tau}_0^{K-1})]-\nabla_{\hat{\boldsymbol{\theta}}^{K-1}}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^{K-1},\boldsymbol{\tau}_0^{K-1})\right\|\right]\\
\leq&\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^{K-1}-\boldsymbol{\theta}^{K-1}\right\|\right]+\alpha c_2\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^{K-1}-\boldsymbol{\theta}^{K-1}\right\|\right]+\\
&\alpha\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\mathbb{E}_{\tau_0^{K-1}}[\nabla_{\hat{\boldsymbol{\theta}}^{K-1}}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^{K-1},\boldsymbol{\tau}_0^{K-1})]-\nabla_{\hat{\boldsymbol{\theta}}^{K-1}}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^{K-1},\boldsymbol{\tau}_0^{K-1})\right\|\right]\\
\leq&(1+\alpha c_2)\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^{K-1}-\boldsymbol{\theta}^{K-1}\right\|\right]+\\
&\alpha\mathbb{E}_{\tau_0^{0:K-2}}\left[\mathbb{E}_{\tau_0^{K-1}}\left[\left\|\mathbb{E}_{\tau_0^{K-1}}[\nabla_{\hat{\boldsymbol{\theta}}^{K-1}}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^{K-1},\boldsymbol{\tau}_0^{K-1})]-\nabla_{\hat{\boldsymbol{\theta}}^{K-1}}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^{K-1},\boldsymbol{\tau}_0^{K-1})\right\|\mid\boldsymbol{\tau}_0^{0:K-2}\right]\right]\\
\leq&(1+\alpha c_2)\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^{K-1}-\boldsymbol{\theta}^{K-1}\right\|\right]+\alpha\mathbb{E}_{\tau_0^{0:K-2}}\left[\sqrt{\frac{\mathbb{V}\left[\nabla_{\hat{\boldsymbol{\theta}}^{K-1}}\hat{J}^{\mathrm{In}}\left(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^{K-1},\boldsymbol{\tau}_0^{K-1}\right)\mid\boldsymbol{\tau}_0^{0:K-2}\right]}{|\boldsymbol{\tau}_0^{K-1}|}}\right]
\end{aligned}
\tag{28}
$$

where $(i)$ follows from Lemma H.3 and Assumption 4.2.

Let $\hat{\sigma}_{\mathrm{In}}=\max_i\sqrt{\mathbb{V}[\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\tau_0^i)]}$, $|\boldsymbol{\tau}|=|\tau_0^i|$, $i\in\{0,\ldots,K-1\}$

$$
\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^K-\boldsymbol{\theta}^K\right\|\right]\leq(1+\alpha c_2)\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^{K-1}-\boldsymbol{\theta}^{K-1}\right\|\right]+\alpha\frac{\hat{\sigma}_{\mathrm{In}}}{\sqrt{|\boldsymbol{\tau}|}}
\tag{29}
$$

Iteratively, we can get

$$
\begin{aligned}
\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^K-\boldsymbol{\theta}^K\right\|\right]&\leq\left(1+\ldots+(1+\alpha c_2)^{K-1}\right)\alpha\frac{\hat{\sigma}_{\mathrm{In}}}{\sqrt{|\boldsymbol{\tau}|}}\\
&=\left((1+\alpha c_2)^K-1\right)\frac{\hat{\sigma}_{\mathrm{In}}}{c_2\sqrt{|\boldsymbol{\tau}|}}
\end{aligned}
\tag{30}
$$

which concludes the proof of Lemma 4.4. □

# G Proof of Theorem in Section 4

## G.1 Proof of Theorem 4.5

**Theorem 4.5** (Upper bound for the bias and the variance)**.** *Suppose that Assumption 4.1 and 4.2 and 4.3 hold. Let $J_{\boldsymbol{\phi},\boldsymbol{\theta}}$ denote $\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}}J^{In}$, $H_{\boldsymbol{\theta},\boldsymbol{\theta}}$ denote $\nabla^2_{\boldsymbol{\theta}}J^{In}$, $\hat{\Delta}^K=\|\mathbb{E}[\nabla_{\boldsymbol{\phi}}\hat{J}^K(\boldsymbol{\phi})]-\nabla_{\boldsymbol{\phi}}J^K(\boldsymbol{\phi})\|$ be the meta-gradient estimation bias, set $B=1+\alpha c_2$. Then the bound of bias hold:*

$$
\hat{\Delta}^K\leq O\left((B+\hat{\Delta}_H)^{K-1}\left(\mathbb{E}[\|\hat{\boldsymbol{\theta}}^K-\boldsymbol{\theta}^K\|]+\hat{\Delta}_J+(K-1)\right)\right).
\tag{9}
$$

*Let* $(\hat{\sigma}^K)^2 = \mathbb{V}\left[\nabla_{\boldsymbol{\phi}} \hat{J}^K(\boldsymbol{\phi})\right]$ *be the meta-gradient estimation variance, set* $V_1 = (1+\alpha c_2)^2$, $V_2 = 2\alpha^2(m_1^2 + 3\sigma_2^2)$ *the estimation variance is given by*

$$
\begin{aligned}
(\hat{\sigma}^K)^2 \leq O\Bigg( & (V_1 + \hat{\Delta}_H^2)^{K-1} \left( \mathbb{E}[\|\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^K\|^2] + (K-1) \right) \\
& + \left( V_2 + (V_1 + \hat{\Delta}_H^2 + \hat{\sigma}_H^2)^{K-1} - (V_1 + \hat{\Delta}_H^2)^{K-1} \right) (\hat{\Delta}_J^2 + \hat{\sigma}_J^2) \Bigg).
\end{aligned} \tag{10}
$$

*where* $\hat{\Delta}_J = \max_{\boldsymbol{\phi}\times\boldsymbol{\theta}} \|\mathbb{E}[\hat{J}_{\boldsymbol{\phi},\boldsymbol{\theta}}] - J_{\boldsymbol{\phi},\boldsymbol{\theta}}\|$. $\hat{\Delta}_H = \max_{\boldsymbol{\theta}} \|\mathbb{E}[\hat{H}_{\boldsymbol{\theta},\boldsymbol{\theta}}] - H_{\boldsymbol{\theta},\boldsymbol{\theta}}\|$. $(\hat{\sigma}_J)^2 = \frac{\max_{\boldsymbol{\phi}\times\boldsymbol{\theta}} \mathbb{V}[\hat{J}_{\boldsymbol{\phi},\boldsymbol{\theta}}]}{|\tau|}$. $(\hat{\sigma}_H)^2 = \frac{\max_{\boldsymbol{\theta}} \mathbb{V}[\hat{H}_{\boldsymbol{\theta},\boldsymbol{\theta}}]}{|\tau|}$.

*Proof.* According to Proposition 3.1, exact meta-gradient $\nabla_{\boldsymbol{\phi}} J^{\mathrm{K}}(\boldsymbol{\phi})$ takes the form

$$
\nabla_{\boldsymbol{\phi}} J^{\mathrm{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}^K) + \alpha \sum_{i=0}^{K-1} \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}^i} J^{\mathrm{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^i) \prod_{j=i+1}^{K-1} \left( I + \alpha \nabla_{\boldsymbol{\theta}^j}^2 J^{\mathrm{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^j) \right) \nabla_{\boldsymbol{\theta}^K} J^{\mathrm{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}^K) \tag{31}
$$

where

$$
\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i + \alpha \nabla_{\boldsymbol{\theta}^i} J^{\mathrm{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^i),\ i = 0, \ldots, K-1 \tag{32}
$$

Acoordingly, in Eq. (4), $K$-step meta-gradient estimator $\nabla_{\boldsymbol{\phi}} \hat{J}^{\mathrm{K}}(\boldsymbol{\phi})$ takes the form

$$
\nabla_{\boldsymbol{\phi}} \hat{J}^{\mathrm{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K, \boldsymbol{\tau}_3) + \alpha \sum_{i=0}^{K-1} \nabla_{\boldsymbol{\phi}} \nabla_{\hat{\boldsymbol{\theta}}^i} \hat{J}^{\mathrm{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^i, \boldsymbol{\tau}_1^i) \prod_{j=i+1}^{K-1} \left( I + \alpha \nabla_{\hat{\boldsymbol{\theta}}^j}^2 \hat{J}^{\mathrm{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \right) \nabla_{\hat{\boldsymbol{\theta}}^K} \hat{J}^{\mathrm{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K, \boldsymbol{\tau}_3) \tag{33}
$$

where

$$
\hat{\boldsymbol{\theta}}^{i+1} = \hat{\boldsymbol{\theta}}^i + \alpha \nabla_{\hat{\boldsymbol{\theta}}^i} \hat{J}^{\mathrm{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^i, \boldsymbol{\tau}_0^i),\ \hat{\boldsymbol{\theta}}^0 = \boldsymbol{\theta}^0,\ i = 0, \ldots, K-1 \tag{34}
$$

Hence the expectation of meta-gradient estimator takes the form

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}, \boldsymbol{\tau}_1^{0:K-1}, \boldsymbol{\tau}_2^{1:K-1}, \boldsymbol{\tau}_3} [\nabla_{\boldsymbol{\phi}} \hat{J}^{\mathrm{K}}(\boldsymbol{\phi})] \\
= & \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}} \Bigg[ \mathbb{E}_{\boldsymbol{\tau}_3} [\nabla_{\boldsymbol{\phi}} \hat{J}^{\mathrm{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K, \boldsymbol{\tau}_3) \mid \boldsymbol{\tau}_0^{0:K-1}] + \alpha \sum_{i=0}^{K-1} \mathbb{E}_{\boldsymbol{\tau}_1^i} [\nabla_{\boldsymbol{\phi}} \nabla_{\hat{\boldsymbol{\theta}}^i} \hat{J}^{\mathrm{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^i, \boldsymbol{\tau}_1^i) \mid \boldsymbol{\tau}_0^{0:i-1}] \times \\
& \prod_{j=i+1}^{K-1} \mathbb{E}_{\boldsymbol{\tau}_2^j} [I + \alpha \nabla_{\hat{\boldsymbol{\theta}}^j}^2 \hat{J}^{\mathrm{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \mid \boldsymbol{\tau}_0^{0:j-1}] \times \mathbb{E}_{\boldsymbol{\tau}_3} [\nabla_{\hat{\boldsymbol{\theta}}^K} \hat{J}^{\mathrm{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K, \boldsymbol{\tau}_3) \mid \boldsymbol{\tau}_0^{0:K-1}] \Bigg]
\end{aligned} \tag{35}
$$

we can then derive meta-gradient bias in $K$-step expected policy gradient setting,

