Journal of Machine Learning Research 25 (2024) 1-49 Submitted 9/22; Published 5/24

Decentralized Natural Policy Gradient with Variance Reduction for Collaborative Multi-Agent Reinforcement Learning

Jinchi Chen[*] JCCHEN@ECUST.EDU.CN
School of Data Science
Fudan University
Shanghai, China

School of Mathematics
East China University of Science and Technology
Shanghai, China

Jie Feng[*] 19110980001@FUDAN.EDU.CN
School of Data Science
Fudan University
Shanghai, China

Weiguo Gao WGGAO@FUDAN.EDU.CN
School of Mathematical Sciences and School of Data Science
Fudan University
Shanghai, China

Ke Wei KEWEI@FUDAN.EDU.CN
School of Data Science
Fudan University
Shanghai, China

Editor: Csaba Szepesvari

Abstract

This paper studies a policy optimization problem arising from collaborative multi-agent reinforcement learning in a decentralized setting where agents communicate with their neighbors over an undirected graph to maximize the sum of their cumulative rewards. A novel decentralized natural policy gradient method, dubbed Momentum-based Decentralized Natural Policy Gradient (MDNPG), is proposed, which incorporates natural gradient, momentum-based variance reduction, and gradient tracking into the decentralized stochastic gradient ascent framework. The $\mathcal{O}(n^{-1}\epsilon^{-3})$ sample complexity for MDNPG to converge to an ϵ-stationary point has been established under standard assumptions, where n is the number of agents. It indicates that MDNPG can achieve the optimal convergence rate for decentralized policy gradient methods and possesses a linear speedup in contrast to centralized optimization methods. Moreover, superior empirical performance of MDNPG over other state-of-the-art algorithms has been demonstrated by extensive numerical experiments.

[*] Jinchi Chen and Jie Feng contribute equally to this work.

Keywords: multi-agent reinforcement learning, natural policy gradient, decentralized optimization, variance reduction

1. Introduction

Reinforcement learning (RL) is a sequential decision-making task in which an agent seeks a strategy that maximizes the long-term return received from the environment via interaction with the system. Recent years have witnessed considerable theoretical and empirical advances in RL, see for example Kaelbling et al. (1996); Arulkumaran et al. (2017); Rajeswaran et al. (2020); Agarwal et al. (2021) and references therein. In particular, when combined with deep learning, RL has achieved the most recent state of the art in various data-driven applications, including robotics (Kober et al., 2013), finance (Liu et al., 2020a) and game playing (Mnih et al., 2013).

Markov decision processes (MDPs) are widely used to model how agents interact with an environment. An MDP can be defined as a tuple $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition model which determines the probability from $(\boldsymbol{s}, \boldsymbol{a})$ to state $\boldsymbol{s}'$, r: $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [-1, 1]$ is the immediate reward function associated with the transition from $(\boldsymbol{s}, \boldsymbol{a})$ to $\boldsymbol{s}'$, and $\gamma \in [0, 1)$ is the discount factor. Moreover, a policy, denoted $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, specifies a decision-making strategy, that is, $\pi(\boldsymbol{a}|\boldsymbol{s})$ is the probability of executing action $\boldsymbol{a}$ at state $\boldsymbol{s}$. Given an initial state distribution $\rho(\boldsymbol{s}_0)$, let $\tau = (\boldsymbol{s}^0, \boldsymbol{a}^0, r^0, \boldsymbol{s}^1, \boldsymbol{a}^1, r^1, \cdots, \boldsymbol{s}^{H-1}, \boldsymbol{a}^{H-1}, r^{H-1}, \boldsymbol{s}^H)$ be a trajectory of time horizon H induced by a policy π, where $r^h = r(\boldsymbol{s}^h, \boldsymbol{a}^h, \boldsymbol{s}^{h+1})$. The overall goal of RL is to find a policy that maximizes the expected discounted cumulative rewards, which can be formulated as the following optimization problem:

$$\max_{\pi \in \Pi} \left\{ V(\pi) := \mathbb{E}_{\tau \sim p(\cdot|\pi)} \left\{ R(\tau) \right\} \right\}, \tag{1}$$

where $R(\tau) = \sum_{h=0}^{H-1} \gamma^h r^h$ is the discounted return obtained from the trajectory τ, $p(\cdot|\pi)$ is the distribution of the trajectories, and Π represents the policy space.

There are several classical categories of RL algorithms. Model-based approaches, such as policy iteration and value iteration (see e.g., Puterman, 2014) find the optimal policy based on the ideas of of fixed point iteration. Whereas in model-free settings, value-based methods, like temporal difference learning and Q-learning (see e.g., Sutton and Barto, 2018; Bertsekas, 2019) solely use reward obtained from the environment to seek the optimal strategy. These methods can be roughly thought of as approximate dynamic programming with Monte Carlo learning. In contrast, policy gradient methods (Williams, 1992; Sutton et al., 1999; Konda and Tsitsiklis, 1999) maximize the objective function in (1) by gradient ascent with a differentiable parameterized policy in the model-free manner. Gradient-based approaches have a few advantages. For example, they can generate stochastic policies, which are more exploratory and are easily extended to continuous control problems. Coupled with neural networks, they have gained tremendous success in many applications due to their flexibility and adaptability. Moreover, the theoretical guarantees for gradient-based methods are relatively more complete, even in conjunction with simple function approximations (Sutton et al., 1999; Xu et al., 2020c; Agarwal et al., 2021).

In this paper, we restrict our attention to policy optimization based methods. Using a parameterized policy $\pi_{\boldsymbol{\theta}}$ where $\boldsymbol{\theta} \in \mathbb{R}^d$, (1) can be expressed as a finite dimensional optimization problem:

$$\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \left\{ V(\boldsymbol{\theta}) := \mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta})} \left\{ R(\tau) \right\} \right\}. \tag{2}$$

After parameterization, the distribution of the trajectories, denoted $p(\tau|\boldsymbol{\theta})$, is given by

$$p(\tau|\boldsymbol{\theta}) := \rho(\boldsymbol{s}^0) \prod_{h=0}^{H-1} \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h) P(\boldsymbol{s}^{h+1}|\boldsymbol{s}^h, \boldsymbol{a}^h), \tag{3}$$

where we recall that $\rho(\boldsymbol{s}^0)$ is the initial state distribution.

A direct method for solving problem (2) is policy gradient (PG). Despite its simplicity, PG is not invariant to reparameterization. As an alternative, natural policy gradient (NPG) methods (Kakade, 2001; Bagnell and Schneider, 2003; Peters and Schaal, 2008; Bhatnagar et al., 2007) utilize the intrinsic distance between policies, i.e., the Kullback-Leibler (KL) divergence, to modify the search direction so that parameterization invariance can be preserved. As two variants of NPG methods, trust region policy optimization (TRPO), see Schulman et al. (2015), combines NPG with a line search procedure to guarantee improvement, whereas proximal policy optimization (PPO), see Schulman et al. (2017), uses a simplified objective with a penalty term or a clipped ratio rather than the KL constraint.

1.1 Collaborative Multi-Agent Reinforcement Learning

More recently, there has been a growing interest in multi-agent reinforcement learning (MARL) which allows agents to address problems simultaneously in more complicated settings, such as fully cooperative, fully competitive, and mixed of the two (Busoniu et al., 2008; Nowé et al., 2012; Zhang et al., 2021b). MARL arises in many applications, including autonomous driving (Shalev-Shwartz et al., 2016), game playing (Vinyals et al., 2019), and wireless networks (Yao and Jia, 2019). In this paper, we study an n-agent fully cooperative setting in which the goal of agents is to cooperatively maximize the global value function defined as follows:

$$\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \left\{ V(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} V_i(\boldsymbol{\theta}) \right\}, \tag{4}$$

where $\boldsymbol{\theta} \in \mathbb{R}^d$ is the parameter of policy and $V_i(\boldsymbol{\theta})$ is the value function of the i-th agent. Let $\tau_i = (\boldsymbol{s}^0, \boldsymbol{a}^0, r_i^0, \boldsymbol{s}^1, \boldsymbol{a}^1, r_i^1, \cdots, \boldsymbol{s}^{H-1}, \boldsymbol{a}^{H-1}, r_i^{H-1}, \boldsymbol{s}^H)$ be the trajectory induced by the policy $\pi_{\boldsymbol{\theta}}$ for the i-th agent[1]. The discounted return $R(\tau_i)$ of the i-th agent over trajectory τ_i is given by

$$R(\tau_i) = \sum_{h=0}^{H-1} \gamma_i^h r_i^h.$$

Therefore, $V_i(\boldsymbol{\theta})$ in (4) has the following expression:

$$V_i(\boldsymbol{\theta}) := \mathbb{E}_{\tau_i\sim p(\cdot|\boldsymbol{\theta})} \left\{ R(\tau_i) \right\}.$$

Problem (4) can be used to model different cooperative MARL settings. Next we give two examples.

[1] For ease of notation, we drop the subscript i for $\boldsymbol{s}_i^h$ and $\boldsymbol{a}_i^h$ but only keep the subscript for r_i^h.

1.1.1 Collaborative Reinforcement Learning

In this setting, agents aim to maximize the sum of their cumulative rewards in a global environment (Zhang et al., 2018; Jiang et al., 2022). Consider an n-agent MDP denoted by a tuple $\langle \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^n, P, \{r_i\}_{i=1}^n, \{\gamma_i\}_{i=1}^n \rangle$, where

- $\mathcal{S}$ is the global state space shared by all agents,
- $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ is the joint action space of all agents,
- $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition model,
- $r_i : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [-1, 1]$ is the reward function of agent i,
- γ_i is the discount factor for the i-th agent.

Let $\boldsymbol{s} \in \mathcal{S}$ be the global state. The action space of the i-th agent is denoted by $\mathcal{A}_i$. Let $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ be the joint action space. In collaborative RL setting, the state $\boldsymbol{s} \in \mathcal{S}$ and joint action $\boldsymbol{a} \in \mathcal{A}$ are globally observable, while the reward is locally observed. The joint policy is denoted as $\pi : \mathcal{S} \to \Delta(\mathcal{A})$. We assume that the joint policy is parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$, specifically represented as $\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})$. Since each agent makes decisions independently, we have that $\pi(\boldsymbol{a}|\boldsymbol{s}) = \prod_{i=1}^n \pi_i(\boldsymbol{a}_i|\boldsymbol{s})$. Further, suppose that the i-th policy is parameterized by $\boldsymbol{\theta}_{[i]} \in \mathbb{R}^{d_i}$, denoted $\pi_{\boldsymbol{\theta}_{[i]}} : \mathcal{S} \to \Delta(\mathcal{A}_i)$. The probability of executing $\boldsymbol{a}$ at sate $\boldsymbol{s}$ can be rewritten as

$$\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s}) := \prod_{i=1}^n \pi_{\boldsymbol{\theta}_{[i]}}(\boldsymbol{a}_i|\boldsymbol{s}). \tag{5}$$

In this scenario, $\boldsymbol{\theta}$ in (4) is given by $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_{[1]}^\top & \cdots & \boldsymbol{\theta}_{[n]}^\top \end{bmatrix}^\top \in \mathbb{R}^d$ and $d = \sum_{i=1}^n d_i$. It is worth noting that the policy is a mapping onto joint action space. Consequently, the action spaces can either be identical, completely disjoint, or partially overlapping, without affecting the formulation of the policy. A trajectory of collaborative RL induced by $\pi_{\boldsymbol{\theta}}$ is given by

$$\tau = \left\{ \boldsymbol{s}^0, \boldsymbol{a}^0, (r_1^0, \cdots, r_n^0), \boldsymbol{s}^1, \boldsymbol{a}^1, (r_1^1, \cdots, r_n^1), \cdots, \boldsymbol{s}^{H-1}, \boldsymbol{a}^{H-1}, (r_1^{H-1}, \cdots, r_n^{H-1}), \boldsymbol{s}^H \right\},$$

where the superscript denotes time in trajectory and subscript indicates agent. The distribution of τ can be written as

$$p(\tau|\boldsymbol{\theta}) := \rho(\boldsymbol{s}^0) \prod_{h=0}^{H-1} \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h) p(\boldsymbol{s}^{h+1}|\boldsymbol{s}^h, \boldsymbol{a}^h).$$

The value function for the i-th agent is given by

$$V_i(\boldsymbol{\theta}) = \mathbb{E}_{\tau \sim p(\cdot|\boldsymbol{\theta})} \left\{ \sum_{h=0}^{H-1} \gamma_i^h r_i^h \right\}.$$

Then, the objective function in collobrative RL is defined as $V(\boldsymbol{\theta}) := n^{-1} \sum_{i=1}^n V_i(\boldsymbol{\theta})$.

1.1.2 Multi-Task Reinforcement Learning

Multi-task reinforcement learning (MTRL) refers to the problem of different agents learning a shared policy in different but similar environments so that the learned policy can perform well in all the environments. MTRL utilizes similarities across different environments to enhance learning efficiency and generalization. Such an approach has received a lot of attention in recent years (Hessel et al., 2019; Yu et al., 2020; Zeng et al., 2021). In the MTRL setting, the MDP for the i-th agent is expressed as $\langle \mathcal{S}_i, \mathcal{A}, P_i, r_i, \gamma_i \rangle$. The setup for different agents can differ in terms of:

- $\mathcal{S}_i$, the state space for the i-th environment,
- $P_i : \mathcal{S}_i \times \mathcal{A} \to \Delta(\mathcal{S}_i)$, the transition model for the i-th environment,
- $r_i : \mathcal{S}_i \times \mathcal{A} \times \mathcal{S}_i \to [-1, 1]$, the reward function for the i-th agent,
- γ_i, the discount factor for the i-th agent.

In this case, the action spaces are constrained to be identical in order for the agents to share a common parameterized policy, while each local state space $\mathcal{S}_i$ can either be identical, completely disjoint, or partially overlapping. Consider $\mathcal{S}_i$ as the state space for the i-th agent, and we define $\mathcal{S}$ as the union of all $\mathcal{S}_i$, i.e., $\mathcal{S} = \cup_i \mathcal{S}_i$. The policies of various agents share a common parameterization denoted as $\boldsymbol{\theta} \in \mathbb{R}^d$, implying that $\pi_i(\boldsymbol{a}|\boldsymbol{s}) = \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})$ holds for all $\boldsymbol{a} \in \mathcal{A}$ and $\boldsymbol{s} \in \mathcal{S}$. The trajectory for the i-th agent is given by

$$\tau_i = \left\{ \boldsymbol{s}_i^0, \boldsymbol{a}_i^0, r_i^0, \cdots, \boldsymbol{s}_i^{H-1}, \boldsymbol{a}_i^{H-1}, r_i^{H-1}, \boldsymbol{s}_i^H \right\},$$

whose distribution is given by

$$p(\tau_i|\boldsymbol{\theta}) := \rho(\boldsymbol{s}_i^0) \prod_{h=0}^{H-1} \pi_{\boldsymbol{\theta}}(\boldsymbol{a}_i^h|\boldsymbol{s}_i^h) p(\boldsymbol{s}_i^{h+1}|\boldsymbol{s}_i^h, \boldsymbol{a}_i^h).$$

The value function for the i-th agent is defined as

$$V_i(\boldsymbol{\theta}) = \mathbb{E}_{\tau_i \sim p(\cdot|\boldsymbol{\theta})} \left\{ \sum_{h=1}^{H-1} \gamma_i^h r_i^h \right\}.$$

Thus the objective function in MTRL can also be written as (4).

Roughly speaking, collaborative RL shares a global state space, while MTRL shares a common action space. For conciseness, we refer to both of the aforementioned settings as collaborative multi-agent reinforcement learning. It should be easy to see whether "collaborative" refers to two tasks or the particular one task from the context.

1.2 Decentralized Optimization Setup

Since each agent only has access to local information, solving problem (4) needs to aggregate all local computations to update the learning parameter. The centralized optimization method uses a central coordinator for data collection and information transmission, inevitably leading to high communication costs. Moreover, the central coordinator does not exist or may be too expensive to deploy in real applications. By contrast, in a decentralized framework that is considered in this paper each agent only communicates with its neighbors through a communication network. Let $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ be the communication network which is indeed an undirected graph, where $\mathcal{N} = \{1, \cdots, n\}$ is the set of agents, and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the collection of edges. Note that a pair $(i, j) \in \mathcal{E}$ represents that i can communicate with j. For the i-th agent, define the set of its neighbors as $\mathcal{N}(i) = \{j \in \mathcal{N} | (i, j) \in \mathcal{E} \text{ or } i = j\}$. In addition, we can associate a weight matrix $\boldsymbol{W} = [W_{ij}] \in \mathbb{R}^{n \times n}$ with the graph $\mathcal{G}$, where $W_{ij} > 0$ if $(i, j) \in \mathcal{E}$, and $W_{ij} = 0$ otherwise. Assuming $\mathcal{G}$ is a connected graph, it is not hard to see that problem (4) is equivalent to

$$\max_{\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n \in \mathbb{R}^d} \quad \frac{1}{n} \sum_{i=1}^{n} V_i(\boldsymbol{\theta}_i) \quad \text{subject to} \quad \boldsymbol{\theta}_i = \boldsymbol{\theta}_j, \quad \text{for all } (i, j) \in \mathcal{E}. \tag{6}$$

1.3 Main Contributions and Outline of This Paper

The main contributions of this work are summarized as follows.

- Roughly speaking, NPG is preconditioned gradient method which is suitable for solving optimization problems over probability distribution. Some classic methods in single-agent RL, such as TRPO (Schulman et al., 2015), PPO (Schulman et al., 2017), and NAC (Peters and Schaal, 2008), are essentially variants of NPG. Thus, it is natural to extend NPG from the single-agent setting to the decentralized multi-agent setting, which motivates our work. Specifically, we develop a Momentum-based Decentralized Natural Policy Gradient (MDNPG) method for the collaborative MARL problem. MDNPG combines natural gradient with momentum-based variance reduction and gradient tracking to solve the decentralized optimization problem. Extensive numerical experiments show that introducing this preconditioning in the decentralized setting is also able to improve the empirical performance for collaborative MARL.

- Theoretical guarantees for MDNPG have been obtained, showing that MDNPG is able to converge to an ϵ-stationary point in $\mathcal{O}(n^{-1}\epsilon^{-3})$ iterations provided a mini-batch initialization. Even though the variance reduced decentralized policy gradient has been studied in the collaborative MARL scenario, the existing proof techniques cannot directly apply to MDNPG due to the complexity in handling the additional precondition matrices. To overcome this difficulty, a novel stochastic ascent inequality (see Lemma 13) has been established to accommodate the preconditioned gradients for the non-convex objective in the decentralized setting. Moreover, the presence of precondition matrices requires a distinct approach to derive the bound of the consensus errors among different agents (see Lemma 14). To the best of our knowledge, this is first work to study the convergence of the NPG method in the decentralized multi-agent setting. Furthermore, these intermediate technical results are of independent interest and may be available for analyzing other preconditioned stochastic gradient methods in decentralized non-convex optimization.

The rest of this paper is organized as follows. In Section 2, we present a complete description of MDNPG and provide theoretical guarantees for it. In addition, more closely related works are reviewed. In Section 3, we compare MDNPG with other state-of-the-art algorithms in single-agent and multi-agent experiments, which demonstrate the efficiency of the proposed method. The proofs of the main results and key lemmas are presented in Sections 4 and 5. Finally, in Section 6, we conclude this paper with future research directions.

