Momentum for the Win: Collaborative Federated Reinforcement Learning across Heterogeneous Environments

Han Wang [1] **Sihong He** [2] **Zhili Zhang** [2] **Fei Miao** [2] **James Anderson** [1]

Abstract

We explore a Federated Reinforcement Learning (FRL) problem where N agents collaboratively learn a common policy without sharing their trajectory data. To date, existing FRL work has primarily focused on agents operating in the same or "similar" environments. In contrast, our problem setup allows for arbitrarily large levels of environment heterogeneity. To obtain the optimal policy which maximizes the average performance across all *potentially completely different* environments, we propose two algorithms: FEDSVRPG-M and FEDHAPG-M. In contrast to existing results, we demonstrate that both FEDSVRPG-M and FEDHAPG-M, both of which leverage momentum mechanisms, can exactly converge to a stationary point of the average performance function, regardless of the magnitude of environment heterogeneity. Furthermore, by incorporating the benefits of variance-reduction techniques or Hessian approximation, both algorithms achieve state-of-the-art convergence results, characterized by a sample complexity of $\mathcal{O}\left(\epsilon^{-\frac{3}{2}}/N\right)$. Notably, our algorithms enjoy linear convergence speedups with respect to the number of agents, highlighting the benefit of collaboration among agents in finding a common policy.

1. Introduction

Recently, there has been a lot of interest applying Federated Learning (FL) algorithms to reinforcement learning (RL) problems in order to solve complex sequential decision-making tasks (Qi et al., 2021; Jin et al., 2022; Liu et al., 2019). Federated reinforcement learning (FRL) has been widely applied as it provides the following advantages: First, FRL protects each agent's privacy by only allowing the model to be shared between the server and agent, while keeping the raw data localized. Secondly, by sharing the model with the server, FRL can reduce the sample complexity and produce a better policy than if each agent learns individually with its own limited data. However, existing work in the FRL framework is limited to either multiple agents interacting with the same environment (Fan et al., 2021; Khodadadian et al., 2022) *or* multiple agents with distinct, *yet similar* environments (Jin et al., 2022; Xie & Song, 2023b; Wang et al., 2023a). It remains an open problem to formally characterize how FRL performs when multiple agents from completely different environments, i.e., with arbitrarily large heterogeneity levels, are allowed to collaborate. In this work, we provide an answer to the following question: *what is the best achievable sample complexity when considering severely heterogeneous environments*?

We focus on developing FRL algorithms that compute an optimal universal policy that ensures uniformly good performance for N agents, despite their operation in disparate environments. The motivation for a shared policy stems from practical applications necessitating uniform approaches for distinct agents. For instance, Spotify, a leading audio streaming company, intends to design a uniform pricing plan that suits the listening habits of all users. Given the substantial variations in listening habits among users, establishing a pricing strategy that aligns with the preferences of all users is of great importance. Similarly, autonomous vehicles navigating diverse settings like urban streets, rural areas, and highways must adapt to varied challenges. A uniform policy that adjusts to this environmental heterogeneity ensures consistent, safe decision-making across all terrains, highlighting the need for robust algorithms capable of handling dynamic driving conditions efficiently. Moreover, a universally optimal policy could serve as a foundational model that can be individually fine-tuned, a concept that has gained a lot of attention in meta- and few-shot RL research (Finn et al., 2017; Yu et al., 2020). This approach underscores the broader necessity of designing a uniform and adaptable policy for heterogeneous settings.

In this work, the environment heterogeneity refers to the

[1] Department of Electrical Engineering, Columbia University, New York, USA. [2] School of Computing, University of Connecticut, Storrs, USA. Correspondence to: Han Wang <hw2786@columbia.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

fact that each agent has a different reward function, state transition kernel, or initial state distribution, while they share common state and action spaces. Notably, compared with the existing work (Jin et al., 2022; Wang et al., 2023a), we do not assume that all the environments are similar, i.e., environmental heterogeneity does not need to be bounded by small constants. Instead, we consider a more general setting where the magnitude of heterogeneity can be arbitrary. With this setup, we aim to answer the following question:

> *Is it possible to design a provably efficient FRL algorithm which can accommodate arbitrary levels of environmental heterogeneity among agents?*

We answer this question affirmatively. Our main contributions are listed below.

• **New momentum-powered federated reinforcement learning algorithms:** We propose two new algorithms FedSVRPG-M and FedHAPG-M for solving heterogeneous FRL problems (formally specified in Eq. (3)). Leveraging momentum, we prove that our algorithms, even with constant local step-sizes, converge to the exact stationary point of the heterogeneous FRL problem, *regardless of the magnitude of environment heterogeneity*. This stands in contrast to the state-of-the-art work, which only show convergence to a ball around the stationary point whose radius depends on the environmental heterogeneity levels. Importantly, our results hold even when different notions of environment heterogeneity are considered such as the heterogeneity in Markov decision processes (MDPs) or policy advantage heterogeneity (Xie & Song, 2023b).

• **State-of-the-art convergence rates:** By integrating variance-reduction techniques and curvature information into the policy gradient estimation, our algorithms achieve sample-efficiency improvement over prior work (Fan et al., 2021). In particular, we reduce the sample complexity from $\mathcal{O}\left(\epsilon^{-\frac{5}{3}}/N^{\frac{2}{3}}\right)$ to $\mathcal{O}\left(\epsilon^{-\frac{3}{2}}/N\right)$ when finding the ϵ-approximate first order stationary point[1] (ϵ-FOSP) (Nesterov, 2003). When only a single agent is included, i.e., $N = 1$, our results align with the best known sample complexity of $\mathcal{O}\left(\varepsilon^{-\frac{3}{2}}\right)$ from Fatkhullin et al. (2023).

• **Practical algorithm structures:** Our algorithms are easy to implement because: (1) *Constant local step-sizes*. This feature reduces the amount of algorithm tuning. In contrast, many FL optimization algorithms (Karimireddy et al., 2020; Wang et al., 2019; Yang et al., 2021) require diminishing local step-sizes preset according to complex schedules in order to counteract the effects of heterogeneity. (2) *Sampling one trajectory per local iteration*. This means our algorithms can address the challenge of poor sample efficiency in RL. Unlike existing variance-reduced policy gradient (PG) algorithms for the single agent setting (Papini et al., 2018; Xu et al., 2019; Gargiani et al., 2022), our approach avoids the need for large batch sizes during certain iterations. (3) *Accommodating multiple local updates*. With this feature, our algorithms become more suitable for real-world applications, where communication latency causes serious bottlenecks.

• **Linear speedup:** Analysis of FedSVRPG-M and FedHAPG-M shows that they can converge N-times faster than the scenario where each agent learns a policy on its own. Essentially, by adopting the FL approach, the sample complexity of our algorithms can be linearly scaled by the number of agents N, i.e., collaboration always helps. To our knowledge, *we are the first to achieve a linear speedup for finding a stationary point of FRL problems using policy-based methods*. Importantly, the linear speedup is established even when considering multiple local updates and without making any assumptions about environment heterogeneity. Compared to prior work, our result outperforms that of Jin et al. (2022); Fan et al. (2021), which at best achieves sublinear speedup, see Table 1.

2. Related Work

Federated RL A comprehensive overview of techniques and open problems in FRL was offered by Qi et al. (2021). Much of the work in FRL has focused on developing federated versions of value-based methods (Khodadadian et al., 2022; Woo et al., 2023; Wang et al., 2023a). Notably, Khodadadian et al. (2022) and Woo et al. (2023) established the benefits of FL in terms of linear speedup, assuming all agents operate in *identical* environment. Wang et al. Wang et al. (2023a) introduced the FedTD(0) algorithm to address the FRL problem with *distinct yet similar* environments demonstrated linear speed up was achievable. On the other hand, Zhang et al. (2024) proposed the FedSARSA algorithm to solve the on-policy FRL problem, but it is applicable only in similar environments. Another major area of FRL research studies federated policy-based algorithms (Jin et al., 2022; Xie & Song, 2023b; Fan et al., 2021; Wang et al., 2023b; Lan et al., 2023). However, Fan et al. (2021) only consider uniform environments and only one local update step. While Xie & Song explored diverse environments, they only showed an asymptotic convergence. Most relevant to our work, Jin et al. (2022) studied *heterogeneous* environments. Nevertheless, the algorithms from Jin et al. (2022) were saddled with a non-vanishing convergence error. This non-vanishing error depended on the environmental heterogeneity levels. Note that none of these papers investigated the FRL problems with *arbitrary environment*

[1] Finding a parameter θ such that $\|\nabla J(\theta)\|^2 \leq \epsilon$, where J is defined in Eq. (3). Note that in work such as (Shen et al., 2019; Fatkhullin et al., 2023), the notion $\|\nabla J(\theta)\|^2 \leq \epsilon^2$ is applied instead.

Table 1. Comparision of the results for policy-based methods in FRL. LU and HETER denote the multiple local updates and environment heterogeneity, respectively.

ALGORITHM	CONVERGENCE	SPEEDUP	LU	HETER
PAVG (Jin et al., 2022)	finite but inexact	No speedup	✓	✓
FEDKL (Xie & Song, 2023b)	asymptotic	No speedup	✗	✓
FEDPG-BR (Fan et al., 2021)	finite and exact	Sublinear: $N^{\frac{2}{3}}$	✗	✗
FAPI (Xie & Song, 2023a)	asymptotic and inexact	No speedup	✗	✓
FEDSVRPG-M (Ours)	finite and exact	Linear: N	✓	✓
FEDHAPG-M (Ours)	finite and exact	Linear: N	✓	✓

heterogeneity. To bridge this gap, our proposed algorithms, FEDSVRPG-M and FEDHAPG-M, utilize policy-based techniques and can converge *exactly*. See Table 1 for a comparison of our results with the existing work in FRL policy-based methods.

3. Preliminaries

Federated Learning. Federated learning (FL) is a machine learning approach where a model is trained across multiple clients. Each client runs several iterations of a learning algorithm on its own dataset. Periodically, clients send their local models to the server. The server aggregates the models and then broadcasts the resulting model to all clients and the process repeats. By performing multiple local updates with its own data, FL can substantially reduce communication costs. Our proposed algorithms align with the structure of standard FL algorithms such as FEDAVG (McMahan et al., 2017): an agent performs multiple local updates (using SGD) between two communication rounds. Nonetheless, such local updates will introduce *"client-drift"* problems (Karimireddy et al., 2020; Charles & Konečnỳ, 2021; Wang et al., 2022), presenting a key challenge in FL regarding the trade-off between communication cost and model accuracy. Additionally, handling data that is not identically distributed across devices, affecting both data modeling and convergence analysis, presents another challenge. These challenges are further amplified in the context of FRL.

3.1. Centralized Reinforcement Learning

A centralized reinforcement learning task[2] is generally modeled as a discrete-time Markov Decision Process (MDP): $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \rho\}$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space and ρ denotes the initial state distribution. Here, $\mathcal{P}(s' \mid s, a)$ denotes the probability that the agent transitions from the state s to s' when taking the action $a \in \mathcal{A}$. The discount factor is $\gamma \in (0,1)$, and $\mathcal{R}(s,a) : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is the reward function for taking action a at state s for some constant $R_{\max} > 0$. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is a mapping from the state space $\mathcal{S}$ to the probability distribution over the action space $\mathcal{A}$.

Under any stationary policy, the agent can collect a trajectory $\tau \triangleq \{s_0, a_0, s_1, a_1, \ldots, s_{H-1}, a_{H-1}, s_H\}$, which is the collection of state-action pairs, where H is the maximum length of all trajectories. Once a trajectory τ is obtained, a cumulative discounted reward can be observed; $\mathcal{R}(\tau) \triangleq \sum_{h=0}^{H-1} \gamma^h \mathcal{R}(s_h, a_h)$.

3.2. Policy Gradients

Given finite state and action spaces, the policy $\pi(a|s)$ can be stored in a $|\mathcal{S}| \times |\mathcal{A}|$ table. However, in practice, both the state and action spaces are large and the tabular approach becomes intractable. Alternatively, the policy is parameterized by an unknown parameter $\theta \in \mathbb{R}^d$, the resulting policy is denoted by π_θ. Given the initial distribution ρ, $p(\tau \mid \theta)$ denotes the probability distribution over trajectory τ, which can be calculated as

$$p(\tau \mid \theta) = \rho(s_0) \prod_{h=0}^{H-1} \pi_\theta(a_h \mid s_h) \mathcal{P}(s_{h+1} \mid s_h, a_h).$$

The goal of RL is to find the optimal policy parameter θ that maximizes the expected discounted trajectory reward:

$$\max_{\theta \in \mathbb{R}^d} J(\theta) \triangleq \mathbb{E}_{\tau \sim p(\tau|\theta)}[\mathcal{R}(\tau)] = \int \mathcal{R}(\tau) p(\tau \mid \theta) d\tau. \quad (1)$$

Note that the underlying distribution p in Eq. (1) depends on the variable θ which varies through the whole optimization procedure. This property, referred to as *non-obliviousness*, highlights a unique challenge in RL and creates a notable distinction from supervised learning problems, where the distribution p is stationary.

To deal with the *non-oblivious* and *non-convex* problem (1), a standard approach is to use the policy gradient (PG)

[2] To distinguish from the federated setting, we refer to the single-agent case as centralized RL or when it's clear from context, simply reinforcement learning.

method (Williams, 1992; Sutton et al., 1999). PG takes the first-order derivative of the objective (1) where $\nabla J(\theta)$ can be expressed as

$$\int \mathcal{R}(\tau)\nabla p(\tau \mid \theta)d\tau = \mathbb{E}_{\tau\sim p(\tau|\theta)}[\nabla \log p(\tau \mid \theta)\mathcal{R}(\tau)].$$

Then, the policy θ can be optimized by running gradient ascent-based algorithms. However, since the distribution $p(\tau \mid \theta)$ is unknown, it is impossible to calculate the full gradient. To address this issue, stochastic gradient ascent is typically used, producing a sequence of the form:

$$\theta \leftarrow \theta + \eta \cdot \frac{1}{B}\sum_{i=1}^{B} g(\tau_i \mid \theta)$$

where $\eta > 0$ denotes the stepsize, B is the number of trajectories, and $g(\tau_i \mid \theta)$ is an estimate of the full gradient $\nabla J(\theta)$ using the trajectory τ_i. The most common unbiased estimators of PG are REINFORCE (Williams, 1992) and GPOMDP (Baxter & Bartlett, 2001). In this paper, $g(\tau \mid \theta)$ is defined as

$$g(\tau \mid \theta) = \sum_{t=0}^{H-1}\left(\sum_{h=t}^{H-1}\gamma^h \mathcal{R}\left(s_h, a_h\right)\right)\nabla \log \pi_\theta\left(a_t \mid s_t\right).$$

Importance Sampling Since problem 1 is *non-oblivious*, we have $\mathbb{E}_{\tau\sim p(\tau|\theta)}\left[g(\tau \mid \theta) - g\left(\tau \mid \theta'\right)\right] \neq \nabla J(\theta) - \nabla J\left(\theta'\right)$. To address this issue of distribution shift, we introduce an importance sampling (IS) weight, denoted by

$$w\left(\tau \mid \theta', \theta\right) \triangleq \frac{p\left(\tau \mid \theta'\right)}{p(\tau \mid \theta)} = \prod_{h=0}^{H-1}\frac{\pi_{\theta'}\left(a_h \mid s_h\right)}{\pi_\theta\left(a_h \mid s_h\right)}. \quad (2)$$

With the definition of the IS weight, we can ensure that $\mathbb{E}_{\tau\sim p(\tau|\theta)}\left[g(\tau \mid \theta) - w\left(\tau \mid \theta', \theta\right) g\left(\tau \mid \theta'\right)\right] = \nabla J(\theta) - \nabla J\left(\theta'\right)$.

4. Problem Formulation

We are now ready to characterize heterogeneity in our N-agent FRL problem. Environmental heterogeneity is modeled by allowing each agent to have its own state transition kernel $\mathcal{P}^{(i)}$, reward function $\mathcal{R}^{(i)}$, or the initial state distribution $\rho^{(i)}$. However, all agents share the same state and action space. These environments are characterized by the MDPs, $\mathcal{M}_i = \left\langle \mathcal{S}, \mathcal{A}, \mathcal{R}^{(i)}, \mathcal{P}^{(i)}, \gamma, \rho^{(i)}\right\rangle$, for $i = 1, \cdots, N$.