$$
\begin{aligned}
& \left\| \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}, \boldsymbol{\tau}_1^{0:K-1}, \boldsymbol{\tau}_2^{0:K-1}, \boldsymbol{\tau}_3} [\nabla_{\boldsymbol{\phi}} \hat{J}^{\mathrm{K}}(\boldsymbol{\phi})] - \nabla_{\boldsymbol{\phi}} J^{\mathrm{K}}(\boldsymbol{\phi}) \right\| \\
\leq & \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}} \Bigg[ \Bigg\| \mathbb{E}_{\boldsymbol{\tau}_3} [\nabla_{\boldsymbol{\phi}} \hat{J}^{\mathrm{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K, \boldsymbol{\tau}_3) \mid \boldsymbol{\tau}_0^{0:K-1}] + \alpha \sum_{i=0}^{K-1} \mathbb{E}_{\boldsymbol{\tau}_1^i} [\nabla_{\boldsymbol{\phi}} \nabla_{\hat{\boldsymbol{\theta}}^i} \hat{J}^{\mathrm{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^i, \boldsymbol{\tau}_1^i) \mid \boldsymbol{\tau}_0^{0:i-1}] \times \\
& \prod_{j=i+1}^{K-1} \mathbb{E}_{\boldsymbol{\tau}_2^j} [I + \alpha \nabla_{\hat{\boldsymbol{\theta}}^j}^2 \hat{J}^{\mathrm{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \mid \boldsymbol{\tau}_0^{0:i-1}] \times \mathbb{E}_{\boldsymbol{\tau}_3} [\nabla_{\hat{\boldsymbol{\theta}}^K} \hat{J}^{\mathrm{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K, \boldsymbol{\tau}_3) \mid \boldsymbol{\tau}_0^{0:K-1}] - \\
& \nabla_{\boldsymbol{\phi}} J^{\mathrm{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}^K) - \alpha \sum_{i=0}^{K-1} \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}^i} J^{\mathrm{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^i) \prod_{j=i+1}^{K-1} \left( I + \alpha \nabla_{\boldsymbol{\theta}^j}^2 J^{\mathrm{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^j) \right) \nabla_{\boldsymbol{\theta}^K} J^{\mathrm{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}^K) \Bigg\| \Bigg]
\end{aligned} \tag{36}
$$

$$
\begin{aligned}
&\leq \mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\mathbb{E}_{\tau_3}[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{Out}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3)\mid\boldsymbol{\tau}_0^{0:K-1}]-\nabla_{\boldsymbol{\phi}}J^{\mathrm{Out}}(\boldsymbol{\phi},\boldsymbol{\theta}^K)\right\|\right]\\
&\quad+\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\alpha\sum_{i=0}^{K-1}\mathbb{E}_{\tau_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\tau_1^i)\mid\boldsymbol{\tau}_0^{0:i-1}]\times\prod_{j=i+1}^{K-1}\mathbb{E}_{\tau_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\tau_2^j)\mid\boldsymbol{\tau}_0^{0:i-1}]\right.\right.\\
&\quad\times\mathbb{E}_{\tau_3}[\nabla_{\hat{\boldsymbol{\theta}}^K}\hat{J}^{\mathrm{Out}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3)\mid\boldsymbol{\tau}_0^{0:K-1}]\\
&\quad\left.\left.-\alpha\sum_{i=0}^{K-1}\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\prod_{j=i+1}^{K-1}\left(I+\alpha\nabla^2_{\boldsymbol{\theta}^j}J^{\mathrm{In}}\left(\boldsymbol{\phi},\boldsymbol{\theta}^j\right)\right)\nabla_{\boldsymbol{\theta}^K}J^{\mathrm{Out}}(\boldsymbol{\phi},\boldsymbol{\theta}^K)\right\|\right]\\
&\overset{(i)}{\leq}\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\nabla_{\boldsymbol{\phi}}J^{\mathrm{Out}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K)-\nabla_{\boldsymbol{\phi}}J^{\mathrm{Out}}(\boldsymbol{\phi},\boldsymbol{\theta}^K)\right\|\mid\boldsymbol{\tau}_0^{0:K-1}\right]\\
&\quad+\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\alpha\sum_{i=0}^{K-1}\mathbb{E}_{\tau_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\tau_1^i)\mid\boldsymbol{\tau}_0^{0:i-1}]\times\prod_{j=i+1}^{K-1}\mathbb{E}_{\tau_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\tau_2^j)\mid\boldsymbol{\tau}_0^{0:i-1}]\right.\right.\\
&\quad\times\mathbb{E}_{\tau_3}[\nabla_{\hat{\boldsymbol{\theta}}^K}\hat{J}^{\mathrm{Out}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3)\mid\boldsymbol{\tau}_0^{0:K-1}]-\\
&\quad\left.\left.\alpha\sum_{i=0}^{K-1}\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\prod_{j=i+1}^{K-1}\left(I+\alpha\nabla^2_{\boldsymbol{\theta}^j}J^{\mathrm{In}}\left(\boldsymbol{\phi},\boldsymbol{\theta}^j\right)\right)\nabla_{\boldsymbol{\theta}^K}J^{\mathrm{Out}}(\boldsymbol{\phi},\boldsymbol{\theta}^K)\right\|\right]
\end{aligned}
\tag{37}
$$

where $(i)$ follows from Assumption 4.2.

$$
\begin{aligned}
&\overset{(ii)}{\leq}\mu_1\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^K-\boldsymbol{\theta}^K\right\|\mid\boldsymbol{\tau}_0^{0:K-1}\right]\\
&\quad+\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\alpha\sum_{i=0}^{K-1}\mathbb{E}_{\tau_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\tau_1^i)\mid\boldsymbol{\tau}_0^{0:i-1}]\times\prod_{j=i+1}^{K-1}\mathbb{E}_{\tau_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\tau_2^j)\mid\boldsymbol{\tau}_0^{0:i-1}]\right.\right.\\
&\quad\left.\left.\times\mathbb{E}_{\tau_3}[\nabla_{\hat{\boldsymbol{\theta}}^K}\hat{J}^{\mathrm{Out}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3)\mid\boldsymbol{\tau}_0^{0:K-1}]-\alpha\sum_{i=0}^{K-1}\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\prod_{j=i+1}^{K-1}\left(I+\alpha\nabla^2_{\boldsymbol{\theta}^j}J^{\mathrm{In}}\left(\boldsymbol{\phi},\boldsymbol{\theta}^j\right)\right)\nabla_{\boldsymbol{\theta}^K}J^{\mathrm{Out}}(\boldsymbol{\phi},\boldsymbol{\theta}^K)\right\|\right]
\end{aligned}
\tag{38}
$$

where $(ii)$ follows from Assumption 4.1 on Lipschitz Continuity of $\nabla_{\boldsymbol{\phi}}J^{\mathrm{Out}}$

$$
\begin{aligned}
&\overset{(iii)}{\leq}\mu_1\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^K-\boldsymbol{\theta}^K\right\|\mid\boldsymbol{\tau}_0^{0:K-1}\right]\\
&\quad+\alpha\sum_{i=0}^{K-1}\mathbb{E}_{\tau_0^{0:K-1}}\left[\left\|\mathbb{E}_{\tau_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\tau_1^i)\mid\boldsymbol{\tau}_0^{0:i-1}]\times\right.\right.\\
&\quad\prod_{j=i+1}^{K-1}\mathbb{E}_{\tau_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\tau_2^j)\mid\boldsymbol{\tau}_0^{0:i-1}]\times\mathbb{E}_{\tau_3}[\nabla_{\hat{\boldsymbol{\theta}}^K}\hat{J}^{\mathrm{Out}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3)\mid\boldsymbol{\tau}_0^{0:K-1}]-\\
&\quad\left.\left.\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\prod_{j=i+1}^{K-1}\left(I+\alpha\nabla^2_{\boldsymbol{\theta}^j}J^{\mathrm{In}}\left(\boldsymbol{\phi},\boldsymbol{\theta}^j\right)\right)\nabla_{\boldsymbol{\theta}^K}J^{\mathrm{Out}}(\boldsymbol{\phi},\boldsymbol{\theta}^K)\right\|\right]
\end{aligned}
\tag{39}
$$

where $(iii)$ follows from Lemma H.3. Using the similar add-minus trick in the proof of Lemma 4.4, we can have

$$
\begin{aligned}
&\overset{(iv)}{\leq} \mu_1 \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^K\right\| \mid \boldsymbol{\tau}_0^{0:K-1}\right] \\
&\quad + \alpha \sum_{i=0}^{K-1} \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\| \prod_{j=i+1}^{K-1} \mathbb{E}_{\tau_2^j}[I + \alpha \nabla^2_{\hat{\boldsymbol{\theta}}^j} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \mid \boldsymbol{\tau}_0^{0:i-1}] \times \mathbb{E}_{\tau_3}[\nabla_{\hat{\boldsymbol{\theta}}^K} \hat{J}^{\text{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K, \tau_3) \mid \boldsymbol{\tau}_0^{0:K-1}] - \right.\right. \\
&\quad \left.\left. \prod_{j=i+1}^{K-1} \left(I + \alpha \nabla^2_{\boldsymbol{\theta}^j} J^{\text{In}}\left(\boldsymbol{\phi}, \boldsymbol{\theta}^j\right)\right) \nabla_{\boldsymbol{\theta}^K} J^{\text{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}^K) \right\|\right] \times \|\nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}^i} J^{\text{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^i)\| + \\
&\quad \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\mathbb{E}_{\tau_1^i}[\nabla_{\boldsymbol{\phi}} \nabla_{\hat{\boldsymbol{\theta}}^i} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^i, \boldsymbol{\tau}_1^i) \mid \boldsymbol{\tau}_0^{0:i-1}] - \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}^i} J^{\text{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^i)\right\|\right] \times \\
&\quad \left\|\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\prod_{j=i+1}^{K-1} \mathbb{E}_{\tau_2^j}[I - \nabla^2_{\hat{\boldsymbol{\theta}}^j} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \mid \boldsymbol{\tau}_0^{0:i-1}] \times \mathbb{E}_{\tau_3}[\nabla_{\hat{\boldsymbol{\theta}}^K} \hat{J}^{\text{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K, \tau_3) \mid \boldsymbol{\tau}_0^{0:K-1}]\right]\right\|
\end{aligned}
\tag{40}
$$

Based on Assumption 4.1 and Assumption 4.2, we can change the expectation of unbiased first-order stochastic estimator to respective first-order gradient function, then we can replace it with Lipschitz constants.