Throughout this paper, we refer to $\boldsymbol{A} \otimes \boldsymbol{B}$ as the Kronecker product. We denote by $\mathbf{1}_n \in \mathbb{R}^n$ the all-one vector (i.e., all entries of $\mathbf{1}_n$ are 1) and by $\boldsymbol{J}_n \in \mathbb{R}^{n\times n}$ the all-one matrix. The $d \times d$ identity matrix is denoted by $\boldsymbol{I}_d$. Additionally, we denote by $\Delta(\mathcal{S})$ (or $\Delta(\mathcal{A})$) the probability simplex over the state (or action) space.

2. MDNPG and Convergence Results

The Momentum-based Decentralized Natural Policy Gradient (MDNPG) algorithm is summarized in Algorithm 1. In the algorithm, agents perform the following steps at each iteration t: gradient estimator calculation, gradient tracking, and parameter update. Notice that each step is simultaneously executed by all agents but is only presented from agent i' s view for simplicity. Overall, there are three pillars in MDNPG, which will be detailed next. Compared with policy gradient based decentralized optimization algorithms (Jiang et al., 2022; Zeng et al., 2021; Zhao et al., 2021) for collaborative MARL, the key difference is in the parameter update step where a natural gradient direction is used for each agent.

Algorithm 1 Momentum-based Decentralized Natural Policy Gradient (MDNPG)

```
Input: number of iterations $T$, horizon $H$, batch size $B$, learning rate $\eta$, momentum parameter $\beta$, initial parameter $\bar{\boldsymbol{\theta}}^0 \in \mathbb{R}^d$, initial estimator $\boldsymbol{v}_i^{-1} = \mathbf{0} \in \mathbb{R}^d$, initial tracker $\boldsymbol{y}_i^0 = \mathbf{0} \in \mathbb{R}^d$.
Initialization: $\boldsymbol{\theta}_i^0 = \bar{\boldsymbol{\theta}}^0$, $\boldsymbol{v}_i^0 = \frac{1}{B}\sum_{b=1}^B \boldsymbol{g}_i(\tau_{i,b}^0|\boldsymbol{\theta}_i^0)$ and $\boldsymbol{y}_i^1 = \sum_{j\in\mathcal{N}(i)} W_{ij}\boldsymbol{v}_j^0$ for $i = 1, \cdots, n$, where $\{\tau_{i,b}^0\}_{b=1}^B$ represents the $B$ trajectories i.i.d sampled from $p(\cdot|\boldsymbol{\theta}_i^0)$.
for $t = 1, 2, \ldots, T$ do
    Generate an estimator $\boldsymbol{v}_i^t$ of $\nabla V_i(\boldsymbol{\theta}^t)$:
        $$\boldsymbol{v}_i^t = \beta \boldsymbol{g}_i(\tau_i^t|\boldsymbol{\theta}_i^t) + (1-\beta)\left(\boldsymbol{v}_i^{t-1} + \boldsymbol{g}_i(\tau_i^t|\boldsymbol{\theta}_i^t) - \omega(\tau_i^t|\boldsymbol{\theta}_i^{t-1}, \boldsymbol{\theta}_i^t)\cdot \boldsymbol{g}_i(\tau_i^t|\boldsymbol{\theta}_i^{t-1})\right).$$
    Gradient Tracking:
        $$\boldsymbol{y}_i^{t+1} = \sum_{j\in\mathcal{N}(i)} W_{ij}\left(\boldsymbol{y}_j^t + \boldsymbol{v}_j^t - \boldsymbol{v}_j^{t-1}\right).$$
    Parameter Update:
        $$\boldsymbol{\theta}_i^{t+1} = \sum_{j\in\mathcal{N}(i)} W_{ij}\left(\boldsymbol{\theta}_j^t + \eta \boldsymbol{H}_j^t \boldsymbol{y}_j^{t+1}\right).$$
end for
Output: $\boldsymbol{\theta}_{\text{out}} \in \mathbb{R}^d$ chooses randomly from $\{\boldsymbol{\theta}_i^t\}_{i=1,\ldots,n,t=0,\ldots,T}$.
```

2.1 Three Pillars in MDNPG Algorithm

2.1.1 Pillar I: Decentralized Optimization

To solve problem (6), each agent can first perform a local gradient update and then seek consensus with its neighbors in order to fulfill the equality constraint. This is the basic idea behind the decentralized gradient ascent method which can be expressed as

$$\boldsymbol{\theta}_i^{t+1} = \sum_{j \in \mathcal{N}(i)} W_{ij}(\boldsymbol{\theta}_j^t + \eta \nabla V_j(\boldsymbol{\theta}_j^t)), \tag{7}$$

where η represents the learning rate, $\nabla V_j(\boldsymbol{\theta}_j^t)$ represents the gradient of $V_j(\boldsymbol{\theta}_j)$ with respect to $\boldsymbol{\theta}_j^t$, and W_{ij} are the elements of the weight matrix $\boldsymbol{W}$ associated with the communication network $\mathcal{G}$. Note that $\boldsymbol{W}$ here plays a role of weighted average for consensus which should satisfy certain properties (see Assumption 1). Despite its simplicity, the original decentralized gradient ascent suffers from slow convergence. To address this issue, a gradient tracking technique has been developed in Li et al. (2020); Pu and Nedić (2021), of which the central idea is to correct biases between local copies of $\boldsymbol{\theta}$ via tracking the average gradient, i.e., $\frac{1}{n}\sum_{i=1}^n \nabla V_i(\boldsymbol{\theta}_i)$. The modified version of update (7) with gradient tracking consists of the following two steps:

$$\begin{aligned} \boldsymbol{y}_i^{t+1} &= \sum_{j \in \mathcal{N}(i)} W_{ij} \left(\boldsymbol{y}_j^t + \nabla V_j(\boldsymbol{\theta}_j^t) - \nabla V_j(\boldsymbol{\theta}_j^{t-1}) \right), \\ \boldsymbol{\theta}_i^{t+1} &= \sum_{j \in \mathcal{N}(i)} W_{ij} \left(\boldsymbol{\theta}_j^t + \eta \boldsymbol{y}_j^{t+1} \right), \end{aligned} \tag{8}$$

where $\boldsymbol{y}_i$ denotes the gradient tracker for agent i. Simple calculation shows that (8) satisfies the dynamic average consensus property:

$$\frac{1}{n}\sum_{i=1}^n \boldsymbol{y}_i^{t+1} = \frac{1}{n}\sum_{i=1}^n \nabla V_i(\boldsymbol{\theta}_i^t), \quad t \geq 1,$$

which implies that the average of $\nabla V_i(\boldsymbol{\theta}_i^t)$ is dynamically tracked by the average of $\boldsymbol{y}_i^{t+1}$. Further, it can be proved that decentralized optimization methods equipped with gradient tracking can achieve better convergence rate (Li et al., 2020).

2.1.2 Pillar II: Variance Reduction

Consider the optimization problem

$$\max_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{E}\left\{ f(\boldsymbol{\theta}; \xi) \right\},$$

where ξ represents a random variable drawn from an unknown distribution $\mathcal{D}$. The stochastic gradient ascent at the t-th iteration is given as

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \eta \cdot \boldsymbol{g}(\boldsymbol{\theta}^t; \xi^t),$$

where η is the learning rate and $\boldsymbol{g}(\boldsymbol{\theta}^t; \xi^t) = \nabla f(\boldsymbol{\theta}^t; \xi^t)$ is the gradient estimator with ξ^t being independently sampled from $\mathcal{D}$. Due to the high variance incurred by the stochastic evaluation of the gradient, vanilla stochastic gradient methods suffer from slow convergence. Thus, in order to accelerate the methods, various variance reduction methods have been proposed and studied in the past decades, such as SVRG (Johnson and Zhang, 2013), SAGA (Defazio et al., 2014), SARAH (Nguyen et al., 2017), and SPIDER (Fang et al., 2018). More recently, a momentum-based variance reduction method (Cutkosky and Orabona, 2019; Tran-Dinh et al., 2019) is proposed, in which the gradient estimator is given by

$$\boldsymbol{v}^t = \beta \underbrace{\boldsymbol{g}(\boldsymbol{\theta}^t; \xi^t)}_{\text{SGD}} + (1-\beta) \underbrace{\left(\boldsymbol{v}^{t-1} + \boldsymbol{g}(\boldsymbol{\theta}^t; \xi^t) - \boldsymbol{g}(\boldsymbol{\theta}^{t-1}; \xi^t)\right)}_{\text{SARAH}}, \tag{9}$$

where $\beta \in (0, 1]$ is the momentum parameter. A key feature of the momentum-based method is that it is a single-loop algorithm which leverages the benefits of both the unbiased SGD estimator (Bottou, 2012) and the novel SARAH estimator (Nguyen et al., 2017). Thus it can avoid the high computational cost of batch gradients to reduce variance.

In this work, we will adopt the momentum-based variance reduction method for the policy gradient estimation. The gradient of $V_i(\boldsymbol{\theta}_i)$ in (6) with respect to $\boldsymbol{\theta}_i$ can be computed as follows

$$\begin{aligned}
\nabla V_i(\boldsymbol{\theta}_i) &= \nabla_{\boldsymbol{\theta}_i} \mathbb{E}_{\tau_i \sim p(\cdot|\boldsymbol{\theta}_i)} \left\{ R(\tau_i) \right\} \\
&= \int_{\tau_i} \nabla_{\boldsymbol{\theta}_i} p(\tau_i|\boldsymbol{\theta}_i) R(\tau_i) d\tau_i \\
&= \int_{\tau_i} p(\tau_i|\boldsymbol{\theta}_i) \frac{\nabla_{\boldsymbol{\theta}_i} p(\tau_i|\boldsymbol{\theta}_i)}{p(\tau_i|\boldsymbol{\theta}_i)} R(\tau_i) d\tau_i \\
&= \mathbb{E}_{\tau_i \sim p(\cdot|\boldsymbol{\theta}_i)} \left\{ \nabla_{\boldsymbol{\theta}_i} \log p(\tau_i|\boldsymbol{\theta}_i) R(\tau_i) \right\}.
\end{aligned} \tag{10}$$

The commonly used gradient estimators of policy gradient include REINFORCE (Williams, 1992) or GPOMDP (Baxter and Bartlett, 2001). For the i-th agent, we adopt REINFORCE with a baseline b_i as the policy gradient estimator:

$$\boldsymbol{g}_i(\tau_i|\boldsymbol{\theta}_i) = \left[\sum_{h=0}^{H-1} \nabla_{\boldsymbol{\theta}_i} \log \pi_{\boldsymbol{\theta}_i}(\boldsymbol{a}^h|\boldsymbol{s}^h) \right] \cdot \left[\sum_{h=0}^{H-1} \gamma^h r_i^h - b_i \right], \tag{11}$$

where τ_i denotes a trajectory generated under policy $\pi_{\boldsymbol{\theta}_i}$. Without loss of generality, we use the same γ for all agents.

Note that in the momentum-based gradient estimator (9) for the ordinary stochastic optimization, the ξ^t sampled from $\mathcal{D}$ is independent of $\boldsymbol{\theta}^t$. However, in (11), the sampled trajectory τ_i^t is determined by the distribution $p(\cdot|\boldsymbol{\theta}_i^t)$. It is easily seen that $\mathbf{g}_i(\tau_i^t|\boldsymbol{\theta}_i^{t-1})$ is a biased estimator for $\nabla V(\boldsymbol{\theta}_i^{t-1})$. To ensure the unbiased property, we can utilize the importance sampling technique,

$$\mathbb{E}_{\tau_i^t \sim p(\cdot|\boldsymbol{\theta}_i^t)} \left\{ \omega(\tau_i^t|\boldsymbol{\theta}_i^{t-1}, \boldsymbol{\theta}_i^t) \mathbf{g}_i(\tau_i^t|\boldsymbol{\theta}_i^{t-1}) \right\} = \nabla V_i(\boldsymbol{\theta}_i^{t-1}),$$

where $\omega(\tau_i^t|\boldsymbol{\theta}_i^{t-1}, \boldsymbol{\theta}_i^t)$ represents the importance weight defined as

$$\omega(\tau_i^t|\boldsymbol{\theta}_i^{t-1}, \boldsymbol{\theta}_i^t) = \frac{p(\tau_i^t|\boldsymbol{\theta}_i^{t-1})}{p(\tau_i^t|\boldsymbol{\theta}_i^t)} = \prod_{h=0}^{H-1} \frac{\pi_{\boldsymbol{\theta}_i^{t-1}}(\boldsymbol{a}^h|\boldsymbol{s}^h)}{\pi_{\boldsymbol{\theta}_i^t}(\boldsymbol{a}^h|\boldsymbol{s}^h)}. \tag{12}$$

Therefore, the momentum-based variance reduction (9) for the policy gradient of the i-th agent, denoted $\boldsymbol{v}_i^t$, is given by Huang et al. (2020):

$$\boldsymbol{v}_i^t = \beta \boldsymbol{g}_i(\tau_i^t|\boldsymbol{\theta}_i^t) + (1-\beta)\left(\boldsymbol{v}_i^{t-1} + \boldsymbol{g}_i(\tau_i^t|\boldsymbol{\theta}_i^t) - \omega(\tau_i^t|\boldsymbol{\theta}_i^{t-1}, \boldsymbol{\theta}_i^t)\cdot \boldsymbol{g}_i(\tau_i^t|\boldsymbol{\theta}_i^{t-1})\right).$$

2.1.3 Pillar III: Natural Policy Gradient

In order to introduce the natural policy gradient method, we consider the policy optimization problem (2):

$$\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \left\{V(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta})}\left\{R(\tau)\right\}\right\},$$

where τ is the H-horizon trajectory and $p(\tau|\boldsymbol{\theta})$ given by (3). The PG update $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \eta\nabla V(\boldsymbol{\theta}^t)$ is a gradient ascent method over the parameter space, which is also the minimizer of the following problem

$$\min_{\boldsymbol{\theta}} \left\langle -\nabla V(\boldsymbol{\theta}^t), \boldsymbol{\theta}-\boldsymbol{\theta}^t\right\rangle + \frac{1}{2\eta}\left\|\boldsymbol{\theta}-\boldsymbol{\theta}^t\right\|_2^2. \tag{13}$$

However, since the objective function essentially relies on the distributions of τ, it is more natural to conduct a search over distribution space, leading to the following sub-problem for updating $\boldsymbol{\theta}^t$:

$$\min_{\boldsymbol{\theta}} \left\langle -\nabla V(\boldsymbol{\theta}^t), \boldsymbol{\theta}-\boldsymbol{\theta}^t\right\rangle + \frac{1}{2\eta}\mathrm{KL}(p(\tau|\boldsymbol{\theta}^t); p(\tau|\boldsymbol{\theta})), \tag{14}$$

where the KL divergence is used to enable the search around $p(\tau|\boldsymbol{\theta}^t)$ over the distribution space.

Since $\mathrm{KL}(p(\tau|\boldsymbol{\theta}^t); p(\tau|\boldsymbol{\theta}^t)) = 0$ and $\nabla_{\boldsymbol{\theta}^t}\mathrm{KL}(p(\tau|\boldsymbol{\theta}^t); p(\tau|\boldsymbol{\theta})) = \mathbf{0}$, one can approximate $\mathrm{KL}(p(\tau|\boldsymbol{\theta}^t); p(\tau|\boldsymbol{\theta}))$ by its second order information and thus approximate (14) by

$$\min_{\boldsymbol{\theta}} \left\langle -\nabla V(\boldsymbol{\theta}^t), \boldsymbol{\theta}-\boldsymbol{\theta}^t\right\rangle + \frac{1}{2\eta}(\boldsymbol{\theta}-\boldsymbol{\theta}^t)^{\mathsf{T}}\boldsymbol{F}(\boldsymbol{\theta}^t)(\boldsymbol{\theta}-\boldsymbol{\theta}^t), \tag{15}$$

where $\boldsymbol{F}(\boldsymbol{\theta}^t) = \mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta}^t)}\left\{\nabla_{\boldsymbol{\theta}}\log p(\tau|\boldsymbol{\theta}^t)\left(\nabla_{\boldsymbol{\theta}}\log p(\tau|\boldsymbol{\theta}^t)\right)^{\mathsf{T}}\right\}$ is the Fisher information matrix (FIM) of $p(\tau|\boldsymbol{\theta}^t)$ and $\boldsymbol{F}(\boldsymbol{\theta}^t)^{\dagger}$ is the Moore-Penrose pseudoinverse of $\boldsymbol{F}(\boldsymbol{\theta}^t)$. It can be easily seen that the optimal solution to (15) is given by

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \eta\boldsymbol{F}(\boldsymbol{\theta}^t)^{\dagger}\nabla V(\boldsymbol{\theta}^t), \tag{16}$$

which yields the natural policy gradient update.

Given the definition of $p(\tau|\boldsymbol{\theta})$ in (3), the FIM can be further expressed as

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{\theta}) &= \mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta})}\left\{\left(\sum_{h=0}^{H-1}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\right)\left(\sum_{h=0}^{H-1}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\right)^{\mathsf{T}}\right\}\\ &= \mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta})}\left\{\sum_{h=0}^{H-1}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\left(\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\right)^{\mathsf{T}}\right\}.\end{aligned} \tag{17}$$

Here the second line has used the fact that the cross term is equal to 0, which can be easily verified. When $H\to\infty$, the FIM may not be well determined. There are two typical ways to deal with this issue:

- Averaged case. Let τ be a trajectory induced by $\pi_{\boldsymbol{\theta}}$ up to horizon H. The FIM in the average case is given by

$$\boldsymbol{F}(\boldsymbol{\theta}) = \lim_{H\to\infty} \frac{1}{H}\mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta})}\left\{\sum_{h=0}^{H-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\right)^{\mathsf{T}}\right\}, \tag{18}$$

It has been shown by Bagnell and Schneider (2003); Peters et al. (2003) that (18) is equivalent to

$$\boldsymbol{F}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{s}\sim d^{\pi_{\boldsymbol{\theta}}}, \boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s})}\left\{\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})\left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})\right)^{\mathsf{T}}\right\},$$

where $d^{\pi_{\boldsymbol{\theta}}}$ is the stationary distribution of state.

- Discounted case. On the other hand, one can consider infinite horizon but introduce a discounted factor $\gamma \in [0, 1)$. In this situation, the FIM is given by

$$\boldsymbol{F}(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta})}\left\{\sum_{h=0}^{+\infty} \gamma^h \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\right)^{\mathsf{T}}\right\}. \tag{19}$$

Moreover, letting τ be the H-horizon trajectory induced by $\pi_{\boldsymbol{\theta}}$, where H obeys the geometric distribution with parameter $1-\gamma$, then $F(\boldsymbol{\theta})$ in (19) is indeed the FIM associated with the random-length trajectory τ. That is (Bagnell and Schneider, 2003; Peters et al., 2003),

$$\begin{aligned}\boldsymbol{F}(\boldsymbol{\theta}) &= \mathbb{E}_{H\sim\mathrm{Geo}(1-\gamma)}\left\{\mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta})}\left\{\sum_{h=0}^{H-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\right)^{\mathsf{T}}\middle| H\right\}\right\}\\ &= \frac{1}{1-\gamma}\mathbb{E}_{\boldsymbol{s}\sim d_\rho^{\pi_{\boldsymbol{\theta}}}, \boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s})}\left\{\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})\left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})\right)^{\mathsf{T}}\right\},\end{aligned}$$

where $d_\rho^{\pi_{\boldsymbol{\theta}}}(\boldsymbol{s}) = \mathbb{E}_{\boldsymbol{s}^0\sim\rho}\left\{(1-\gamma)\sum_{h=0}^{\infty}\gamma^h P(\boldsymbol{s}^h = \boldsymbol{s}|\boldsymbol{s}^0, \pi_{\boldsymbol{\theta}})\right\}$ is the discounted state visitation distribution under the initial distribution ρ. Such formulation has been widely used in the literature (Kakade, 2001; Bhatnagar et al., 2007; Agarwal et al., 2021).