The objective of FRL is to enable N agents to collaboratively learn a common policy function or a value function that uniformly performs well across all environments. To preserve privacy, agents are not allowed to exchange their raw observations (i.e., their rewards, states, or actions). In particular, we consider solving the following optimization problem:

$$\max_{\theta} \quad \left\{J(\theta) \triangleq \frac{1}{N}\sum_{i=1}^{N} J_i(\theta)\right\}$$

$$\text{where} \quad J_i(\theta) \triangleq \mathbb{E}\left[\sum_{h=0}^{H-1}\gamma^h \mathcal{R}^{(i)}\left(s_h, a_h\right) \mid s_0 \sim \rho^{(i)},\right.$$

$$\left. a_h \sim \pi_\theta\left(\cdot \mid s_h\right), s_{h+1} \sim \mathcal{P}^{(i)}\left(\cdot \mid s_h, a_h\right)\right]. \quad (3)$$

Objective. For solving the optimization problem (3), we aim to find the ϵ-FOSP, i.e., a parameter θ such that $\|\nabla J(\theta)\|^2 \le \varepsilon$. There exists work that leverages the "gradient domination" condition (Agarwal et al., 2020; Liu et al., 2020; Ding et al., 2021; Fatkhullin et al., 2023) for finding a global optimal policy in the centralized RL setting. The gradient domination condition is useful as it guarantees that every stationary policy is globally optimal. However, as shown in Zeng et al. (2021), we cannot expect this condition to hold in general for FL or multi-agent problems. Specifically, even if a single performance function, $J_i(\theta)$, satisfies the "gradient domination" condition, the average function $J(\theta) = \frac{1}{N}\sum_{i=1}^{N} J_i(\theta)$ might not. Zeng et al. (2021) resolved this issue by introducing strong assumptions into the problem. For instance, Assumption 2 in their paper requires that the joint states between the environments are equally explored, which is difficult to verify in real-world applications.

Difference in the problem setup. Our setting is more general than existing work (Jin et al., 2022; Wang et al., 2023a). In our work, each MDP can have a distinct initial state distribution, a feature not addressed in Jin et al.. Furthermore, our framework does not require the bounded heterogeneity assumption of Wang et al. and thus can handle arbitrary environment heterogeneity.

5. Algorithms

To solve problem (3), we present two federated momentum-based algorithms: FEDSVRPG-M and FEDHAPG-M. FEDSVRPG-M is based on a variance reduction method, while FEDHAPG-M leverages a fast Hessian-aided technique. Since FEDSVRPG-M only uses the first-order information (gradient), it is computationally cheaper than FEDHAPG-M, which aims to approximate second-order information (Hessians). Conversely, FEDHAPG-M, with its use of second-order information, is more robust than FEDSVRPG-M.

In the centralized RL setting, momentum-based PG methods (Yuan et al., 2020; Huang et al., 2020) are proposed to reduce the variance of stochastic gradients. In contrast, our algorithms integrate momentum within a federated context, achieving dual benefits: it not only accelerates the convergence and stabilizes oscillations, but also mitigates the impact of environment heterogeneity. Consequently, our algorithms can exactly converge to the ϵ-FOSP of problem (3), no matter how large the environment heterogeneity is. This represents a significant improvement upon (Jin et al., 2022; Xie & Song, 2023a), which only show the convergence to the neighborhood around the stationary point of problem. The size of the neighborhood in their papers is determined by the environment heterogeneity.

5.1. FEDSVRPG-M

We now describe the federated stochastic variance-reduced PG with momentum algorithm (FEDSVRPG-M for short). We outline its steps in Algorithm 1.

FEDSVRPG-M initializes all agents and the server with a common model θ_0. In Algorithm 1, we use the superscript (i) to index the i-th agent and the subscript r and k to denote the r-th communication round and k-th local iteration. In each communication round r, each agent $i \in [N]$ is initiated from a common model θ_r and samples a single trajectory from its own environment to perform K local iterations. At each local iteration k, instead of using PG, FEDSVRPG-M uses the following momentum-based variance-reduced stochastic PG estimator:

$$\begin{aligned} u_{r,k}^{(i)} = \beta g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) + (1-\beta)\Big[u_r + g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) \\ -w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}\right)\Big], \end{aligned} \tag{4}$$

where $\beta \in (0,1]$ and $w^{(i)}$ is the importance sampling weight, which is defined as:

$$w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) \triangleq \frac{p^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}\right)}{p^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right)}.$$

When $\beta = 1$, Eq. (4) reduces to the stochastic PG direction. When $\beta = 0$, it reduces to the variance-reduced PG direction. Notably, compared to the IS-MBPG algorithm of Huang et al. (2020) for the centralized RL setting, the local updating rule in Algorithm 1 differs in that we estimate the PG directions locally, $\theta_{r,k}^{(i)}$, and globally θ_{r-1}, instead of two consecutive local policies. Furthermore, FEDSVRPG-M only requires constant local step-sizes, in contrast to the decreasing step-sizes in Huang et al. (2020). Moreover, FEDSVRPG-M only samples *one trajectory* per iterate, i.e., not does not require very large batch sizes, which is often necessary for centralized variance-reduced PG methods (Xu et al., 2019; Yuan et al., 2020). For more discussion on the variance-reduced PG-type algorithms, we refer readers to Gargiani et al. (2022).

A notable feature of FEDSVRPG-M is communication efficiency and data locality. To save the communication costs and preserve privacy, all agents upload their local model's difference $\Delta_r^{(i)}$, instead of the raw trajectories, to the server only after K local iterations (line 10). Following this step, the server aggregates all the differences to update the global model θ_{r+1} using the global step-size λ and then broadcasts it to all agents. Note that FEDSVRPG-M follows the same structure of the vanilla FEDAVG and achieves the same communication cost per communication round as FEDAVG.

Comparison with prior work. Note that the algorithms in Fan et al. (2021) require the server to own its own environment (an MDP). They utilized the variance-reduced PG method for updating global models on the server side and applied the stochastic PG method to update the local model *only once* on the agent side. In contrast, our algorithms eliminate the need for the server to own its environment, enhancing its applicability in real-world scenarios. This is crucial as, in numerous cases, the server may function as a third-party entity without access to the environment.

Challenges. Most importantly, our algorithms *accommodate multiple local updates*, a crucial step for reducing the communication costs in FL. Thus, it is important for us to mitigate the common *"client-drift"* problems due to heterogeneity among agents. Notably, even for the standard FL algorithms in the supervised setting, it takes a substantial effort for the FL community to tackle this problem, such as FEDPROX (Li et al., 2020), FEDNOVA (Wang et al., 2020), SCAFFOLD (Karimireddy et al., 2020) and FEDLIN (Mitra et al., 2021). This challenge is further exacerbated in FRL, where the *non-oblivious* nature of problems makes it uncertain whether the bounded gradient heterogeneity assumption, commonly employed in FL optimization literature, remains applicable. Consequently, achieving a balance between communication cost and convergence rate is challenging. We analyze the performance of FEDSVRPG-M in Section 6.

5.2. FEDHAPG-M

Recently, HAPG (Shen et al., 2019) has been proposed for the centralized RL to reduce the sample complexity from $\mathcal{O}\left(1/\epsilon^4\right)$ to $\mathcal{O}\left(1/\epsilon^3\right)$ to obtain the ϵ-FOSP. The main success of HAPG comes from that it utilizes the stochastic approximation of the second-order policy differential. While HAPG uses curvature information, the computation cost of HAPG is still *linear* per iteration with respect to the parameter dimension d, as it avoids computing the Hessian explicitly.

We now provide a federated variant of HAPG; Federated Hessian Aided Policy Gradient with Momentum (FEDHAPG-M). As discussed in FEDSVRPG-M, the usage of momentum in FEDHAPG-M primarily serves to offer an "anchoring" direction that encodes PG estimates

Algorithm 1 Description of FEDSVRPG-M

```
Input: initial model $\theta_{-1} = \theta_0$, gradient estimate $u_0$, local step-size $\eta$, global step-size $\lambda$ and momentum $\beta$.
for $r = 0, 1, \ldots, R-1$ do
  ▷ Agent side
  for each agent $i \in [N]$ do
    Initial local model $\theta_{r,0}^{(i)} = \theta_r$
    for $k = 0, 1, \ldots, K-1$ do
      Sample a trajectory $\tau_{r,k}^{(i)} \sim p^{(i)}\left(\tau \mid \theta_{r,k}^{(i)}\right)$ and compute $u_{r,k}^{(i)}$ using Eq. (4).
      Update local model $\theta_{r,k+1}^{(i)} = \theta_{r,k}^{(i)} + \eta u_{r,k}^{(i)}$
    end for
    Send $\Delta_r^{(i)} = \theta_{r,K}^{(i)} - \theta_r$ to the server
  end for
  ▷ Server side
  Aggregate $u_{r+1} = \frac{1}{\eta NK} \sum_{i=1}^{N} \Delta_r^{(i)}$
  Update global model $\theta_{r+1} = \theta_r + \lambda u_{r+1}$
end for
```

from all agents. Consequently, it eliminates the need for bounded environment heterogeneity assumption in existing FRL literature (Jin et al., 2022; Wang et al., 2023a; Xie & Song, 2023b). Moreover, FEDHAPG-M employs a second-order approximation instead of computing the difference between two consecutive stochastic gradients. As a result, FEDHAPG-M obtains an improved sample complexity akin to that of FEDSVRPG-M.

Note that FEDHAPG-M follows the same structure of the vanilla FEDAVG and FEDSVRPG-M, differing only in the local update procedure. In FEDHAPG-M, we replace the local update direction in FEDAVG with a variant of HAPG, see line 7 $\sim$ 9 in Algorithm 2. It is worth noting that the uniform sampling step in line 7 guarantees that $\Lambda_{r,k}^{(i)}$ is an unbiased estimator of $\nabla J(\theta_{r,k}^{(i)}) - \nabla J(\theta_{r-1})$. To estimate the term $\Lambda_{r,k}^{(i)}$, as in Furmston et al. (2016); Shen et al. (2019), we first assume that the function $J_i(\theta)$ is twice differentiable for all $i \in [N]$. Then we compute it as:

$$\Lambda_{r,k}^{(i)} \triangleq \left\langle \nabla \log p\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right), v_{r,k}^{(i)} \right\rangle g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right) + \nabla \left\langle g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right), v_{r,k}^{(i)} \right\rangle \tag{5}$$

where $v_{r,k}^{(i)} \triangleq \theta_{r,k}^{(i)} - \theta_{r-1}$. The variable θ_{r-1} represents the last-iterate global policy maintained in the server. As mentioned in Fatkhullin et al. (2023), the computation of the second term in Eq (5) can be simplified through via automatic differentiation of the scalar quantity $g\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right)$. Thus, the computation cost of FEDHAPG-M does not increase and remains at $\mathcal{O}(Hd)$.

Algorithm 2 Description of FEDHAPG-M

```
Input: initial model $\theta_{-1} = \theta_0$ and gradient estimate $u_0$, local step-size $\eta$, global step-size $\lambda$ and momentum $\beta$.
for $r = 0, \cdots, R-1$ do
  ▷ Agent side
  for each agent $i \in [N]$ do
    Initial local model $\theta_{r,0}^{(i)} = \theta_r$
    for $k = 0, \cdots, K-1$ do
      Choose $\alpha$ uniformly at random from $[0, 1]$, and compute $\theta_{r,k}^{(i)}(\alpha) = \alpha\theta_{r-1} + (1-\alpha)\theta_{r,k}^{(i)}$
      Sample a trajectory $\tau_{r,k}^{(i)}$ from the density $p^{(i)}\left(\tau \mid \theta_{r,k}^{(i)}(\alpha)\right)$ and compute $u_{r,k}^{(i)} = \beta w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}, \theta_{r,k}^{(i)}(\alpha)\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) + (1-\beta)\left[u_r + \Lambda_{r,k}^{(i)}\right]$, where $\Lambda_{r,k}^{(i)}$ can be computed by using Eq. (5)
      Update local model $\theta_{r,k+1}^{(i)} = \theta_{r,k}^{(i)} + \eta u_{r,k}^{(i)}$
    end for
    Send $\Delta_r^{(i)} = \theta_{r,K}^{(i)} - \theta_r$ back to the server
  end for
  ▷ Server side
  Aggregate $u_{r+1} = \frac{1}{\eta NK} \sum_{i=1}^{N} \Delta_r^{(i)}$
  Update global model $\theta_{r+1} = \theta_r + \lambda u_{r+1}$
end for
```

Discussion. Same as FEDSVRPG-M, FEDHAPG-M enjoys the following favorable features: (1) Only sampling one trajectory per local iteration; (2) No need for the server to have its own environment; (3) Multiple local updates. Such features were not simultaneously addressed in Fan et al. (2021); Xie & Song (2023a).

6. Convergence Analysis

First, we introduce some standard assumptions.

Assumption 6.1. Let $\pi_\theta^{(i)}(a \mid s)$ be the policy of the i-th agent at state s. There exist constants $G, M > 0$ such that the log-density of the policy function satisfies

$$\left\|\nabla_\theta \log \pi_\theta^{(i)}(a \mid s)\right\| \leq G, \quad \left\|\nabla_\theta^2 \log \pi_\theta^{(i)}(a \mid s)\right\|_2 \leq M,$$

for all $a \in \mathcal{A}$ and $s \in \mathcal{S}$ and $i \in [N]$.

Assumption 6.2. For each agent $i \in [N]$, the variance of stochastic gradient $g_i(\tau \mid \theta)$ is bounded, i.e., there exists a constant $\sigma > 0$, for all policies π_θ such that $\mathrm{Var}(g_i(\tau \mid \theta)) = \mathbb{E}\|g_i(\tau \mid \theta) - \nabla J_i(\theta)\|^2 \leq \sigma^2$.

Assumption 6.3. For each agent $i \in [N]$, the variance of importance sampling weight $w^{(i)}(\tau \mid \theta_1, \theta_2)$ is bounded, i.e., there exists a constant $W > 0$ such that $\mathrm{Var}\left(w^{(i)}(\tau \mid \theta_1, \theta_2)\right) \leq W$ holds for any $\theta_1, \theta_2 \in \mathbb{R}^d$ and $\tau \sim p^{(i)}(\cdot \mid \theta_2)$.

Assumption 6.1, 6.2 and 6.3 are commonly made in the convergence analysis of PG algorithms and their variance-reduced variants (Papini et al., 2018; Xu et al., 2019; Shen et al., 2019; Liu et al., 2020). They can be easily verified for Gaussian policies (Cortes et al., 2010; Pirotta et al., 2013; Papini et al., 2018). With these assumptions, we are ready to present the convergence guarantees for our FEDSVRPG-M algorithms.

Theorem 6.4. *(*FEDSVRPG-M*) Under Assumption 6.1–6.3, let $u_0 = \frac{1}{NB}\sum_{i=1}^{N}\sum_{b=1}^{B} g_i\left(\tau_b^{(i)}|\theta_0\right)$ with $B = \left\lceil \frac{K}{R\beta^2} \right\rceil$ and $\left\{\tau_b^{(i)}\right\}_{b=1}^{B} \overset{iid}{\sim} p^{(i)}(\tau|\theta_0)$. There exists a constant local step-size η, a proper global step-size λ and momentum coefficient β, such that the output of* FEDSVRPG-M *after R rounds satisfies:*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\|\nabla J(\theta_r)\|^2\right] \lesssim \left(\frac{\bar{L}\Delta\sigma}{NKR}\right)^{2/3} + \frac{\bar{L}\Delta}{R} \tag{6}$$

where $\Delta \triangleq J(\theta^*) - J(\theta_0)$, $G_0 \triangleq \frac{1}{N}\sum_{i=1}^{N}\|\nabla J_i(\theta_0)\|^2$.

Note that $\bar{L}$ in Theorem 6.4 is a constant depending on the constants $G, M, W, H, R_{\max}$ and $\frac{1}{(1-\gamma)^2}$. See Appendix for details. The notation $\lesssim$ denotes that inequalities hold up to some numeric number.

Comparison with prior work in FRL. FEDSVRPG-M surpasses all existing results in FRL in convergence, as shown in Table 1. Specifically, the results in Theorem 6 from Jin et al. (2022) achieve only inexact convergence to a suboptimal solution, depending on the heterogeneity levels among N agents. In contrast, FEDSVRPG-M exactly converges to the ϵ-FOSP of Problem (3), with no heterogeneity term observed in Eq. (6). Fan et al. (2021) exclusively considered the homogeneous environment. However, their results are limited to the sublinear result. i.e., the stationary point optimality can be scaled by $N^{\frac{2}{3}}$. In contrast, the dominant term $\left(\frac{\bar{L}\Delta\sigma}{NKR}\right)^{2/3}$ in the right-hand side of FiEq. (6) demonstrates that our algorithm provides a N-fold linear speedup over the single-agent scenario. Unique to our algorithm is the fact that this speed up is agnostic to the heterogeneity levels, unlike Woo et al. (2023) and Wang et al. (2023a) which obtain a speedup in the no and low heterogeneity regimes respectively.

Comparison with prior work in RL. Compared to the centralized RL, i.e., $N = 1$, FEDSVRPG-M exhibits a convergence rate of $\mathcal{O}\left(1/(KR)^{\frac{2}{3}}\right)$, which aligns with the near-optimal convergence rate in Fatkhullin et al. (2023). In contrast, Huang et al. (2020), utilizing diminishing step-sizes, achieves a slower convergence rate of $\mathcal{O}\left(\log(KR)/(KR)^{\frac{2}{3}}\right)$.

Comparison with prior work in FL optimization. To appreciate the tightness of our results, we note that our results align with the state-of-the-art convergence rates Cheng et al. (2024); Huang et al. (2024) in the FL optimization literature. However, our results are established for a more complex RL setting. In contrast to the supervised learning scenario, where the distribution of τ is fixed over all iterations, our problem is *non-oblivious*. Furthermore, FEDSVRPG-M allows for the constant local step-sizes. In contrast, many FL optimization algorithms (Yang et al., 2021; Khodadadian et al., 2022) require the decreasing local step-sizes to mitigate heterogeneity among agents.

Now, we analyze the convergence of FEDHAPG-M.