$$
\begin{aligned}
&\leq \mu_1 \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^K\right\| \mid \boldsymbol{\tau}_0^{0:K-1}\right] \\
&\quad + \alpha \sum_{i=0}^{K-1} \left[\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\| \prod_{j=i+1}^{K-1} \mathbb{E}_{\tau_2^j}[I + \alpha \nabla^2_{\hat{\boldsymbol{\theta}}^j} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \mid \boldsymbol{\tau}_0^{0:i-1}] - \prod_{j=i+1}^{K-1} \left(I + \alpha \nabla^2_{\boldsymbol{\theta}^j} J^{\text{In}}\left(\boldsymbol{\phi}, \boldsymbol{\theta}^j\right)\right) \right\|\right] \times \right. \\
&\quad \left. \left\|\nabla_{\boldsymbol{\theta}^K} J^{\text{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}^K)\right\| + \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\mathbb{E}_{\tau_3}[\nabla_{\hat{\boldsymbol{\theta}}^K} \hat{J}^{\text{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K, \tau_3) \mid \boldsymbol{\tau}_0^{0:K-1}] - \nabla_{\boldsymbol{\theta}^K} J^{\text{Out}}(\boldsymbol{\phi}, \boldsymbol{\theta}^K)\right\|\right] \times \right. \\
&\quad \left. \left\|\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\prod_{j=i+1}^{K-1} \mathbb{E}_{\tau_2^j}[I + \alpha \nabla^2_{\hat{\boldsymbol{\theta}}^j} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \mid \boldsymbol{\tau}_0^{0:i-1}]\right]\right\|\right] \times \|\nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}^i} J^{\text{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^i)\| + \\
&\quad \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\mathbb{E}_{\tau_1^i}[\nabla_{\boldsymbol{\phi}} \nabla_{\hat{\boldsymbol{\theta}}^i} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^i, \boldsymbol{\tau}_1^i) \mid \boldsymbol{\tau}_0^{0:i-1}] - \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}^i} J^{\text{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^i)\right\|\right] \times \\
&\quad \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\| \prod_{j=i+1}^{K-1} \mathbb{E}_{\tau_2^j}[I + \alpha \nabla^2_{\hat{\boldsymbol{\theta}}^j} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \mid \boldsymbol{\tau}_0^{0:i-1}]\right\| \times \left\|\mathbb{E}_{\tau_3}[\nabla_{\hat{\boldsymbol{\theta}}^K} \hat{J}^{\text{Out}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^K, \tau_3) \mid \boldsymbol{\tau}_0^{0:K-1}]\right\|\right]
\end{aligned}
\tag{41}
$$

$$
\begin{aligned}
&\leq \mu_1 \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^K\right\|\right] \\
&+ \alpha \sum_{i=0}^{K-1} c_1 \left[ m_2 \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\| \prod_{j=i+1}^{K-1} \mathbb{E}_{\boldsymbol{\tau}_2^j}[I + \alpha \nabla^2_{\hat{\boldsymbol{\theta}}^j} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \mid \boldsymbol{\tau}_0^{0:i-1}] - \prod_{j=i+1}^{K-1} \left(I + \alpha \nabla^2_{\boldsymbol{\theta}^j} J^{\text{In}}\left(\boldsymbol{\phi}, \boldsymbol{\theta}^j\right)\right) \right\|\right] \right. \\
&\left. + \mu_2 \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^K\right\|\right] \times \left\| \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[ \prod_{j=i+1}^{K-1} \mathbb{E}_{\boldsymbol{\tau}_2^j}[I + \alpha \nabla^2_{\hat{\boldsymbol{\theta}}^j} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \mid \boldsymbol{\tau}_0^{0:i-1}] \right] \right\| \right] + \\
&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\| \mathbb{E}_{\boldsymbol{\tau}_1^i}[\nabla_{\boldsymbol{\phi}} \nabla_{\hat{\boldsymbol{\theta}}^i} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^i, \boldsymbol{\tau}_1^i) \mid \boldsymbol{\tau}_0^{0:i-1}] - \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}^i} J^{\text{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^i) \right\|\right] \times \\
&m_2 \left\| \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[ \prod_{j=i+1}^{K-1} \mathbb{E}_{\boldsymbol{\tau}_2^j}[I + \alpha \nabla^2_{\hat{\boldsymbol{\theta}}^j} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \mid \boldsymbol{\tau}_0^{0:i-1}] \right] \right\|
\end{aligned}
\tag{42}
$$

$$
\begin{aligned}
&\leq \mu_1 \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^K\right\|\right] \\
&+ \alpha \sum_{i=0}^{K-1} c_1 m_2 \underbrace{\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\| \prod_{j=i+1}^{K-1} \mathbb{E}_{\boldsymbol{\tau}_2^j}[I + \alpha \nabla^2_{\hat{\boldsymbol{\theta}}^j} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \mid \boldsymbol{\tau}_0^{0:i-1}] - \prod_{j=i+1}^{K-1} \left(I + \alpha \nabla^2_{\boldsymbol{\theta}^j} J^{\text{In}}\left(\boldsymbol{\phi}, \boldsymbol{\theta}^j\right)\right) \right\|\right]}_{\textbf{Term (i)}} \\
&+ \left[ c_1 \mu_2 \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^K\right\|\right] + m_2 \underbrace{\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\| \mathbb{E}_{\boldsymbol{\tau}_1^i}[\nabla_{\boldsymbol{\phi}} \nabla_{\hat{\boldsymbol{\theta}}^i} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^i, \boldsymbol{\tau}_1^i) \mid \boldsymbol{\tau}_0^{0:i-1}] - \nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}^i} J^{\text{In}}(\boldsymbol{\phi}, \boldsymbol{\theta}^i) \right\|\right]}_{\textbf{Term (ii)}} \right] \times \\
&\underbrace{\left\| \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[ \prod_{j=i+1}^{K-1} \mathbb{E}_{\boldsymbol{\tau}_2^j}[I + \alpha \nabla^2_{\hat{\boldsymbol{\theta}}^j} \hat{J}^{\text{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j) \mid \boldsymbol{\tau}_0^{0:i-1}] \right] \right\|}_{\textbf{Term (iii)}}
\end{aligned}
\tag{43}
$$

Let $J_{\boldsymbol{\phi},\boldsymbol{\theta}}$ denote $\nabla_{\boldsymbol{\phi}} \nabla_{\boldsymbol{\theta}} J^{\text{In}}$, $H_{\boldsymbol{\theta},\boldsymbol{\theta}}$ denote $\nabla^2_{\boldsymbol{\theta}} J^{\text{In}}$ ,$\hat{\Delta}_J = \max \|\mathbb{E}[\hat{J}_{\boldsymbol{\phi},\boldsymbol{\theta}}] - J_{\boldsymbol{\phi},\boldsymbol{\theta}}\|$. $\hat{\Delta}_H = \max \|\mathbb{E}[\hat{H}_{\boldsymbol{\theta},\boldsymbol{\theta}}] - H_{\boldsymbol{\theta},\boldsymbol{\theta}}\|$. $(\hat{\sigma}_J)^2 = \frac{\max \mathbb{V}[\hat{J}_{\boldsymbol{\phi},\boldsymbol{\theta}}]}{|\tau|}$. $(\hat{\sigma}_H)^2 = \frac{\max \mathbb{V}[\hat{H}_{\boldsymbol{\theta},\boldsymbol{\theta}}]}{|\tau|}$. We upper bound terms (i)-(ii) in Eq. (43) respectively, that is,

**Term (i).** According to

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)\mid\boldsymbol{\tau}_0^{0:i-1}]-\prod_{j=i+1}^{K-1}\left(I+\alpha\nabla^2_{\boldsymbol{\theta}^j}J^{\mathrm{In}}\left(\boldsymbol{\phi},\boldsymbol{\theta}^j\right)\right)\right\|\right]\\
\leq&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\prod_{j=i+1}^{K-2}(I+\alpha c_2+\alpha\hat{\Delta}_H)\right]\left(\alpha\hat{\Delta}_H+\alpha\rho_2\left((1+\alpha c_2)^{K-1}-1\right)\frac{\hat{\sigma}_{\mathrm{In}}}{c_2\sqrt{|\boldsymbol{\tau}|}}\right)\\
&+(1+\alpha c_2)\times\\
&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\prod_{j=i+1}^{K-2}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{In}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)\mid\boldsymbol{\tau}_0^{0:i-1}]-\prod_{j=i+1}^{K-2}\left(I+\alpha\nabla^2_{\boldsymbol{\theta}^j}J^{In}\left(\boldsymbol{\phi},\boldsymbol{\theta}^j\right)\right)\right\|\right]\\
\leq&\alpha\left[(1+\alpha c_2+\alpha\hat{\Delta}_H)^{K-i-1}-(1+\alpha c_2)^{K-i-1}\right]\\
&+\frac{\rho_2}{c_2}\left((1+\alpha c_2+\alpha\hat{\Delta}_H)^{K-i-1}-1\right)(1+\alpha c_2)^{K-1}\frac{\hat{\sigma}_{\mathrm{In}}}{c_2\sqrt{|\boldsymbol{\tau}|}}
\end{aligned}
\tag{44}
$$

**Term (ii).**

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\mathbb{E}_{\boldsymbol{\tau}_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)\mid\boldsymbol{\tau}_0^{0:i-1}]-\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\right\|\right]\\
\leq&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\mathbb{E}_{\boldsymbol{\tau}_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)\mid\boldsymbol{\tau}_0^{0:i-1}]-\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i)\right\|\right]+\\
&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i)-\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\mid\boldsymbol{\tau}_0^{0:i-1}]\right\|\right]
\end{aligned}
\tag{45}
$$

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\mathbb{E}_{\boldsymbol{\tau}_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)\mid\boldsymbol{\tau}_0^{0:i-1}]-\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^i)\right\|\right]\\
\leq&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\hat{\Delta}_J\right]+\lambda_2\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^i-\boldsymbol{\theta}^i\right\|\right]\\
\leq&\hat{\Delta}_J+\lambda_2\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\hat{\boldsymbol{\theta}}^i-\boldsymbol{\theta}^i\right\|\right]
\end{aligned}
\tag{46}
$$

**Term (iii).**

$$
\begin{aligned}
&\left\|\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)\mid\boldsymbol{\tau}_0^{0:i-1}]\right]\right\|\\
\leq&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\prod_{j=i+1}^{K-1}\left(I+\alpha\left\|\mathbb{E}_{\boldsymbol{\tau}_2^j}[\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)\mid\boldsymbol{\tau}_0^{0:i-1}]-\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j)\right\|+\alpha\left\|\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j)\right\|\right)\right]
\end{aligned}
\tag{47}
$$

$$
\begin{aligned}
&\left\|\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)\mid\boldsymbol{\tau}_0^{0:i-1}]\right]\right\|\\
\leq&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\prod_{j=i+1}^{K-1}(1+\alpha c_2+\alpha\hat{\Delta}_H)\mid\boldsymbol{\tau}_0^{0:i-1}]\right]\\
\leq&(1+\alpha c_2+\alpha\hat{\Delta}_H)^{K-i-1}
\end{aligned}
\tag{48}
$$