With a slight abuse of notation, we will use the following definition of FIM in this paper:

$$\boldsymbol{F}(\boldsymbol{\theta}) = \mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta})}\left\{\frac{1}{H}\sum_{h=0}^{H-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\right)^{\mathsf{T}}\right\}, \tag{20}$$

which agrees with (17) up to a scale.

In contrast to PG, NPG can be approximately viewed as a second-order method since the FIM serves as a structured precondition based on the underlying structure of the parameterized policy space (Amari, 1996; Martens, 2020). Such a precondition can adaptively adjust the update direction to improve the convergence rate. In Algorithm 1, we have extended the NPG update (16) to the decentralized multi-agent setting (6). For the i-th agent at the t-th iteration, we have

$$\boldsymbol{\theta}_i^{t+1} = \sum_{j\in\mathcal{N}(i)} W_{ij}\left(\boldsymbol{\theta}_j^t + \eta \boldsymbol{H}_j^t \nabla V_j(\boldsymbol{\theta}_j^t)\right),$$

where $\boldsymbol{H}_j^t \in \mathbb{R}^{d\times d}$ denotes the Moore-Penrose pseudoinverse of $\boldsymbol{F}_j(\boldsymbol{\theta}_j^t)$. Namely, each agent searches along the (preconditioned) natural gradient direction of its own before the consensus.

In addition, the following lemma establishes that the FIM in the collaborative RL setting mentioned in Section 1.1 is indeed a block diagonal matrix for each agent due to the product structure of the joint policy, see (5).

Lemma 1 *In collaborative RL, let τ_i be the H-horizon trajectory induced by the policy $\pi_{\boldsymbol{\theta}_i}$ for agent i. The FIM of the i-th agent $\boldsymbol{F}_i(\boldsymbol{\theta}_i) \in \mathbb{R}^{d\times d}$ is given by*

$$\boldsymbol{F}_i(\boldsymbol{\theta}_i) = \operatorname{diag}\left(\boldsymbol{F}_i(\boldsymbol{\theta}_{[1]}), \cdots, \boldsymbol{F}_i(\boldsymbol{\theta}_{[n]})\right),$$

where

$$\boldsymbol{F}_i(\boldsymbol{\theta}_{[j]}) = \frac{1}{H} E_{\tau_i\sim p(\cdot|\boldsymbol{\theta})} \left\{ \sum_{h=0}^{H-1} \nabla_{\boldsymbol{\theta}_{[j]}} \log \pi_{\boldsymbol{\theta}_{[j]}}(\boldsymbol{a}_j^h|\boldsymbol{s}^h) \left(\nabla_{\boldsymbol{\theta}_{[j]}} \log \pi_{\boldsymbol{\theta}_{[j]}}(\boldsymbol{a}_j^h|\boldsymbol{s}^h)\right)^T \right\} \in \mathbb{R}^{d_j\times d_j} \quad (21)$$

for $j = 1, \cdots, n$.

2.2 Theoretical Result

Before stating the main convergence result in Theorem 9, we first introduce some standard assumptions.

Assumption 1 *The weight matrix $\boldsymbol{W} \in \mathbb{R}^{n\times n}$ associated with the communication graph $\mathcal{G}$ is doubly stochastic, i.e., $\boldsymbol{W}\mathbf{1}_n = \mathbf{1}_n$ and $\mathbf{1}_n^\top \boldsymbol{W} = \mathbf{1}_n^\top$.*

Remark 2 *Assumption 1 can be easily satisfied and is commonly used in the convergence analysis of decentralized optimization methods (see for example Boyd et al. (2006); Tang et al. (2018); Nedić et al. (2018); Xin et al. (2021)). Under this assumption, one can show that*

$$\rho := \left\| \boldsymbol{W} - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top \right\| \in [0, 1). \quad (22)$$

Assumption 2 *Let $\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})$ be the policy parameterized by $\boldsymbol{\theta} \in \mathbb{R}^d$. There are constants G and M such that the gradient and Hessian of the log-density of the policy function satisfy*

$$\|\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})\|_2^2 \le G \quad \text{and} \quad \left\|\nabla_{\boldsymbol{\theta}}^2 \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})\right\| \le M$$

for any $\boldsymbol{a} \in \mathcal{A}$ and $\boldsymbol{s} \in \mathcal{S}$.

Remark 3 *Assumption 2 is widely used in the studies of policy gradient methods (Pirotta et al., 2015; Papini et al., 2018; Liu et al., 2020b; Ding et al., 2021; Fatkhullin et al., 2023) and can be satisfied for simple policy parameterization such as softmax tabular policy, log-linear policy with bounded feature vectors. Moreover, it also holds for Gaussian policy with bounded action and bounded mean parameterization (Fatkhullin et al., 2023).*

Assumption 3 *The variance of $\omega(\tau|\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$, the importance sampling weight defined in* (12)*, is bounded,*

$$\mathrm{Var}(\omega(\tau|\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})) \leq W,$$

for any $\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}} \in \mathbb{R}^d$ and $\tau \sim p_i(\cdot|\boldsymbol{\theta})$.

Remark 4 *Assumption 3 is commonly adopted in the analysis of variance reduced policy gradient methods (Papini et al., 2018; Xu et al., 2020b,a; Shen et al., 2019; Liu et al., 2020b; Huang et al., 2020; Ding et al., 2021). For two Gaussian policies $\pi_{\boldsymbol{\theta}_1}(\cdot|s) = \mathcal{N}(\boldsymbol{\theta}_1, \sigma_1^2)$ and $\pi_{\boldsymbol{\theta}_2}(\cdot|s) = \mathcal{N}(\boldsymbol{\theta}_2, \sigma_2^2)$, it is shown in (Cortes et al., 2010) that the variance of the importance sampling weight is bounded provided $\sigma_2 > \frac{\sqrt{2}}{2}\sigma_1$. That being said, it is worth noting that such an assumption might be violated in practice (Huang et al., 2020; Ding et al., 2021). To deal with the challenge, many tricks have been proposed, e.g., clipping the importance sampling weights (Huang et al., 2020), applying truncated policy gradient (Zhang et al., 2021a).*

The following two auxiliary lemmas can be derived from the above assumptions directly.

Lemma 5 *(Xu et al., 2020a, Proposition 5.2) Under Assumption 2, one has the following facts:*

- *The objective function $V(\boldsymbol{\theta})$ is L-smooth with $L = H(M + HG)/(1 - \gamma)$.*
- *Let $\boldsymbol{g}_i(\tau; \boldsymbol{\theta})$ be the gradient estimator defined in* (11)*. Then for all $\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}} \in \mathbb{R}^d$, one has*

$$\left\|\boldsymbol{g}_i(\tau; \boldsymbol{\theta}) - \boldsymbol{g}_i(\tau; \widetilde{\boldsymbol{\theta}})\right\|_2 \leq L_g \left\|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}\right\|_2$$

and $\|\boldsymbol{g}_i(\tau; \boldsymbol{\theta})\|_2 \leq C_g$ for all $i \in [n]$, where $L_g = HM(1 + |b|)/(1 - \gamma)$ and $C_g = HG^{1/2}(1 + |b|)/(1 - \gamma)$.

Lemma 6 *(Jiang et al., 2022, Lemma 3) Under Assumption 2 and 3, one has*

$$\mathrm{Var}(\omega(\tau|\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})) = \mathbb{E}_{\tau \sim p(\cdot|\boldsymbol{\theta})}\left\{\left(\omega(\tau|\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) - 1\right)^2\right\} \leq C_\omega^2 \left\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|_2^2, \tag{23}$$

where $C_\omega^2 = H(2HG + M)(W + 1)$.

The next lemma states that the variance of the gradient estimator (11) is bounded, which is a direct consequence of Assumption 2.

Lemma 7 *(Yuan et al., 2022, Lemma 4.2). Under Assumption 2, consider the gradient estimator* (11)*. One has*

$$\mathrm{Var}(g_i(\tau; \boldsymbol{\theta})) = \mathbb{E}\left\{\|\boldsymbol{g}_i(\tau; \boldsymbol{\theta}) - \nabla V_i(\boldsymbol{\theta})\|_2^2\right\} \leq \nu_i^2 \tag{24}$$

for all policy $\pi_{\boldsymbol{\theta}}$, where $\nu_i = \frac{\sqrt{HG}}{1-\gamma}$ and $\tau \sim p(\cdot|\boldsymbol{\theta})$. Define $\bar{\nu}^2 = \frac{1}{n}\sum_{i=1}^n \nu_i^2$.

Assumption 4 *Let $\boldsymbol{F}(\boldsymbol{\theta}) \in \mathbb{R}^{d\times d}$ be the FIM defined in (20). There exists a constant $\mu_F > 0$ such that $\boldsymbol{F}(\boldsymbol{\theta}) \succeq \mu_F \boldsymbol{I}_d$ for all $\boldsymbol{\theta} \in \mathbb{R}^d$.*

Remark 8 *Assumption 4 indicates the positive definiteness of FIM, which is fairly standard in the analysis of single-agent NPG algorithm, e.g., Liu et al. (2020b); Ding et al. (2021); Fatkhullin et al. (2023). The assumpition on the positive definiteness of preconditioned matrices is also needed for establishing the convergence of preconditioned algorithms in both convex and nonconvex optimizations (Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970). In addition to some Gaussian polices, this assumption can also be satisfied by a subclass of exponential family policies and certain neural policies. We refer to Liu et al. (2020b); Ding et al. (2021); Fatkhullin et al. (2023) for more discussions. Moreover, this assumption can be always satisfied if we use $\boldsymbol{F} + \epsilon \boldsymbol{I}_d$ for $\epsilon > 0$ instead of $\boldsymbol{F}$ as the precondition matrix.*

Let $\boldsymbol{H}_i$ be the inverse FIM $\boldsymbol{F}(\boldsymbol{\theta}_i)$ for the i-th agent with policy parameter $\boldsymbol{\theta}_i \in \mathbb{R}^d$. Then Assumptions 2 and 4 imply that

$$\frac{1}{G}\boldsymbol{I}_d \preccurlyeq \boldsymbol{H}_i^t \preccurlyeq \frac{1}{\mu_F}\boldsymbol{I}_d, \tag{25}$$

where the lower bound holds since the fact $\|\boldsymbol{F}(\boldsymbol{\theta})\| \le G$. Moreover, this fact can be proved as follows:

$$\begin{aligned}\|\boldsymbol{F}(\boldsymbol{\theta})\| &= \left\| \mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta})} \left\{ \frac{1}{H}\sum_{h=0}^{H-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h) \left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\right)^{\mathsf{T}} \right\} \right\| \\ &\le \frac{1}{H}\sum_{h=0}^{H-1} \mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta})} \left\{ \left\| \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h) \left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\right)^{\mathsf{T}} \right\| \right\} \\ &\le G,\end{aligned}$$

where the last line is due to Assumption 2.

We are in position to present the main result of this paper.

Theorem 9 *Let $\boldsymbol{\theta}_{\text{out}} \in \mathbb{R}^d$ be the output of Algorithm 1. Suppose that*

$$0 < \eta < \frac{\mu_F(1-\rho^2)^3}{\kappa_F\sqrt{1632000(L^2+\Phi^2)}}$$

and choose β such that $\frac{1632000(L^2+\Phi^2)\kappa_F^2\eta^2}{n\mu_F^2(1-\rho^2)^6} \le \beta < \frac{1}{n}$, where $\Phi^2 = L_g^2 + C_g^2 C_\omega^2$ and $\kappa_F = G/\mu_F$. Then under Assumptions 1-4, one has

$$\begin{aligned}\mathbb{E}\left\{\|\nabla V(\boldsymbol{\theta}_{\text{out}})\|_2^2\right\} \le{}& \frac{8G^2\Delta}{T\eta\mu_F} + \frac{76\bar{\nu}^2\kappa_F^2}{nT\beta B} + \frac{152\beta\bar{\nu}^2\kappa_F^2}{n} \\ &+ \frac{44\rho^2\bar{\nu}^2\kappa_F^2}{TB(1-\rho^2)} + \frac{352\beta^2\bar{\nu}^2\kappa_F^2}{(1-\rho^2)^2} + \frac{352\beta\bar{\nu}^2\kappa_F^2}{TB(1-\rho^2)^2} \\ &+ \frac{704\bar{\nu}^2\kappa_F^2\beta^3}{(1-\rho^2)^2} + \frac{44\rho^2\kappa_F^2}{nT(1-\rho^2)}\left\|\widetilde{\nabla V}(\boldsymbol{\theta}^0)\right\|_2^2.\end{aligned} \tag{26}$$

Remark 10 *If we choose η and β according to Theorem 9, then the mean squared stationary gap $\mathbb{E}\left\{\|\nabla V(\boldsymbol{\theta}_{\text{out}})\|_2^2\right\}$ converges to a steady-state error as $T \to \infty$ at a rate of $\mathcal{O}(1/T)$, i.e.,*

$$\mathbb{E}\left\{\|\nabla V(\boldsymbol{\theta}_{\text{out}})\|_2^2\right\} \to \mathcal{O}\left(\frac{\beta\bar{\nu}^2}{n} + \frac{\beta^2\bar{\nu}^2}{(1-\rho^2)^2}\right), \quad T \to \infty.$$

It can be seen that the steady-state error will decrease as the number of agents increases. Moreover, the second term in the steady-state error indicates that the impact of communication graph $\mathcal{G}$ through ρ can be reduced with small β.

Corollary 11 *Choose step size η, momentum parameter β, and batch size B in initialization such that $\eta = \frac{\mu_F n^{2/3}}{\kappa_F\sqrt{L^2+\Phi^2}T^{1/3}}, \beta = \frac{n^{1/3}}{T^{2/3}}$ and $B = \left\lceil \frac{T^{1/3}}{n^{2/3}} \right\rceil$. Then for all $T > \frac{1632000^{3/2}n^2}{(1-\rho^2)^9}$, one has*

$$\mathbb{E}\left\{\|\nabla V(\boldsymbol{\theta}_{\text{out}})\|_2^2\right\} \leq \frac{8\Delta\kappa_F^3\sqrt{L^2+\Phi} + 228\bar{\nu}^2\kappa_F^2}{(nT)^{2/3}} \\ + \frac{44\rho^2\kappa_F^2\left\|\widetilde{\nabla V}(\boldsymbol{\theta}^0)\right\|_2^2}{(1-\rho^2)} \cdot \frac{1}{nT} + \frac{396\kappa_F^2\bar{\nu}^2}{(1-\rho^2)^2} \cdot \frac{n^{2/3}}{T^{4/3}} + \frac{1056\bar{\nu}^2\kappa_F^2}{(1-\rho^2)^2} \cdot \frac{n}{T^2}.$$

Remark 12 *Corollary 11 implies that*

$$\mathbb{E}\left\{\|\nabla V(\boldsymbol{\theta}_{\text{out}})\|_2^2\right\} = \mathcal{O}((nT)^{-2/3})$$

when T is large enough. Thus one can achieve ϵ-stationary, i.e., $\mathbb{E}\left\{\|\nabla V(\boldsymbol{\theta}_{\text{out}})\|_2^2\right\} \lesssim \epsilon^2$, in $\mathcal{O}(n^{-1}\epsilon^{-3})$ iteration complexity, which shows that MDNPG also enjoys the linear speedup convergence rate.

2.3 Related Work

In this section, we discuss the recent progress that is mostly related to our work, especially those gradient based methods in reinforcement learning and decentralized stochastic optimization.

Single-agent policy gradient and natural policy gradient. Inspired by stochastic optimization, there has been extensive research in designing variance reduction methods for policy gradient estimator (Papini et al., 2018; Xu et al., 2020b,a; Shen et al., 2019; Huang et al., 2020). For instance, Papini et al. (2018) show that SVRPG achieves an ϵ-stationary point given $\mathcal{O}(\epsilon^{-4})$ trajectories. Xu et al. (2020a) improve this sample complexity to $\mathcal{O}(\epsilon^{-10/3})$. Moreover, SRVR-PG (Xu et al., 2020b) and HAPG (Shen et al., 2019) can obtain an ϵ-stationary point provided $\mathcal{O}(\epsilon^{-3})$ trajectories, both of which are nearly optimal in the sample complexity (Arjevani et al., 2022). However, these methods require large batches or double-loop updates. Recently, Huang et al. (2020) incorporate the momentum-based variance reduction technique in policy gradient methods and achieve the sample complexity of $\mathcal{O}(\epsilon^{-3})$ with a single trajectory at each iteration. The global convergence of policy gradients with variance reduction has also been studied in Ding et al. (2021); Liu et al. (2020b). As already mentioned, NPG (Kakade, 2001) and its generalizations, such as TRPO (Schulman et al., 2015) and PPO (Schulman et al., 2017), are widely used in RL. There has been a lot of interest in understanding the theoretical performance of this class of methods, see Agarwal et al. (2021); Cen et al. (2021); Lan (2022) and the references therein. Different variance reduction techniques have also been utilized in NPG. For example, SRVR-NPG is proposed in Liu et al. (2020b) which reaches a sample complexity of $\mathcal{O}(\epsilon^{-3})$. In addition, two variance reduced mirror ascent methods, named VRMPO and VR-BGPO, are developed in Yang et al. (2022) and Huang et al. (2021), respectively. These methods reduces to NPG with variance reduction if a special mirror mapping is used.

Multi-agent policy gradient. For collaborative RL problem, many decentralized policy gradient algorithms have been developed. Lu et al. (2021) study a decentralized policy gradient method in safe MARL and show that an ϵ-stationary point can be achieved from $\mathcal{O}(\epsilon^{-4})$ iterations. Zhao et al. (2021) study the convergence of decentralized policy gradient with variance reduction and gradient tracking in collaborative RL and establish the sample complexity of $\mathcal{O}(\epsilon^{-3})$. However, the method in Zhao et al. (2021) requires very large batch gradients to obtain this optimal complexity. In contrast, Jiang et al. (2022) adopt the momentum-based variance reduction technique for decentralized policy gradient which also achieves the optimal sample complexity but only uses a single trajectory in each iteration. For the MTRL problem, various policy gradient methods have been developed and studied. In Espeholt et al. (2018); Hessel et al. (2019), a distributed framework is used to solve the learning problem. However, in these works, each agent collects local data, which are then shared to a centralized coordination. In a subsequent work, a decentralized policy gradient method is proposed in Zeng et al. (2021). However, the proposed method only adopts the vanilla gradient ascent without gradient tracking and variance reduction, thus resulting in a sample complexity of $\mathcal{O}(\epsilon^{-4})$. In addition, decentralized optimization methods have also been studied in the framework of policy evaluation (Qu et al., 2019; Doan et al., 2019; Lin and Ling, 2021).