Theorem 6.5. *(*FEDHAPG-M*) Under Assumption 6.1–6.3, choose the same u_0 as Theorem 6.4. There exists a constant local step-size η, a proper global step-size λ and momentum coefficient β, such that the output of* FEDHAPG-M *after R rounds satisfies*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\|\nabla J(\theta_r)\|^2\right] \lesssim \left(\frac{\hat{L}\Delta\sigma}{NKR}\right)^{2/3} + \frac{\hat{L}\Delta}{R} \tag{7}$$

where $\Delta \triangleq J(\theta^*) - J(\theta_0)$, $G_0 \triangleq \frac{1}{N}\sum_{i=1}^{N}\|\nabla J_i(\theta_0)\|^2$

From Theorem 6.5, we remark that FEDHAPG-M enjoys the same worst-case convergence rate, i.e., $\mathcal{O}(1/(NKR)^{2/3})$, as FEDSVRPG-M, except for the differences in the constant $\hat{L}$ and parameter selection. Interested readers are referred to Appendix for details.

Based on Theorem 6.4 and 6.5, we can now translate the convergence results to the total sample complexity of each agent, which is shown in the following corollary.

Corollary 6.6. *Under Assumption 6.1–6.3, the sample complexity of* FEDSVRPG-M *and* FEDHAPG-M *is $\mathcal{O}\left(\epsilon^{-\frac{3}{2}}/N\right)$ per agent to find an ϵ-FOSP.*

7. Experiments

We first use tabular environments to verify our theories on the proposed FEDSVRPG-M algorithm. It is important to note that FEDHAPG-M algorithm can not be assessed in the tabular setting due to the objective function $J_i(\theta)$ not being twice differentiable. We then evaluate both FEDSVRPG-M and FEDHAPG-M's performance on MuJoCo (Todorov et al., 2012) with a deep RL extension. The baseline algorithm is the PAVG algorithm (Jin et al., 2022).

Tabular Case. We evaluate the performance of our algorithms in the environment of random MDPs, where both state transitions and reward functions are generated randomly. We use the same method as Jin et al. (2022) to control the environment heterogeneity. First, we randomly sample a nominal state transition kernel $\mathcal{P}_0$ and then generate the environments $\left\{\mathcal{P}^{(i)} = \kappa\mathcal{P}_i + (1-\kappa)\mathcal{P}_0\right\}_{i=1}^{N}$. Each entry

Table 2. Impact of environment heterogeneity κ and momentum coefficient β. We evaluate FEDSVRPG-M with various κ and various momentum coefficient β in $\{0.1, 0.2, 0.5, 0.8\}$. The baseline method is denoted by $\beta = 1$. Larger κ denotes larger environment heterogeneity. Each setting was run with 16,000 random seeds.

	RANDOM MDPS					
	$\kappa = 0$	$\kappa = 0.2$	$\kappa = 0.4$	$\kappa = 0.6$	$\kappa = 0.8$	$\kappa = 1.0$
$\beta = 0.1$	$\mathbf{8.013}_{\pm\mathbf{0.07}}$	$\mathbf{7.957}_{\pm\mathbf{0.07}}$	$\mathbf{7.968}_{\pm\mathbf{0.06}}$	$\mathbf{7.961}_{\pm\mathbf{0.06}}$	$\mathbf{7.964}_{\pm\mathbf{0.07}}$	$\mathbf{7.981}_{\pm\mathbf{0.06}}$
$\beta = 0.2$	$7.876_{\pm 0.06}$	$7.877_{\pm 0.06}$	$7.851_{\pm 0.06}$	$7.837_{\pm 0.06}$	$7.841_{\pm 0.06}$	$7.824_{\pm 0.07}$
$\beta = 0.5$	$7.561_{\pm 0.07}$	$7.208_{\pm 0.06}$	$7.529_{\pm 0.07}$	$7.525_{\pm 0.06}$	$7.536_{\pm 0.07}$	$7.525_{\pm 0.06}$
$\beta = 0.8$	$7.211_{\pm 0.07}$	$7.203_{\pm 0.07}$	$7.201_{\pm 0.06}$	$7.192_{\pm 0.06}$	$7.193_{\pm 0.06}$	$7.184_{\pm 0.06}$
$\beta = 1.0$	$6.965_{\pm 0.07}$	$6.951_{\pm 0.06}$	$6.955_{\pm 0.06}$	$6.936_{\pm 0.06}$	$6.940_{\pm 0.06}$	$6.937_{\pm 0.07}$

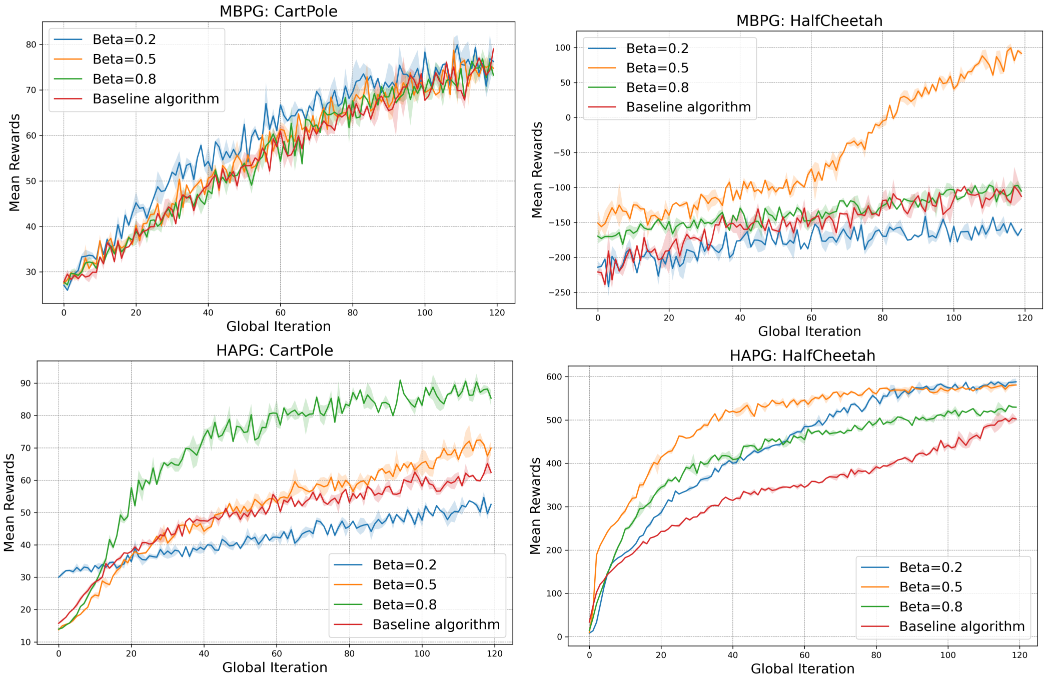

Figure 1. Mean rewards over global iterations for the CartPole and HalfCheetah tasks: (**Top**): FEDSVRPG-M; (**Bottom**): FEDHAPG-M.

of the kernels $\{\mathcal{P}_i\}_{i=1}^N$ are uniformly sampled between 0 and 1 and then normalized. Then, we can evaluate the impact of environment heterogeneity by varying κ. We compare the performance of FEDSVRPG-M with the existing baseline algorithm (PAVG). The results are shown in Table 2. The performance is measured by the average performance function in Eq. (3). We observe that FEDSVRPG-M with $\beta = 0.1$ outperforms the baseline algorithm ($\beta = 1$). Furthermore, the performance of FEDSVRPG-M is agnostic to the environment heterogeneity level κ. These trends are expected and consistent with theoretical analysis in Sec. 6.

Deep RL Case. We evaluate the performance of our algorithms across two benchmark RL tasks: CartPole and HalfCheetah. While CartPole is a classic control task with discrete actions, HalfCheetah represents a continuous RL task. Both are widely recognized tasks in the MuJoCo simulation environment (Todorov et al., 2012). Comprehensive details of the experimental setups can be found in the appendix. To introduce environment heterogeneity, we change the initial state distribution parameters in both tasks. We use Categorical Policy for CartPole, and Gaussian Policy for HalfCheetah. All policies are parameterized by the fully connected neural network which has two hidden layers and a hyperbolic tangent activation function. The hidden layers neural network sizes are 32 for Gaussian policies and 8 for Categorical policies. In Figure 1, we show how the mean rewards change over the global iterations for our proposed algorithms and baseline algorithm. In both tasks, as the number of iterations increases, all algorithms exhibit a rising trend in mean rewards. There exist a $\beta \neq 1$ that our proposed algorithms outperform the baseline algorithm. In particular, FEDSVRPG-M exhibits optimal performance at $\beta = 0.2$ for CartPole and $\beta = 0.5$ for HalfCheetah. In contrast, FEDHAPG-M performs optimally with $\beta = 0.8$ for CartPole and $\beta = 0.5$ for HalfCheetah. FEDHAPG-M, which uses second-order information, shows smaller variance than FEDSVRPG-M, as indicated by the narrower color-shaded regions in the figure. Overall, our algorithms demonstrated superior performance compared to the baseline. See Appendix for more experiments evaluating the linear speedup in the number of agents N.

8. Conclusion

We introduced FEDSVRPG-M and FEDHAPG-M, overcoming the limitation of bounded environment heterogeneity assumed in prior FRL research. Our results demonstrate the best known convergence for these algorithms and highlight the benefits of collaboration in FRL, even in scenarios with conflicting rewards across different environments. In the future, we plan to focus on algorithms that facilitate downstream fine-tuning or personalization, aiming to discover each MDP's optimal policy through FRL, rather than seeking a universally optimal policy.

Acknowledgements

Han Wang and James Anderson are supported by NSF awards NSF awards ECCS 2144634 and ECCS 2047213.

Sihong He, Zhili Zhang, and Fei Miao are supported by the National Science Foundation under Grant CNS-2047354 (CAREER).

We thank Xinmeng Huang for many helpful comments and suggestions.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in Markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *journal of artificial intelligence research*, 15: 319–350, 2001.

Charles, Z. and Konečnỳ, J. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2575–2583. PMLR, 2021.

Cheng, Z., Huang, X., and Yuan, K. Momentum benefits non-iid federated learning simply and provably. *International Conference on Learning Representations*, 2024.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

Ding, Y., Zhang, J., and Lavaei, J. Beyond exact gradients: Convergence of stochastic soft-max policy gradient methods with entropy regularization. *arXiv preprint arXiv:2110.10117*, 2021.

Fan, X., Ma, Y., Dai, Z., Jing, W., Tan, C., and Low, B. K. H. Fault-tolerant federated reinforcement learning with theoretical guarantee. *Advances in Neural Information Processing Systems*, 34:1007–1021, 2021.

Fatkhullin, I., Barakat, A., Kireeva, A., and He, N. Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies. *arXiv preprint arXiv:2302.01734*, 2023.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Furmston, T., Lever, G., and Barber, D. Approximate Newton methods for policy search in markov decision processes. *Journal of Machine Learning Research*, 17, 2016.

Gargiani, M., Zanelli, A., Martinelli, A., Summers, T., and Lygeros, J. PAGE-PG: A simple and loopless variance-reduced policy gradient method with probabilistic gradient estimation. In *International Conference on Machine Learning*, pp. 7223–7240. PMLR, 2022.

Huang, F., Gao, S., Pei, J., and Huang, H. Momentum-based policy gradient methods. In *International conference on machine learning*, pp. 4422–4433. PMLR, 2020.

Huang, X., Li, P., and Li, X. Stochastic controlled averaging for federated learning with communication compression. *International Conference on Learning Representations*, 2024.

Jin, H., Peng, Y., Yang, W., Wang, S., and Zhang, Z. Federated reinforcement learning with environment heterogeneity. In *International Conference on Artificial Intelligence and Statistics*, pp. 18–37. PMLR, 2022.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020.

Khodadadian, S., Sharma, P., Joshi, G., and Maguluri, S. T. Federated reinforcement learning: Linear speedup under Markovian sampling. In *International Conference on Machine Learning*, pp. 10997–11057. PMLR, 2022.

Lan, G., Wang, H., Anderson, J., Brinton, C., and Aggarwal, V. Improved communication efficiency in federated natural policy gradient via admm-based gradient updates. *arXiv preprint arXiv:2310.19807*, 2023.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Liu, B., Wang, L., and Liu, M. Lifelong federated reinforcement learning: a learning architecture for navigation in cloud robotic systems. *IEEE Robotics and Automation Letters*, 4(4):4555–4562, 2019.

Liu, Y., Zhang, K., Basar, T., and Yin, W. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Mitra, A., Jaafar, R., Pappas, G. J., and Hassani, H. Linear convergence in federated learning: Tackling client heterogeneity and sparse gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

Nesterov, Y. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. Stochastic variance-reduced policy gradient. In *International conference on machine learning*, pp. 4026–4035. PMLR, 2018.

Pirotta, M., Restelli, M., and Bascetta, L. Adaptive stepsize for policy gradient methods. *Advances in Neural Information Processing Systems*, 26, 2013.

Qi, J., Zhou, Q., Lei, L., and Zheng, K. Federated reinforcement learning: Techniques, applications, and open challenges. *arXiv preprint arXiv:2108.11887*, 2021.

Shen, Z., Ribeiro, A., Hassani, H., Qian, H., and Mi, C. Hessian aided policy gradient. In *International conference on machine learning*, pp. 5729–5738. PMLR, 2019.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Todorov, E., Erez, T., and Tassa, Y. MuJoCo: A physics engine for model-based control. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pp. 5026–5033. IEEE, 2012.

Wang, H., Marella, S., and Anderson, J. Fedadmm: A federated primal-dual algorithm allowing partial participation. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pp. 287–294. IEEE, 2022.

Wang, H., Mitra, A., Hassani, H., Pappas, G. J., and Anderson, J. Federated temporal difference learning with linear function approximation under environmental heterogeneity. *arXiv preprint arXiv:2302.02212*, 2023a.

Wang, H., Toso, L. F., Mitra, A., and Anderson, J. Model-free learning with heterogeneous dynamical systems: A federated lqr approach. *arXiv preprint arXiv:2308.11743*, 2023b.

Wang, J., Tantia, V., Ballas, N., and Rabbat, M. SlowMo: Improving communication-efficient distributed sgd with slow momentum. *arXiv preprint arXiv:1910.00643*, 2019.

Wang, J., Liu, Q., Liang, H., Joshi, G., and Poor, H. V. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Woo, J., Joshi, G., and Chi, Y. The blessing of heterogeneity in federated Q-learning: Linear speedup and beyond. *arXiv preprint arXiv:2305.10697*, 2023.

Xie, Z. and Song, S. Client selection for federated policy optimization with environment heterogeneity. *arXiv preprint arXiv:2305.10978*, 2023a.

Xie, Z. and Song, S. FedKL: Tackling data heterogeneity in federated reinforcement learning by penalizing KL divergence. *IEEE Journal on Selected Areas in Communications*, 41(4):1227–1242, 2023b.

Xu, P., Gao, F., and Gu, Q. Sample efficient policy gradient methods with recursive variance reduction. *arXiv preprint arXiv:1909.08610*, 2019.

Xu, P., Gao, F., and Gu, Q. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pp. 541–551. PMLR, 2020.

Yang, H., Fang, M., and Liu, J. Achieving linear speedup with partial worker participation in non-iid federated learning. *arXiv preprint arXiv:2101.11203*, 2021.

Yu, T., Quillen, D., He, Z., Julian, R., Hausman, K., Finn, C., and Levine, S. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020.

Yuan, H., Lian, X., Liu, J., and Zhou, Y. Stochastic recursive momentum for policy gradient methods. *arXiv preprint arXiv:2003.04302*, 2020.

Zeng, S., Anwar, M. A., Doan, T. T., Raychowdhury, A., and Romberg, J. A decentralized policy gradient approach to multi-task reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 1002–1012. PMLR, 2021.

Zhang, C., Wang, H., Mitra, A., and Anderson, J. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. *International Conference on Learning Representations*, 2024.

A. Notation

We denote $\mathcal{F}_0 = \emptyset$ and $\mathcal{F}_{r,k}^{(i)} := \sigma\left(\left\{\theta_{r,j}^{(i)}\right\}_{0 \leq j \leq k} \cup \mathcal{F}_r\right)$ and $\mathcal{F}_{r+1} := \sigma\left(\cup_i \mathcal{F}_{r,K}^{(i)}\right)$ for all $r \geq 0$ where $\sigma(\cdot)$ indicates the σ-algebra. Let $\mathbb{E}_r[\cdot] := \mathbb{E}\left[\cdot \mid \mathcal{F}_r\right]$ be the expectation, conditioned on the filtration $\mathcal{F}_r$, with respect to the random variables $\left\{\tau_{r,k}^{(i)}\right\}_{1 \leq i \leq N, 0 \leq k < K}$ in the r-th iteration. Moreover, we use $\mathbb{E}[\cdot]$ to denote the global expectation over all randomness in algorithms. For all $r \geq 0$, we define the following notations to simpify the proof:

$$\begin{aligned}
\Sigma_r &:= \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right) - u_{r+1}\right\|^2\right], \\
\mathcal{D}_r &:= \frac{1}{NK}\sum_i \sum_k \mathbb{E}\left[\left\|\theta_{r,k}^{(i)} - \theta_r\right\|\right]^2, \\
c_{r,k}^{(i)} &:= \mathbb{E}\left[\theta_{r,k+1}^{(i)} - \theta_{r,k}^{(i)} \mid \mathcal{F}_{r,k}^{(i)}\right], \\
\mathcal{M}_r &:= \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}\left[\left\|c_{r,0}^{(i)}\right\|^2\right].
\end{aligned}$$

Throughout the appendix, we denote

$$\Delta := J\left(\theta^*\right) - J(\theta_0),\ G_0 := \frac{1}{N}\sum \left\|\nabla J_i\left(\theta_0\right)\right\|^2,\ \theta_{-1} := \theta_0 \text{ and } \Sigma_{-1} := \mathbb{E}\left[\left\|\nabla J\left(\theta_0\right) - u_0\right\|^2\right].$$

and θ^* denotes the optimal policy of the optimization problem (3).