Then combine terms (i)-(iii) together, that is

$$
\begin{aligned}
&\left\|\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1},\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{0:K-1},\boldsymbol{\tau}_3}[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi})] - \nabla_{\boldsymbol{\phi}}J^{\mathrm{K}}(\boldsymbol{\phi})\right\| \\
\leq&\mu_1\left((1+\alpha c_2)^K-1\right)\frac{\hat{\sigma}_{\mathrm{In}}}{c_2\sqrt{|\boldsymbol{\tau}|}} \\
&+\alpha\sum_{i=0}^{K-1}\alpha c_1 m_2\left[(1+\alpha c_2+\alpha\hat{\Delta}_H)^{K-i-1}-(1+\alpha c_2)^{K-i-1}\right] \\
&+c_1 m_2\frac{\rho_2}{c_2}\left((1+\alpha c_2+\hat{\Delta}_H)^{K-i-1}-1\right)(1+\alpha c_2)^{K-1}\frac{\hat{\sigma}_{\mathrm{In}}}{c_2\sqrt{|\boldsymbol{\tau}|}} \\
&+(1+\alpha c_2+\hat{\Delta}_H)^{K-i-1}\left[c_1\mu_2\left((1+\alpha c_2)^K-1\right)\frac{\hat{\sigma}_{\mathrm{In}}}{c_2\sqrt{|\boldsymbol{\tau}|}}+m_2\hat{\Delta}_J+m_2\lambda_2\left((1+\alpha c_2)^i-1\right)\frac{\hat{\sigma}_{\mathrm{In}}}{c_2\sqrt{|\boldsymbol{\tau}|}}\right] \\
\leq&(\mu_1+\alpha(c_1 m_2\frac{\rho_2}{c_2}+c_1\mu_2+m_2\lambda_2))\left(\left((1+\alpha c_2)^K-1\right)\left((1+\alpha c_2+\alpha\hat{\Delta}_H)^{K-1}-1\right)\frac{\hat{\sigma}_{In}}{c_2\sqrt{|\boldsymbol{\tau}|}}\right. \\
&+(\alpha m_2)\left((1+\alpha c_2+\alpha\hat{\Delta}_H)^{K-1}-1\right)\hat{\Delta}_J \\
&+(\alpha^2 c_1 m_2)\left((1+\alpha c_2+\alpha\hat{\Delta}_H)^{K-1}-(1+\alpha c_2)^{K-1}\right)
\end{aligned}
\tag{49}
$$

$$
\begin{aligned}
&\left\|\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1},\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{0:K-1},\boldsymbol{\tau}_3}[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi})] - \nabla_{\boldsymbol{\phi}}J^{\mathrm{K}}(\boldsymbol{\phi})\right\| \\
\leq&\mathcal{O}\left((1+\alpha c_2+\alpha\hat{\Delta}_H)^{K-1}\left(\mathbb{E}[\|\hat{\boldsymbol{\theta}}^K-\boldsymbol{\theta}^K\|]+\hat{\Delta}_J+(K-1)\right)\right)
\end{aligned}
\tag{50}
$$

which concludes the proof of upper bound of meta-gradient bias.

According to Lemma H.7,

$$
\mathbb{V}\left[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi})\right]=\underbrace{\mathbb{V}\left[\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\boldsymbol{\tau}_3}\left[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi})\mid\boldsymbol{\tau}_0^{0:K-1}\right]\right]}_{\textbf{Term (i)}}+\underbrace{\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\mathbb{V}\left[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi})\mid\boldsymbol{\tau}_0^{0:K-1}\right]\right]}_{\textbf{Term (ii)}} \tag{51}
$$

We upper bound terms (i)-(ii) in Eq. (51) respectively, that is,

**Term (i).**

$$
\begin{aligned}
&\mathbb{V}\left[\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\boldsymbol{\tau}_3}\left[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi})\mid\boldsymbol{\tau}_0^{0:K-1}\right]\right] \\
=&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\boldsymbol{\tau}_3}\left[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi})\mid\boldsymbol{\tau}_0^{0:K-1}\right]-\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1},\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\boldsymbol{\tau}_3}\left[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi})\right]\right\|^2\right] \\
\leq&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\left\|\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\boldsymbol{\tau}_3}\left[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi})\mid\boldsymbol{\tau}_0^{0:K-1}\right]-\nabla_{\boldsymbol{\phi}}J^{\mathrm{K}}(\boldsymbol{\phi})\right\|^2\right]
\end{aligned}
\tag{52}
$$

According to proof of upper bound of bias term, together with Lemma H.5 (ii).

$$
\begin{aligned}
&\left\|\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1},\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{0:K-1},\boldsymbol{\tau}_3}[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi})] - \nabla_{\boldsymbol{\phi}}J^{\mathrm{K}}(\boldsymbol{\phi})\right\| \\
\leq&\mu_1\left((1+\alpha c_2)^K-1\right)\frac{\hat{\sigma}_{\mathrm{In}}}{c_2\sqrt{|\boldsymbol{\tau}|}} \\
&+\alpha\sum_{i=0}^{K-1}\alpha c_1 m_2\left[(1+\alpha c_2+\hat{\Delta}_H)^{K-i-1}-(1+\alpha c_2)^{K-i-1}\right] \\
&+c_1 m_2\frac{\rho_2}{c_2}\left((1+\alpha c_2+\hat{\Delta}_H)^{K-i-1}-1\right)(1+\alpha c_2)^{K-1}\frac{\hat{\sigma}_{\mathrm{In}}}{c_2\sqrt{|\boldsymbol{\tau}|}} \\
&+(1+\alpha c_2+\hat{\Delta}_H)^{K-i-1}\left[c_1\mu_2\left((1+\alpha c_2)^K-1\right)\frac{\hat{\sigma}_{\mathrm{In}}}{c_2\sqrt{|\boldsymbol{\tau}|}}+m_2\hat{\Delta}_J+m_2\lambda_2\left((1+\alpha c_2)^i-1\right)\frac{\hat{\sigma}_{\mathrm{In}}}{c_2\sqrt{|\boldsymbol{\tau}|}}\right]
\end{aligned}
\tag{53}
$$

$$
\begin{aligned}
&\mathbb{V}\left[\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\boldsymbol{\tau}_3}\left[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi})\mid\boldsymbol{\tau}_0^{0:K-1}\right]\right] \\
\leq&4^K\left(2\mu_1^2+2K\alpha^2(c_1m_2\frac{\rho_2}{c_2}+c_1\mu_2+m_2\lambda_2)^2\right)\times \\
&\left(\left((1+\alpha c_2)^2+\alpha^2c_2^2\right)^K-1\right)\left(\left((1+\alpha c_2)^2+\alpha^2(\hat{\Delta}_H)^2\right)^{K-1}-1\right)\frac{(\hat{\sigma}_{\mathrm{In}})^2}{c_2^2|\boldsymbol{\tau}|} \\
&+(2K\alpha^2m_2^2)\left(\left((1+\alpha c_2)^2+\alpha^2(\hat{\Delta}_H)^2\right)^{K-1}-1\right)(\hat{\Delta}_J)^2 \\
&+(2K\alpha^4c_1^2m_2^2)\left(\left((1+\alpha c_2)^2+\alpha^2(\hat{\Delta}_H)^2\right)^{K-1}-\left((1+\alpha c_2)^2+\alpha^2c_2^2\right)^{K-1}\right)
\end{aligned}
\tag{54}
$$

**Term (ii).**

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\mathbb{V}\left[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi}) \mid \boldsymbol{\tau}_0^{0:K-1}\right]\right] \\
=&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\tau_3}\left[\left\|\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi}) - \mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\tau_3}\left[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{K}}(\boldsymbol{\phi})\right]\right\|^2 \mid \boldsymbol{\tau}_0^{0:K-1}\right]\right] \\
\leq&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[2\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\tau_3}\left[\left\|\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{Out}}\left(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3\right) - \mathbb{E}_{\tau_3}\left[\nabla_{\boldsymbol{\phi}}\hat{J}^{\mathrm{Out}}\left(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3\right)\right]\right\|^2\right] + \right. \\
&2\alpha^2\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\tau_3}\left[\left\|\sum_{i=0}^{K-1}\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)\prod_{j=i+1}^{K-1}\left(I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)\right)\nabla_{\hat{\boldsymbol{\theta}}^K}\hat{J}^{\mathrm{Out}}\left(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3\right)\right. \\
&\left.\left.-\sum_{i=0}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)]\prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)]\mathbb{E}_{\tau_3}[\nabla_{\hat{\boldsymbol{\theta}}^K}\hat{J}^{\mathrm{Out}}\left(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3\right)]\right\|^2 \mid \boldsymbol{\tau}_0^{0:K-1}\right] \\
\leq&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[2(\sigma_1)^2 + 2K\alpha^2\sum_{i=0}^{K-1}\left\|\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)\right\|^2\times\right. \\
&\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\tau_3}\left[\left\|\prod_{j=i+1}^{K-1}\left(I-\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)\right)\nabla_{\hat{\boldsymbol{\theta}}^K}\hat{J}^{\mathrm{Out}}\left(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3\right) - \right.\right. \\
&\left.\prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)]\mathbb{E}_{\tau_3}[\nabla_{\hat{\boldsymbol{\theta}}^K}\hat{J}^{\mathrm{Out}}\left(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3\right)]\right\|^2\Big] + \\
&\left\|\prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)]\mathbb{E}_{\tau_3}[\nabla_{\hat{\boldsymbol{\theta}}^K}\hat{J}^{\mathrm{Out}}\left(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3\right)]\right\|^2 \\
&\left.\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\tau_3}\left[\left\|\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i) - \mathbb{E}_{\boldsymbol{\tau}_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)]\right\|^2\right] \mid \boldsymbol{\tau}_0^{0:K-1}\right]
\end{aligned}
\tag{55}
$$

$$
\begin{aligned}
\leq&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\left[2(\sigma_1)^2 + 2K\alpha^2\sum_{i=0}^{K-1}\left[\left\|\nabla_{\hat{\boldsymbol{\theta}}^K}\hat{J}^{\mathrm{Out}}\left(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^K,\tau_3\right)\right\|^2\left\|\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)\right\|^2\times\right.\right. \\
&\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\tau_3}\left[\left\|\prod_{j=i+1}^{K-1}\left(I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)\right) - \prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)]\right\|^2\right] + \\
&(\sigma_2)^2\left\|\prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)]\right\|^2\left\|\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)\right\|^2\right] + \\
&m_2^2\left\|\prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)]\right\|^2\times \\
&\left.\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\tau_3}\left[\left\|\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i) - \mathbb{E}_{\boldsymbol{\tau}_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)]\right\|^2\right] \mid \boldsymbol{\tau}_0^{0:K-1}\right]
\end{aligned}
\tag{56}
$$

$$
\begin{aligned}
\leq & \mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}}\Bigg[2(\sigma_1)^2+2K\alpha^2\sum_{i=0}^{K-1}\Bigg[\left(2m_2^2+2(\sigma_2)^2\right)\times\left\|\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)\right\|^2\times \\
& \underbrace{\mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\boldsymbol{\tau}_3}\left[\left\|\prod_{j=i+1}^{K-1}\left(I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)\right)-\prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)]\right\|^2\right]}_{\text{Part (I)}}+ \\
& (\sigma_2)^2\left\|\prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)]\right\|^2\times\underbrace{\left\|\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)\right\|^2}_{\text{Part (II)}}+ \\
& \underbrace{m_2^2\left\|\prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)]\right\|^2}_{\text{Part (III)}}\times \\
& \mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\boldsymbol{\tau}_3}\left[\left\|\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)-\mathbb{E}_{\boldsymbol{\tau}_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)]\right\|^2\right]\mid\boldsymbol{\tau}_0^{0:K-1}\Bigg]
\end{aligned}
\tag{57}
$$