Decentralized optimization. In general, decentralized optimization has been extensively studied for non-convex problems. There are many algorithms developed toward this line of research, including DSGD (Lian et al., 2017), EXTRA (Shi et al., 2015), and Exact Diffusion (Yuan et al., 2018). To achieve the lower oracle complexity, various kinds of variance reduced techniques have been utilized. The D-GET proposed in Sun et al. (2020) is built upon gradient tracking and the SARAH gradient estimator and achieves an oracle complexity of $\mathcal{O}(\epsilon^{-3})$. The same oracle complexity is also obtained by D-SPIDER-SFO (Pan et al., 2020) which uses SPIDER in the variance reduction step. Recently, built on the hybrid variance reduction scheme introduced in Cutkosky and Orabona (2019); Tran-Dinh et al. (2022), the GT-HSGD is developed in Xin et al. (2021) which can achieve an ϵ-approximate first-order stationary point within $\mathcal{O}(n^{-1}\epsilon^{-3})$ samples for each node.

Our work distinguishes from the existing analysis in common decentralized optimization in two aspects. First and foremost, the introduction of the precondition matrix (i.e., the Fisher information matrix) proposes new challenges in the establishment of the convergence of the algorithm. To this end, a novel stochastic gradient ascent inequality (see Lemma 13) is derived, where the effect of preconditioning is carefully analyzed. In addition, a distinct approach is needed to derive the bound of the consensus errors among different agents (see Lemma 14). Secondly, the randomness of data samples is independent of the optimized parameters in common decentralized optimization, so the unbiased property of the gradient estimation based on the last iteration is satisfied automatically. In contrast, for the RL problem, the importance sampling weight is needed to guarantee the unbiased property, which causes new difficulties in the analysis.

3. Numerical Experiments

In this section, we empirically compare MDNPG with other state-of-the-art algorithms in several typical RL environments[2]. In our implementations, we use the sample version of (20) with one random trajectory to approximately compute the FIM for the single-agent experiments as well as the experiments about multi-task GridWorld:

$$\boldsymbol{F}(\boldsymbol{\theta}) \approx \frac{1}{H} \sum_{h=0}^{H-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h) \left(\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}^h|\boldsymbol{s}^h)\right)^{\mathsf{T}} .$$

Regrading the experiments for the collaborative RL setting on cooperative navigation, the FIM $\boldsymbol{F}_i(\boldsymbol{\theta}_{[j]})$ is computed via (21) of Lemma 1,

$$\boldsymbol{F}_i(\boldsymbol{\theta}_{[j]}) \approx \frac{1}{H} \sum_{h=0}^{H-1} \nabla_{\boldsymbol{\theta}_{[j]}} \log \pi_{\boldsymbol{\theta}_{[j]}}(\boldsymbol{a}_j^h|\boldsymbol{s}^h) \left(\nabla_{\boldsymbol{\theta}_{[j]}} \log \pi_{\boldsymbol{\theta}_{[j]}}(\boldsymbol{a}_j^h|\boldsymbol{s}^h)\right)^{T} .$$

3.1 Single-Agent Experiments

When $n = 1$, MDNPG reduces to the single-agent NPG with momentum-based variance reduction, which can also be viewed as BGPO (Huang et al., 2021) with special mirror mappings. In this subsection, we compare the single-agent version of MDNPG with the momentum-based policy gradient (Huang et al., 2020), PPO (Schulman et al., 2017), and SRVR-NPG (Liu et al., 2020b) over two single-agent environments: GridWorld and MountainCar.

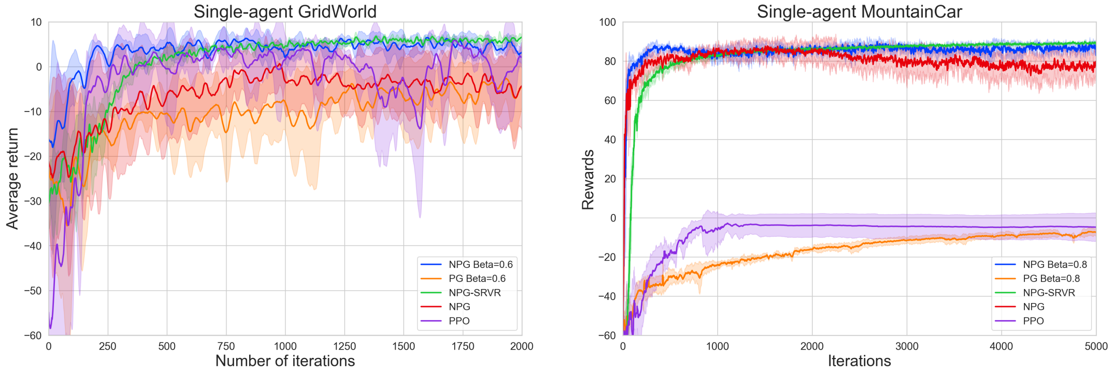
In a 10×10 GridWorld, an agent at a random initial position try to reach the grid labeled as “goal” and at the same time avoid grids labeled as “obstacle” in a minimum number of steps. Five obstacle grids are set up in our experiments. The agent can select one of four discrete actions (up, down, right, and left) to move to another grid, and the state is simply the location of the agent. The received reward is $-0.1 \times$ (distance to the goal) $\pm$ 10, up to whether the goal is reached or the agent falls into an obstacle. The other environment, called MountainCar, is a continuous control task from OpenAI Gym (Brockman et al., 2016), in which the goal of an agent is to reach the top of the hill. A detailed description of the environment is provided in Brockman et al. (2016).

In our implementation, a one-hidden-layer (of size 128) neural network with ReLU activation function is used to parameterize the policy. For the GirdWorld task, the parameterized policy can be obtained via a softmax layer. For the MountainCar task, the outputs of the network are $(\mu_{\boldsymbol{\theta}}(s), \sigma_{\boldsymbol{\theta}}(s))$, the mean and standard deviation of a Gaussian distribution. Moreover, we also use a value network (one hidden layer of size 128, with ReLU

[2] Codes for reproducing the computational results in this section are available at `https://github.com/fccc0417/mdnpg`.

as the activation function) for the estimation of value functions. All the parameters in the algorithms are finely tuned for the pursuit of better performance.

The plots of average return and standard deviation over five random instances against the number of iterations are presented in Figure 1. It can be observed that overall the momentum-based NPG method displays better convergence and stability than the other algorithms for both tasks. Note that even though SRVR-NPG is competitive with the momentum-based NPG method, the former one costs significantly more time and memory due to its double-loop nature for variance reduction.



(a)

(b)

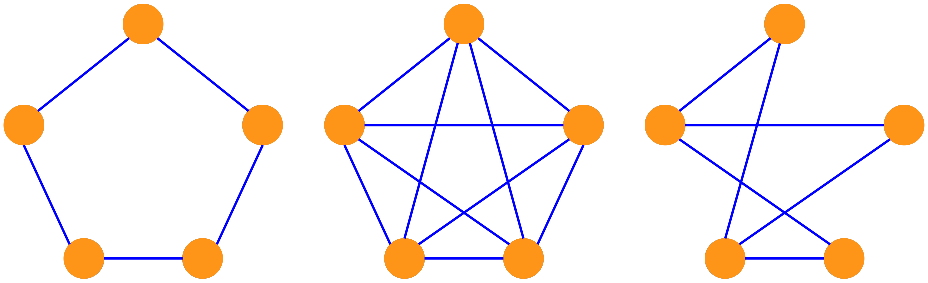
Figure 1: Average return and standard deviation over five random instances against the number of iterations. The parameter β in GridWorld (a) and MountainCar (b) is set to 0.6 and 0.8 for the momentum-based NPG and PG methods.

3.2 Multi-Agent Experiments

3.2.1 Cooperative Navigation

For the collaborative RL setting in Section 1.1, we compare MDNPG with other state-of-the-art algorithms such as MDPGT (Jiang et al., 2022) and value propagation (Qu et al., 2019) on a simulated cooperative navigation environment introduced by Lowe et al. (2017). As a benchmark multi-agent environment, it has been modified in several previous works such as Zhang et al. (2018); Qu et al. (2019); Jiang et al. (2022) to be compatible with the collaborative RL setting. In the n-agent cooperative navigation, each agent at a randomly initialized location needs to find its specific landmark and avoid collisions with other agents in a rectangle region of size 2×2. Agents can move up, down, right, left, or keep still at each step. The globally observed state consists of the positions of all agents as well as their landmarks. The received reward of each agents is $-$(distance to the landmark) $-$ $\sum \mathbb{1}_{\{\text{if colliding with an agent}\}}$.

More precisely, there are 5 agents in our experiments. The policy-based methods, MDNPG as well as MDPGT, utilize a policy network and a value network, both of which have two hidden layers with 64 and 128 units and use ReLU as the activation function. Additionally, the value propagation method utilizes another auxiliary network to approximate the dual function. Since MDNPG and MDPGT are both on-policy algorithms, the on-policy version of value propagation is implemented here.



(a) Ring (b) Fully-connected (c) Bi-partite

Figure 2: Three network topologies.

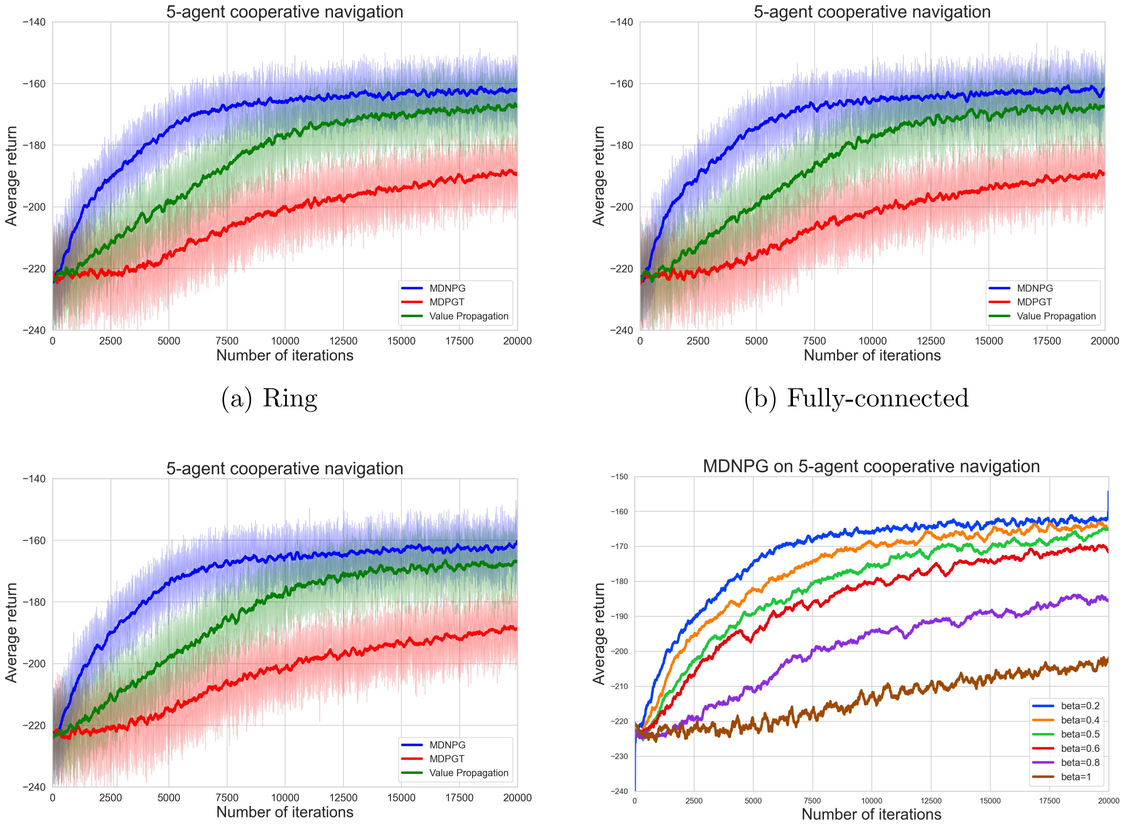
In order to demonstrate the influence of the communication network on the algorithms' performance, we follow the work in Jiang et al. (2022) and test three network topologies (see Figure 2): ring, fully-connected, and bi-partite. The empirical results are displayed in Figure 3. It is evident from Figure 3a—3c that the performance of MDNPG is superior to the other two test methods in all the three network topologies. We have also tested the influence of the momentum parameter β on the performance of MDNPG. Since similar trend has been observed for different topologies, only the results for the ring topology is presented in Figure 3d.

3.2.2 Multi-Task GridWorld

For the MTRL setting in Section 1.1, experiments have been conducted on a multi-task GridWorld problem, whose setup is overall similar to the single-agent case in Section 3.1 but with multiple individual environments. Each agent has its own environment but uses the same policy. By doing so, it is expected to obtain a policy with better generalization.

Five different yet similar environments are considered in our experiments and we compare the proposed MDNPG with MDPGT (Jiang et al., 2022) and PG with entropy regularization (Zeng et al., 2021). As with the single-agent case, one-hidden-layer (of size 128) policy network and value network with ReLU have been utilized, and all of the hyperparameters are properly tuned. Again, three network topologies have been tested and the empirical results are presented in Figure 4a—4c, which clearly shows that MDNPG outperforms the other two test methods. The influence of β on the performance of MDNPG for the ring topology is presented in Figure 4d.

To evaluate the generalization effect of the learned policies in the multi-task experiments, we compare them with the policies learned by training each agent separately (that is, by solving $\max_{\boldsymbol{\theta}} V_i(\boldsymbol{\theta})$ for each i instead of solving $\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} V_i(\boldsymbol{\theta})$ as in (4)). Table 1 contains the average returns over 100 random trajectories computed from the policies trained in the multi-task environment (for the ring topology) as well as in each single environment. Note that even though the single-agent versions of MDNPG, MDPGT, and PG with entropy regularization are all tested for training each agent separately, only results for the single-agent MDNPG are presented due to its superior performance. It is clear that the shared



(a) Ring

(b) Fully-connected

(c) Bi-partite

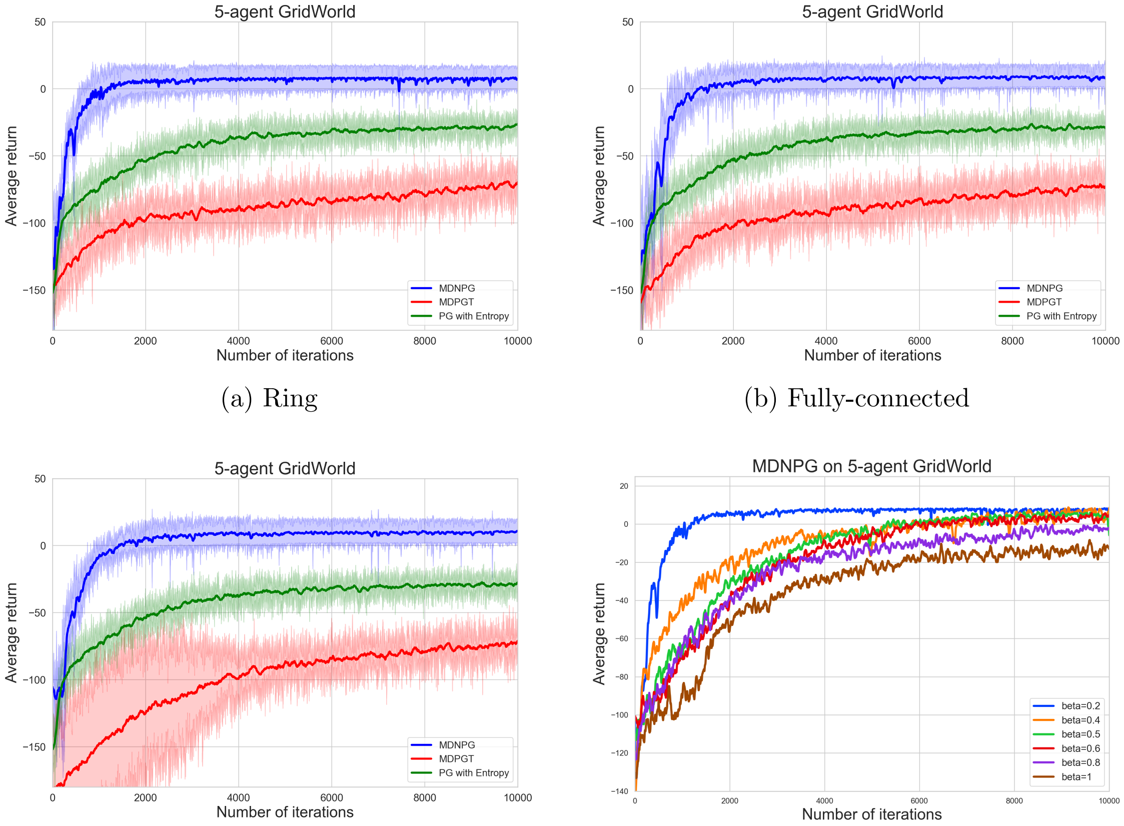
(d) Influence of β for the ring topology

Figure 3: Empirical results on cooperative navigation. In (a)—(c), plots of average return and standard deviation against number of iterations over five random instances, where $\beta = 0.2$ for MDNPG and MDPGT. In (d), influence of β on the performance MDNPG for the ring topology.

policy learned by MDNPG overall generalizes better than the other two methods, and it is competitive with (or better than) the policy learned by training the individual agents in 4 out of 5 environments.

3.2.3 Influence of Variance Reduction and Gradient Tracking

Here we empirically study the influence of variance reduction and gradient tracking in MDNPG using the aforementioned two tasks based on the ring topology. To this end, we compare MDNPG with DNPG (where variance reduction is missing), MDNPG-noGT (where gradient tracking is missing), and DNPG-noGT (where both ingredients are missing). As illustrated in Figure 5, the numerical results demonstrate the superior performance of MDNPG over DNPG, MDNPG-noGT, and DNPG-noGT, thus suggesting the substantial effect of variance reduction and gradient tracking.



(a) Ring

(b) Fully-connected

(c) Bi-partite

(d) Influence of β for the ring topology

Figure 4: Empirical results on multi-task GridWorld. In (a)—(c), plots of average return and standard deviation against number of iterations over five random instances where $\beta = 0.2$ for MDNPG and MDPGT. In (d), influence of β on the performance MDNPG for the ring topology.