B. Useful Lemmas and Inequalities

We make repeated use throughout the appendix (often without explicitly stating so) of the following inequalities:

- Given any two vectors $x, y \in \mathbb{R}^d$, for any $a > 0$, we have

$$\|x + y\|^2 \leq (1 + a)\|x\|^2 + \left(1 + \frac{1}{a}\right)\|y\|^2. \tag{8}$$

- Given any two vectors $x, y \in \mathbb{R}^d$, for any constant $a > 0$, we have

$$\langle x, y\rangle \leq \frac{a}{2}\|x\|^2 + \frac{1}{2a}\|y\|^2. \tag{9}$$

  This inequality goes by the name of Young's inequality.

- Given m vectors $x_1, \ldots, x_m \in \mathbb{R}^d$, the following is a simple application of Jensen's inequality:

$$\left\|\sum_{i=1}^{m} x_i\right\|^2 \leq m\sum_{i=1}^{m}\left\|x_i\right\|^2. \tag{10}$$

Proposition 1. (Proposition 5.2 in (Xu et al., 2020)) Under Assumption 1, both $J(\theta)$ and $\{J_i(\theta)\}_{i=1}^N$ are L-smooth with $L = HR_{\max}\left(M + HG^2\right)/(1 - \lambda)$. In addition, for all $\theta_1, \theta_2 \in \mathbb{R}^d$, we have

$$\left\|g_i\left(\tau \mid \theta_1\right) - g_i\left(\tau \mid \theta_2\right)\right\|_2 \leq L_g\left\|\theta_1 - \theta_2\right\|_2$$

and $\|g_i(\tau \mid \theta)\|_2 \leq C_g$ for all $\theta \in \mathbb{R}^d$ and $i \in [N]$, where $L_g = HMR_{\max}/(1 - \lambda), C_g = HGR_{\max}/(1 - \lambda)$.

Lemma B.1. *If $\lambda L \leq \frac{1}{24}$, the following inequality holds for all $r \geq 0$:*

$$\mathbb{E}\left[J\left(\theta_{r+1}\right)\right] \geq \mathbb{E}\left[J\left(\theta_r\right)\right]+\frac{11 \lambda}{24} \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]-\frac{13 \lambda}{24} \Sigma_r .$$

Proof. Since J is L-smooth, we have

$$\begin{aligned} J\left(\theta_{r+1}\right) & \geq J\left(\theta_r\right)+\left\langle\nabla J\left(\theta_r\right), \theta_{r+1}-\theta_r\right\rangle-\frac{L}{2}\left\|\theta_{r+1}-\theta_r\right\|^2 \\ & =J\left(\theta_r\right)+\lambda\left\|\nabla J\left(\theta_r\right)\right\|^2+\lambda\left\langle\nabla J\left(\theta_r\right), u_{r+1}-\nabla J\left(\theta_r\right)\right\rangle-\frac{L \lambda^2}{2}\left\|u_{r+1}\right\|^2 . \end{aligned}$$

where we use the fact that $\theta_{r+1}=\theta_r+\eta u_{r+1}$. By using Young's inequality, we have

$$\begin{aligned} & J\left(\theta_{r+1}\right) \\ \geq & J\left(\theta_r\right)+\frac{\lambda}{2}\left\|\nabla J\left(\theta_r\right)\right\|^2-\frac{\lambda}{2}\left\|\nabla J\left(\theta_r\right)-u_{r+1}\right\|^2-L \lambda^2\left(\left\|\nabla J\left(\theta_r\right)\right\|^2+\left\|\nabla J\left(\theta_r\right)-u_{r+1}\right\|^2\right) \\ \geq & J\left(\theta_r\right)+\frac{11 \lambda}{24}\left\|\nabla J\left(\theta_r\right)\right\|^2-\frac{13 \lambda}{24}\left\|\nabla J\left(\theta_r\right)-u_{r+1}\right\|^2, \end{aligned}$$

where the last inequality holds due to $\lambda L \leq \frac{1}{24}$. Taking the global expectation completes the proof. □

Lemma B.2. *(Lemma 6.1 in (Xu et al., 2020)) Under Assumptions 6.1 and 6.3, we have*

$$\operatorname{Var}\left(w^{(i)}(\tau \mid \theta_1, \theta_2)\right) \leq C_w\left\|\theta_1-\theta_2\right\|^2$$

holds for any $\theta_1, \theta_2 \in \mathcal{R}^d$ and any $i \in [N]$, where $C_\omega=H\left(2 H G^2+M\right)(W+1)$.

C. Federated Stochastic Variance-Reduced Policy Gradient with Momentum

According to the updating rule of FEDSVRPG-M, we have

$$\mathbb{E}[u_{r+1}]=\frac{1}{NK} \sum_{i,k} \mathbb{E}\left[\nabla J_i(\theta_{r,k}^{(i)})+(1-\beta)\left(u_r-\nabla J_i(\theta_r)\right)\right].$$

Lemma C.1. *If $\lambda \leq \sqrt{\frac{16\beta NK}{\widetilde{L_2}^2}}$, we have*

$$\Sigma_r \leq (1-\frac{8\beta}{9})\Sigma_{r-1}+\frac{\widetilde{L_1}^2}{\beta}\mathcal{D}_r+\frac{3\beta^2\sigma^2}{NK}+18\lambda^2\frac{\widetilde{L_2}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2$$

holds for $r \geq 1$, where $\widetilde{L_1}^2 := L^2+24C_wC_g^2+6L_g^2$ and $\widetilde{L_2}^2 := L_g^2+2C_wC_g^2$. When $r=0$, we have

$$\Sigma_0 \leq (1-\beta)\Sigma_{-1}+\frac{\widetilde{L_1}^2}{\beta}\mathcal{D}_0+\frac{3\beta^2\sigma^2}{NK}$$

Proof.

$$\Sigma_r=\mathbb{E}\left[\left\|u_{r+1}-\nabla J(\theta_r)\right\|^2\right]$$

$$= \mathbb{E}\Bigg[\Big\|\frac{1}{NK}\sum_{i,k}\beta g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) + (1-\beta)\Big[u_r + g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}\right)\Big]$$

$$-\nabla J(\theta_r)\Big\|^2\Bigg]$$

$$= \mathbb{E}\Bigg[\Big\|(1-\beta)(u_r - \nabla J(\theta_{r-1})) + \frac{1}{NK}\sum_{i,k}\Big[g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - \nabla J(\theta_r)\Big]$$

$$-(1-\beta)\left[\frac{1}{NK}\sum_{i,k} w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}\right) - \nabla J(\theta_{r-1})\right]\Big\|^2\Bigg]$$

$$= (1-\beta)^2\Sigma_{r-1} + \underbrace{2\mathbb{E}\left[\Big\langle (1-\beta)(u_r - \nabla J(\theta_{r-1})), \frac{1}{NK}\sum_{i,k}\Big[\nabla J_i(\theta_{r,k}^{(i)}) - \nabla J(\theta_r)\Big]\Big\rangle\right]}_{T_1}$$

$$+ \underbrace{\mathbb{E}\left\|\frac{1}{NK}\sum_{i,k}\Big[g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - \nabla J(\theta_r)\Big] - (1-\beta)\left[\frac{1}{NK}\sum_{i,k} w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}\right) - \nabla J(\theta_{r-1})\right]\right\|^2}_{T_2}$$

Using Young's inequality to bound T_1, we have

$$\begin{aligned} T_1 &\leq \beta(1-\beta)^2\mathbb{E}\Big\|u_r - \nabla J(\theta_{r-1})\Big\|^2 + \frac{1}{\beta}\mathbb{E}\Big\|\frac{1}{NK}\sum_{i,k}\nabla J_i(\theta_{r,k}^{(i)}) - \nabla J(\theta_r)\Big\|^2 \\ &\leq \beta(1-\beta)^2\Sigma_{r-1} + \frac{L^2}{\beta}\underbrace{\frac{1}{NK}\sum_{i,k}\mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2}_{\mathcal{D}_r} \end{aligned} \tag{11}$$

Further bounding T_2, we have

$$\begin{aligned} T_2 \leq \mathbb{E}\Big\|&\frac{1}{NK}\sum_{i,k}\left(g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_r\right)\right) \\ &+ \beta\left(\frac{1}{NK}\sum_{i,k} w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_r\right) - \nabla J(\theta_r)\right) \\ &+ (1-\beta)\Big(\frac{1}{NK}\sum_{i,k}\left(w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i(\tau_{r,k}^{(i)} \mid \theta_r) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i(\tau_{r,k}^{(i)} \mid \theta_{r-1})\right) \\ &- \nabla J(\theta_r) + \nabla J(\theta_{r-1})\Big)\Big\|^2 \\ &\leq 3\underbrace{\mathbb{E}\Big\|\frac{1}{NK}\sum_{i,k}\left(g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_r\right)\right)\Big\|^2}_{T_{21}} \\ &+ 3\frac{\beta^2\sigma^2}{NK} + 3(1-\beta)^2\,\underbrace{\mathbb{E}\left[\left\|\frac{1}{NK}\sum_{i,k} w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i(\tau_{r,k}^{(i)} \mid \theta_r) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i(\tau_{r,k}^{(i)} \mid \theta_{r-1})\right\|^2\right]}_{T_{22}} \end{aligned} \tag{12}$$

where we use the Young's inequality in the last equality and the fact that $\mathbb{E}[\|X-\mathbb{E}[X]\|^2] \leq \mathbb{E}[\|X\|^2]$ holds for any random variable X.

To precede, we continue to bound T_{21} and have that

$$\begin{aligned}
T_{21} &= \mathbb{E}\left\| \frac{1}{NK} \sum_{i,k} \left(g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_r\right) \right) \right\|^2 \\
&\leq 2\mathbb{E}\left\| \frac{1}{NK} \sum_{i,k} \left(1 - w^{(i)}(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)})\right) g_i(\tau_{r,k}^{(i)} \mid \theta_r) \right\|^2 \\
&+ 2\mathbb{E}\left\| \frac{1}{NK} \sum_{i,k} \left[g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) - g_i(\tau_{r,k}^{(i)} \mid \theta_r) \right] \right\|^2 \\
&\leq \frac{2C_w C_g^2}{NK} \sum_{i,k} \mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2 + 2\frac{L_g^2}{NK} \sum_{i,k} \mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2 \\
&= (2C_w C_g^2 + 2L_g^2)\mathcal{D}_r
\end{aligned} \tag{13}$$

where we use the fact that $\|g^{(i)}(\tau \mid \theta)\|_2 \leq C_g$ for all $\theta \in \mathbb{R}^d$ and $i \in [N]$.

To bound T_{22}, we have

$$\begin{aligned}
T_{22} &= \mathbb{E}\left[\left\| \frac{1}{NK} \sum_{i,k} w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) g_i(\tau_{r,k}^{(i)} \mid \theta_r) - w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) g_i(\tau_{r,k}^{(i)} \mid \theta_{r-1}) \right\|^2 \right] \\
&\leq 3\mathbb{E}\left[\left\| \frac{1}{NK} \sum_{i,k} \left[w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_r, \theta_{r,k}^{(i)}\right) - 1 \right] g_i(\tau_{r,k}^{(i)} \mid \theta_r) \right\|^2 \right] \\
&+ 3\frac{1}{N^2K^2} \sum_{i,k} \mathbb{E}\left\| g_i(\tau_{r,k}^{(i)} \mid \theta_r) - g_i(\tau_{r,k}^{(i)} \mid \theta_{r-1}) \right\|^2 + 3\mathbb{E}\left[\left\| \frac{1}{NK} \sum_{i,k} \left[w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r-1}, \theta_{r,k}^{(i)}\right) - 1 \right] g_i(\tau_{r,k}^{(i)} \mid \theta_{r-1}) \right\|^2 \right] \\
&\leq 3C_g^2 C_w \frac{1}{N^2K^2} \sum_{i,k} \mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2 + 3\frac{L_g^2}{NK}\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2 + 3C_g^2 C_w \frac{1}{N^2K^2} \sum_{i,k} \mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_{r-1}\right\|^2 \\
&\leq 6C_g^2 C_w \frac{1}{NK}\mathcal{D}_r + \frac{3L_g^2 + 6C_w C_g^2}{NK}\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2
\end{aligned} \tag{14}$$

Combining the upper bound of T_{21} and T_{22} (i.e., (13) and (14)) into T_2 in Eq. (12), we have

$$T_2 \leq (24C_w C_g^2 + 6L_g^2)\mathcal{D}_r + \frac{3\beta^2\sigma^2}{NK} + 9(1-\beta)^2 \frac{L_g^2 + 2C_w C_g^2}{NK}\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2 \tag{15}$$

Therefore, for $r \geq 1$, we have

$$\begin{aligned}
\Sigma_r &\leq (1-\beta)\Sigma_{r-1} + \frac{L^2 + 24C_w C_g^2 + 6L_g^2}{\beta}\mathcal{D}_r + \frac{3\beta^2\sigma^2}{NK} + 9(1-\beta)^2 \frac{L_g^2 + 2C_w C_g^2}{NK}\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2 \qquad (16) \\
&\leq (1-\beta)\Sigma_{r-1} + \frac{L^2 + 24C_w C_g^2 + 6L_g^2}{\beta}\mathcal{D}_r + \frac{3\beta^2\sigma^2}{NK} \\
&+ 18\lambda^2 \frac{L_g^2 + 2C_w C_g^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 + 18\lambda^2 \frac{L_g^2 + 2C_w C_g^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1}) - u_r\right\|^2 \\
&= \left(1 - \beta + 18\lambda^2 \frac{L_g^2 + 2C_w C_g^2}{NK}\right)\Sigma_{r-1} + 18\lambda^2 \frac{L_g^2 + 2C_w C_g^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2
\end{aligned}$$

$$+\frac{L^2+24C_wC_g^2+6L_g^2}{\beta}\mathcal{D}_r+\frac{3\beta^2\sigma^2}{NK} \tag{17}$$

By choosing λ such that $18\lambda^2\frac{L_g^2+2C_wC_g^2}{NK}\le\frac{8\beta}{9}$, which holds when $\lambda\le\sqrt{\frac{16\beta NK}{L_g^2+2C_wC_g^2}}$, we have

$$\Sigma_r\le(1-\frac{8\beta}{9})\Sigma_{r-1}+\frac{L^2+24C_wC_g^2+6L_g^2}{\beta}\mathcal{D}_r+\frac{3\beta^2\sigma^2}{NK}+18\lambda^2\frac{L_g^2+2C_wC_g^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 \tag{18}$$

holds for $r>0$. When $r=0$, we have that

$$\Sigma_0\le(1-\beta)\Sigma_{-1}+\frac{L^2+24C_wC_g^2+6L_g^2}{\beta}\mathcal{D}_0+\frac{3\beta^2\sigma^2}{NK} \tag{19}$$

which can be derived from Eq.(16). □

Lemma C.2. *(Bounding drift-term) If the local step-size satisfies* $\eta\le\min\{\frac{L}{32e^2\widetilde{L_3}^2K},\frac{1}{KL}\}$*, the drift-term can be upper bounded as:*

$$\mathcal{D}_r\le4eK^2\mathcal{M}_r+(16\eta^4K^4L^2+8\eta^2K)\left(\beta^2\sigma^2+2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)$$

where $\widetilde{L_3}^2:=2C_wC_g^2+2L_g^2$.