**Part (I)** According to

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{\tau}_1^{0:K-1},\boldsymbol{\tau}_2^{1:K-1},\boldsymbol{\tau}_3}\left[\left\|\prod_{j=i+1}^{K-1}\left(I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)\right)-\prod_{j=i+1}^{K-1}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)]\right\|^2\right] \\
\leq & 2\prod_{j=i+1}^{K-2}\left(3(1+\alpha c_2)^2+3\alpha^2(\hat{\Delta}_H)^2+3\alpha^2(\hat{\sigma}_H)^2\right)\times\alpha^2(\hat{\sigma}_H)^2+ \\
& 4\left((1+\alpha c_2)^2+\alpha^2(\hat{\Delta}_H)^2\right)\times \\
& \mathbb{E}_{\boldsymbol{\tau}_2^{1:K-1}}\left[\left\|\left(\prod_{j=i+1}^{K-2}\left(I+\alpha\nabla^2_{\hat{\boldsymbol{\theta}}^j}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)\right)-\prod_{j=i+1}^{K-2}\mathbb{E}_{\boldsymbol{\tau}_2^j}[I+\alpha\nabla^2_{\boldsymbol{\theta}}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^j,\boldsymbol{\tau}_2^j)]\right)\right\|^2\right] \\
\leq & 6^{K-i-1}\left[\left((1+\alpha c_2)^2+\alpha^2(\hat{\Delta}_H)^2+\alpha^2(\hat{\sigma}_H)^2\right)^{K-i-1}-\left((1+\alpha c_2)^2+\alpha^2(\hat{\Delta}_H)^2\right)^{K-i-1}\right]
\end{aligned}
\tag{58}
$$

**Part (II)** According to the supporting Lemma

$$
\begin{aligned}
& \left\|\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^0}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^0,\tau_1)\right\|^2 \\
\leq & \Big\|\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i)-\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}J^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i)+\mathbb{E}_{\boldsymbol{\tau}_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}\left(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i\right)] \\
& -\mathbb{E}_{\boldsymbol{\tau}_1^i}[\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)]+\nabla_{\boldsymbol{\phi}}\nabla_{\hat{\boldsymbol{\theta}}^i}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\hat{\boldsymbol{\theta}}^i,\boldsymbol{\tau}_1^i)\Big\|^2 \\
\leq & 3c_1^2+3\left((\hat{\Delta}_J)^2+(\hat{\sigma}_J)^2\right)
\end{aligned}
\tag{59}
$$

$$
\left\|\nabla_{\boldsymbol{\phi}}\nabla_{\boldsymbol{\theta}^0}\hat{J}^{\mathrm{In}}(\boldsymbol{\phi},\boldsymbol{\theta}^0,\boldsymbol{\tau}_1^i)\right\|^2\leq 3c_1^2+3\left((\hat{\Delta}_J)^2+(\hat{\sigma}_J)^2\right)
\tag{60}
$$

**Part (III)**

$$
\begin{aligned}
&\left\| \prod_{j=i+1}^{K-1} \mathbb{E}_{\tau_2^j}[I + \alpha \nabla^2_{\hat{\boldsymbol{\theta}}^j} \hat{J}^{\mathrm{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j)] \right\|^2 \\
&\leq \prod_{j=i+1}^{K-1} \left\| I - \alpha \nabla^2_{\hat{\boldsymbol{\theta}}^j} J^{\mathrm{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j) + \alpha \nabla^2_{\hat{\boldsymbol{\theta}}^j} J^{\mathrm{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j) - \alpha \mathbb{E}_{\tau_2^j}[\nabla^2_{\hat{\boldsymbol{\theta}}^j} \hat{J}^{\mathrm{In}}(\boldsymbol{\phi}, \hat{\boldsymbol{\theta}}^j, \boldsymbol{\tau}_2^j)] \right\|^2 \\
&\leq \left( (1 + \alpha c_2)^2 + \alpha^2 (\hat{\Delta}_H)^2 \right)^{K-i-1}
\end{aligned} \tag{61}
$$

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}} \left[ \mathbb{V} \left[ \nabla_{\boldsymbol{\phi}} \hat{J}^{\mathrm{K}}(\boldsymbol{\phi}) \mid \boldsymbol{\tau}_0^{0:K-1} \right] \right] \\
\leq &\mathbb{E}_{\boldsymbol{\tau}_0^{0:K-1}} \Bigg[ 2(\sigma_1)^2 + 2K\alpha^2 \sum_{i=0}^{K-1} \Bigg[ \left( 2m_2^2 + 2\sigma_2^2 \right) \times \left( 3c_1^2 + 3 \left( (\hat{\Delta}_J)^2 + (\hat{\sigma}_J)^2 \right) \right) \times \\
&6^{K-i-1} \Bigg[ \left( (1 + \alpha c_2)^2 + \alpha^2 (\hat{\Delta}_H)^2 + \alpha^2 (\hat{\sigma}_H)^2 \right)^{K-i-1} - \left( (1 + \alpha c_2)^2 + \alpha^2 (\hat{\Delta}_H)^2 \right)^{K-i-1} \Bigg] + \\
&(\sigma_2)^2 \left( (1 + \alpha c_2)^2 + \alpha^2 (\hat{\Delta}_H)^2 \right)^{K-i-1} \times \left( 3c_1^2 + 3 \left( (\hat{\Delta}_J)^2 + (\hat{\sigma}_J)^2 \right) \right) + \\
&m_2^2 \left( (1 + \alpha c_2)^2 + \alpha^2 (\hat{\Delta}_H)^2 \right)^{K-i-1} \times (\hat{\sigma}_J)^2 \mid \boldsymbol{\tau}_0^{0:K-1} \Bigg] \Bigg] \\
\leq &2(\sigma_1)^2 \\
&+ (6K\alpha^2 \sigma_2^2) \left( \left( (1 + \alpha c_2)^2 + \alpha^2 (\hat{\Delta}_H)^2 \right)^{K-1} - 1 \right) \left( c_1^2 + (\hat{\Delta}_J)^2 + (\hat{\sigma}_J)^2 \right) \\
&+ (2K\alpha^2 m_2^2) \left( \left( (1 + \alpha c_2)^2 + \alpha^2 (\hat{\Delta}_H)^2 \right)^{K-1} - 1 \right) (\hat{\sigma}_J)^2 \\
&+ 2\alpha^2 (m_1^2 + 3\sigma_2^2) \left( c_1^2 + (\hat{\Delta}_J)^2 + (\hat{\sigma}_J)^2 \right) \\
&+ (6^K K\alpha^2 (12m_2^2 + 12\sigma_2^2)) \left( c_1^2 + (\hat{\Delta}_J)^2 + (\hat{\sigma}_J)^2 \right) \times \\
&\left( \left( (1 + \alpha c_2)^2 + \alpha^2 (\hat{\Delta}_H)^2 + \alpha^2 (\hat{\sigma}_H)^2 \right)^{K-1} - \left( (1 + \alpha c_2)^2 + \alpha^2 (\hat{\Delta}_H)^2 \right)^{K-1} \right)
\end{aligned} \tag{62}
$$

Then combine terms (i)-(ii) together, that is

$$
\begin{aligned}
&\mathbb{V} \left[ \nabla_{\boldsymbol{\phi}} \hat{J}^{\mathrm{K}}(\boldsymbol{\phi}) \right] \\
\leq &O \Bigg( (V_1 + \hat{\Delta}_H^2)^{K-1} \left( \mathbb{E}[\|\hat{\boldsymbol{\theta}}^K - \boldsymbol{\theta}^K\|^2] + (K-1) \right) \\
&+ \left( V_2 + (V_1 + \hat{\Delta}_H^2 + \hat{\sigma}_H^2)^{K-1} - (V_1 + \hat{\Delta}_H^2)^{K-1} \right) (\hat{\Delta}_J^2 + \hat{\sigma}_J^2) \Bigg)
\end{aligned} \tag{63}
$$

which concludes the proof of Theorem 4.5. □

## H Supporting Lemmas

In this section, we present the supporting lemmas.

**Definition H.1.** Let $X$ be a random vector in $\mathbb{R}^d$. Then the norm of $X$ is

$$
\|X\| := \sqrt{\sum_i X_i^2} \tag{64}
$$

**Lemma H.2.** *Let $X$ be a random vector in $\mathbb{R}^d$ with finite second moment, where $\mathbb{E}[\|X\|^2] \leq +\infty$. Then $\|\mathbb{E}[X]\| \leq \mathbb{E}[\|X\|]$, $\|\mathbb{E}[X]\|^2 \leq \mathbb{E}[\|X\|^2]$.*

*Proof.* Due to the convexity of norm operator, we can have $\|\mathbb{E}[X]\| \leq \mathbb{E}[\|X\|]$ using Jensen's inequality. Further we can get $\|\mathbb{E}[X]\|^2 \leq (\mathbb{E}[\|X\|^2])^2 \leq \mathbb{E}[\|X\|^2]$ and the statement follows. □

**Lemma H.3.** *Let $X$ and $Y$ be two random variables in $\mathbb{R}^d$ with finite second moment. Then $\mathbb{E}[\|X+Y\|] \leq \mathbb{E}[\|X\|] + \mathbb{E}[\|Y\|]$.*

*Proof.* According to Minkowski's inequality that $(\mathbb{E}[\|X+Y\|^p])^{1/p} \leq (\mathbb{E}[\|X\|^p])^{1/p} + (\mathbb{E}[\|Y\|^p])^{1/p}$, set $p=1$ and the statement follows. □

**Definition H.4.** Let $X$ be a random vector with values in $\mathbb{R}^d$. Then the variance of $X$ is

$$\mathbb{V}[X] := \mathbb{E}[\|X - \mathbb{E}[X]\|^2] \tag{65}$$

**Lemma H.5** (Properties of the variance)**.** *Let $X$ and $Y$ be two independent random variables in $\mathbb{R}^d$. We also assume that $X, Y$, have finite second moment. Then the following hold.*

(i) $\mathbb{V}[X] = \mathbb{E}[\|X\|^2] - \|\mathbb{E}[X]\|^2$,

(ii) *For every $x \in \mathbb{R}^d$, $\mathbb{E}[\|X-x\|^2] = \mathbb{V}[X] + \|\mathbb{E}[X]-x\|^2$. Hence, $\mathbb{V}[X] = \min_{x\in\mathbb{R}^d} \mathbb{E}[\|X-x\|^2]$,*

(iii) $\mathbb{V}[X+Y] = \mathbb{V}[X] + \mathbb{V}[Y]$.

*Proof.* (i)-(ii): Let $x \in \mathbb{R}^d$. Then, $\|X-x\|^2 = \|X-\mathbb{E}[X]\|^2 + \|\mathbb{E}[X]-x\|^2 + 2(X-\mathbb{E}[X])^\top(\mathbb{E}[X]-x)$. Hence, taking the expectation we get $\mathbb{E}[\|X-x\|^2] = \mathbb{V}[X] + \|\mathbb{E}[X]-x\|^2$. Therefore, $\mathbb{E}[\|X-x\|^2] \geq \mathbb{V}[X]$ and for $x = \mathbb{E}[X]$ we get $\mathbb{E}[\|X-x\|^2] = \mathbb{V}[X]$. Finally, for $x=0$ we get (i).