4. Proof of Main Results

We first introduce some convenient notations. Letting $\boldsymbol{\theta}_i \in \mathbb{R}^d$ be the local variable for i-th agent, we define the aggregated variable $\boldsymbol{\theta} \in \mathbb{R}^{nd}$ by

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_1^\intercal & \cdots & \boldsymbol{\theta}_n^\intercal \end{bmatrix}^\intercal .$$

Let $\boldsymbol{H}^t = \mathrm{diag}(\boldsymbol{H}_1^t, \cdots, \boldsymbol{H}_n^t) \in \mathbb{R}^{nd\times nd}$ be the block diagonal matrix and define $\boldsymbol{d}_i^t = \boldsymbol{H}_i^t \boldsymbol{y}_i^{t+1} \in \mathbb{R}^d$. We apply the same aggregation rules to obtain other concatenated variables $\boldsymbol{y}, \boldsymbol{v}, \boldsymbol{d} \in \mathbb{R}^{nd}$. Using these notations, the key steps in Algorithm 1 can be rewritten in a more compact form:

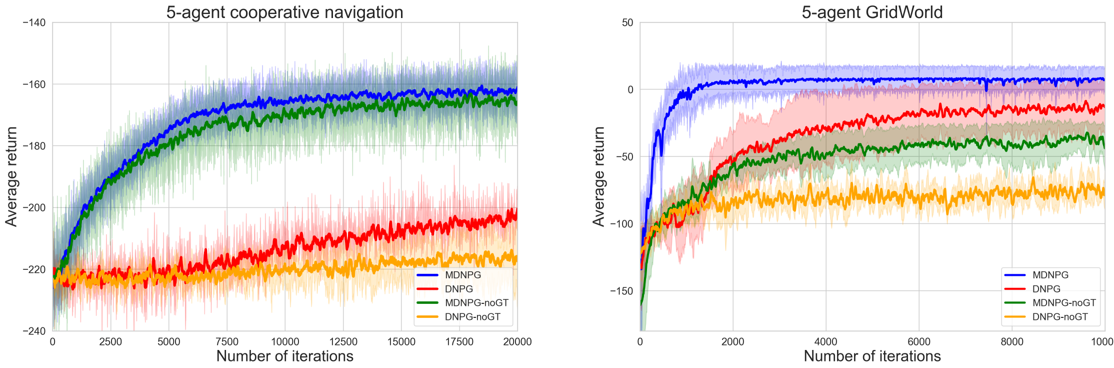
$$\boldsymbol{y}^{t+1} = (\boldsymbol{W} \otimes \boldsymbol{I}_d)\,(\boldsymbol{y}^t + \boldsymbol{v}^t - \boldsymbol{v}^{t-1}) \text{ and } \boldsymbol{\theta}^{t+1} = (\boldsymbol{W} \otimes \boldsymbol{I}_d)\,(\boldsymbol{\theta}^t + \eta \boldsymbol{d}^t). \tag{27}$$

Let $\bar{\boldsymbol{\theta}} \in \mathbb{R}^d$ be the average of $\bar{\boldsymbol{\theta}}_i$ over all the agents, i.e,

$$\bar{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\theta}_i = \frac{1}{n}(\mathbf{1}_n^\intercal \otimes \boldsymbol{I}_d)\boldsymbol{\theta}.$$

Table 1: Average returns over 100 random trajectories based on different learned policies. Agent i means that the policy is learned by training the i-th agent in the i-th grid using the single-agent MDNPG (i.e., the momentum-based NPG). In contrast, MDNPG, MDPGT, and PG with entropy regularization learn a shared policy by training the multi-task environment.

	Grid 1	Grid 2	Grid 3	Grid 4	Grid 5	Sum
Agent 1	**7.64**	-13.24	-81.66	-136.32	-14.1	-237.68
Agent 2	-10.95	**5.93**	-81.09	-11.94	-15.99	-114.03
Agent 3	-10.92	-17.97	**6.76**	-97.5	-67.04	-186.68
Agent 4	-10.92	-126.05	-14.24	**9.25**	-14.04	-156.01
Agent 5	-11.02	-18.01	-3.95	-57.9	**2.5**	-88.37
MDNPG	**7.94**	-17.93	**6.74**	**9.25**	**4.42**	10.43
MDPGT	-7.5	-18.82	-7.04	-3.37	-20.25	-56.98
PG with entropy	0.51	**-11.82**	-0.22	2.89	-19.46	-28.1



(a) Empirical results on cooperative navigation

(b) Empirical results on multi-task GridWorld

Figure 5: Influence of gradient tracking and variance reduction. In (a) and (b), plots of average return and standard deviation against number of iterations over five random instances where $\beta = 0.2$ for MDNPG and MDNPG-noGT.

Similarly, $\bar{\boldsymbol{y}}, \bar{\boldsymbol{v}}, \bar{\boldsymbol{d}} \in \mathbb{R}^d$ also denote the averages of related variables. By the update described in (27), it is straightforward to obtain that

$$\bar{\boldsymbol{y}}^{t+1} = \bar{\boldsymbol{v}}^t \text{ and } \bar{\boldsymbol{\theta}}^{t+1} = \bar{\boldsymbol{\theta}}^t + \eta \bar{\boldsymbol{d}}^t.$$

Moreover, we define the aggregated gradient and averaged gradient by

$$\widetilde{\nabla V}(\boldsymbol{\theta}) = \begin{bmatrix} \nabla V_1(\boldsymbol{\theta}_1)^\mathsf{T} & \cdots & \nabla V_n(\boldsymbol{\theta}_n)^\mathsf{T} \end{bmatrix}^\mathsf{T} \in \mathbb{R}^{nd}, \text{ and } \overline{\nabla V}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \nabla V_i(\boldsymbol{\theta}_i) \in \mathbb{R}^d.$$

Throughout this work, we will frequently use the following relationship

$$\left\| \left((\boldsymbol{W} - n^{-1}\boldsymbol{J}) \otimes \boldsymbol{I}_d \right) \boldsymbol{a} \right\|_2 = \left\| \left((\boldsymbol{W} - n^{-1}\boldsymbol{J}) \otimes \boldsymbol{I}_d \right) (\boldsymbol{a} - \mathbf{1}_n \otimes \bar{\boldsymbol{a}}) \right\|_2 \tag{28}$$

for any $\boldsymbol{a} = \begin{bmatrix}\boldsymbol{a}_1^\mathsf{T} & \cdots & \boldsymbol{a}_n^\mathsf{T}\end{bmatrix}^\mathsf{T} \in \mathbb{R}^{nd}$, where $\bar{\boldsymbol{a}} = n^{-1}\sum_{i=1}^n \boldsymbol{a}_i \in \mathbb{R}^d$ and $\boldsymbol{J} = \mathbf{1}_n\mathbf{1}_n^\mathsf{T} \in \mathbb{R}^{n\times n}$.

The following key lemma establishes the ascent property of MDNPG. This lemma may be of broader interest in analyzing preconditioned stochastic first order methods in decentralized non-convex optimization.

Lemma 13 *Let $\{\boldsymbol{\theta}_i^t\}$ be generated by Algorithm 1 and $\Delta = V^\star - V(\boldsymbol{\theta}^0)$ where $V^\star := \sup_{\boldsymbol{\theta}\in\mathbb{R}^d} V(\boldsymbol{\theta})$. Suppose $0 < \eta \le \frac{\mu_F}{8L}$. Under Assumption 2, one has*

$$\frac{1}{n}\sum_{t=0}^{T}\sum_{i=1}^{n}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 \le \frac{8G^2\Delta}{\eta\mu_F} - \frac{G^2}{n}\sum_{t=0}^{T}\left\|\boldsymbol{d}^t\right\|_2^2 + \frac{76G^2}{\mu_F^2}\sum_{t=0}^{T}\left\|\overline{\nabla V}(\boldsymbol{\theta}^t) - \bar{\boldsymbol{v}}^t\right\|_2^2 + \frac{10G^2}{n\mu_F^2}\sum_{t=0}^{T}\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2 + \frac{82G^2L^2}{n\mu_F^2}\sum_{t=0}^{T}\left\|\boldsymbol{\theta}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2.$$

Next, we will look for conditions on the step size η and weight factor β such that

$$-\frac{G^2}{nT}\sum_{t=0}^{T}\left\|\boldsymbol{d}^t\right\|_2^2 + \frac{76G^2}{T\mu_F^2}\sum_{t=0}^{T}\left\|\overline{\nabla V}(\boldsymbol{\theta}^t) - \bar{\boldsymbol{v}}^t\right\|_2^2 + \frac{10G^2}{nT\mu_F^2}\sum_{t=0}^{T}\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2 + \frac{82G^2L^2}{nT\mu_F^2}\sum_{t=0}^{T}\left\|\boldsymbol{\theta}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2 = \mathcal{O}\left(\eta, \beta, \frac{1}{B}, \frac{1}{T}\right).$$

Assuming this holds, then the application of Lemma 13 will yield

$$\frac{1}{n}\sum_{t=0}^{T}\sum_{i=1}^{n}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 \le \frac{8G^2\Delta}{\eta\mu_F} + \mathcal{O}\left(\eta, \beta, \frac{1}{B}, \frac{1}{T}\right),$$

which implies the convergence of Algorithm 1. To achieve this goal, we need several lemmas whose proofs are either deferred to Section 5 or already given in the literature.

Lemma 14 *Under Assumptions 1, 2 and 4, for all $t \ge 0$, one has*

$$\left\|\boldsymbol{\theta}^{t+1} - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t+1}\right\|_2^2 \le \frac{1+\rho^2}{2}\left\|\boldsymbol{\theta}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2 + \frac{4\eta^2}{\mu_F^2(1-\rho^2)}\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2 + \frac{G^2\eta^2}{\mu_F^2(1-\rho^2)}\left\|\boldsymbol{d}^t\right\|_2^2. \tag{29}$$

Moreover, one has

$$\left\|\boldsymbol{\theta}^{t+1} - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t+1}\right\|_2^2 \le 2\rho^2\left\|\boldsymbol{\theta}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2 + \frac{4\eta^2\rho^2}{\mu_F^2}\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2 + \frac{G^2\eta^2\rho^2}{\mu_F^2}\left\|\boldsymbol{d}^t\right\|_2^2. \tag{30}$$

Lemma 15 *Let $\{\boldsymbol{y}^t\}$ be generated by Algorithm 1. Under Assumptions 1 and 2-4, we have*

$$\mathbb{E}\left\{\left\|\boldsymbol{y}^1 - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^1\right\|_2^2\right\} \le \frac{n\rho^2\bar{\nu}^2}{B} + \rho^2\sum_{i=1}^{n}\left\|\nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\|_2^2. \tag{31}$$

Furthermore, if $\eta \leq \frac{\mu_F(1-\rho^2)}{24\sqrt{2}\Phi}$ *and* $0 \leq \beta \leq 1$*, then for all* $t \geq 1$*, one has*

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2\right\} \\
&\leq \frac{3+\rho^2}{4}\mathbb{E}\left\{\left\|\boldsymbol{y}^{t} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t}\right\|_2^2\right\} + \frac{8n\beta^2\bar{\nu}^2}{1-\rho^2} + \frac{144\Phi^2G^2\eta^2}{\mu_F^2(1-\rho^2)}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\} \\
&\quad + \frac{8\beta^2}{1-\rho^2}\mathbb{E}\left\{\left\|\widetilde{\nabla V}(\boldsymbol{\theta}^{t-1}) - \boldsymbol{v}^{t-1}\right\|_2^2\right\} + \frac{216\Phi^2}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\}. \qquad (32)
\end{aligned}$$

Lemma 16 (Lemma 6 in Xin et al. (2021)) *Let* $\{a_t\}, \{b_t\}$ *and* $\{c_t\}$ *be nonnegative sequences and* $d > 0$ *be some constant such that* $a_t \leq \gamma a_{t-1} + \gamma b_{t-1} + c_t + d$ *for* $t \geq 1$*, where* $\gamma \in (0,1)$*. Then for* $T \geq 1$*, we have*

$$\sum_{t=0}^{T} a_t \leq \frac{1}{1-\gamma}a_0 + \frac{1}{1-\gamma}\sum_{t=0}^{T-1} b_t + \frac{1}{1-\gamma}\sum_{t=1}^{T} c_t + \frac{dT}{1-\gamma}. \qquad (33)$$

Moreover, if $a_{t+1} \leq \gamma a_t + b_{t-1} + d$ *for* $t \geq 1$*, then for* $T \geq 2$*, one has*

$$\sum_{t=1}^{T} a_t \leq \frac{1}{1-\gamma}a_1 + \frac{1}{1-\gamma}\sum_{t=0}^{T-2} b_t + \frac{dT}{1-\gamma}. \qquad (34)$$

Lemma 17 *Let*

$$A_1 = \frac{4n\rho^2\bar{\nu}^2}{B(1-\rho^2)} + \frac{32nT\beta^2\bar{\nu}^2}{(1-\rho^2)^2} + \frac{32n\beta\bar{\nu}^2}{B(1-\rho^2)^2} + \frac{64nT\bar{\nu}^2\beta^3}{(1-\rho^2)^2} + \frac{4\rho^2}{1-\rho^2}\left\|\widetilde{\nabla V}(\boldsymbol{\theta}^0)\right\|_2^2.$$

Suppose $\eta \leq \frac{\mu_F(1-\rho^2)}{24\sqrt{2}\Phi}$ *and* $\beta < 1$*. Then one has*

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left\{\left\|\boldsymbol{y}^{t} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t}\right\|_2^2\right\} \\
&\leq A_1 + \frac{1632\Phi^2}{(1-\rho^2)^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t} - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^{t}\right\|_2^2\right\} + \frac{960\Phi^2\kappa_F^2\eta^2}{(1-\rho^2)^2}\sum_{t=0}^{T-1}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t}\right\|_2^2\right\}. \qquad (35)
\end{aligned}$$

Lemma 18 *Suppose* $0 < \eta < \frac{\mu_F(1-\rho^2)^3}{\kappa_F\sqrt{1632000(L^2+\Phi^2)}}$ *and* $\beta < 1$*. Then for any* $T \geq 1$*, one has*

$$\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t} - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^{t}\right\|_2^2\right\} \leq \frac{16A_1\eta^2}{\mu_F^2(1-\rho^2)^2} + \frac{10G^2\eta^2}{\mu_F^2(1-\rho^2)^3}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t}\right\|_2^2\right\}.$$

Lemma 19 (Lemma 8 in Jiang et al. (2022)) *Let* $\boldsymbol{v}^t$ *and* $\boldsymbol{\theta}^t$ *be generated by Algorithm 1 and let* $\Phi^2 = L_g^2 + C_g^2C_\omega^2$*. Then under Assumption 2, 7 and 3, for any* $t \geq 1$*, one has*

$$\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\bar{\boldsymbol{v}}^{t} - \overline{\nabla V}(\boldsymbol{\theta}^{t})\right\|_2^2\right\} \leq \frac{\bar{\nu}^2}{n\beta B} + \frac{2\beta\bar{\nu}^2T}{n}$$

$$
+ \frac{12\Phi^2\eta^2}{n\beta} \sum_{t=0}^{T-1} \mathbb{E}\left\{ \left\| \bar{\boldsymbol{d}}^t \right\|_2^2 \right\} + \frac{24\Phi^2}{\beta n^2} \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t \right\|_2^2 \right\} \tag{36}
$$

and

$$
\begin{aligned}
\sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{v}^t - \widetilde{\nabla V}(\boldsymbol{\theta}^t) \right\|_2^2 \right\} \le \frac{n\bar{\nu}^2}{\beta B} + 2n\beta T\bar{\nu}^2 \\
+ \frac{12n\eta^2\Phi^2}{\beta} \sum_{t=0}^{T-1} \mathbb{E}\left\{ \left\| \bar{\boldsymbol{d}}^t \right\|_2^2 \right\} + \frac{24\Phi^2}{\beta} \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t \right\|_2^2 \right\}.
\end{aligned} \tag{37}
$$

4.1 Proof of Theorem 9

Lemma 13 implies that

$$
\begin{aligned}
&\frac{1}{n} \sum_{t=0}^{T} \sum_{i=1}^{n} \mathbb{E}\left\{ \left\| \nabla V(\boldsymbol{\theta}_i^t) \right\|_2^2 \right\} \\
&\le \frac{8G^2\Delta}{\eta\mu_F} - \frac{G^2}{n} \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{d}^t \right\|_2^2 \right\} + \frac{10G^2}{n\mu_F^2} \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1} \right\|_2^2 \right\} \\
&\quad + \frac{82G^2L^2}{n\mu_F^2} \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t \right\|_2^2 \right\} + \frac{76G^2}{\mu_F^2} \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \overline{\nabla V}(\boldsymbol{\theta}^t) - \bar{\boldsymbol{v}}^t \right\|_2^2 \right\} \\
&\overset{(a)}{\le} \frac{8G^2\Delta}{\eta\mu_F} - \frac{G^2}{n} \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{d}^t \right\|_2^2 \right\} + \frac{10G^2}{n\mu_F^2} \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1} \right\|_2^2 \right\} \\
&\quad + \frac{82G^2L^2}{n\mu_F^2} \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t \right\|_2^2 \right\} + \frac{76G^2}{\mu_F^2} \Bigg(\frac{\bar{\nu}^2}{n\beta B} + \frac{2\beta\bar{\nu}^2 T}{n} \\
&\qquad + \frac{12\Phi^2\eta^2}{n\beta} \sum_{t=0}^{T-1} \mathbb{E}\left\{ \left\| \bar{\boldsymbol{d}}^t \right\|_2^2 \right\} + \frac{24\Phi^2}{\beta n^2} \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t \right\|_2^2 \right\} \Bigg) \\
&= \frac{8G^2\Delta}{\eta\mu_F} + \frac{76\bar{\nu}^2 G^2}{n\beta B\mu_F^2} + \frac{152\beta T\bar{\nu}^2 G^2}{n\mu_F^2} - \frac{G^2}{n} \mathbb{E}\left\{ \sum_{t=0}^{T} \left\| \boldsymbol{d}^t \right\|_2^2 \right\} + \frac{912\Phi^2\eta^2 G^2}{n\beta\mu_F^2} \sum_{t=0}^{T-1} \mathbb{E}\left\{ \left\| \bar{\boldsymbol{d}}^t \right\|_2^2 \right\} \\
&\quad + \left(\frac{1824\Phi^2 G^2}{\beta n^2 \mu_F^2} + \frac{82G^2L^2}{n\mu_F^2} \right) \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t \right\|_2^2 \right\} \\
&\quad + \frac{10G^2}{n\mu_F^2} \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1} \right\|_2^2 \right\} \\
&\overset{(b)}{\le} \frac{8G^2\Delta}{\eta\mu_F} + \frac{76\bar{\nu}^2 G^2}{n\beta B\mu_F^2} + \frac{152\beta T\bar{\nu}^2 G^2}{n\mu_F^2} - \frac{G^2}{n} \sum_{t=0}^{T} \mathbb{E}\left\{ \left\| \boldsymbol{d}^t \right\|_2^2 \right\} + \frac{912\Phi^2\eta^2 G^2}{n^2\beta\mu_F^2} \sum_{t=0}^{T-1} \mathbb{E}\left\{ \left\| \boldsymbol{d}^t \right\|_2^2 \right\}
\end{aligned}
$$

$$+\left(\frac{1824\Phi^2G^2}{\beta n^2\mu_F^2}+\frac{82G^2L^2}{n\mu_F^2}\right)\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\right\}$$
$$+\frac{10G^2}{n\mu_F^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{y}^{t+1}-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2\right\}, \tag{38}$$

where step (a) follows from (36) and step (b) is due to $\left\|\bar{\boldsymbol{d}}\right\|_2^2 \le \frac{1}{n}\left\|\boldsymbol{d}\right\|_2^2$. Since $0<\eta<\frac{\mu_F(1-\rho^2)^3}{\kappa_F\sqrt{1632000(L^2+\Phi^2)}}$ and $\frac{1632000(L^2+\Phi^2)\kappa_F^2\eta^2}{n\mu_F^2(1-\rho^2)^6}\le\beta<\frac{1}{n}$, it implies that

$$\begin{aligned}
\frac{912\Phi^2\eta^2G^2}{n^2\beta\mu_F^2} &\le \frac{912\Phi^2\eta^2G^2}{n^2\mu_F^2}\cdot\frac{n\mu_F^2(1-\rho^2)^6}{1632000(L^2+\Phi^2)\kappa_F^2\eta^2}\le\frac{G^2}{2n},\\
\frac{1824\Phi^2G^2}{\beta n^2\mu_F^2}+\frac{82G^2L^2}{n\mu_F^2} &\le \frac{1824(L^2+\Phi^2)G^2}{n\mu_F^2}\left(1+\frac{1}{\beta n}\right)\\
&\le \frac{1824(L^2+\Phi^2)G^2}{n\mu_F^2}\cdot\frac{2}{\beta n}\\
&\le \frac{1824(L^2+\Phi^2)G^2}{n\mu_F^2}\cdot\frac{2(1-\rho^2)^6\mu_F^2}{1632000(L^2+\Phi^2)\kappa_F^2\eta^2}\\
&\le \frac{(1-\rho^2)^6G^2}{100n\kappa_F^2\eta^2}.
\end{aligned}$$