Proof. Define $c_{r,k}^{(i)}:=-\eta\left(\nabla J_i(\theta_{r,k}^{(i)})+(1-\beta)(u_r-\nabla J_i(\theta_{r-1}))\right)$. For any $1\le j\le k-1\le K-2$, we have:

$$\begin{aligned}\mathbb{E}\left\|c_{r,j}^{(i)}-c_{r,j-1}^{(i)}\right\|^2&\le\eta^2L^2\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_{r,j-1}^{(i)}\right\|^2\\&=\eta^2L^2\left(\mathbb{E}\left\|c_{r,j-1}^{(i)}\right\|^2+\mathbb{E}\left[\mathrm{Var}\left[\theta_{r,j}^{(i)}-\theta_{r,j-1}^{(i)}\mid\mathcal{F}_{r,j-1}^{(i)}\right]\right]\right).\end{aligned} \tag{20}$$

where we use the bias-variance decomposition in the last inequality.

$$\begin{aligned}&\mathbb{E}\left[\mathrm{Var}\left[\theta_{r,j}^{(i)}-\theta_{r,j-1}^{(i)}\mid\mathcal{F}_{r,j-1}^{(i)}\right]\right]\\&=\eta^2\mathbb{E}\Bigg\|g_i\left(\tau_{r,j-1}^{(i)}\mid\theta_{r,j-1}^{(i)}\right)-\nabla J_i(\theta_{r,j-1}^{(i)})\\&-(1-\beta)\left[w^{(i)}\left(\tau_{r,j-1}^{(i)}\mid\theta_{r-1},\theta_{r,j-1}^{(i)}\right)g_i\left(\tau_{r,j-1}^{(i)}\mid\theta_{r-1}\right)-\nabla J_i(\theta_{r-1})\right]\Bigg\|^2\\&=\eta^2\mathbb{E}\Bigg\|\beta\left[g_i\left(\tau_{r,j-1}^{(i)}\mid\theta_{r,j-1}^{(i)}\right)-\nabla J_i(\theta_{r,j-1}^{(i)})\right]\\&+(1-\beta)\Big[g_i\left(\tau_{r,j-1}^{(i)}\mid\theta_{r,j-1}^{(i)}\right)-w^{(i)}\left(\tau_{r,j-1}^{(i)}\mid\theta_{r-1},\theta_{r,j-1}^{(i)}\right)g_i\left(\tau_{r,j-1}^{(i)}\mid\theta_{r-1}\right)\\&\quad-\left(\nabla J_i(\theta_{r,j-1}^{(i)})-\nabla J_i(\theta_{r-1})\right)\Big]\Bigg\|^2\\&\le2\eta^2\beta^2\sigma^2\\&+2\eta^2(1-\beta)^2\underbrace{\mathbb{E}\left\|g_i\left(\tau_{r,j-1}^{(i)}\mid\theta_{r,j-1}^{(i)}\right)-w^{(i)}\left(\tau_{r,j-1}^{(i)}\mid\theta_{r-1},\theta_{r,j-1}^{(i)}\right)g_i\left(\tau_{r,j-1}^{(i)}\mid\theta_{r-1}\right)\right\|^2}_{T_3}\end{aligned} \tag{21}$$

where Eq.(21) holds due to the Young's inequality and the fact that $\mathbb{E}[\|X-\mathbb{E}[X]\|^2]\le\mathbb{E}[\|X\|^2]$.

To precede, we bound T_3 as

$$\begin{aligned}
T_3 &= \mathbb{E}\left\| g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r,j-1}^{(i)}\right) - w^{(i)}\left(\tau_{r,j-1}^{(i)} \mid \theta_{r-1}, \theta_{r,j-1}^{(i)}\right) g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r-1}\right) \right\|^2 \\
&\leq 2\mathbb{E}\left\| \left(1 - w^{(i)}(\tau_{r,j-1}^{(i)} \mid \theta_r, \theta_{r,j-1}^{(i)})\right) g_i(\tau_{r,j-1}^{(i)} \mid \theta_r) \right\|^2 \\
&+ 2\mathbb{E}\left\| g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r,j-1}^{(i)}\right) - g_i(\tau_{r,j-1}^{(i)} \mid \theta_{r-1}) \right\|^2 \\
&\leq 2C_w C_g^2 \mathbb{E}\left\| \theta_{r,j-1}^{(i)} - \theta_{r-1} \right\|^2 + 2L_g^2 \mathbb{E}\left\| \theta_{r,j-1}^{(i)} - \theta_{r-1} \right\|^2 \\
&= (2C_w C_g^2 + 2L_g^2)\mathbb{E}\left\| \theta_{r,j-1}^{(i)} - \theta_{r-1} \right\|^2 \qquad (22)
\end{aligned}$$

where we use the fact that $\|g^{(i)}(\tau \mid \theta)\|_2 \leq C_g$ for all $\theta \in \mathbb{R}^d$ and $i \in [N]$.

With the upper bound of T_3 and $\widetilde{L_3}^2 := 2C_w C_g^2 + 2L_g^2$, we have

$$\begin{aligned}
\mathbb{E}\left\| c_{r,j}^{(i)} - c_{r,j-1}^{(i)} \right\|^2 &\leq \eta^2 L^2 \left(\mathbb{E}\left\| c_{r,j-1}^{(i)} \right\|^2 + 2\eta^2\beta^2\sigma^2 + 2\eta^2(1-\beta)^2 \widetilde{L_3}^2 \mathbb{E}\left\| \theta_{r,j-1}^{(i)} - \theta_{r-1} \right\|^2 \right) \\
&\leq \eta^2 L^2 \left(\mathbb{E}\left\| c_{r,j-1}^{(i)} \right\|^2 + 2\eta^2\beta^2\sigma^2 + 4\eta^2 \widetilde{L_3}^2 \mathbb{E}\left\| \theta_{r,j-1}^{(i)} - \theta_r \right\|^2 + 4\eta^2 \widetilde{L_3}^2 \mathbb{E}\left\| \theta_{r-1} - \theta_r \right\|^2 \right). \qquad (23)
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathbb{E}\left\| c_{r,j}^{(i)} \right\|^2 \leq (1 + \frac{1}{q})\mathbb{E}\left\| c_{r,j-1}^{(i)} \right\|^2 + (1+q)\mathbb{E}\left\| c_{r,j}^{(i)} - c_{r,j-1}^{(i)} \right\|^2 \\
&\leq (1 + \frac{2}{q})\mathbb{E}\left\| c_{r,j-1}^{(i)} \right\|^2 + (1+q)\eta^2 L^2 \left(2\eta^2\beta^2\sigma^2 + 4\eta^2 \widetilde{L_3}^2 \mathbb{E}\left\| \theta_{r,j-1}^{(i)} - \theta_r \right\|^2 + 4\eta^2 \widetilde{L_3}^2 \mathbb{E}\left\| \theta_{r-1} - \theta_r \right\|^2 \right) \qquad (24)
\end{aligned}$$

where we use the fact that $\eta L \leq \frac{1}{K} \leq \frac{1}{q+1}$ and let $q = k - 1$. By unrolling this recurrence, we have

$$\begin{aligned}
\mathbb{E}\left\| c_{r,j}^{(i)} \right\|^2 &\leq (1 + \frac{2}{k-1})^j \mathbb{E}\left\| c_{r,0}^{(i)} \right\|^2 + k\eta^2 L^2 \sum_{i=0}^{j-1} (2\eta^2\beta^2\sigma^2 + 4\eta^2 \widetilde{L_3}^2 \mathbb{E}\left\| \theta_{r-1} - \theta_r \right\|^2) \Pi_{j'=i+1}^{j-1} (1 + \frac{2}{k-1}) \\
&+ k\eta^2 L^2 \sum_{s=0}^{j-1} (4\eta^2 \widetilde{L_3}^2 \mathbb{E}\left\| \theta_{r,s}^{(i)} - \theta_r \right\|^2) \Pi_{j'=s+1}^{j-1} (1 + \frac{2}{k-1}) \\
&\leq (1 + \frac{2}{k-1})^{k-1} \mathbb{E}\left\| c_{r,0}^{(i)} \right\|^2 + k\eta^2 L^2 \sum_{i=0}^{k-1} (2\eta^2\beta^2\sigma^2 + 4\eta^2 \widetilde{L_3}^2 \mathbb{E}\left\| \theta_{r-1} - \theta_r \right\|^2)(1 + \frac{2}{k-1})^{k-1} \\
&+ k\eta^2 L^2 \sum_{j'=0}^{j-1} (4\eta^2 \widetilde{L_3}^2 \mathbb{E}\left\| \theta_{r,j'}^{(i)} - \theta_r \right\|^2)(1 + \frac{2}{k-1})^{k-1} \qquad (25)
\end{aligned}$$

Based on the inequality $(1 + \frac{2}{K-1}^{k-1}) \leq e^2 \leq 8$, we have

$$\mathbb{E}\left\| c_{r,j}^{(i)} \right\|^2 \leq e^2 \mathbb{E}\left\| c_{r,0}^{(i)} \right\|^2 + 8k^2\eta^4 L^2 \left(2\beta^2\sigma^2 + 4\widetilde{L_3}^2 \mathbb{E}\left\| \theta_{r-1} - \theta_r \right\|^2 \right) + 4e^2 k\eta^4 L^2 \widetilde{L_3}^2 \sum_{j'=0}^{j-1} \mathbb{E}\left\| \theta_{r,j'}^{(i)} - \theta_r \right\|^2 \qquad (26)$$

By Lemma A.3, we have

$$\mathbb{E}\left\| \theta_{r,k}^{(i)} - \theta_r \right\|^2 \leq 2\mathbb{E}\left\| \sum_{j=0}^{k-1} c_{r,j}^{(i)} \right\|^2 + 2\sum_{j=0}^{k-1} \mathbb{E}\left[\mathrm{Var}\left[\theta_{r,j+1}^{(i)} - \theta_{r,j}^{(i)} \mid \mathcal{F}_{r,j}^{(i)} \right] \right]$$

$$\overset{(a)}{\leq} 2k\sum_{j=0}^{k-1}\mathbb{E}\left\|c_{r,j}^{(i)}\right\|^2 + 2\sum_{j=0}^{k-1}\left(2\beta^2\eta^2\sigma^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right) \tag{27}$$

where (a) is due to Eq.(21) and Eq.(22). Plugging Eq.(26) into Eq.(27), we have

$$\begin{aligned}
&\mathbb{E}\left\|\theta_{r,k}^{(i)}-\theta_r\right\|^2 \leq \\
&2k\sum_{j=0}^{k-1}\left\{e^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + 8k^2\eta^4L^2\left(2\beta^2\sigma^2 + 4\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right) + 4e^2k\eta^4L^2\widetilde{L_3}^2\sum_{j'=0}^{j-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)}-\theta_r\right\|^2\right\} \\
&+ 2\sum_{j=0}^{k-1}\left(2\beta^2\eta^2\sigma^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2 + 4\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)
\end{aligned} \tag{28}$$

Summing up the above equation over $k = 0, \cdots, K-1$, we have

$$\begin{aligned}
\sum_{k=0}^{K-1}\mathbb{E}\left\|\theta_{r,k}^{(i)}-\theta_r\right\|^2 &\leq \sum_{k=0}^{K-1}\left\{2k^2e^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + 16k^4\eta^4L^2\left(2\beta^2\sigma^2 + 4\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)\right\} \\
&+ \sum_{k=0}^{K-1} 8e^2k^2\eta^4L^2\widetilde{L_3}^2\sum_{j=0}^{k-1}\sum_{j'=0}^{j-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)}-\theta_r\right\|^2 \\
&+ \sum_{k=0}^{K-1}\left(4k\beta^2\eta^2\sigma^2 + 8k\eta^2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2 + 8\eta^2\widetilde{L_3}^2\sum_{j=0}^{k-1}\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2\right) \\
&\leq 2eK^3\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + (8\eta^4K^5L^2 + 4\eta^2K^2)\left(\beta^2\sigma^2 + 2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right) \\
&+ K^2\sum_{k=0}^{K-1}8e^2\eta^4L^2\widetilde{L_3}^2\sum_{j=0}^{K-1}\sum_{j'=0}^{K-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)}-\theta_r\right\|^2 + \sum_{k=0}^{K-1}8\eta^2\widetilde{L_3}^2\sum_{j=0}^{K-1}\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2 \\
&= 2eK^3\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + (8\eta^4K^5L^2 + 4\eta^2K^2)\left(\beta^2\sigma^2 + 2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right) \\
&+ (8e^2\eta^4K^4L^2\widetilde{L_3}^2 + 8\eta^2\widetilde{L_3}^2K)\sum_{j=0}^{K-1}\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2
\end{aligned} \tag{29}$$

With the choice of step-size η satisfying $8e^2\eta^4K^4L^2\widetilde{L_3}^2 + 8\eta^2\widetilde{L_3}^2K \leq \frac{1}{2}$, after some rearrangement, we have

$$\frac{1}{2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\theta_{r,k}^{(i)}-\theta_r\right\|^2 \leq 2eK^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + (8\eta^4K^4L^2 + 4\eta^2K)\left(\beta^2\sigma^2 + 2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)$$

In summary, we can bound the drift-term as

$$\mathcal{D}_r \leq 4eK^2\underbrace{\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2}_{\mathcal{M}_r} + (16\eta^4K^4L^2 + 8\eta^2K)\left(\beta^2\sigma^2 + 2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)$$

□

Lemma C.3. *If* $\lambda L \leq \frac{1}{24}$ *and* $\eta^2\left[\frac{289}{72}(1-\beta)^2 + 8e(\lambda\beta LR)^2\right] \leq \frac{\beta^2}{288eK^2\widetilde{L_1}^2}$, *we have*

$$\sum_{r=0}^{R-1}\mathcal{M}_r = \frac{1}{N}\sum_{r=0}^{R-1}\sum_{i=1}^{N}\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 \leq \frac{\beta^2}{288eK^2\widetilde{L_1}^2}\sum_{r=-1}^{R-2}\left(\Sigma_r + \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]\right) + 4\eta^2\beta^2 eRG_0. \tag{30}$$

where $G_0 := \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\|\nabla J_i(\theta_0)\|^2\right]$ *and* $\widetilde{L_1}^2$ *is defined in Lemma C.1.*

Proof. Recall that $c_{r,0}^{(i)} := -\eta\left(\nabla J_i(\theta_r) + (1-\beta)(u_r - \nabla J_i(\theta_{r-1}))\right)$. Then, it is straightforward to have

$$
\begin{aligned}
\left\|c_{r,0}^{(i)}\right\|^2 &\le 2\eta^2\left((1-\beta)^2\|u_r\|^2 + \|\nabla J_i(\theta_r) - (1-\beta)\nabla J_i(\theta_{r-1})\|^2\right)\\
&\le 2\eta^2(1-\beta)^2\|u_r\|^2 + 4\eta^2(1-\beta)^2\|\nabla J_i(\theta_r) - \nabla J_i(\theta_{r-1})\|^2 + 4\eta^2\beta^2\|\nabla J_i(\theta_r)\|^2\\
&\le 2\eta^2(1-\beta)^2\left(1 + 2(\lambda L)^2\right)\|u_r\|^2 + 4\eta^2\beta^2\|\nabla J_i(\theta_r)\|^2\\
&\overset{(a)}{\le} \frac{289}{144}\eta^2(1-\beta)^2\|u_r\|^2 + 4\eta^2\beta^2\|\nabla J_i(\theta_r)\|^2 . \qquad (31)
\end{aligned}
$$

where (a) is due to the choice of λ such that $\lambda L \le \frac{1}{24}$.

Using the Young's inequality, we have that for any $\zeta > 0$,

$$
\begin{aligned}
\mathbb{E}\left[\|\nabla J_i(\theta_r)\|^2\right] &\le (1+\zeta)\mathbb{E}\left[\|\nabla J_i(\theta_{r-1})\|^2\right] + \left(1 + \frac{1}{\zeta}\right)\mathbb{E}\|\nabla J_i(\theta_r) - \nabla J_i(\theta_{r-1})\|^2\\
&\le (1+\zeta)\mathbb{E}\left[\|\nabla J_i(\theta_{r-1})\|^2\right] + \left(1 + \frac{1}{\zeta}\right)L^2\mathbb{E}\|\theta_r - \theta_{r-1}\|^2\\
&\le (1+\zeta)\mathbb{E}\left[\|\nabla J_i(\theta_{r-1})\|^2\right] + 2\left(1 + \frac{1}{\zeta}\right)(\lambda L)^2\left(\mathbb{E}\|u_r - \nabla J(\theta_{r-1})\|^2 + \mathbb{E}\|\nabla J(\theta_r)\|^2\right)\\
&= (1+\zeta)\mathbb{E}\left[\|\nabla J_i(\theta_{r-1})\|^2\right] + 2\left(1 + \frac{1}{\zeta}\right)(\lambda L)^2\left(\Sigma_{r-1} + \mathbb{E}\|\nabla J(\theta_r)\|^2\right)
\end{aligned}
$$

By unrolling the recursive bound, we have

$$
\mathbb{E}\left[\|\nabla J_i(\theta_r)\|^2\right] \le (1+\zeta)^r\mathbb{E}\left[\|\nabla J_i(\theta_0)\|^2\right] + \frac{2}{\zeta}(\lambda L)^2\sum_{j=0}^{r-1}\left(\Sigma_j + \mathbb{E}\left[\|\nabla J(\theta_j)\|^2\right]\right)(1+\zeta)^{r-j}
$$

By choosing $\zeta = \frac{1}{r}$, we have

$$
\mathbb{E}\left[\|\nabla J_i(\theta_r)\|^2\right] \le e\mathbb{E}\left[\|\nabla J_i(\theta_0)\|^2\right] + 2e(r+1)(\lambda L)^2\sum_{j=0}^{r-1}\left(\Sigma_j + \mathbb{E}\left[\|\nabla J(\theta_j)\|^2\right]\right) \qquad (32)
$$