(iii): Let $\bar{X} := \mathbb{E}[X]$ and $\bar{Y} := \mathbb{E}[Y]$, we have

$$\begin{aligned}
\mathbb{V}[X+Y] &= \mathbb{E}[\|X - \bar{X} + Y - \bar{Y}\|^2] \\
&= \mathbb{E}[\|X-\bar{X}\|^2] + \mathbb{E}[\|Y-\bar{Y}\|^2] + 2\mathbb{E}[X-\bar{X}]^\top \mathbb{E}[Y-\bar{Y}] \\
&= \mathbb{E}[\|X-\bar{X}\|^2] + \mathbb{E}[\|Y-\bar{Y}\|^2]
\end{aligned}$$

Recalling the definition of $\mathbb{V}[X]$ the statement follows. □

**Definition H.6.** *(Conditional Variance).* Let $X$ be a random variable with values in $\mathbb{R}^d$ and $Y$ be a random variable with values in a measurable space $\mathcal{Y}$. We call *conditional variance* of $X$ given $Y$ the quantity

$$\mathbb{V}[X \mid Y] := \mathbb{E}[\|X - \mathbb{E}[X \mid Y]\|^2 \mid Y].$$

**Lemma H.7.** *(Law of total variance) Let $X$ and $Y$ be two random variables, we can prove that*

$$\mathbb{V}[X] = \mathbb{E}[\mathbb{V}[X \mid Y]] + \mathbb{V}[\mathbb{E}[X \mid Y]] \tag{66}$$

*Proof.*

$$\begin{aligned}
\mathbb{V}[X] &= \mathbb{E}[\|X - \mathbb{E}[X]\|^2] \\
&= \mathbb{E}[\|X\|^2] - \|\mathbb{E}[X]\|^2 \\
&= \mathbb{E}[\mathbb{E}[\|X\|^2 \mid Y]] - \|\mathbb{E}[\mathbb{E}[X \mid Y]]\|^2 \\
&= \mathbb{E}[\mathbb{V}[X \mid Y] + \|\mathbb{E}[X \mid Y]\|^2] - \|\mathbb{E}[\mathbb{E}[X \mid Y]]\|^2 \\
&= \mathbb{E}[\mathbb{V}[X \mid Y]] + \left(\mathbb{E}[\|\mathbb{E}[X \mid Y]\|^2] - \|\mathbb{E}[\mathbb{E}[X \mid Y]]\|^2\right)
\end{aligned}$$

recognizing that the term inside the parenthesis is the conditional variance of $\mathbb{E}[X|Y]$ gives the result. □

**Lemma H.8.** *Let $\zeta$ and $\eta$ be two independent random variables with values in $\mathcal{Z}$ and $\mathcal{Y}$ respectively. Let $\psi : \mathcal{Y} \to \mathbb{R}^{m\times n}$, $\phi : \mathcal{Z} \to \mathbb{R}^{n\times p}$, and $\varphi : \mathcal{Y} \to \mathbb{R}^{p\times q}$ matrix-valued measurable functions. Then*

$$\mathbb{E}[\psi(\eta)(\phi(\zeta) - \mathbb{E}[\phi(\zeta)])\varphi(\eta)] = 0 \tag{67}$$

*Proof.* Since, for every $y \in \mathcal{Y}$, $B \mapsto \psi(y)B\varphi(y)$ is linear and $\zeta$ and $\eta$ are independent, we have

$$\mathbb{E}[\psi(\eta)(\psi(\zeta) - \mathbb{E}[\psi(\zeta)])\varphi(\eta)\,|\eta] = \psi(\eta)\mathbb{E}\big[\phi(\zeta) - \mathbb{E}[\phi(\zeta)]\big]\varphi(\eta) = 0.$$

Taking the expectation the statement follows. □

# I Experiment

**Computational resources.** For compute resources, We used one internal compute servers which consists consisting of 2$x$ Tesla A100 cards and 256 CPUs, however each model is trained on at most 1 card.

## I.1 Tabular MDP

### I.1.1 Experimental Settings

We adopt the tabular random MDP setting presented in [33]. The dimension is 20 for state space and 5 for action space, so we have the reward matrix $R \in \mathbb{R}^{20\times 5}$. The transition probability matrix is generated from independent Dirichlet distributions. The policy is a matrix $\theta^0 \in \mathbb{R}^{20\times 5}$. The final policy $\pi_\theta$ is obtained by adopting Softmax activation on this policy matrix: $\pi_\theta(a \mid s) = \exp(\theta(s,a))/\sum_b \exp(\theta(s,b))$. We set the initial policy as the uniform policy (by setting $\theta^0$ as zero matrix) in MAML and LIRPG experiment. We conduct the inner-loop update stating from the same point for several times when estimating the meta-gradient correlation and variance. For accuracy measurement between estimation $x \in \mathbb{R}^L$ and ground truth $y \in \mathbb{R}^L$, we use the following equation:

$$\mathrm{Acc}(x, y) := \frac{x^T y}{\sqrt{x^T x}\sqrt{y^T y}}. \tag{68}$$

### I.1.2 Implementation for decomposing Gradient estimation

To decompose the gradient estimation effects brought by different sources, such as outer estimation variance and inner estimation bias (compositional bias, hessian estimation error), we utilise the following implementation trick: Using estimator I to estimate $\theta'_c = \theta + \alpha\nabla J(\theta)$, estimator II to estimate $\theta'_h = \theta + \alpha\nabla J(\theta)$ and finally combine them with: $\theta' =\perp \theta'_c + \theta'_h - \perp \theta'_h$, where $\perp$ is the "stop gradient" operator. By this implementation trick, we can have the following property: $\theta' \to \theta'_c$ and $\nabla_\theta\theta' = \nabla_\theta\theta'_h$, where $\to$ is the "evaluates to" operator. "Evaluates to" operator $\to$ is in contrast with =, which also brings the equality of gradients. By "Evaluates to" operator, the "stop gradient" operator means that $\perp (f_\theta(x)) \to f_\theta(x)$ but $\nabla_\theta \perp (f_\theta(x)) \to 0$. This property guarantee that the compositional bias is only influenced by estimator I while hessian estimation error is controlled by estimator II. Besides estimator I and estimator II, an extra estimator III is used for outer-loop policy gradient $\nabla_{\theta^k} J^{\text{out}}(\pi^k)$ estimation, which helps us understand the effect of outer-loop policy gradient.

### I.1.3 Additional Experimental Results on Tabular MAML-RL

We offer additional experimental results on more estimators (DiCE/ Loaded-DiCE)/settings (All 7 permutations)/metrics (variance of Meta-gradient estimation).

**Ablation study on sample size and estimator.** Additional experimental results are shown in Fig. 5. The comparison between $SSS, SES, ESS$ and $SSE, SEE, ESE$ reveals the importance of the outer-loop gradient estimation. Accurate outer-loop policy gradient estimation brings more significant improvement over the correlation compared with the correction of Hessian error or compositional bias. In addition, with estimated outer-loop policy gradient, the correction of these two terms also helps ($EES > SES > ESS > SSS$).

Next we discuss the comparison between different estimators. The DiCE estimator have real high variance on first-order and second-order, and its first-order gradient corresponds to the REINFORCE algorithm [38] while the rest 3 estimators' first-order gradient corresponds to the Actor-critic algorithm. That is why DiCE performs the worst in all cases. With stochastic outer-loop estimation, the LVC and Loaded-DiCE estimator have comparable correlation while the variance of LVC is smaller than Loaded-DiCE. The AD estimator performs worse than LVC and Loaded-DiCE when the Hessian is estimated (SSE, ESE, SSS, ESS). This corresponds to the conclusion in [29] that

the LVC estimator introduces low-bias and low-variance Hessian estimation while AD estimator has large-bias and low-varaince Hessian estimation. With exact outer-loop estimation, the LVC has relatively great Hessian estimation so the correction of compositional bias has the same effect with Hessian correction ($ESE = SEE > SSE$), while the Hessian correction is still important in Loaded-DiCE ($SEE > ESE > SSE$).

**Ablation study on inner learning rate, step and estimator.** Additional ablation study on inner learning rate and number of steps are shown in Fig. 6, 7. The results show that: With more steps and larger learning rates, the inner-loop estimation can become more important than outer-loop policy gradient (the correlation decreases a lot in $SSE$ in all estimators). Also in multi-step and large learning rate setting, the importance of Hessian estimation and compositional bias become comparable in LVC and Loaded-DiCE ($SEE \approx ESE, SES \approx ESS$).

**Meta-gradient variance** In all three plots Fig. 5, 6, 7, we report additional metric on variance of the meta-gradient estimation. We observe that the correction of compositional bias increases the variance especially when outer-loop policy gradient estimator is poor (estimator III uses stochastic samples) or Hessian variance is large (in DiCE and Loaded-DiCE). Only with low Hessian variance (LVC/AD) and great outer-loop policy gradient (estimator III uses analytical solution), the correction of compositional bias can decrease the variance.

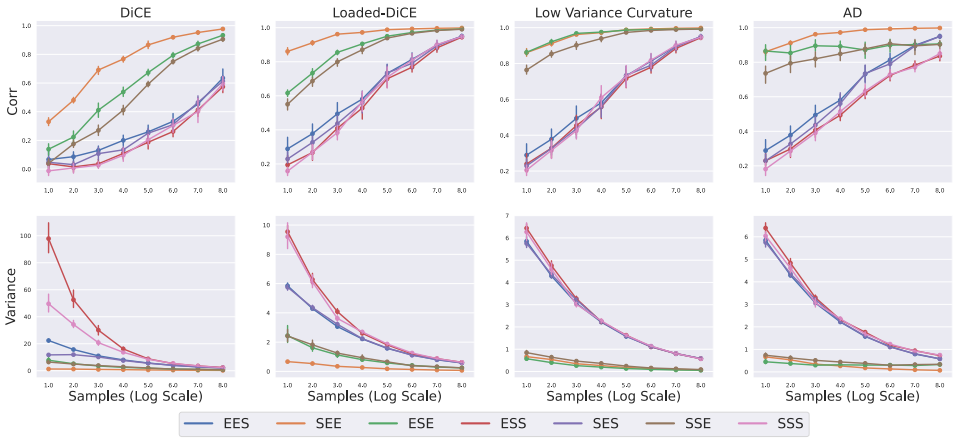



Figure 5: Ablation study on sample size and estimator in 1-step inner-loop setting. (1) Outer-loop policy gradient is important for estimation (2) Compositional bias correction helps increase the correlation (3) The LVC and Loaded-DiCE can achieve higher correlation compared with AD when the Hessian matrix is estimated.

### I.1.4 Additional Experimental Results on Tabular LIRPG

In Fig. 8 we offer additional experimental Results on estimation variance. Basically the AD based estimation in LIRPG setting tend to have higher variance.

### I.1.5 Hyperparameters

We offer the hyperparameter settings for our Tabular MDP experiment in Table 2.