Plugging these inequalities into (38) yields that

$$\begin{aligned}
&\frac{1}{n}\sum_{t=0}^{T}\sum_{i=1}^{n}\mathbb{E}\left\{\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2\right\}\\
&\le \frac{8G^2\Delta}{\eta\mu_F}+\frac{76\bar{\nu}^2G^2}{n\beta B\mu_F^2}+\frac{152\beta T\bar{\nu}^2G^2}{n\mu_F^2}-\frac{G^2}{2n}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\}\\
&\quad+\frac{(1-\rho^2)^6G^2}{100n\kappa_F^2\eta^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\right\}+\frac{10G^2}{n\mu_F^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{y}^{t+1}-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2\right\}\\
&\overset{(a)}{\le} \frac{8G^2\Delta}{\eta\mu_F}+\frac{76\bar{\nu}^2G^2}{n\beta B\mu_F^2}+\frac{152\beta T\bar{\nu}^2G^2}{n\mu_F^2}-\frac{G^2}{2n}\mathbb{E}\left\{\sum_{t=0}^{T}\left\|\boldsymbol{d}^t\right\|_2^2\right\}\\
&\quad+\frac{(1-\rho^2)^6G^2}{100n\kappa_F^2\eta^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\right\}\\
&\quad+\frac{10G^2}{n\mu_F^2}\left(A_1+\frac{1632\Phi^2}{(1-\rho^2)^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\right\}+\frac{960\Phi^2\kappa_F^2\eta^2}{(1-\rho^2)^2}\sum_{t=0}^{T-1}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\}\right)\\
&=\frac{8G^2\Delta}{\eta\mu_F}+\frac{76\bar{\nu}^2G^2}{n\beta B\mu_F^2}+\frac{152\beta T\bar{\nu}^2G^2}{n\mu_F^2}+\frac{10A_1G^2}{n\mu_F^2}\\
&\quad-\left(\frac{G^2}{2n}-\frac{10G^2}{n\mu_F^2}\cdot\frac{960\Phi^2\kappa_F^2\eta^2}{(1-\rho^2)^2}\right)\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\}
\end{aligned}$$

$$+\left(\frac{(1-\rho^2)^6G^2}{100n\kappa_F^2\eta^2}+\frac{16320G^2\Phi^2}{n\mu_F^2(1-\rho^2)^2}\right)\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\right\}, \tag{39}$$

where step (a) follows from (35). Using the conditions for η and β again gives that

$$\frac{10G^2}{n\mu_F^2}\cdot\frac{960\Phi^2\kappa_F^2\eta^2}{(1-\rho^2)^2}\le\frac{10G^2}{n\mu_F^2}\cdot\frac{960\Phi^2\kappa_F^2}{(1-\rho^2)^2}\cdot\frac{\mu_F^2(1-\rho^2)^6}{\kappa_F^2\cdot 1632000(L^2+\Phi^2)}\le\frac{G^2}{4n},$$

$$\begin{aligned}
\frac{(1-\rho^2)^6G^2}{100n\kappa_F^2\eta^2}+\frac{16320G^2\Phi^2}{n\mu_F^2(1-\rho^2)^2}&=\frac{(1-\rho^2)^6G^2}{100n\kappa_F^2\eta^2}+\frac{16320\kappa_F^2\Phi^2}{n(1-\rho^2)^2}\\
&\le\frac{(1-\rho^2)^6G^2}{100n\kappa_F^2\eta^2}+\frac{16320\kappa_F^2(L^2+\Phi^2)}{n(1-\rho^2)^2}\\
&\le\frac{(1-\rho^2)^6G^2}{100n\kappa_F^2\eta^2}+\frac{16320}{n(1-\rho^2)^2}\cdot\frac{\mu_F^2(1-\rho^2)^6}{1632000\eta^2}\\
&=\frac{(1-\rho^2)^6G^2}{100n\kappa_F^2\eta^2}+\frac{G^2(1-\rho^2)^4}{100n\kappa_F^2\eta^2}\\
&\le\frac{(1-\rho^2)^4G^2}{50n\kappa_F^2\eta^2}.
\end{aligned}$$

Substituting these inequalities into (39) leads to

$$\begin{aligned}
&\frac{1}{n}\sum_{t=0}^{T}\sum_{i=1}^{n}\mathbb{E}\left\{\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2\right\}\\
&\le\frac{8G^2\Delta}{\eta\mu_F}+\frac{76\bar{\nu}^2G^2}{n\beta B\mu_F^2}+\frac{152\beta T\bar{\nu}^2G^2}{n\mu_F^2}+\frac{10A_1G^2}{n\mu_F^2}-\frac{G^2}{4n}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\}\\
&\quad+\frac{(1-\rho^2)^4G^2}{50n\kappa_F^2\eta^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\right\}\\
&\le\frac{8G^2\Delta}{\eta\mu_F}+\frac{76\bar{\nu}^2G^2}{n\beta B\mu_F^2}+\frac{152\beta T\bar{\nu}^2G^2}{n\mu_F^2}+\frac{10A_1G^2}{n\mu_F^2}-\frac{G^2}{4n}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\}\\
&\quad+\frac{(1-\rho^2)^4G^2}{50n\kappa_F^2\eta^2}\left(\frac{16A_1\eta^2}{\mu_F^2(1-\rho^2)^2}+\frac{10G^2\eta^2}{\mu_F^2(1-\rho^2)^3}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\}\right)\\
&\le\frac{8G^2\Delta}{\eta\mu_F}+\frac{76\bar{\nu}^2G^2}{n\beta B\mu_F^2}+\frac{152\beta T\bar{\nu}^2G^2}{n\mu_F^2}+\frac{10A_1G^2}{n\mu_F^2}+\frac{16A_1G^2}{50n\mu_F^2}\\
&\quad-\left(\frac{G^2}{4n}-\frac{G^2}{5n}\right)\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\}\\
&\le\frac{8G^2\Delta}{\eta\mu_F}+\frac{76\bar{\nu}^2G^2}{n\beta B\mu_F^2}+\frac{152\beta T\bar{\nu}^2G^2}{n\mu_F^2}+\frac{11A_1G^2}{n\mu_F^2},
\end{aligned}$$

where the second inequality follows from Lemma 18. Since $\boldsymbol{\theta}_{\text{out}}$ is sampled uniformly from $\{\boldsymbol{\theta}_i^t\}_{i=1,\ldots,n;t=0,\ldots,T}$, we have

$$\mathbb{E}\left\{\left\|\nabla V(\boldsymbol{\theta}_{\text{out}})\right\|_2^2\right\}$$

$$
\begin{aligned}
&= \frac{1}{n(T+1)} \sum_{t=0}^{T} \sum_{i=1}^{n} \mathbb{E}\left\{ \left\| \nabla V(\boldsymbol{\theta}_i^t) \right\|_2^2 \right\} \\
&\leq \frac{1}{nT} \sum_{t=0}^{T} \sum_{i=1}^{n} \mathbb{E}\left\{ \left\| \nabla V(\boldsymbol{\theta}_i^t) \right\|_2^2 \right\} \\
&\leq \frac{8G^2\Delta}{T\eta\mu_F} + \frac{76\bar{\nu}^2 G^2}{nT\beta B \mu_F^2} + \frac{152\beta\bar{\nu}^2 G^2}{n\mu_F^2} + \frac{11A_1G^2}{nT\mu_F^2} \\
&= \frac{8G^2\Delta}{T\eta\mu_F} + \frac{76\bar{\nu}^2\kappa_F^2}{nT\beta B} + \frac{152\beta\bar{\nu}^2\kappa_F^2}{n} \\
&\quad + \frac{11G^2}{nT\mu_F^2}\left(\frac{4n\rho^2\bar{\nu}^2}{B(1-\rho^2)} + \frac{32nT\beta^2\bar{\nu}^2}{(1-\rho^2)^2} + \frac{32n\beta\bar{\nu}^2}{B(1-\rho^2)^2} + \frac{64nT\bar{\nu}^2\beta^3}{(1-\rho^2)^2} + \frac{4\rho^2}{1-\rho^2}\left\| \widetilde{\nabla V}(\boldsymbol{\theta}^0) \right\|_2^2 \right) \\
&= \frac{8G^2\Delta}{T\eta\mu_F} + \frac{76\bar{\nu}^2\kappa_F^2}{nT\beta B} + \frac{152\beta\bar{\nu}^2\kappa_F^2}{n} \\
&\quad + \frac{44\rho^2\bar{\nu}^2\kappa_F^2}{TB(1-\rho^2)} + \frac{352\beta^2\bar{\nu}^2\kappa_F^2}{(1-\rho^2)^2} + \frac{352\beta\bar{\nu}^2\kappa_F^2}{TB(1-\rho^2)^2} + \frac{704\bar{\nu}^2\kappa_F^2\beta^3}{(1-\rho^2)^2} + \frac{44\rho^2\kappa_F^2}{nT(1-\rho^2)}\left\| \widetilde{\nabla V}(\boldsymbol{\theta}^0) \right\|_2^2,
\end{aligned}
$$

which completes the proof of the main result.

4.2 Proof of Corollary 11

Since $\eta = \frac{\mu_F n^{2/3}}{\kappa_F\sqrt{L^2+\Phi}T^{1/3}}, \beta = \frac{n^{1/3}}{T^{2/3}}$ and $B = \left\lceil \frac{T^{1/3}}{n^{2/3}} \right\rceil$, we have

$$
\begin{aligned}
\frac{8\Delta G^2}{\eta T\mu_F} &= \frac{8\Delta G^2}{T\mu_F} \cdot \frac{\kappa_F\sqrt{L^2+\Phi}T^{1/3}}{\mu_F n^{2/3}} = \frac{8\Delta\kappa_F^3\sqrt{L^2+\Phi}}{(nT)^{2/3}}, \\
\frac{76\bar{\nu}^2\kappa_F^2}{n\beta TB} &\leq 76\bar{\nu}^2\kappa_F^2 \cdot \frac{1}{nT} \cdot \frac{T^{2/3}}{n^{1/3}} \cdot \frac{n^{2/3}}{T^{1/3}} = \frac{76\bar{\nu}^2\kappa_F^2}{(nT)^{2/3}}, \\
\frac{152\beta\bar{\nu}^2\kappa_F^2}{n} &= \frac{152\bar{\nu}^2\kappa_F^2}{(nT)^{2/3}}, \\
\frac{44\rho^2\bar{\nu}^2\kappa_F^2}{BT(1-\rho^2)} &\leq \frac{44\rho^2\bar{\nu}^2\kappa_F^2}{(1-\rho^2)^2} \cdot \frac{n^{2/3}}{T^{4/3}}, \\
\frac{352\beta^2\bar{\nu}^2\kappa_F^2}{(1-\rho^2)^2} &\leq \frac{352\bar{\nu}^2\kappa_F^2}{(1-\rho^2)^2} \cdot \frac{n^{2/3}}{T^{4/3}}, \\
\frac{352\beta\bar{\nu}^2\kappa_F^2}{BT(1-\rho^2)^2} &\leq \frac{352\bar{\nu}^2\kappa_F^2}{(1-\rho^2)^2} \cdot \frac{1}{T} \cdot \frac{n^{1/3}}{T^{2/3}} \cdot \frac{n^{2/3}}{T^{1/3}} = \frac{352\bar{\nu}^2\kappa_F^2}{(1-\rho^2)^2} \cdot \frac{n}{T^2}, \\
\frac{704\bar{\nu}^2\kappa_F^2\beta^3}{(1-\rho^2)^2} &= \frac{704\bar{\nu}^2\kappa_F^2}{(1-\rho^2)^2} \cdot \frac{n}{T^2}.
\end{aligned}
$$

Thus it can be seen that

$$
\begin{aligned}
\mathbb{E}\left\{ \left\| \nabla V(\boldsymbol{\theta}_{\text{out}}) \right\|_2^2 \right\} \leq \frac{8\Delta\kappa_F^3\sqrt{L^2+\Phi} + 228\bar{\nu}^2\kappa_F^2}{(nT)^{2/3}} &+ \frac{44\rho^2\kappa_F^2\left\| \widetilde{\nabla V}(\boldsymbol{\theta}^0) \right\|_2^2}{(1-\rho^2)} \cdot \frac{1}{nT} \\
&+ \frac{396\kappa_F^2\bar{\nu}^2}{(1-\rho^2)^2} \cdot \frac{n^{2/3}}{T^{4/3}} + \frac{1056\bar{\nu}^2\kappa_F^2}{(1-\rho^2)^2} \cdot \frac{n}{T^2},
\end{aligned}
$$

which completes the proof of the corollary.

5. Proofs

5.1 Proof of Lemma 1

For simplicity, let $\boldsymbol{\theta}_i = [\boldsymbol{x}^{1\mathsf{T}}, \cdots, \boldsymbol{x}^{n\mathsf{T}}]^\mathsf{T} \in \mathbb{R}^d$, where $\boldsymbol{x}^j \in \mathbb{R}^{d_j}$ and $d = \sum_{j=1}^n d_j$. From the definition (5) of the policy in collaborative RL, we have

$$\nabla_{\boldsymbol{\theta}_i} \log \pi_{\boldsymbol{\theta}_i}(\boldsymbol{a}^h|\boldsymbol{s}^h) = \nabla_{\boldsymbol{\theta}_i} \sum_{j=1}^n \log \pi_{\boldsymbol{x}^j}(\boldsymbol{a}_j^h|\boldsymbol{s}^h) = \begin{bmatrix} \nabla_{\boldsymbol{x}^1} \log \pi_{\boldsymbol{x}^1}(\boldsymbol{a}_1^h|\boldsymbol{s}^h) \\ \vdots \\ \nabla_{\boldsymbol{x}^n} \log \pi_{\boldsymbol{x}^n}(\boldsymbol{a}_n^h|\boldsymbol{s}^h) \end{bmatrix} \in \mathbb{R}^{d\times 1}.$$

Then the (j,ℓ)-th block of $\boldsymbol{F}_i(\boldsymbol{\theta}_i)$ is given by

$$[\boldsymbol{F}_i(\boldsymbol{\theta}_i)]_{j,\ell} = \mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta}_i)} \left\{ \frac{1}{H} \sum_{h=0}^{H-1} \nabla_{\boldsymbol{x}^j} \log \pi_{\boldsymbol{x}^j}(\boldsymbol{a}_j^h|\boldsymbol{s}^h) \left(\nabla_{\boldsymbol{x}^\ell} \log \pi_{\boldsymbol{x}^\ell}(\boldsymbol{a}_\ell^h|\boldsymbol{s}^h)\right)^\mathsf{T} \right\} \in \mathbb{R}^{d_j \times d_\ell}.$$

We will show that $[\boldsymbol{F}_i(\boldsymbol{\theta}_i)]_{j,\ell} = \boldsymbol{0}$ for any $j \neq \ell$. To this end, for any $\alpha \in [d_j], \beta \in [d_\ell]$, one has

$$\begin{aligned}
&\left[\mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta}_i)} \left\{ \frac{1}{H} \sum_{h=0}^{H-1} \nabla_{\boldsymbol{x}^j} \log \pi_{\boldsymbol{x}^j}(\boldsymbol{a}_j^h|\boldsymbol{s}^h) \left(\nabla_{\boldsymbol{x}^\ell} \log \pi_{\boldsymbol{x}^\ell}(\boldsymbol{a}_\ell^h|\boldsymbol{s}^h)\right)^\mathsf{T} \right\}\right]_{\alpha,\beta} \\
=&\frac{1}{H} \sum_{h=0}^{H-1} \left[\mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta}_i)} \left\{ \nabla_{\boldsymbol{x}^j} \log \pi_{\boldsymbol{x}^j}(\boldsymbol{a}_j^h|\boldsymbol{s}^h) \left(\nabla_{\boldsymbol{x}^\ell} \log \pi_{\boldsymbol{x}^\ell}(\boldsymbol{a}_\ell^h|\boldsymbol{s}^h)\right)^T \right\}\right]_{\alpha,\beta} \\
=&\frac{1}{H} \sum_{h=0}^{H-1} \mathbb{E}_{\tau\sim p(\cdot|\boldsymbol{\theta}_i)} \left\{ \frac{\partial \log \pi_{\boldsymbol{x}^j}(\boldsymbol{a}_j^h|\boldsymbol{s}^h)}{\partial \boldsymbol{x}_\alpha^j} \frac{\partial \log \pi_{\boldsymbol{x}^\ell}(\boldsymbol{a}_\ell^h|\boldsymbol{s}^h)}{\partial \boldsymbol{x}_\beta^\ell} \right\} \\
=&\frac{1}{H} \sum_{h=0}^{H-1} \int p(\tau|\boldsymbol{\theta}_i) \frac{\partial \log \pi_{\boldsymbol{x}^j}(\boldsymbol{a}_j^h|\boldsymbol{s}^h)}{\partial \boldsymbol{x}_\alpha^j} \frac{\partial \log \pi_{\boldsymbol{x}^\ell}(\boldsymbol{a}_\ell^h|\boldsymbol{s}^h)}{\partial \boldsymbol{x}_\beta^\ell} d\tau \\
=&\frac{1}{H} \sum_{h=0}^{H-1} \int p(\tau^{-h}) \cdot \pi_{\boldsymbol{\theta}_i}(\boldsymbol{a}^h|\boldsymbol{s}^h) \frac{\partial \log \pi_{\boldsymbol{x}^j}(\boldsymbol{a}_j^h|\boldsymbol{s}^h)}{\partial \boldsymbol{x}_\alpha^j} \frac{\partial \log \pi_{\boldsymbol{x}^\ell}(\boldsymbol{a}_\ell^h|\boldsymbol{s}^h)}{\partial \boldsymbol{x}_\beta^\ell} d\tau \\
=&\frac{1}{H} \sum_{h=0}^{H-1} \int p(\tau^{-h}) \prod_{i=1}^n \pi_{\boldsymbol{x}^i}(\boldsymbol{a}_i^h|\boldsymbol{s}^h) \frac{\partial \log \pi_{\boldsymbol{x}^j}(\boldsymbol{a}_j^h|\boldsymbol{s}^h)}{\partial \boldsymbol{x}_\alpha^j} \frac{\partial \log \pi_{\boldsymbol{x}^\ell}(\boldsymbol{a}_\ell^h|\boldsymbol{s}^h)}{\partial \boldsymbol{x}_\beta^\ell} d\tau \\
=&\frac{1}{H} \sum_{h=0}^{H-1} \int p(\tau^{-h}) \prod_{i\neq j,\ell}^n \pi_{\boldsymbol{x}^i}(\boldsymbol{a}_i^h|\boldsymbol{s}^h) \cdot \pi_{\boldsymbol{x}^j}(\boldsymbol{a}_j^h|\boldsymbol{s}^h) \frac{\partial \log \pi_{\boldsymbol{x}^j}(\boldsymbol{a}_j^h|\boldsymbol{s}^h)}{\partial \boldsymbol{x}_\alpha^j} \\
&\qquad \cdot \pi_{\boldsymbol{x}^\ell}(\boldsymbol{a}_\ell^h|\boldsymbol{s}^h) \frac{\partial \log \pi_{\boldsymbol{x}^\ell}(\boldsymbol{a}_\ell^h|\boldsymbol{s}^h)}{\partial \boldsymbol{x}_\beta^\ell} d\tau \\
=&\frac{1}{H} \sum_{h=0}^{H-1} \int p(\tau^{-h}) \prod_{i\neq j,\ell}^n \pi_{\boldsymbol{x}^i}(\boldsymbol{a}_i^h|\boldsymbol{s}^h) \cdot \frac{\partial \pi_{\boldsymbol{x}^j}(\boldsymbol{a}_j^h|\boldsymbol{s}^h)}{\partial \boldsymbol{x}_\alpha^j} \cdot \frac{\partial \pi_{\boldsymbol{x}^\ell}(\boldsymbol{a}_\ell^h|\boldsymbol{s}^h)}{\partial \boldsymbol{x}_\beta^\ell} d\tau
\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{H}\sum_{h=0}^{H-1}\frac{\partial^2}{\partial \boldsymbol{x}_\alpha^j \partial \boldsymbol{x}_\beta^\ell}\int p(\tau) d\tau \\
=&0,
\end{aligned}
$$

where $p(\tau^{-h}) := \rho(\boldsymbol{s}^0)\prod_{h'\neq h}\pi_{\boldsymbol{\theta}}(\boldsymbol{a}^{h'}|\boldsymbol{s}^{h'})P(\boldsymbol{s}^{h'+1}|\boldsymbol{s}^{h'},\boldsymbol{a}^{h'})\cdot P(\boldsymbol{s}^{h+1}|\boldsymbol{s}^h,\boldsymbol{a}^h)$. Thus we complete the proof.