Summing up Eq. (31) over $r = 0, 1, \cdots, R-1$ and then averaing Eq. (31) over all $i \in N$, we have

$$
\begin{aligned}
\sum_{r=0}^{R-1}\mathcal{M}_r &\le \sum_{r=0}^{R-1}\mathbb{E}\left[\frac{289}{144}\eta^2(1-\beta)^2\|u_r\|^2 + 4\eta^2\beta^2\frac{1}{N}\sum_{i=1}^{N}\|\nabla J_i(\theta_r)\|^2\right]\\
&\le \sum_{r=0}^{R-1}\frac{289}{72}\eta^2(1-\beta)^2\left(\Sigma_{r-1} + \mathbb{E}[\|\nabla J(\theta_{r-1})\|^2]\right)\\
&\overset{(b)}{+} 4\eta^2\beta^2\sum_{r=0}^{R-1}\left(\frac{e}{N}\sum_{i=1}^{N}\mathbb{E}\left[\|\nabla J_i(\theta_0)\|^2\right] + 2e(r+1)(\lambda L)^2\sum_{j=0}^{r-1}\left(\Sigma_j + \mathbb{E}\left[\|\nabla J(\theta_j)\|^2\right]\right)\right) \qquad (33)\\
&\le \frac{289}{72}\eta^2(1-\beta)^2\sum_{r=0}^{R-1}\left(\Sigma_{r-1} + \mathbb{E}\left[\|\nabla J(\theta_{r-1})\|^2\right]\right)\\
&+ 4\eta^2\beta^2\left(eRG_0 + 2e(\lambda LR)^2\sum_{r=0}^{R-2}\left(\Sigma_r + \mathbb{E}\left[\|\nabla J(\theta_r)\|^2\right]\right)\right)\\
&\overset{(c)}{\le} \frac{\beta^2}{288eK^2\widetilde{L_1}^2}\sum_{r=-1}^{R-2}\left(\Sigma_r + \mathbb{E}\left[\|\nabla J(\theta_r)\|^2\right]\right) + 4\eta^2\beta^2 eRG_0.
\end{aligned}
$$

where (b) is due to the upper bound of $\mathbb{E}\left[\left\|\nabla J_i\left(\theta_r\right)\right\|^2\right]$ in Eq.(32) and (c) is due to the choice of η such that $\eta^2\left[\frac{289}{72}(1-\beta)^2+8e(\lambda\beta LR)^2\right] \leq \frac{\beta^2}{288eK^2\widetilde{L_1}^2}$. □

C.1. Proof of Theorem 6.4

Theorem C.4. *(Complete version of Theorem 6.4) Under Assumptions 6.1–6.3, by setting* $u_0 = \frac{1}{NB}\sum_{i=1}^N\sum_{b=1}^B g_i\left(\tau_b^{(i)}|\theta_0\right)$ *with* $\left\{\tau_b^{(i)}\right\}_{b=1}^B \overset{iid}{\sim} p^{(i)}(\tau|\theta_0)$ *and choosing* $\beta = \min\left\{1, \left(\frac{NK\bar{L}^2\Delta^2}{\sigma^4R^2}\right)^{1/3}\right\}$, $\lambda = \min\left\{\frac{1}{24\bar{L}}, \sqrt{\frac{\beta NK}{162\bar{L}^2}}\right\}$, $B = \left\lceil\frac{K}{R\beta^2}\right\rceil$, *and*

$$\eta K\bar{L} \lesssim \min\left\{\left(\frac{\bar{L}\Delta}{G_0\lambda\bar{L}R}\right)^{1/2}, \left(\frac{\beta}{N}\right)^{1/2}, \left(\frac{\beta}{NK}\right)^{1/4}\right\}$$

in Algorithm 1, then the output of FEDSVRPG-M *after R rounds satisfies:*

$$\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right] \lesssim \left(\frac{\bar{L}\Delta\sigma}{NKR}\right)^{2/3}+\frac{\bar{L}\Delta}{R}, \tag{34}$$

where $\bar{L} := \max\{L, \widetilde{L_1}, \widetilde{L_2}, \widetilde{L_3}\}$ *and* $L, \widetilde{L_1}, \widetilde{L_2}, \widetilde{L_3}$ *are defined in Proposition 1, Lemma C.1 and Lemma C.2, respectively.*

Proof. Based on Lemma C.1, we have for any $r \geq 1$

$$\begin{aligned}\Sigma_r &\leq (1-\frac{8\beta}{9})\Sigma_{r-1}+\frac{\widetilde{L_1}^2}{\beta}\mathcal{D}_r+\frac{3\beta^2\sigma^2}{NK}+18\lambda^2\frac{\widetilde{L_2}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 \\ &\leq (1-\frac{8\beta}{9})\Sigma_{r-1}+18\lambda^2\frac{\widetilde{L_2}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2+\frac{3\beta^2\sigma^2}{NK} \\ &\quad+\frac{\widetilde{L_1}^2}{\beta}\left(4eK^2\mathcal{M}_r+(16\eta^4K^4L^2+8\eta^2K)\right)\left(\beta^2\sigma^2+2\widetilde{L_3}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)\end{aligned} \tag{35}$$

where the last inequality is due to Lemma C.2. When $r=0$, we have

$$\Sigma_0 \leq (1-\beta)\Sigma_{-1}+\frac{3\beta^2\sigma^2}{NK}+\frac{\widetilde{L_1}^2}{\beta}\left(4eK^2\mathcal{M}_0+(16\eta^4K^4L^2+8\eta^2K)\right)\beta^2\sigma^2$$

Summing up the above equation over r from 0 to $R-1$, we have

$$\begin{aligned}\sum_{r=0}^{R-1}\Sigma_r &\leq \left(1-\frac{8\beta}{9}\right)\sum_{r=-1}^{R-2}\Sigma_r+\frac{18(\lambda\widetilde{L_2})^2}{NK}\mathbb{E}\left[\sum_{r=0}^{R-2}\left\|\nabla J\left(\theta_r\right)\right\|^2\right]+\frac{3\beta^2\sigma^2}{NK}R \\ &\quad+\frac{\widetilde{L_1}^2}{\beta}\left(4eK^2\sum_{r=0}^{R-1}\mathcal{M}_r+8(\eta K)^2\left(2(\eta KL)^2+\frac{1}{K}\right)\left(R\beta^2\sigma^2+2L^2\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\theta_r-\theta_{r-1}\right\|^2\right]\right)\right)\end{aligned}$$

By incorporating Lemma C.3 into the inequality above, we have

$$\begin{aligned}\sum_{r=0}^{R-1}\Sigma_r &\leq \left(1-\frac{8\beta}{9}\right)\sum_{r=-1}^{R-2}\Sigma_r+\frac{18(\lambda\widetilde{L_2})^2}{NK}\mathbb{E}\left[\sum_{r=0}^{R-2}\left\|\nabla J\left(\theta_r\right)\right\|^2\right]+\frac{3\beta^2\sigma^2}{NK}R \\ &\quad+\frac{\widetilde{L_1}^2}{\beta}8(\eta K)^2\left(2(\eta KL)^2+\frac{1}{K}\right)\left(R\beta^2\sigma^2+2L^2\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\theta_r-\theta_{r-1}\right\|^2\right]\right)\end{aligned}$$

$$\begin{aligned}
&+ \frac{\widetilde{L_1}^2}{\beta} 4eK^2 \left\{ \frac{\beta^2}{288eK^2\widetilde{L_1}^2} \sum_{r=-1}^{R-2} \left(\Sigma_r + \mathbb{E}\left[\left\| \nabla J\left(\theta_r\right) \right\|^2 \right] \right) + 4\eta^2\beta^2 eRG_0 \right\} \\
&\leq \left[1 - \frac{8\beta}{9} + \frac{\beta}{72} + \frac{32(\eta K\widetilde{L_1})^2}{\beta} (2(\eta KL)^2 + \frac{1}{K})(\lambda L)^2 \right] \sum_{r=-1}^{R-2} \Sigma_r \\
&+ \left[\frac{18(\lambda\widetilde{L_2})^2}{NK} + \frac{32(\eta K\widetilde{L_1})^2}{\beta} (2(\eta KL)^2 + \frac{1}{K})(\lambda L)^2 + \frac{\beta}{72} \right] \sum_{r=-1}^{R-2} \mathbb{E}\left[\left\| \nabla J\left(\theta_r\right) \right\|^2 \right] \\
&+ \left[8\beta\widetilde{L_1}^2 (\eta K)^2 (2(\eta KL)^2 + \frac{1}{K}) + \frac{3\beta^2}{NK} \right] R\sigma^2 + 16\beta(e\eta K\widetilde{L_1})^2 RG_0
\end{aligned} \tag{36}$$

Where the last inequality is derived by $\left\|\theta_r - \theta_{r-1}\right\|^2 \leq 2\lambda^2 \left(\left\| \nabla J\left(\theta_{r-1}\right) \right\|^2 + \left\| u_r - \nabla J\left(\theta_{r-1}\right) \right\|^2 \right)$. We require the following inequalities to hold,

$$\begin{cases}
\frac{32(\eta K\widetilde{L_1})^2}{\beta} (2(\eta KL)^2 + \frac{1}{K})(\lambda L)^2 \leq \frac{\beta}{18} \\
8\widetilde{L_1}^2 (\eta K)^2 (2(\eta KL)^2 + \frac{1}{K}) \leq \frac{\beta^2}{NK} \\
\lambda\widetilde{L_2} \leq \sqrt{\frac{\beta NK}{162}}.
\end{cases} \tag{37}$$

Then, we have that

$$\begin{aligned}
\sum_{r=0}^{R-1} \Sigma_r &\leq \left[1 - \frac{8\beta}{9} + \frac{\beta}{72} + \frac{\beta}{18} \right] \sum_{r=-1}^{R-2} \Sigma_r + \left[\frac{\beta}{9} + \frac{\beta}{18} + \frac{\beta}{72} \right] \sum_{r=-1}^{R-2} \mathbb{E}\left[\left\| \nabla J\left(\theta_r\right) \right\|^2 \right] \\
&+ \left[\frac{\beta^2}{NK} + \frac{3\beta^2}{NK} \right] R\sigma^2 + 16\beta(e\eta K\widetilde{L_1})^2 RG_0 \\
&\leq (1 - \frac{7\beta}{9}) \sum_{r=-1}^{R-2} \Sigma_r + \frac{2\beta}{9} \sum_{r=-1}^{R-2} \mathbb{E}\left[\left\| \nabla J\left(\theta_r\right) \right\|^2 \right] + \frac{4R\beta^2\sigma^2}{NK} + 16\beta(e\eta K\widetilde{L_1})^2 RG_0
\end{aligned}$$

After some rearrangement, we have

$$\sum_{r=0}^{R-1} \Sigma_r \leq \frac{9}{7\beta}\Sigma_{-1} + \frac{2}{7} \sum_{r=-1}^{R-2} \mathbb{E}\left[\left\| \nabla J\left(\theta_r\right) \right\|^2 \right] + \frac{36R\beta\sigma^2}{7NK} + \frac{144}{7}(e\eta K\widetilde{L_1})^2 RG_0$$

Based on Lemma B.1, we have

$$\frac{1}{\lambda}\mathbb{E}[J(\theta_R) - J(\theta_0)] \geq \frac{2}{7} \sum_{r=0}^{R-1} \mathbb{E}\left[\left\| \nabla J\left(\theta_r\right) \right\|^2 \right] - \frac{1}{35\beta}\Sigma_{-1} - \frac{39R\beta\sigma^2}{14NK} - \frac{78}{7}(e\eta K\widetilde{L_1})^2 RG_0$$

Notice that $u_0 = \frac{1}{NB}\sum_i \sum_{b=1}^{B} g_i\left(\tau_b^{(i)} | \theta_0\right)$ implies $\Sigma_{-1} = \mathbb{E}\|u_0 - \nabla J(\theta_0)\|^2 \leq \frac{\sigma^2}{NB} \leq \frac{\beta^2\sigma^2 R}{NK}$. Define $\bar{L} := \max\{L, \widetilde{L_1}, \widetilde{L_2}, \widetilde{L_3}\}$ and after some rearrangement, we have

$$\begin{aligned}
\frac{1}{R} \sum_{r=0}^{R-1} \mathbb{E}\left[\left\| \nabla J\left(\theta_r\right) \right\|^2 \right] &\lesssim \frac{\bar{L}\Delta}{\lambda\bar{L}R} + \frac{\Sigma_{-1}}{\beta R} + (\eta K\widetilde{L_1})^2 G_0 + \frac{\beta\sigma^2}{NK} \\
&\overset{(a)}{\lesssim} \frac{\bar{L}\Delta}{\lambda\bar{L}R} + \frac{\beta\sigma^2}{NK} \\
&\overset{(b)}{\lesssim} \frac{\bar{L}\Delta}{R} + \frac{\bar{L}\Delta}{\sqrt{\beta NK}} + \frac{\beta\sigma^2}{NK} \\
&\overset{(c)}{\lesssim} \frac{\bar{L}\Delta}{R} + \left(\frac{\bar{L}\Delta\sigma}{NKR} \right)^{2/3}
\end{aligned}$$

where (a) is due to the fact $\eta K\bar{L} \lesssim \left(\frac{\bar{L}\Delta}{G_0\lambda LR}\right)^{\frac{1}{2}}$; For (b), it holds because $\lambda\bar{L} \leq \min\{\frac{1}{24}, \sqrt{\frac{\beta NK}{162}}\}$; For (c), it holds because $\beta = \min\left\{1, \left(\frac{NK\bar{L}^2\Delta^2}{\sigma^4 R^2}\right)^{1/3}\right\}$. □

D. Federated Hessian Aided Policy Gradient with Momentum

According to the updating rule of FEDHAPG-M, we can rewrite $\Lambda_{r,k}^{(i)}$ as

$$\Lambda_{r,k}^{(i)} = \left(\nabla \log p^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right)^T v_{r,k}^{(i)}\right) \nabla\Phi_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right) + \nabla^2\Phi_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right) v_{r,k}^{(i)} \tag{38}$$

where $\Phi_i(\tau \mid \theta) = \sum_{h=0}^{H-1}\sum_{i=h}^{H-1} \lambda^i \mathcal{R}^{(i)}\left(s_i, a_i\right) \log \pi_\theta\left(a_h, s_h\right)$ and $v_{r,k}^{(i)} = \theta_{r,k}^{(i)} - \theta_{r-1}$. Note that

$$\mathbb{E}_{\alpha \sim U[0,1], \tau \sim p^{(i)}\left(\tau \mid \theta_{r,k}^{(i)}(\alpha)\right)}\left[\Lambda_{r,k}^{(i)}\right] = \nabla J\left(\theta_{r,k}^{(i)}\right) - \nabla J\left(\theta_{r-1}\right).$$

Moreover, we have $\Lambda_{r,k}^{(i)} := \hat{\nabla}_i^2\left(\theta_{r,k}^{(i)}(\alpha), \tau_{r,k}^{(i)}\right) v_{r,k}^{(i)}$ where

$$\begin{aligned}\hat{\nabla}_i^2\left(\theta_{r,k}^{(i)}(\alpha), \tau_{r,k}^{(i)}\right) =& \nabla\Phi_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right) \nabla \log p^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right)^T \\ &+ \nabla^2\Phi_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}(\alpha)\right).\end{aligned}$$

and $\mathbb{E}_{\tau \sim p^{(i)}\left(\tau \mid \theta_{r,k}^{(i)}(\alpha)\right)}\left[\hat{\nabla}^2\left(\theta_{r,k}^{(i)}(\alpha), \tau\right)\right] = \nabla^2 J_i\left(\theta_{r,k}^{(i)}(\alpha)\right)$.

Proposition 2. (Lemma 4.1 in (Shen et al., 2019)) Under Assumption 6.1, we have for all θ and $i \in [N]$

$$\left\|\hat{\nabla}_i^2(\theta, \tau)\right\|^2 \leq \frac{H^2 G^4 R_{\max}^2 + M^2 R^2}{(1-\lambda)^4} = \widetilde{L_4}^2.$$

where τ is a trajectory sampled according to $p^{(i)}(\tau|\theta)$.