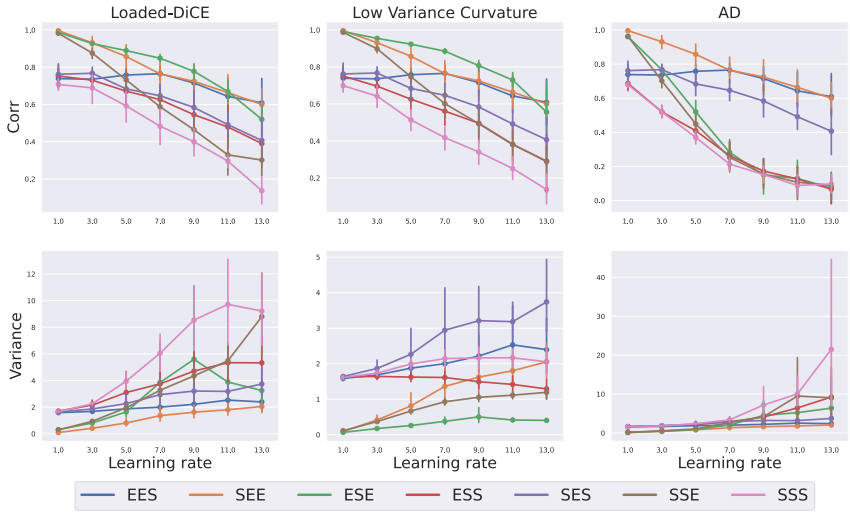



Figure 6: Ablation study on inner learning rate and estimator. (1) In Loaded-DiCE and LVC, With larger learning rate, the compositional bias basically shares the same importance with Hessian estimation error. (2) With larger learning rate, the Hessian estimation problem in AD largely decreases the correlation.

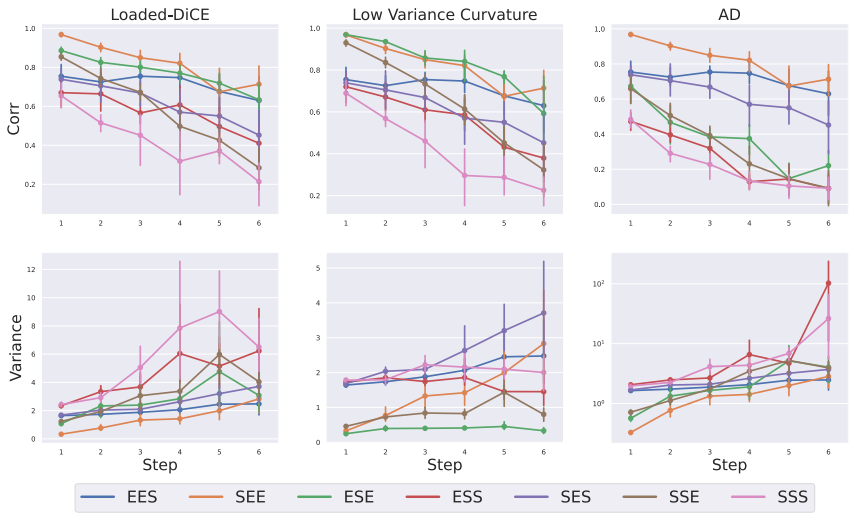



Figure 7: Ablation study on inner step and estimator. Results of larger steps show similar phenomenon with larger inner-loop learning rate.

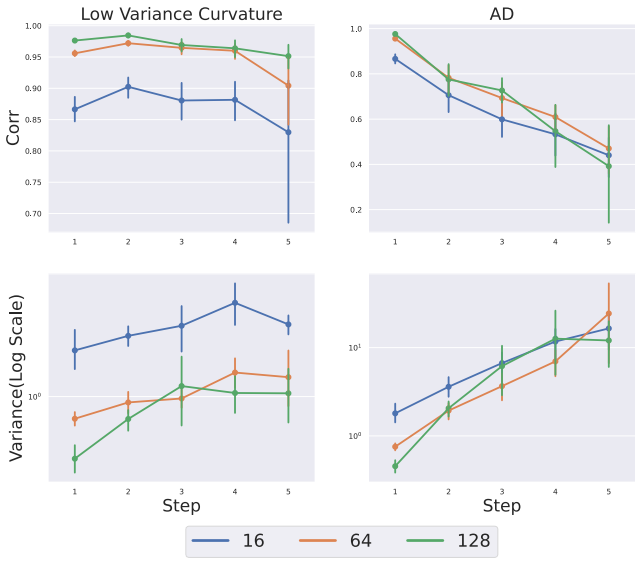



Figure 8: Additional experiment results on. Different color refers to different trajectory sample size.

Table 2: Hyper-parameter settings for Tabular MDP.

| SETTINGS | VALUE | DESCRIPTION |
|---|---|---|
| TRAJECTORY LENGTH | 20 | RL TRAJECOTRY LENGTH |
| DISCOUNT FACTOR | 0.8 | LEARNING RATE FOR META-SOLVER UPDATES |
| INNER LEARNING RATE | 10 | LEARNING RATE FOR INNER-LOOP UPDATE |
| INNER STEP | 1 | STEP NUMBER OF INNER-UPDATE |
| INDEPENDENT TRIALS | 10 | NUMBER OF INDEPENDENT TRIALS ON ENVIRONMENTS |
| SAME TRIALS | 20 | NUMBER OF INDEPENDENT TRIALS ON THE SAME POINT |
| DIMENSION OF STATE | 20 | DIMENSION OF STATE |
| DIMENSION OF ACTION | 5 | DIMENSION OF ACTION |
| NOISE COEFFICIENT | 1.0 | NOISE FACTOR FOR SIMULATING ESTIMATED VALUE FUNCTION |
| DENSITY | 0.001 | PARAMETERS OF DIRICHILET DISTRIBUTION |

### I.2 LOLA-DiCE on Iterated Prisoner Dilemma (IPD)

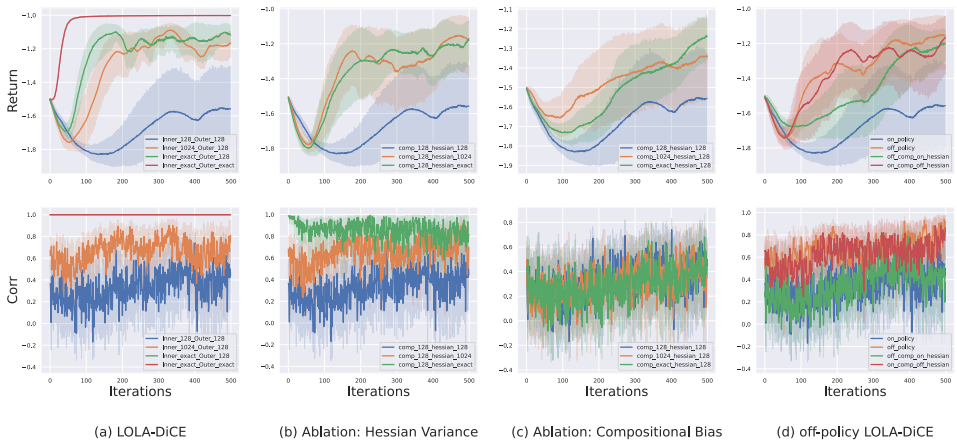
#### I.2.1 Experimental Settings

In Iterated Prisoner Dilemma, the Prisoner Dilemma game is played repeatedly by the same players. The payoffs of Prisoner Dilemma for players are shown as follows.

$$\mathbf{R}^1 = \begin{bmatrix} -2 & 0 \\ -3 & -1 \end{bmatrix} \quad \mathbf{R}^2 = \begin{bmatrix} -2 & -3 \\ 0 & -1 \end{bmatrix},$$

where the action 0 (correpsonds to column/row 0) as "cooperation" (don't confess) and the action 1 (correpsonds to column/row 1) as "defection" (confess). Agent in Iterated Prisoner Dilemma aims at maximising the cumulative Discounted reward. By LOLA-DiCE algorithm, it is possible for both agent to reach social welfare: (-1, -1). Refer to Appendix A.2 for how the algorithm is formulated.

We conduct our experiment by adapting code from the official codebase*. The official code only conducts the experiment using one fixed seed and the performance is highly sensitive to different random seeds using default hyperparameters. To evaluate the performance reliably, we conduct all the experiments for 10 random seeds and report the average result.




(a) LOLA-DiCE (b) Ablation: Hessian Variance (c) Ablation: Compositional Bias (d) off-policy LOLA-DiCE

Figure 9: Experiment result of LOLA-DiCE. (a) Poor inner-loop estimation can fail the LOLA-DiCE algorithm. (b) Hessian estimation variance is the main problem in LOLA-DiCE. (c) The correction of compositional bias also helps increase the average return. (d) The off-policy correction can both decrease the compositional bias and Hessian estimation variance, which largely increases the final return.

### I.2.2 Additional Experimental Results

**Ablation on LOLA-DiCE inner/outer estimation.** We report the correlation result of conducting ablation study for different inner/outer-loop estimation of LOLA-DiCE in the Fig. 9(a). Higher correlation does not guarantee higher return. The bonus brought by setting inner-loop as exact solution have a really large improvement over correlation (from 0.7 to 1.0) but have limited improvement on return. We believe it is because the outer-loop gradient estimation becomes the main issue when inner-loop estimation is really well.

**Ablation on LOLA-DiCE Hessian variance and compositional bias.** We show additional experimental result in Fig. 9(c). An interesting thing is that we find out the gradient correlation of these three settings are comparable. An possible explanation is that the main issue here is the hessian variance and this is why the performance gain by lowering hessian variance is larger that lowering compositional bias. Though by correcting compositional bias LOLA can have better estimation with performance gain, the gain is not obvious in the aspect of gradient correlation because the hessian variance is still large.

**Off-policy DiCE and ablation study** The correlation gain for off-policy comp&on-policy hessian is still limited like that in Fig. 9(c). But the performance gain verifies the bonus brought by correcting compositional bias.

### I.2.3 Hyperparameters

We offer the hyperparameter settings for our LOLA experiment in Table 3.

*https://github.com/alexis-jacq/LOLA_DiCE

Table 3: Hyper-parameter settings for LOLA-DiCE.

| SETTINGS | VALUE | DESCRIPTION |
|---|---|---|
| OUTER LEARNING RATE | 0.1 | OUTER LEARNING RATE |
| INNER LEARNING RATE | 0.3 | INNER LEARNING RATE |
| DISCOUNT FACTOR | 0.96 | DISCOUNT FACTOR |
| UPDATE | 500 | STEP NUMBER OF META-UPDATE |
| ROLLOUT LENGTH | 100 | LENGTH OF IPD ROLLOUT |
| INNER STEP | 1 | NUMBER OF VIRTUAL INNER-STEP LOOK-AHEAD |
| VALUE FUNCTION LEARNING RATE | 0.1 | VALUE FUNCTION LEARNING RATE |
| OFF-POLICY BUFFER SIZE | 1024 | BUFFER SIZE |
| SAMPLE BATCH SIZE | 128 | COMP/HESSIAN/OUTER SAMPLE BATCH SIZE |

### I.3 MGRL on Atari games

#### I.3.1 Experimental setting

We reimplement the MGRL algorithm based on A2C baseline. In this case, Meta-parameters $\phi$ involves 4 hyperparameters: Discount factor, value loss coefficient, entropy loss coefficient and GAE ratio. The procedures of 'discard' strategy we use is summarized as follows: Starting from the inner-policy parameters $\theta^0$, we utilise take 3 A2C updates and get the 3-step updated policy $\theta^3$. Then we can calculate the meta-gradient by backpropogating from $R(\theta^3)$ to the meta parameters. Finally we reset the inner-loop policy parameters back to $\theta^1$ so the rest 2 updates are in fact virtual update. It is only used for the meta-gradient estimation.