5.2 Proof of Lemma 13

Since the objective function V is L-smooth, one has

$$
\begin{aligned}
V(\bar{\boldsymbol{\theta}}^{t+1}) &\geq V(\bar{\boldsymbol{\theta}}^t) + \left\langle \nabla V(\bar{\boldsymbol{\theta}}^t), \bar{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^t \right\rangle - \frac{L}{2}\left\|\bar{\boldsymbol{\theta}}^{t+1} - \bar{\boldsymbol{\theta}}^t\right\|_2^2 \\
&= V(\bar{\boldsymbol{\theta}}^t) + \eta\left\langle \nabla V(\bar{\boldsymbol{\theta}}^t), \bar{\boldsymbol{d}}^t\right\rangle - \frac{L\eta^2}{2}\left\|\bar{\boldsymbol{d}}^t\right\|_2^2, \qquad (40)
\end{aligned}
$$

where the second line follows from $\bar{\boldsymbol{\theta}}^{t+1} = \bar{\boldsymbol{\theta}}^t + \eta\bar{\boldsymbol{d}}^t$. Moreover, for any $i \in [n]$, one has

$$
\begin{aligned}
\eta\mu_F\left\|\boldsymbol{d}_i^t\right\|_2^2 &\overset{(a)}{\leq} \eta\left\langle \boldsymbol{H}_i^{t-1}\boldsymbol{d}_i^t, \boldsymbol{d}_i^t\right\rangle \\
&= \eta\left\langle \boldsymbol{y}_i^{t+1}, \boldsymbol{d}_i^t\right\rangle \\
&= \eta\left\langle \boldsymbol{y}_i^{t+1} - \bar{\boldsymbol{y}}^{t+1}, \boldsymbol{d}_i^t\right\rangle + \eta\left\langle \bar{\boldsymbol{y}}^{t+1}, \boldsymbol{d}_i^t\right\rangle \\
&\leq \eta\left\|\boldsymbol{y}_i^{t+1} - \bar{\boldsymbol{y}}^{t+1}\right\|_2\cdot\left\|\boldsymbol{d}_i^t\right\|_2 + \eta\left\langle \bar{\boldsymbol{y}}^{t+1}, \boldsymbol{d}_i^t\right\rangle \\
&\overset{(b)}{\leq} \frac{\eta}{2\mu_F}\left\|\boldsymbol{y}_i^{t+1} - \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 + \frac{\eta\mu_F}{2}\left\|\boldsymbol{d}_i^t\right\|_2^2 + \eta\left\langle \bar{\boldsymbol{y}}^{t+1}, \boldsymbol{d}_i^t\right\rangle \\
&= \frac{\eta}{2\mu_F}\left\|\boldsymbol{y}_i^{t+1} - \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 + \frac{\eta\mu_F}{2}\left\|\boldsymbol{d}_i^t\right\|_2^2 + \eta\left\langle \bar{\boldsymbol{v}}^t, \boldsymbol{d}_i^t\right\rangle, \qquad (41)
\end{aligned}
$$

where step (a) is due to (25), step (b) uses the elementary inequality that $x\cdot y \leq \frac{1}{2\alpha}x^2 + \frac{\alpha}{2}y^2$ with $\alpha = \mu_F$, and the last line follows from $\bar{\boldsymbol{y}}^{t+1} = \bar{\boldsymbol{v}}^t$. Rearranging (41) yields that

$$
0 \geq -\eta\left\langle \bar{\boldsymbol{v}}^t, \boldsymbol{d}_i^t\right\rangle + \frac{\eta\mu_F}{2}\left\|\boldsymbol{d}_i^t\right\|_2^2 - \frac{\eta}{2\mu_F}\left\|\boldsymbol{y}_i^{t+1} - \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 \qquad (42)
$$

holds for any fixed $i \in [n]$. Taking an average over i from 1 to n yields that

$$
\begin{aligned}
0 &\geq -\frac{\eta}{n}\sum_{i=1}^n\left\langle \bar{\boldsymbol{v}}^t, \boldsymbol{d}_i^t\right\rangle + \frac{\eta\mu_F}{2n}\sum_{i=1}^n\left\|\boldsymbol{d}_i^t\right\|_2^2 - \frac{\eta}{2n\mu_F}\sum_{i=1}^n\left\|\boldsymbol{y}_i^{t+1} - \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 \\
&= -\eta\left\langle \bar{\boldsymbol{v}}^t, \bar{\boldsymbol{d}}^t\right\rangle + \frac{\eta\mu_F}{2n}\left\|\boldsymbol{d}^t\right\|_2^2 - \frac{\eta}{2n\mu_F}\left\|\boldsymbol{y}^{t+1} - \boldsymbol{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2. \qquad (43)
\end{aligned}
$$

Summing up (40) and (43), we obtain that

$$
\begin{aligned}
V(\bar{\boldsymbol{\theta}}^{t+1}) \geq\ & V(\bar{\boldsymbol{\theta}}^t) + \eta\left\langle \nabla V(\bar{\boldsymbol{\theta}}^t) - \bar{\boldsymbol{v}}^t, \bar{\boldsymbol{d}}^t\right\rangle - \frac{\eta}{2n\mu_F}\left\|\boldsymbol{y}^{t+1} - \boldsymbol{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2 \\
&+ \frac{\eta\mu_F}{2n}\left\|\boldsymbol{d}^t\right\|_2^2 - \frac{L\eta^2}{2}\left\|\bar{\boldsymbol{d}}^t\right\|_2^2
\end{aligned}
$$

$$\begin{aligned}
&\overset{(a)}{\geq} V(\bar{\boldsymbol{\theta}}^t) - \frac{\eta}{2\gamma}\left\|\nabla V(\bar{\boldsymbol{\theta}}^t) - \bar{\boldsymbol{v}}^t\right\|_2^2 - \frac{\gamma\eta}{2}\left\|\bar{\boldsymbol{d}}^t\right\|_2^2 - \frac{\eta}{2n\mu_F}\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 \\
&\quad + \frac{\eta\mu_F}{2n}\left\|\boldsymbol{d}^t\right\|_2^2 - \frac{L\eta^2}{2}\left\|\bar{\boldsymbol{d}}^t\right\|_2^2 \\
&\overset{(b)}{\geq} V(\bar{\boldsymbol{\theta}}^t) - \frac{\eta}{2\gamma}\left\|\nabla V(\bar{\boldsymbol{\theta}}^t) - \bar{\boldsymbol{v}}^t\right\|_2^2 - \frac{\eta}{2n\mu_F}\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 \\
&\quad - \frac{\gamma\eta + L\eta^2}{2n}\left\|\boldsymbol{d}^t\right\|_2^2 + \frac{\eta\mu_F}{2n}\left\|\boldsymbol{d}^t\right\|_2^2 \\
&= V(\bar{\boldsymbol{\theta}}^t) - \frac{\eta}{2\gamma}\left\|\nabla V(\bar{\boldsymbol{\theta}}^t) - \bar{\boldsymbol{v}}^t\right\|_2^2 - \frac{\eta}{2n\mu_F}\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 \\
&\quad + \frac{\eta\mu_F - 2\gamma\eta - 2L\eta^2}{4n}\left\|\boldsymbol{d}^t\right\|_2^2 + \frac{\eta\mu_F}{4n}\left\|\boldsymbol{d}^t\right\|_2^2 \\
&\overset{(c)}{\geq} V(\bar{\boldsymbol{\theta}}^t) - \frac{4\eta}{\mu_F}\left\|\nabla V(\bar{\boldsymbol{\theta}}^t) - \bar{\boldsymbol{v}}^t\right\|_2^2 - \frac{\eta}{2n\mu_F}\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 \\
&\quad + \frac{\eta\mu_F}{8n}\left\|\boldsymbol{d}^t\right\|_2^2 + \frac{\eta\mu_F}{4n}\left\|\boldsymbol{d}^t\right\|_2^2, \qquad (44)
\end{aligned}$$

where step (a) follows from the elementary inequality that $\langle \boldsymbol{a}, \boldsymbol{b}\rangle \leq \frac{1}{2\gamma}\|\boldsymbol{a}\|_2^2 + \frac{\gamma}{2}\|\boldsymbol{b}\|_2^2$ with $\gamma > 0$ for any $\boldsymbol{a}$ and $\boldsymbol{b}$, step (b) is due to $\left\|\bar{\boldsymbol{d}}^t\right\|_2^2 \leq \frac{1}{n}\left\|\boldsymbol{d}^t\right\|_2^2$ and step (c) holds by choosing $\gamma = \frac{\mu_F}{8}$ and assuming $0 < \eta \leq \frac{\mu_F}{8L}$ (i.e, $\frac{\eta\mu_F - 2\gamma\eta - 2L\eta^2}{4n} = \frac{\eta\mu_F - \frac{1}{4}\eta\mu_F - 2L\eta^2}{4n} = \frac{\eta}{4n}\left(\frac{3\mu_F}{4} - 2L\eta\right) \geq \frac{\eta\mu_F}{8n}$). Moreover, the fact used in step (b) can be proved as follows:

$$\left\|\bar{\boldsymbol{d}}^t\right\|_2^2 = \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{d}_i^t\right\|_2^2 \leq \frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{d}_i^t\right\|_2^2 = \frac{1}{n}\left\|\boldsymbol{d}^t\right\|_2^2,$$

where we have used the Jensen's inequality.

Notice that

$$\frac{1}{G^2}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 \leq \left\|\boldsymbol{H}_i^t \nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 \leq 2\left\|\boldsymbol{H}_i^t \nabla V(\boldsymbol{\theta}_i^t) - \boldsymbol{d}_i^t\right\|_2^2 + 2\left\|\boldsymbol{d}_i^t\right\|_2^2$$

holds for any $i \in [n]$. A direct computation yields that

$$\begin{aligned}
&\frac{\eta\mu_F}{4n}\left\|\boldsymbol{d}^t\right\|_2^2 \\
&= \frac{\eta\mu_F}{4n}\sum_{i=1}^n \left\|\boldsymbol{d}_i^t\right\|_2^2 \\
&\geq \frac{\eta\mu_F}{4n}\sum_{i=1}^n \left(\frac{1}{2G^2}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 - \left\|\boldsymbol{H}_i^t \nabla V(\boldsymbol{\theta}_i^t) - \boldsymbol{d}_i^t\right\|_2^2\right) \\
&= \frac{\eta\mu_F}{8nG^2}\sum_{i=1}^n \left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 - \frac{\eta\mu_F}{4n}\sum_{i=1}^n \left\|\boldsymbol{H}_i^t\left(\boldsymbol{y}_i^{t+1} - \nabla V(\boldsymbol{\theta}_i^t)\right)\right\|_2^2 \\
&\overset{(a)}{\geq} \frac{\eta\mu_F}{8nG^2}\sum_{i=1}^n \left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 - \frac{\eta\mu_F}{4n}\sum_{i=1}^n \frac{1}{\mu_F^2}\left\|\boldsymbol{y}_i^{t+1} - \nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\eta\mu_F}{8nG^2}\sum_{i=1}^{n}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 \\
&\quad - \frac{\eta}{4n\mu_F}\sum_{i=1}^{n}\left\|\boldsymbol{y}_i^{t+1} - \bar{\boldsymbol{y}}^{t+1} + \bar{\boldsymbol{y}}^{t+1} - \nabla V(\bar{\boldsymbol{\theta}}^t) + \nabla V(\bar{\boldsymbol{\theta}}^t) - \nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 \\
&\overset{(b)}{\geq} \frac{\eta\mu_F}{8nG^2}\sum_{i=1}^{n}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 \\
&\quad - \frac{\eta}{4n\mu_F}\sum_{i=1}^{n}\left(3\left(\left\|\boldsymbol{y}_i^{t+1} - \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 + \left\|\bar{\boldsymbol{v}}^t - \nabla V(\bar{\boldsymbol{\theta}}^t)\right\|_2^2 + L^2\left\|\bar{\boldsymbol{\theta}}^t - \boldsymbol{\theta}_i^t\right\|_2^2\right)\right) \\
&= \frac{\eta\mu_F}{8nG^2}\sum_{i=1}^{n}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 - \frac{3\eta}{4n\mu_F}\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2 - \frac{3\eta}{4\mu_F}\left\|\bar{\boldsymbol{v}}^t - \nabla V(\bar{\boldsymbol{\theta}}^t)\right\|_2^2 \\
&\quad - \frac{3L^2\eta}{4n\mu_F}\left\|\boldsymbol{\theta}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2, \qquad (45)
\end{aligned}
$$

where step (a) follows from (25) and step (b) is due to the L-smoothness of V. Then plugging (45) into (44) yields that

$$
\begin{aligned}
&V(\bar{\boldsymbol{\theta}}^{t+1}) \\
&\geq V(\bar{\boldsymbol{\theta}}^t) - \frac{4\eta}{\mu_F}\left\|\nabla V(\bar{\boldsymbol{\theta}}^t) - \bar{\boldsymbol{v}}^t\right\|_2^2 - \frac{\eta}{2n\mu_F}\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2 \\
&\quad + \frac{\eta\mu_F}{8n}\left\|\boldsymbol{d}^t\right\|_2^2 + \frac{\eta\mu_F}{8nG^2}\sum_{i=1}^{n}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 \\
&\quad - \frac{3\eta}{4n\mu_F}\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2 - \frac{3\eta}{4\mu_F}\left\|\bar{\boldsymbol{v}}^t - \nabla V(\bar{\boldsymbol{\theta}}^t)\right\|_2^2 - \frac{3L^2\eta}{4n\mu_F}\left\|\boldsymbol{\theta}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2 \\
&= V(\bar{\boldsymbol{\theta}}^t) - \frac{19\eta}{4\mu_F}\left\|\nabla V(\bar{\boldsymbol{\theta}}^t) - \bar{\boldsymbol{v}}^t\right\|_2^2 + \frac{\eta\mu_F}{8n}\left\|\boldsymbol{d}^t\right\|_2^2 + \frac{\eta\mu_F}{8nG^2}\sum_{i=1}^{n}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 \\
&\quad - \frac{5\eta}{4n\mu_F}\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2 - \frac{3L^2\eta}{4n\mu_F}\left\|\boldsymbol{\theta}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2. \qquad (46)
\end{aligned}
$$

Furthermore, it can be seen that

$$
\begin{aligned}
\left\|\nabla V(\bar{\boldsymbol{\theta}}^t) - \bar{\boldsymbol{v}}^t\right\|_2^2 &= \left\|\nabla V(\bar{\boldsymbol{\theta}}^t) - \overline{\nabla V}(\boldsymbol{\theta}^t) + \overline{\nabla V}(\boldsymbol{\theta}^t) - \bar{\boldsymbol{v}}^t\right\|_2^2 \\
&\leq 2\left\|\nabla V(\bar{\boldsymbol{\theta}}^t) - \overline{\nabla V}(\boldsymbol{\theta}^t)\right\|_2^2 + 2\left\|\overline{\nabla V}(\boldsymbol{\theta}^t) - \bar{\boldsymbol{v}}^t\right\|_2^2 \\
&\leq \frac{2L^2}{n}\left\|\boldsymbol{\theta}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2 + 2\left\|\overline{\nabla V}(\boldsymbol{\theta}^t) - \bar{\boldsymbol{v}}^t\right\|_2^2, \qquad (47)
\end{aligned}
$$

where the last line follows from the fact that $\left\|\nabla V(\bar{\boldsymbol{\theta}}^t) - \overline{\nabla V}(\boldsymbol{\theta}^t)\right\|_2^2 \leq \frac{L^2}{n}\left\|\boldsymbol{\theta}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2$. Indeed, this fact can be proved as follows:

$$
\left\|\nabla V(\bar{\boldsymbol{\theta}}^t) - \overline{\nabla V}(\boldsymbol{\theta}^t)\right\|_2^2 = \left\|\frac{1}{n}\sum_{i=1}^{n}\left(\nabla V_i(\bar{\boldsymbol{\theta}}^t) - \nabla V_i(\boldsymbol{\theta}_i^t)\right)\right\|_2^2
$$

$$\begin{aligned}
&\leq \frac{1}{n}\sum_{i=1}^{n}\left\|\nabla V_i(\bar{\boldsymbol{\theta}}^t)-\nabla V_i(\boldsymbol{\theta}_i^t)\right\|_2^2\\
&\leq \frac{L^2}{n}\sum_{i=1}^{n}\left\|\boldsymbol{\theta}_i^t-\bar{\boldsymbol{\theta}}^t\right\|_2^2\\
&= \frac{L^2}{n}\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2,
\end{aligned}$$

where the second line is due to Jensen's inequality and the third line is due to L-smoothness of V_i. Thus, plugging (47) into (46) yields that

$$\begin{aligned}
V(\bar{\boldsymbol{\theta}}^{t+1}) \geq V(\bar{\boldsymbol{\theta}}^t) &- \frac{19\eta}{4\mu_F}\left(\frac{2L^2}{n}\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2+2\left\|\overline{\nabla V}(\boldsymbol{\theta}^t)-\bar{\boldsymbol{v}}^t\right\|_2^2\right)+\frac{\eta\mu_F}{8n}\left\|\boldsymbol{d}^t\right\|_2^2\\
&+\frac{\eta\mu_F}{8nG^2}\sum_{i=1}^{n}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2\\
&-\frac{5\eta}{4n\mu_F}\left\|\boldsymbol{y}^{t+1}-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2-\frac{3L^2\eta}{4n\mu_F}\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\\
= V(\bar{\boldsymbol{\theta}}^t) &- \frac{19\eta}{2\mu_F}\left\|\overline{\nabla V}(\boldsymbol{\theta}^t)-\bar{\boldsymbol{v}}^t\right\|_2^2+\frac{\eta\mu_F}{8n}\left\|\boldsymbol{d}^t\right\|_2^2+\frac{\eta\mu_F}{8nG^2}\sum_{i=1}^{n}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2\\
&-\frac{5\eta}{4n\mu_F}\left\|\boldsymbol{y}^{t+1}-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2-\frac{41L^2\eta}{4n\mu_F}\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2. \qquad (48)
\end{aligned}$$

Rearranging (48) yields that

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 \leq \frac{8G^2}{\eta\mu_F}\left(V(\bar{\boldsymbol{\theta}}^{t+1})-V(\bar{\boldsymbol{\theta}}^t)\right)-\frac{G^2}{n}\left\|\boldsymbol{d}^t\right\|_2^2+\frac{76G^2}{\mu_F^2}\left\|\overline{\nabla V}(\boldsymbol{\theta}^t)-\bar{\boldsymbol{v}}^t\right\|_2^2\\
+\frac{10G^2}{n\mu_F^2}\left\|\boldsymbol{y}^{t+1}-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2+\frac{82G^2L^2}{n\mu_F^2}\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2. \qquad (49)
\end{aligned}$$

Taking the telescoping sum of (49) over t from 0 to T for any $T \geq 0$, one has

$$\begin{aligned}
\frac{1}{n}\sum_{t=0}^{T}\sum_{i=1}^{n}\left\|\nabla V(\boldsymbol{\theta}_i^t)\right\|_2^2 &\leq \frac{8G^2}{\eta\mu_F}(V(\bar{\boldsymbol{\theta}}^{T+1})-V(\bar{\boldsymbol{\theta}}^0))+\frac{76G^2}{\mu_F^2}\sum_{t=0}^{T}\left\|\overline{\nabla V}(\boldsymbol{\theta}^t)-\bar{\boldsymbol{v}}^t\right\|_2^2\\
&\quad -\frac{G^2}{n}\sum_{t=0}^{T}\left\|\boldsymbol{d}^t\right\|_2^2+\frac{10G^2}{n\mu_F^2}\sum_{t=0}^{T}\left\|\boldsymbol{y}^{t+1}-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2\\
&\quad +\frac{82G^2L^2}{n\mu_F^2}\sum_{t=0}^{T}\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\\
&\leq \frac{8G^2}{\eta\mu_F}(V^\star-V(\bar{\boldsymbol{\theta}}^0))+\frac{76G^2}{\mu_F^2}\sum_{t=0}^{T}\left\|\overline{\nabla V}(\boldsymbol{\theta}^t)-\bar{\boldsymbol{v}}^t\right\|_2^2\\
&\quad -\frac{G^2}{n}\sum_{t=0}^{T}\left\|\boldsymbol{d}^t\right\|_2^2+\frac{10G^2}{n\mu_F^2}\sum_{t=0}^{T}\left\|\boldsymbol{y}^{t+1}-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2
\end{aligned}$$

$$+\frac{82G^2L^2}{n\mu_F^2}\sum_{t=0}^{T}\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2,$$

where the last line uses the fact that $V^\star<+\infty$ as we assume bounded rewards.