Lemma D.1. *If the step-size satisfies* $\lambda \leq \sqrt{\frac{\beta NK}{72\widetilde{L_4}^2}}$, *we have*

$$\Sigma_r \leq (1 - \frac{8\beta}{\beta})\Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\mathcal{D}_r + \frac{2\beta^2\sigma^2}{NK} + \frac{8\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 \tag{39}$$

holds for $r \geq 1$. *When* $r = 0$, *we have*

$$\Sigma_r \leq (1 - \frac{8\beta}{\beta})\Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\mathcal{D}_r + \frac{2\beta^2\sigma^2}{NK}. \tag{40}$$

Proof.

$$\begin{aligned}\Sigma_r &= \mathbb{E}\left[\left\|u_{r+1} - \nabla J(\theta_r)\right\|^2\right] \\ &= \mathbb{E}\left[\left\|\frac{1}{NK}\sum_{i,k} \beta w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}, \theta_{r,k}^{(i)}(\alpha)\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) + (1-\beta)\left[u_r + \Lambda_{r,k}^{(i)}\right] - \nabla J(\theta_r)\right\|^2\right] \\ &= \mathbb{E}\Bigg[\Bigg\|(1-\beta)(u_r - \nabla J(\theta_{r-1})) \\ &+ \frac{1}{NK}\sum_{i,k}\left\{\beta w^{(i)}\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}, \theta_{r,k}^{(i)}(\alpha)\right) g_i\left(\tau_{r,k}^{(i)} \mid \theta_{r,k}^{(i)}\right) + (1-\beta)\left(\Lambda_{r,k}^{(i)} + \nabla J(\theta_{r-1}\right) - \nabla J(\theta_{r,k}^{(i)})\right\}\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{NK}\sum_{i,k}\Big[\nabla J(\theta^{(i)}_{r,k})-\nabla J(\theta_r)\Big]\Big\|^2\Bigg]\\
&=(1-\beta)^2\Sigma_{r-1}+\underbrace{\mathbb{E}\Bigg\|\frac{1}{NK}\sum_{i,k}\Big[\nabla J(\theta^{(i)}_{r,k})-\nabla J(\theta_r)\Big]\Bigg\|^2}_{H_1}\\
&+\underbrace{2\mathbb{E}\Bigg[\Big\langle(1-\beta)(u_r-\nabla J(\theta_{r-1})),\frac{1}{NK}\sum_{i,k}\Big[\nabla J_i(\theta^{(i)}_{r,k})-\nabla J(\theta_r)\Big]\Big\rangle\Bigg]}_{H_2}\\
&+\underbrace{\mathbb{E}\Bigg\|\frac{1}{NK}\sum_{i,k}\Big\{\beta w^{(i)}\Big(\tau^{(i)}_{r,k}\mid\theta^{(i)}_{r,k},\theta^{(i)}_{r,k}(\alpha)\Big)g_i\Big(\tau^{(i)}_{r,k}\mid\theta^{(i)}_{r,k}\Big)+(1-\beta)\Big(\Lambda^{(i)}_{r,k}+\nabla J(\theta_{r-1})\Big)-\nabla J(\theta^{(i)}_{r,k})\Big\}\Bigg\|^2}_{H_3}
\end{aligned}
\tag{41}
$$

To precede, we bound H_1 as

$$
\begin{aligned}
H_1&=\mathbb{E}\Bigg\|\frac{1}{NK}\sum_{i,k}\Big[\nabla J(\theta^{(i)}_{r,k})-\nabla J(\theta_r)\Big]\Bigg\|^2\\
&\le\frac{L^2}{NK}\sum_{i,k}\mathbb{E}\Big\|\theta^{(i)}_{r,k}-\theta_r\Big\|^2=L^2\mathcal{D}_r
\end{aligned}
\tag{42}
$$

Using Young's inequality to bound H_2, we have

$$
\begin{aligned}
H_2&\le\beta(1-\beta)^2\mathbb{E}\Big\|u_r-\nabla J(\theta_{r-1})\Big\|^2+\frac{1}{\beta}\mathbb{E}\Big\|\frac{1}{NK}\sum_{i,k}\nabla J_i(\theta^{(i)}_{r,k})-\nabla J(\theta_r)\Big\|^2\\
&\le\beta(1-\beta)^2\Sigma_{r-1}+\frac{L^2}{\beta}\underbrace{\frac{1}{NK}\sum_{i,k}\mathbb{E}\Big\|\theta^{(i)}_{r,k}-\theta_r\Big\|^2}_{\mathcal{D}_r}
\end{aligned}
\tag{43}
$$

For H_3, we bound it as

$$
\begin{aligned}
H_3&=\mathbb{E}\Bigg\|\frac{1}{NK}\sum_{i,k}\Big\{\beta w^{(i)}\Big(\tau^{(i)}_{r,k}\mid\theta^{(i)}_{r,k},\theta^{(i)}_{r,k}(\alpha)\Big)g_i\Big(\tau^{(i)}_{r,k}\mid\theta^{(i)}_{r,k}\Big)+(1-\beta)\Big(\Lambda^{(i)}_{r,k}+\nabla J(\theta_{r-1})\Big)-\nabla J(\theta^{(i)}_{r,k})\Big\}\Bigg\|^2\\
&\le2\beta^2\frac{\sigma^2}{NK}+2(1-\beta)^2\frac{1}{N^2K^2}\sum_{i,k}\mathbb{E}\Big\|\Lambda^{(i)}_{r,k}+\nabla J(\theta_{r-1})-\nabla J(\theta^{(i)}_{r,k})\Big\|^2\\
&\overset{(a)}{\le}\frac{2\beta^2\sigma^2}{NK}+2(1-\beta)^2\frac{1}{N^2K^2}\sum_{i,k}\mathbb{E}\Big\|\Lambda^{(i)}_{r,k}\Big\|^2\\
&\overset{(b)}{=}\frac{2\beta^2\sigma^2}{NK}+2(1-\beta)^2\frac{1}{N^2K^2}\sum_{i,k}\mathbb{E}\Big\|\hat{\nabla}^2\Big(\theta^{(i)}_{r,k},\tau^{(i)}_{r,k}\Big)v^{(i)}_{r,k}\Big\|^2\overset{(b)}{\le}\frac{2\beta^2\sigma^2}{NK}+2(1-\beta)^2\frac{1}{N^2K^2}\sum_{i,k}\widetilde{L_4}^2\mathbb{E}\Big\|v^{(i)}_{r,k}\Big\|^2\\
&\le\frac{2\beta^2\sigma^2}{NK}+4(1-\beta)^2\widetilde{L_4}^2\underbrace{\frac{1}{NK}\sum_{i,k}\mathbb{E}\Big\|\theta^{(i)}_{r,k}-\theta_r\Big\|^2}_{\mathcal{D}_r}+4(1-\beta)^2\frac{\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2
\end{aligned}
\tag{44}
$$

where we use the fact that $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \leq \mathbb{E}[\|X\|^2]$ for (a); for (b), it holds due to Proposition 2.

Plugging the upper bound of H_1 (Eq. (42)), H_2(Eq. (43)) and H_3 (Eq. (44))into Eq.(41), we have

$$\begin{aligned}
\Sigma_r &\leq (1-\beta)\Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\mathcal{D}_r + \frac{2\beta^2\sigma^2}{NK} + 4\frac{\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2 \\
&= (1-\beta)\Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\mathcal{D}_r + \frac{2\beta^2\sigma^2}{NK} + 4\frac{\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|u_r\right\|^2 \\
&\leq (1-\beta)\Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\mathcal{D}_r + \frac{2\beta^2\sigma^2}{NK} + 8\frac{\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|u_r - \nabla J(\theta_{r-1})\right\|^2 + 8\frac{\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 \\
&\overset{(a)}{\leq} (1 - \frac{8\beta}{\beta})\Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\mathcal{D}_r + \frac{2\beta^2\sigma^2}{NK} + \frac{8\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 \qquad (45)
\end{aligned}$$

where (a) is due to the choice of λ such that $\frac{8\lambda^2\widetilde{L_4}^2}{NK} \leq \frac{\beta}{9}$, which holds when $\lambda \leq \sqrt{\frac{\beta NK}{72\widetilde{L_4}^2}}$. □

Lemma D.2. *(Bounding drift-term) If the local step-size satisfies $\eta \leq \min\{\frac{L}{32e^2\widetilde{L_4}^2K}, \frac{1}{KL}\}$, the drift-term can be upper bounded as:*

$$\mathcal{D}_r \leq 4eK^2\mathcal{M}_r + (16\eta^4K^4L^2 + 8\eta^2K)\left(\beta^2\sigma^2 + 2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right)$$

Proof. Define $c_{r,k}^{(i)} := -\eta\left(\nabla J_i(\theta_{r,k}^{(i)}) + (1-\beta)(u_r - \nabla J_i(\theta_{r-1}))\right)$. For any $1 \leq j \leq k-1 \leq K-2$, we have:

$$\begin{aligned}
\mathbb{E}\left\|c_{r,j}^{(i)} - c_{r,j-1}^{(i)}\right\|^2 &\leq \eta^2L^2\mathbb{E}\left\|\theta_{r,j}^{(i)} - \theta_{r,j-1}^{(i)}\right\|^2 \\
&= \eta^2L^2\left(\mathbb{E}\left\|c_{r,j-1}^{(i)}\right\|^2 + \mathbb{E}\left[\operatorname{Var}\left[\theta_{r,j}^{(i)} - \theta_{r,j-1}^{(i)} \mid \mathcal{F}_{r,j-1}^{(i)}\right]\right]\right). \qquad (46)
\end{aligned}$$

where we use the bias-variance decomposition in the last inequality. To precede, we bound the variance term as:

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{Var}\left[\theta_{r,j}^{(i)} - \theta_{r,j-1}^{(i)} \mid \mathcal{F}_{r,j-1}^{(i)}\right]\right] \\
&= \eta^2\mathbb{E}\Bigg\|\beta\left[w^{(i)}\left(\tau_{r,j-1}^{(i)} \mid \theta_{r,j-1}^{(i)}, \theta_{r,j-1}^{(i)}(\alpha)\right)g_i\left(\tau_{r,j-1}^{(i)} \mid \theta_{r,j-1}^{(i)}\right) - \nabla J_i(\theta_{r,j-1}^{(i)})\right] \\
&+ (1-\beta)\left[\Lambda_{r,j-1}^{(i)} - \nabla J_i(\theta_{r,j-1}^{(i)}) + \nabla J_i(\theta_{r-1})\right]\Bigg\|^2 \\
&\leq 2\eta^2\beta^2\sigma^2 + 2\eta^2(1-\beta)^2\mathbb{E}\left\|\Lambda_{r,j-1}^{(i)} - \nabla J_i(\theta_{r,j-1}^{(i)}) + \nabla J_i(\theta_{r-1})\right\|^2 \\
&\overset{(a)}{\leq} 2\eta^2\beta^2\sigma^2 + 2\eta^2(1-\beta)^2\mathbb{E}\left\|\Lambda_{r,j-1}^{(i)}\right\|^2 \\
&\leq 2\eta^2\beta^2\sigma^2 + 2\eta^2(1-\beta)^2\mathbb{E}\left\|\hat{\nabla}_i^2\left(\theta_{r,j-1}^{(i)}, \tau_{r,j-1}^{(i)}\right)v_{r,j-1}^{(i)}\right\|^2 \\
&\leq 2\eta^2\beta^2\sigma^2 + 4\eta^2(1-\beta)^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,j-1}^{(i)} - \theta_r\right\|^2 + 4\eta^2(1-\beta)^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2 \qquad (47)
\end{aligned}$$

where we use the fact that $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \leq \mathbb{E}[\|X\|^2]$ for (a). Plugging the upper bound of variance into Eq.(46), we have

$$\mathbb{E}\left\|c_{r,j}^{(i)} - c_{r,j-1}^{(i)}\right\|^2 \leq \eta^2L^2\left(\mathbb{E}\left\|c_{r,j-1}^{(i)}\right\|^2 + 2\eta^2\beta^2\sigma^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,j-1}^{(i)} - \theta_r\right\|^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right).$$

Then for any $1 \leq j \leq k-1 \leq K-2$, we have

$$\begin{aligned}
\mathbb{E}\left\|c_{r,j}^{(i)}\right\|^2 &\leq (1+\frac{1}{q})\mathbb{E}\left\|c_{r,j-1}^{(i)}\right\|^2 + (1+q)\mathbb{E}\left\|c_{r,j}^{(i)} - c_{r,j-1}^{(i)}\right\|^2 \\
&\leq (1+\frac{2}{q})\mathbb{E}\left\|c_{r,j-1}^{(i)}\right\|^2 + (1+q)\eta^2 L^2 \left(2\eta^2\beta^2\sigma^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,j-1}^{(i)} - \theta_r\right\|^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right) \qquad (48)
\end{aligned}$$

where we use the fact that $\eta L \leq \frac{1}{K} \leq \frac{1}{q+1}$ and let $q = k-1$. By unrolling this recurrence, for any $1 \leq j \leq k-1 \leq K-2$, we have

$$\begin{aligned}
\mathbb{E}\left\|c_{r,j}^{(i)}\right\|^2 \leq{}& (1+\frac{2}{k-1})^j\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + k\eta^2L^2\sum_{i=0}^{j-1}(2\eta^2\beta^2\sigma^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2)\Pi_{j'=i+1}^{j-1}(1+\frac{2}{k-1}) \\
&+ k\eta^2L^2\sum_{s=0}^{j-1}(4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,s}^{(i)} - \theta_r\right\|^2)\Pi_{j'=s+1}^{j-1}(1+\frac{2}{k-1}) \\
\leq{}& (1+\frac{2}{k-1})^{k-1}\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + k\eta^2L^2\sum_{i=0}^{k-1}(2\eta^2\beta^2\sigma^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2)(1+\frac{2}{k-1})^{k-1} \\
&+ k\eta^2L^2\sum_{j'=0}^{j-1}(4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,j'}^{(i)} - \theta_r\right\|^2)(1+\frac{2}{k-1})^{k-1} \qquad (49)
\end{aligned}$$

Based on the inequality $(1+\frac{2}{K-1}^{k-1}) \leq e^2 \leq 8$, we have

$$\mathbb{E}\left\|c_{r,j}^{(i)}\right\|^2 \leq e^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + 8k^2\eta^4L^2\left(2\beta^2\sigma^2 + 4\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right) + 4e^2k\eta^4L^2\widetilde{L_4}^2\sum_{j'=0}^{j-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)} - \theta_r\right\|^2 \qquad (50)$$

By Lemma A.3, we have

$$\begin{aligned}
\mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2 &\leq 2\mathbb{E}\left\|\sum_{j=0}^{k-1}c_{r,j}^{(i)}\right\|^2 + 2\sum_{j=0}^{k-1}\mathbb{E}\left[\mathrm{Var}\left[\theta_{r,j+1}^{(i)} - \theta_{r,j}^{(i)} \mid \mathcal{F}_{r,j}^{(i)}\right]\right] \\
&\overset{(a)}{\leq} 2k\sum_{j=0}^{k-1}\mathbb{E}\left\|c_{r,j}^{(i)}\right\|^2 + 2\sum_{j=0}^{k-1}\left(2\beta^2\eta^2\sigma^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,j}^{(i)} - \theta_r\right\|^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right) \qquad (51)
\end{aligned}$$

where (a) is due to Eq.(47). Plugging Eq.(50) into Eq.(51), we have

$$\begin{aligned}
&\mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2 \leq \\
&2k\sum_{j=0}^{k-1}\left\{e^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + 8k^2\eta^4L^2\left(2\beta^2\sigma^2 + 4\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right) + 4e^2k\eta^4L^2\widetilde{L_4}^2\sum_{j'=0}^{j-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)} - \theta_r\right\|^2\right\} \\
&+ 2\sum_{j=0}^{k-1}\left(2\beta^2\eta^2\sigma^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r,j}^{(i)} - \theta_r\right\|^2 + 4\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right) \qquad (52)
\end{aligned}$$

Summing up the above equation over $k = 0, \cdots, K-1$, we have

$$\begin{aligned}
\sum_{k=0}^{K-1}\mathbb{E}\left\|\theta_{r,k}^{(i)} - \theta_r\right\|^2 \leq{}& \sum_{k=0}^{K-1}\left\{2k^2e^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2 + 16k^4\eta^4L^2\left(2\beta^2\sigma^2 + 4\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right)\right\} \\
&+ \sum_{k=0}^{K-1}8e^2k^2\eta^4L^2\widetilde{L_4}^2\sum_{j=0}^{k-1}\sum_{j'=0}^{j-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)} - \theta_r\right\|^2
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{k=0}^{K-1}\left(4k\beta^2\eta^2\sigma^2+8k\eta^2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2+8\eta^2\widetilde{L_4}^2\sum_{j=0}^{k-1}\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2\right)\\
&\le 2eK^3\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2+(8\eta^4K^5L^2+4\eta^2K^2)\left(\beta^2\sigma^2+2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)\\
&+K^2\sum_{k=0}^{K-1}8e^2\eta^4L^2\widetilde{L_4}^2\sum_{j=0}^{K-1}\sum_{j'=0}^{K-1}\mathbb{E}\left\|\theta_{r,j'}^{(i)}-\theta_r\right\|^2+\sum_{k=0}^{K-1}8\eta^2\widetilde{L_4}^2\sum_{j=0}^{K-1}\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2\\
&=2eK^3\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2+(8\eta^4K^5L^2+4\eta^2K^2)\left(\beta^2\sigma^2+2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)\\
&+(8e^2\eta^4K^4L^2\widetilde{L_4}^2+8\eta^2\widetilde{L_4}^2K)\sum_{j=0}^{K-1}\mathbb{E}\left\|\theta_{r,j}^{(i)}-\theta_r\right\|^2
\end{aligned}
\tag{53}
$$

With the choice of step-size η satisfying $8e^2\eta^4K^4L^2\widetilde{L_4}^2+8\eta^2\widetilde{L_4}^2K\le\frac{1}{2}$, after some rearrangement, we have

$$
\frac{1}{2K}\sum_{k=0}^{K-1}\mathbb{E}\left\|\theta_{r,k}^{(i)}-\theta_r\right\|^2\le 2eK^2\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2+(8\eta^4K^4L^2+4\eta^2K)\left(\beta^2\sigma^2+2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)
\tag{54}
$$

In summary, we can bound the drift-term as

$$
\mathcal{D}_r\le 4eK^2\mathcal{M}_r+(16\eta^4K^4L^2+8\eta^2K)\left(\beta^2\sigma^2+2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1}-\theta_r\right\|^2\right)
\tag{55}
$$

□

Lemma D.3. *If $\lambda L\le\frac{1}{24}$ and $\eta^2\left[\frac{289}{72}(1-\beta)^2+8e(\lambda\beta LR)^2\right]\le\frac{\beta^2}{288eK^2\left(2L^2+\widetilde{4L_4}^2\right)}$, we have*

$$
\sum_{r=0}^{R-1}\mathcal{M}_r=\frac{1}{N}\sum_{r=0}^{R-1}\sum_{i=1}^{N}\mathbb{E}\left\|c_{r,0}^{(i)}\right\|^2\le\frac{\beta^2}{288eK^2\left(2L^2+\widetilde{4L_4}^2\right)}\sum_{r=-1}^{R-2}\left(\Sigma_r+\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]\right)+4\eta^2\beta^2eRG_0.
\tag{56}
$$

where $G_0:=\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\left\|\nabla J_i\left(\theta_0\right)\right\|^2\right]$.