#### I.3.2 Discussion on the 'Discard' strategy

In the MGRL experiment, we follow previous work [4] for conducting multi-step MGRL. So the inner-loop policy will take multi-step virtual updates for meta-parameters update. As mentioned in Section 4.2 in their paper, this strategy can only keep the RL update times unchanged among different algorithms and is not particularly sample efficient because they need to take virtual look-ahead for the update of meta-parameters. However, one benefit of adopting such strategy is that we can keep the amount of meta-update large enough to verify the effect brought by the LVC correction. We also take some experiments on another setting where we take meta-update after each 3-step inner-loop update. Note that they are no longer virtual inner-loop updates. However, we find out that in many environment this setting largely decrease the meta-update times and make the comparison of different meta-gradient estimation less meaningful.

#### I.3.3 Additional Experimental results

We offer the full experiments results on all 8 Atari games: Asteroids, Qbert, Tennis, BeamRider, Alien, Assault, DoubleDunk, Seaquest. The reward performance is shown in Fig. 10. We also offer trajectories for all 4 meta-parameters on these experiments in Fig. 11. From Fig. 10 it can show that MGRL with LVC correction can achieve comparable or better performance in almost all 8 environments. Note that we need to clarify that in some RL experiments the MGRL cannot achieve better performance compared with A2C baseline. This also corresponds to the experimental results in original MGRL paper [39]. However, since our main comparison only happens between MGRL and MGRL with LVC correction, it is still a fair comparison to verify the effectiveness of LVC hessian correction. Fig. 11 reveals that even we have only 4 meta-parameters, different meta-gradient estimation can still results in large gap between the meta optimisation trajectory and final GMRL performance.

#### I.3.4 Implementations and hyperparameters

We adopt the codebase of A2C from [21] and differentiable optimization library [28] to implement MGRL algorithms. We use a shared CNN network (3 Conv layers and one fully connected (FC) layer) for the policy network and critic network. The (out-channel, filters, stride) for each Conv layer is (32, $8 \times 8$, 4), (64, $4 \times 4$, 2) and (64, $3 \times 3$, 1) respectively while the hidden size is 512 for the FC layer. For the training loss, we adopt additional entropy regularisation for policy loss and Mean Square Error (MSE) for the value loss. We adopt the Generalized Advantage estimation (GAE) for

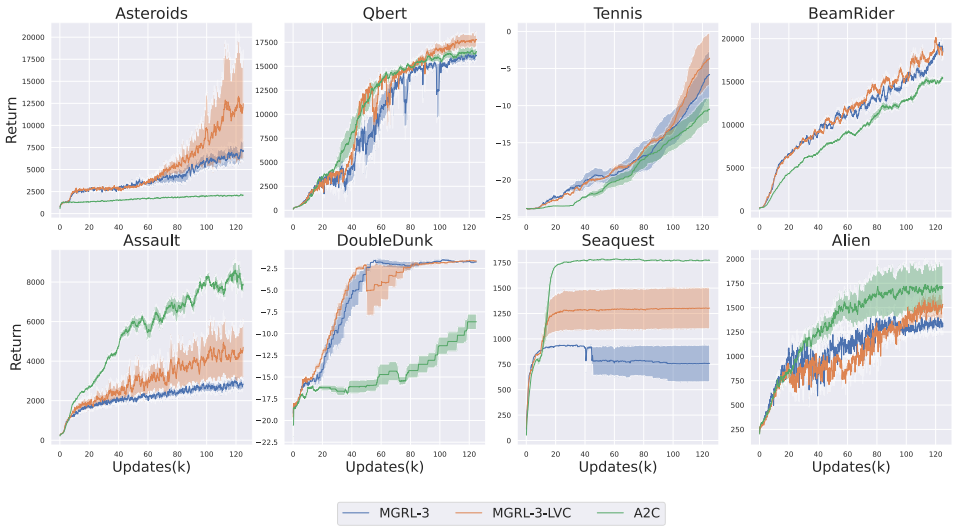



Figure 10: Experimental results on Atari game over 5 seeds. 3-step MGRL with LVC correction can achieve at least the same performance compared with 3-step MGRL in basically all environments.

advantage estimation. We offer the hyperparameter settings for our experiment in Table 4. We tun our algorithm for 125k inner updates, which corresponds to 40M environment steps for baseline A2C.

Table 4: Hyper-parameter settings for MGRL.

| SETTINGS | VALUE | DESCRIPTION |
|---|---|---|
| INNER LEARNING RATE | 7E-4 | INNER LEARNING RATE |
| LEARNING RATE SCHEDULING | LINEAR DECAY | LINEARLY DECREASE TO 0 |
| DISCOUNT FACTOR | 0.99 | DISCOUNT FACTOR |
| GAE LAMBDA | 0.95 | RATIO OF GENERALIZED ADVANTAGE ESTIMATION |
| VALUE COEF | 0.5 | COEFFICIENT OF VALUE LOSS |
| ENTROPY COEF | 0.01 | COEFFICIENT OF ENTROPY LOSS |
| UPDATE | 125K | NUMBER OF INNER UPDATE |
| NUMBER OF PROCESS | 64 | NUMBER OF MULTI PROCESS |
| NUMBER OF STEP PER UPDATE | 5 | NUMBER OF STEP PER UPDATE |
| META UPDATE | 3 | NUMBER OF INNER-UPDATE FOR CONDUCTING META-UPDATE |
| META LEARNING RATE | 0.001 | META LEARNING RATE |
| INNER OPTIMIZER | ADAM | INNER-LOOP OPTIMIZER |
| OUTER OPTIMIZER | ADAM | OUTER-LOOP OPTIMIZER |

## J Author contribution

We summarise the main contributions from each of the authors as follows:

**Bo Liu**: Algorithm design, main theoretical proof, some code implementation and experiments running (on tabular MDP and MGRL), and paper writing.

**Xidong Feng**: Idea proposing, algorithm design, part of theoretical proof, main code implementation and experiments running (on tabular MDP, LOLA and MGRL), and paper writing.

**Jie Ren**: Code implementation and experiments running for MGRL.

**Luo Mai**: Project discussion and paper writing.

**Rui Zhu**: Project discussion and paper writing.

**Haifeng Zhang**: Computational resource sponsor.

**Jun Wang**: Project discussion and overall project supervision.

**Yaodong Yang**: Project lead, idea proposing, experiment supervision, and whole manuscript writing.

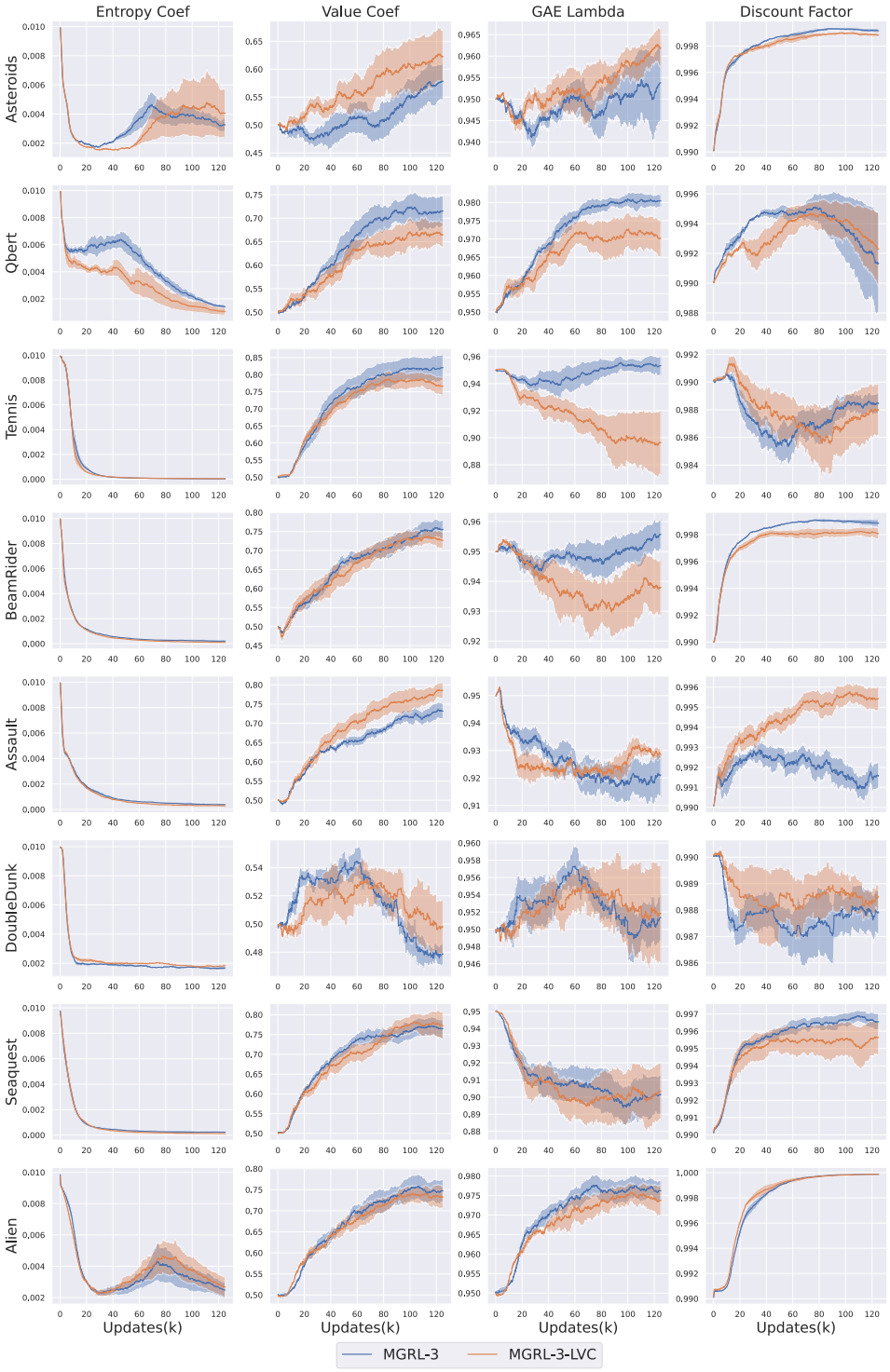



Figure 11: 4 Meta parameters trajectories on Atari game for 3-step MGRL and 3-step MGRL with LVC correction.