Proof. The proof is the same as that of Lemma C.3. □

D.1. Proof of Theorem 6.5

Theorem D.4. *(Complete version of Theorem 6.5) Under Assumption 6.1–6.3, by setting $u_0=\frac{1}{NB}\sum_{i=1}^{N}\sum_{b=1}^{B}g_i\left(\tau_b^{(i)}|\theta_0\right)$ with $\left\{\tau_b^{(i)}\right\}_{b=1}^{B}\overset{iid}{\sim}p^{(i)}(\tau|\theta_0)$ and choosing $\beta=\min\left\{1,\left(\frac{NK\hat{L}^2\Delta^2}{\sigma^4R^2}\right)^{1/3}\right\}$, $\lambda=\min\left\{\frac{1}{24\hat{L}},\sqrt{\frac{\beta NK}{72\hat{L}^2}}\right\}$, $B=\left\lceil\frac{K}{R\beta^2}\right\rceil$, and*

$$
\eta K\hat{L}\lesssim\min\left\{\left(\frac{\hat{L}\Delta}{G_0\lambda\hat{L}R}\right)^{1/2},\left(\frac{\beta}{N}\right)^{1/2},\left(\frac{\beta}{NK}\right)^{1/4}\right\}
$$

in Algorithm 2, then the output of FEDHAPG-M *after R rounds satisfies*

$$
\frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]\lesssim\left(\frac{\hat{L}\Delta\sigma}{NKR}\right)^{2/3}+\frac{\hat{L}\Delta}{R}
\tag{57}
$$

where $\hat{L} := \sqrt{2L^2 + \widetilde{4L_4}^2}$ *and* $L, \widetilde{L_4}$ *are defined in Proposition 1 and Proposition 2, respectively.*

Proof. Based on Lemma D.1, we have for any $r \geq 1$

$$\begin{aligned}
\Sigma_r &\leq (1 - \frac{8\beta}{\beta})\Sigma_{r-1} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\mathcal{D}_r + \frac{2\beta^2\sigma^2}{NK} + \frac{8\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 \\
&\leq (1 - \frac{8\beta}{\beta})\Sigma_{r-1} + \frac{2\beta^2\sigma^2}{NK} + \frac{8\lambda^2\widetilde{L_4}^2}{NK}\mathbb{E}\left\|\nabla J(\theta_{r-1})\right\|^2 \qquad (58) \\
&+ \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\left[4eK^2\mathcal{M}_r + (16\eta^4K^4L^2 + 8\eta^2K)\left(\beta^2\sigma^2 + 2\widetilde{L_4}^2\mathbb{E}\left\|\theta_{r-1} - \theta_r\right\|^2\right)\right] \qquad (59)
\end{aligned}$$

where the last inequality is due to Lemma D.2. When $r = 0$, we have

$$\Sigma_0 \leq (1-\beta)\Sigma_{-1} + \frac{2\beta^2\sigma^2}{NK} + \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\left[4eK^2\mathcal{M}_0 + (16\eta^4K^4L^2 + 8\eta^2K)\right]\beta^2\sigma^2$$

Summing up the above equation over r from 0 to $R-1$, we have

$$\begin{aligned}
\sum_{r=0}^{R-1}\Sigma_r \leq &\left(1 - \frac{8\beta}{9}\right)\sum_{r=-1}^{R-2}\Sigma_r + \frac{8(\lambda\widetilde{L_4})^2}{NK}\mathbb{E}\left[\sum_{r=0}^{R-2}\left\|\nabla J\left(\theta_r\right)\right\|^2\right] + \frac{2\beta^2\sigma^2}{NK}R \\
&+ \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}\left[4eK^2\sum_{r=0}^{R-1}\mathcal{M}_r + 8(\eta K)^2(2(\eta KL)^2 + \frac{1}{K})\right]\left(R\beta^2\sigma^2 + 2L^2\sum_{r=0}^{R-1}\mathbb{E}\left\|\theta_r - \theta_{r-1}\right\|^2\right)
\end{aligned}$$

By incorporating Lemma D.3 into the inequality above, we have

$$\begin{aligned}
\sum_{r=0}^{R-1}\Sigma_r \leq &\left(1 - \frac{8\beta}{9}\right)\sum_{r=-1}^{R-2}\Sigma_r + \frac{8(\lambda\widetilde{L_4})^2}{NK}\mathbb{E}\left[\sum_{r=0}^{R-2}\left\|\nabla J\left(\theta_r\right)\right\|^2\right] + \frac{2\beta^2\sigma^2}{NK}R \\
&+ \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}8(\eta K)^2\left(2(\eta KL)^2 + \frac{1}{K}\right)\left(R\beta^2\sigma^2 + 2L^2\sum_{r=0}^{R-1}\mathbb{E}\left[\left\|\theta_r - \theta_{r-1}\right\|^2\right]\right) \\
&+ \frac{2L^2 + 4\widetilde{L_4}^2}{\beta}4eK^2\left\{\frac{\beta^2}{288eK^2\left(2L^2 + \widetilde{4L_4}^2\right)}\sum_{r=-1}^{R-2}\left(\Sigma_r + \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right]\right) + 4\eta^2\beta^2eRG_0\right\} \\
\leq &\left[1 - \frac{8\beta}{9} + \frac{\beta}{72} + \frac{32(\eta K)^2(2L^2 + \widetilde{4L_4})^2}{\beta}(2(\eta KL)^2 + \frac{1}{K})(\lambda L)^2\right]\sum_{r=-1}^{R-2}\Sigma_r \\
&+ \left[\frac{8(\lambda\widetilde{L_4})^2}{NK} + \frac{32(\eta K)^2(2L^2 + \widetilde{4L_4})^2}{\beta}(2(\eta KL)^2 + \frac{1}{K})(\lambda L)^2 + \frac{\beta}{72}\right]\sum_{r=-1}^{R-2}\mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right] \\
&+ \left[8\beta\left(2L^2 + \widetilde{4L_4}^2\right)(\eta K)^2(2(\eta KL)^2 + \frac{1}{K}) + \frac{2\beta^2}{NK}\right]R\sigma^2 + 16\beta\left(2L^2 + \widetilde{4L_4}^2\right)(e\eta K)^2RG_0 \qquad (60)
\end{aligned}$$

Where the last inequality is derived by $\left\|\theta_r - \theta_{r-1}\right\|^2 \leq 2\lambda^2\left(\left\|\nabla J\left(\theta_{r-1}\right)\right\|^2 + \left\|u_r - \nabla J\left(\theta_{r-1}\right)\right\|^2\right)$. Note that $\hat{L}^2 = 2L^2 + \widetilde{4L_4}^2$. We require the following inequalities to hold,

$$\begin{cases}
\frac{32(\eta K\hat{L})^2}{\beta}(2(\eta KL)^2 + \frac{1}{K})(\lambda L)^2 \leq \frac{\beta}{18} \\
8\hat{L}^2(\eta K)^2(2(\eta KL)^2 + \frac{1}{K}) \leq \frac{\beta^2}{NK} \\
\lambda\hat{L} \leq \sqrt{\frac{\beta NK}{72}}.
\end{cases} \qquad (61)$$

Then, we have that

$$\begin{aligned}\sum_{r=0}^{R-1} \Sigma_r \leq& \left[1 - \frac{8\beta}{9} + \frac{\beta}{72} + \frac{\beta}{18}\right] \sum_{r=-1}^{R-2} \Sigma_r + \left[\frac{\beta}{9} + \frac{\beta}{18} + \frac{\beta}{72}\right] \sum_{r=-1}^{R-2} \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right] \\ &+ \left[\frac{\beta^2}{NK} + \frac{2\beta^2}{NK}\right] R\sigma^2 + 16\beta(e\eta K\widetilde{L_1})^2 RG_0 \\ &\leq (1 - \frac{7\beta}{9}) \sum_{r=-1}^{R-2} \Sigma_r + \frac{2\beta}{9} \sum_{r=-1}^{R-2} \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right] + \frac{4R\beta^2\sigma^2}{NK} + 16\beta(e\eta K\hat{L})^2 RG_0\end{aligned}$$

After some rearrangement, we have

$$\sum_{r=0}^{R-1} \Sigma_r \leq \frac{9}{7\beta}\Sigma_{-1} + \frac{2}{7} \sum_{r=-1}^{R-2} \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right] + \frac{36R\beta\sigma^2}{7NK} + \frac{144}{7}(e\eta K\hat{L})^2 RG_0$$

Based on Lemma B.1, we have

$$\frac{1}{\lambda}\mathbb{E}[J(\theta_R) - J(\theta_0)] \geq \frac{2}{7} \sum_{r=0}^{R-1} \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right] - \frac{1}{35\beta}\Sigma_{-1} - \frac{39R\beta\sigma^2}{14NK} - \frac{78}{7}(e\eta K\hat{L})^2 RG_0$$

Notice that $u_0 = \frac{1}{NB}\sum_i \sum_{b=1}^{B} g_i\left(\tau_b^{(i)}|\theta_0\right)$ implies $\Sigma_{-1} = \mathbb{E}\|u_0 - \nabla J(\theta_0)\|^2 \leq \frac{\sigma^2}{NB} \leq \frac{\beta^2\sigma^2 R}{NK}$. After some rearrangement, we have

$$\begin{aligned}\frac{1}{R}\sum_{r=0}^{R-1} \mathbb{E}\left[\left\|\nabla J\left(\theta_r\right)\right\|^2\right] &\lesssim \frac{\hat{L}\Delta}{\lambda\hat{L}R} + \frac{\Sigma_{-1}}{\beta R} + (\eta K\hat{L})^2 G_0 + \frac{\beta\sigma^2}{NK} \\ &\overset{(a)}{\lesssim} \frac{\hat{L}\Delta}{\lambda\hat{L}R} + \frac{\beta\sigma^2}{NK} \\ &\overset{(b)}{\lesssim} \frac{\hat{L}\Delta}{R} + \frac{\hat{L}\Delta}{\sqrt{\beta NK}} + \frac{\beta\sigma^2}{NK} \\ &\overset{(c)}{\lesssim} \frac{\hat{L}\Delta}{R} + \left(\frac{\hat{L}\Delta\sigma}{NKR}\right)^{2/3}\end{aligned}$$

where (a) is due to the fact $\eta K\hat{L} \lesssim \left(\frac{\hat{L}\Delta}{G_0\lambda\hat{L}R}\right)^{\frac{1}{2}}$; For (b), it holds because $\lambda\hat{L} \leq \min\{\frac{1}{24}, \sqrt{\frac{\beta NK}{72}}\}$; For (c), it holds because $\beta = \min\left\{1, \left(\frac{NK\hat{L}^2\Delta^2}{\sigma^4 R^2}\right)^{1/3}\right\}$. □

E. Additional Experiments and Implementation Details

E.1. Details of Tabular Case.

Random MDPs consist of $N = 20$ environments. In each MDP, both the state and action spaces have a size of 5. We choose $R_{\max} = 1$. The discounted factor λ is 0.9. The state transition kernel is generated randomly (element-wisely Bernoulli distributed). The number of local updates is set as $K = 32$. Additionally, the local step-size is chosen to be $\eta = 0.05$.

E.2. Details of DRL Case

Experiments Setup We adopted a local step-size of 0.75 and a global step-size of 0.6. We experimented with momentum coefficients, denoted as β, ranging from 0.2, 0.5, to 0.8. Additional parameters were set as follows: $N = 5$, $R_{\max} = 120$, and $K = 10$. All experiments are conducted in a host machine that is equipped with an Intel(R) Core(TM) i9-10900X CPU that operates at a base frequency of 3.70GHz. This processor boasts 10 cores and 20 threads, with a maximum turbo frequency of 4300 MHz. It has a total of 125GB of RAMA and 4 NVIDIA GeForce RTX 2080 GPU, compatible with CUDA Version 11.0. The source code is provided in the supplementary materials.

Experimental Environments The **CartPole** environment, often referred to as the ”inverted pendulum” problem, is a classic task in the field of reinforcement learning. In this environment, a pole is attached to a cart, which moves along a frictionless track. The primary objective is to balance the pole upright by moving the cart left or right, without the pole falling over or the cart moving too far off the track. At the start of the experiment, the pole is slightly tilted, and the goal is to prevent it from falling over by applying force to the cart. The environment provides a reward at each time step for keeping the pole upright. The episode ends when the pole tilts beyond a certain angle from the vertical or the cart moves out of a defined boundary on the track.

The **HalfCheetah** environment is another popular benchmark in reinforcement learning, especially within the continuous control domain. It’s designed to emulate the challenges of agile and efficient locomotion. The agent in this environment is a two-dimensional, simplified robotic model inspired by the anatomy of a cheetah, albeit it only represents the ”half” body, often from the waist down, thus the name ”HalfCheetah.” The robotic agent comprises multiple joints and segments, representing the limbs of the cheetah. The primary goal in the HalfCheetah environment is to control and coordinate the movements of these joints to make the robot run as fast as possible on a flat surface. At each timestep, the agent receives a reward based on how fast it’s moving forward minus a small cost for the actions taken (to prevent erratic behaviors). The challenge lies in efficiently propelling the HalfCheetah forward, optimizing for speed and stability.

The **Walker** environment is a more complex task that simulates a bipedal agent which needs to learn to walk. Unlike CartPole, where the challenge is to balance a single pole, the Walker environment involves controlling multiple joints and limbs of a simulated agent to achieve locomotion. The agent receives rewards based on its forward movement and is penalized for falling or performing awkward movements. More information about these environments can be found in Todorov et al. (2012).

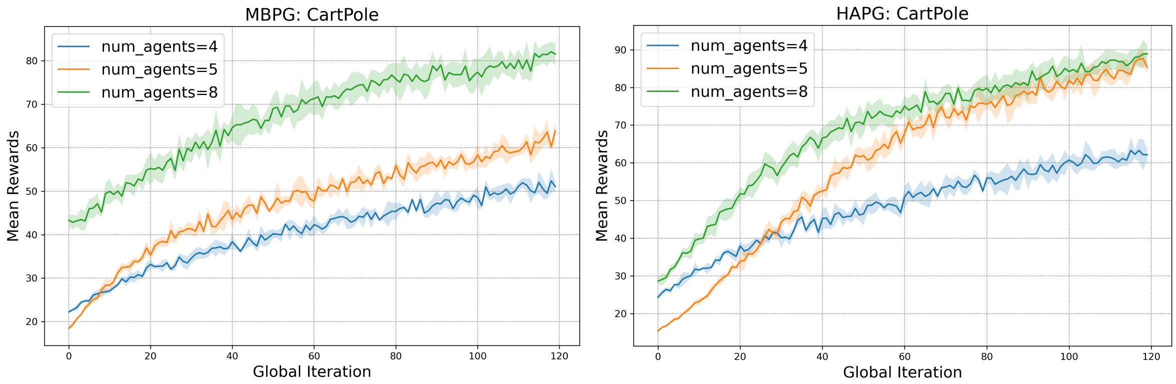

Figure 2. Mean rewards over global iterations for the CartPole task under different values of N (agent number): (**Left**): FedSVRPG-M; (**Right**): FedHAPG-M. The shaded areas represent the variance of rewards. Complying with theory, increasing N will increase the rewards. For both algorithms, the local step-size η is 0.05, global step-size λ satisfies $\lambda = \eta K$ and the number of local updates K is 10.

Ablation Study on Agent Number N. We further provide the ablation study of our FEDSVRPG-M and FEDHAPG-M algorithms on N (agent number). With large N, environment heterogeneity level increases. We choose $\beta = 0.2$ to train policies in the ablation study. Figure 2 illustrates how different N values ($N = 4, 5$, and 8) influence the average rewards in the CartPole task as the number of iterations increases. We find that all policies with larger N values report better performance throughout the iterations. The color-shaded regions indicate the variance in rewards. Such phenomenon observed in Figure 2 complies with our theoretical analysis about linear speedup.

Experiments on FEDHAPG-M Algorihtm The table 3 presents the mean testing rewards and variances for the policies trained by the FedHAPG-M algorithm with various β values and the baseline algorithm (Jin et al., 2022) across two tasks: CartPole and Walker. For both tasks, the FedHAPG-M algorithm with $\beta = 0.8$ outperforms the other configurations in terms of mean rewards.

Table 3. Mean Rewards and Variances of Policy Trained by FEDHAPG-M with Different Beta Values and Baseline Algorithm

Algorithms	CartPole	Walker
FEDHAPG-M with $\beta = 0.2$	83.46 ± 7.92	130.93 ± 7.72
FEDHAPG-M with $\beta = 0.5$	86.54 ± 12.99	287.14 ± 72.26
FEDHAPG-M with $\beta = 0.8$	**86.58** ± 11.21	**301.57** ± 28.04
Baseline algorithm	85.92 ± 12.17	299.69 ± 3.02