5.3 Proof of Lemma 14

A sample computation yields that

$$\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t+1}=\mathbf{1}_n\otimes\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\theta}_i^{t+1}\right)=\frac{1}{n}\mathbf{1}_n\otimes\left(\mathbf{1}_n^\mathsf{T}\otimes\boldsymbol{I}_d\right)\bar{\boldsymbol{\theta}}^{t+1}=\frac{1}{n}\left(\boldsymbol{J}_n\otimes\boldsymbol{I}_d\right)\boldsymbol{\theta}^{t+1}.$$

Thus by the update rule described in (27), it is straightforward to obtain that

$$\begin{aligned}
&\left\|\boldsymbol{\theta}^{t+1}-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t+1}\right\|_2^2\\
&=\left\|\left(\boldsymbol{W}\otimes\boldsymbol{I}_d\right)\left(\boldsymbol{\theta}^t+\eta\boldsymbol{d}^t\right)-\frac{1}{n}(\boldsymbol{J}_n\otimes\boldsymbol{I}_d)(\boldsymbol{\theta}^t+\eta\boldsymbol{d}^t)\right\|_2^2\\
&=\left\|\left(\left(\boldsymbol{W}-\frac{1}{n}\boldsymbol{J}_n\right)\otimes\boldsymbol{I}_d\right)\boldsymbol{\theta}^t+\eta\left(\left(\boldsymbol{W}-\frac{1}{n}\boldsymbol{J}_n\right)\otimes\boldsymbol{I}_d\right)\boldsymbol{d}^t\right\|_2^2\\
&\overset{(a)}{\le}\left(1+\frac{1-\rho^2}{2\rho^2}\right)\left\|\left(\left(\boldsymbol{W}-\frac{1}{n}\boldsymbol{J}_n\right)\otimes\boldsymbol{I}_d\right)\boldsymbol{\theta}^t\right\|_2^2\\
&\quad+\eta^2\left(1+\frac{2\rho^2}{1-\rho^2}\right)\left\|\left(\left(\boldsymbol{W}-\frac{1}{n}\boldsymbol{J}_n\right)\otimes\boldsymbol{I}_d\right)\boldsymbol{d}^t\right\|_2^2\\
&\overset{(b)}{=}\left(1+\frac{1-\rho^2}{2\rho^2}\right)\left\|\left(\left(\boldsymbol{W}-\frac{1}{n}\boldsymbol{J}_n\right)\otimes\boldsymbol{I}_d\right)(\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t)\right\|_2^2\\
&\quad+\eta^2\left(1+\frac{2\rho^2}{1-\rho^2}\right)\left\|\left(\left(\boldsymbol{W}-\frac{1}{n}\boldsymbol{J}_n\right)\otimes\boldsymbol{I}_d\right)(\boldsymbol{d}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{d}}^t)\right\|_2^2\\
&\le\left(1+\frac{1-\rho^2}{2\rho^2}\right)\cdot\left\|\boldsymbol{W}-\frac{1}{n}\boldsymbol{J}_n\right\|^2\cdot\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\\
&\quad+\eta^2\left(1+\frac{2\rho^2}{1-\rho^2}\right)\cdot\left\|\boldsymbol{W}-\frac{1}{n}\boldsymbol{J}_n\right\|^2\cdot\left\|\boldsymbol{d}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{d}}^t\right\|_2^2\\
&=\left(1+\frac{1-\rho^2}{2\rho^2}\right)\rho^2\cdot\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2+\eta^2\left(1+\frac{2\rho^2}{1-\rho^2}\right)\rho^2\left\|\boldsymbol{d}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{d}}^t\right\|_2^2\\
&=\frac{1+\rho^2}{2}\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2+\frac{(1+\rho^2)\rho^2\eta^2}{1-\rho^2}\left\|\boldsymbol{d}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{d}}^t\right\|_2^2\\
&\overset{(c)}{\le}\frac{1+\rho^2}{2}\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2+\frac{2\eta^2}{1-\rho^2}\left\|\boldsymbol{d}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{d}}^t\right\|_2^2\\
&\overset{(d)}{\le}\frac{1+\rho^2}{2}\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2+\frac{2\eta^2}{1-\rho^2}\cdot\left(\frac{G^2}{2\mu_F^2}\left\|\boldsymbol{d}^t\right\|_2^2+\frac{2}{\mu_F^2}\left\|\boldsymbol{y}^{t+1}-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2\right)\\
&=\frac{1+\rho^2}{2}\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2+\frac{4\eta^2}{\mu_F^2(1-\rho^2)}\left\|\boldsymbol{y}^{t+1}-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2+\frac{G^2\eta^2}{\mu_F^2(1-\rho^2)}\left\|\boldsymbol{d}^t\right\|_2^2,
\end{aligned}$$

where step (a) follows from the element inequality that $\left\|\boldsymbol{a}+\boldsymbol{b}\right\|_2^2 \leq (1+\gamma)\left\|\boldsymbol{a}\right\|_2^2 + (1+\gamma^{-1})\left\|\boldsymbol{b}\right\|_2^2$ with $\gamma = \frac{1-\rho^2}{2\rho^2}$, step (b) is due to the fact that $((\boldsymbol{W} - n^{-1}\boldsymbol{J}_n)\otimes \boldsymbol{I}_d)(\boldsymbol{1}_n \otimes \boldsymbol{a}) = (\boldsymbol{W}\boldsymbol{1}_n - n^{-1}\boldsymbol{J}_n\boldsymbol{1}_n)\otimes \boldsymbol{a} = \boldsymbol{0}$ for any $\boldsymbol{a}\in\mathbb{R}^d$, step (c) is due to $(1+\rho^2)\rho^2 \leq 2$, step (d) follows from the fact that

$$\left\|\boldsymbol{d}^t - \boldsymbol{1}_n \otimes \bar{\boldsymbol{d}}^t\right\|_2^2 \leq \frac{G^2}{2\mu_F^2}\left\|\boldsymbol{d}^t\right\|_2^2 + \frac{2}{\mu_F^2}\left\|\boldsymbol{y}^{t+1} - \boldsymbol{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2. \tag{50}$$

Thus we complete the first part of this lemma. For the second part, using the same argument yields that

$$\begin{aligned}
\left\|\boldsymbol{\theta}^{t+1} - \boldsymbol{1}_n \otimes \bar{\boldsymbol{\theta}}^{t+1}\right\|_2^2 &\leq 2\left\|\left(\left(\boldsymbol{W} - \frac{1}{n}\boldsymbol{J}_n\right)\otimes \boldsymbol{I}_d\right)\boldsymbol{\theta}^t\right\|_2^2 + 2\left\|\eta\left(\left(\boldsymbol{W} - \frac{1}{n}\boldsymbol{J}_n\right)\otimes \boldsymbol{I}_d\right)\boldsymbol{d}^t\right\|_2^2 \\
&\leq 2\rho^2\left\|\boldsymbol{\theta}^t - \boldsymbol{1}_n \otimes \bar{\boldsymbol{\theta}}^t\right\|_2^2 + 2\eta^2\rho^2\left\|\boldsymbol{d}^t - \boldsymbol{1}_n \otimes \bar{\boldsymbol{d}}^t\right\|_2^2 \\
&\leq 2\rho^2\left\|\boldsymbol{\theta}^t - \boldsymbol{1}_n \otimes \bar{\boldsymbol{\theta}}^t\right\|_2^2 + \frac{4\eta^2\rho^2}{\mu_F^2}\left\|\boldsymbol{y}^{t+1} - \boldsymbol{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 + \frac{G^2\eta^2\rho^2}{\mu_F^2}\left\|\boldsymbol{d}^t\right\|_2^2.
\end{aligned}$$

It only remains to prove the fact (50) used in step (d). Firstly, we have

$$\begin{aligned}
&\left\|\boldsymbol{d}^t - \boldsymbol{1}_n \otimes \bar{\boldsymbol{d}}^t\right\|_2 \\
&= \left\|\left(\boldsymbol{I}_{nd} - \frac{1}{n}\left(\boldsymbol{J}_n \otimes \boldsymbol{I}_d\right)\right)\boldsymbol{d}^t\right\|_2 \\
&= \left\|\left(\boldsymbol{I}_{nd} - \frac{1}{n}\left(\boldsymbol{J}_n \otimes \boldsymbol{I}_d\right)\right)\left(\boldsymbol{H}^t\boldsymbol{y}^{t+1}\right)\right\|_2 \\
&= \left\|\left(\boldsymbol{I}_{nd} - \frac{1}{n}\left(\boldsymbol{J}_n \otimes \boldsymbol{I}_d\right)\right)\left(\boldsymbol{H}^t\boldsymbol{y}^{t+1} - \left(\frac{1}{2\mu_F} + \frac{1}{2G}\right)\boldsymbol{y}^{t+1} + \left(\frac{1}{2\mu_F} + \frac{1}{2G}\right)\boldsymbol{y}^{t+1}\right)\right\|_2 \\
&\leq \left\|\left(\boldsymbol{I}_{nd} - \frac{1}{n}\left(\boldsymbol{J}_n \otimes \boldsymbol{I}_d\right)\right)\left(\boldsymbol{H}^t\boldsymbol{y}^{t+1} - \left(\frac{1}{2\mu_F} + \frac{1}{2G}\right)\boldsymbol{y}^{t+1}\right)\right\|_2 \\
&\qquad + \left(\frac{1}{2\mu_F} + \frac{1}{2G}\right)\left\|\left(\boldsymbol{I}_{nd} - \frac{1}{n}\left(\boldsymbol{J}_n \otimes \boldsymbol{I}_d\right)\right)\boldsymbol{y}^{t+1}\right\|_2 \\
&\leq \left\|\boldsymbol{I}_{nd} - \frac{1}{n}\left(\boldsymbol{J}_n \otimes \boldsymbol{I}_d\right)\right\| \cdot \left\|\boldsymbol{H}^t - \left(\frac{1}{2\mu_F} + \frac{1}{2G}\right)\boldsymbol{I}_{nd}\right\| \cdot \left\|\boldsymbol{y}^{t+1}\right\|_2 \\
&\qquad + \left(\frac{1}{2\mu_F} + \frac{1}{2G}\right)\left\|\left(\boldsymbol{I}_{nd} - \frac{1}{n}\left(\boldsymbol{J}_n \otimes \boldsymbol{I}_d\right)\right)\boldsymbol{y}^{t+1}\right\|_2 \\
&= \left\|\boldsymbol{I}_{nd} - \frac{1}{n}\left(\boldsymbol{J}_n \otimes \boldsymbol{I}_d\right)\right\| \cdot \left\|\boldsymbol{H}^t - \left(\frac{1}{2\mu_F} + \frac{1}{2G}\right)\boldsymbol{I}_{nd}\right\| \cdot \left\|\boldsymbol{y}^{t+1}\right\|_2 \\
&\qquad + \left(\frac{1}{2\mu_F} + \frac{1}{2G}\right)\left\|\boldsymbol{y}^{t+1} - \boldsymbol{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2 \\
&\overset{(a)}{\leq} \frac{1}{2}\left(\frac{1}{\mu_F} - \frac{1}{G}\right)\left\|\boldsymbol{y}^{t+1}\right\|_2 + \frac{1}{2}\left(\frac{1}{\mu_F} + \frac{1}{G}\right)\left\|\boldsymbol{y}^{t+1} - \boldsymbol{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2,
\end{aligned}$$

where the last line follows from that $\left\|\boldsymbol{I}_{nd} - \frac{1}{n}\left(\boldsymbol{J}_n \otimes \boldsymbol{I}_d\right)\right\| = \left\|(\boldsymbol{I}_d - n^{-1}\boldsymbol{I}_d)\otimes \boldsymbol{I}_n\right\| \leq 1 - \frac{1}{n} \leq 1$ and

$$\left\|\boldsymbol{H}^t - \left(\frac{1}{2\mu_F} + \frac{1}{2G}\right)\boldsymbol{I}_{nd}\right\| \leq \frac{1}{2}\left(\frac{1}{\mu_F} - \frac{1}{G}\right).$$

Then a direct computation yields that

$$\begin{aligned}\left\|\boldsymbol{d}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{d}}^t\right\|_2^2 &\le \frac{1}{2}\left(\frac{1}{\mu_F} - \frac{1}{G}\right)^2 \left\|\boldsymbol{y}^{t+1}\right\|_2^2 + \frac{1}{2}\left(\frac{1}{\mu_F} + \frac{1}{G}\right)^2 \left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 \\ &\overset{(a)}{\le} \frac{1}{2}\left(\frac{1}{\mu_F} - \frac{1}{G}\right)^2 G^2 \left\|\boldsymbol{d}^t\right\|_2^2 + \frac{1}{2}\left(\frac{1}{\mu_F} + \frac{1}{G}\right)^2 \left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 \\ &= \frac{1}{2}\left(\frac{G}{\mu_F} - 1\right)^2 \left\|\boldsymbol{d}^t\right\|_2^2 + \frac{1}{2}\left(\frac{1}{\mu_F} + \frac{1}{G}\right)^2 \left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 \\ &\overset{(b)}{\le} \frac{G^2}{2\mu_F^2} \left\|\boldsymbol{d}^t\right\|_2^2 + \frac{1}{2}\left(\frac{1}{\mu_F} + \frac{1}{G}\right)^2 \left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2 \\ &\le \frac{G^2}{2\mu_F^2} \left\|\boldsymbol{d}^t\right\|_2^2 + \frac{2}{\mu_F^2} \left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2,\end{aligned}$$

where step (a) is due to the fact $\left\|\boldsymbol{y}^{t+1}\right\|_2^2 = \sum_{i=1}^n \left\|\boldsymbol{y}_i^{t+1}\right\|_2^2 = \sum_{i=1}^n \left\|\boldsymbol{H}_i^{t^{-1}} \boldsymbol{d}_i^t\right\|_2^2 \le G^2 \left\|\boldsymbol{d}^t\right\|_2^2$, and step (b) is due to $\frac{G}{\mu_F} \ge 1$.

5.4 Proof of Lemma 15

5.4.1 Proof of (31)

Recall the initialization in Algorithm 1 that $\boldsymbol{y}_i^0 = \mathbf{0}, \boldsymbol{v}_i^{-1} = \mathbf{0}$, and $\boldsymbol{v}_i^0 = \frac{1}{B}\sum_{b=1}^B \boldsymbol{g}_i(\tau_{i,b}^0 | \boldsymbol{\theta}_i^0)$. A direct computation yields that

$$\begin{aligned}\mathbb{E}\left\{\left\|\boldsymbol{y}^1 - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^1\right\|_2\right\}^2 &= \mathbb{E}\left\{\left\|\left(\boldsymbol{W} \otimes \boldsymbol{I}_d\right)\boldsymbol{v}^0 - \mathbf{1}_n \otimes \bar{\boldsymbol{v}}^0\right\|_2^2\right\} \\ &= \mathbb{E}\left\{\left\|\left(\boldsymbol{W} \otimes \boldsymbol{I}_d\right)\boldsymbol{v}^0 - \frac{1}{n}\left(\boldsymbol{J}_n \otimes \boldsymbol{I}_d\right)\boldsymbol{v}^0\right\|_2^2\right\} \\ &\le \left\|\boldsymbol{W} - \frac{1}{n}\boldsymbol{J}_n\right\|^2 \cdot \mathbb{E}\left\{\left\|\boldsymbol{v}^0\right\|_2^2\right\} \\ &= \rho^2 \sum_{i=1}^n \mathbb{E}\left\{\left\|\boldsymbol{v}_i^0 - \nabla V_i(\bar{\boldsymbol{\theta}}^0) + \nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\|_2^2\right\} \\ &= \rho^2 \sum_{i=1}^n \mathbb{E}\left\{\left\|\boldsymbol{v}_i^0 - \nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\|_2^2\right\} + \rho^2 \sum_{i=1}^n \left\|\nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\|_2^2 \\ &\quad + 2\rho^2 \sum_{i=1}^n \left\langle \mathbb{E}\left\{\boldsymbol{v}_i^0\right\} - \nabla V_i(\bar{\boldsymbol{\theta}}^0), \nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\rangle \\ &= \rho^2 \sum_{i=1}^n \mathbb{E}\left\{\left\|\boldsymbol{v}_i^0 - \nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\|_2^2\right\} + \rho^2 \sum_{i=1}^n \left\|\nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\|_2^2, \qquad (51)\end{aligned}$$

where the last line follows from $\mathbb{E}\left\{\boldsymbol{v}_i^0\right\} = \nabla V_i(\boldsymbol{\theta}_i^0) = \nabla V_i(\bar{\boldsymbol{\theta}}^0)$. Moreover, for any $i \in [n]$, it can be seen that

$$\mathbb{E}\left\{\left\|\boldsymbol{v}_i^0 - \nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\|_2^2\right\}$$

$$\begin{aligned}
&= \mathbb{E}\left\{\left\|\frac{1}{B}\sum_{b=1}^{B}\boldsymbol{g}_i(\tau_{i,b}^0|\bar{\boldsymbol{\theta}}^0) - \nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\|_2^2\right\} \\
&= \mathbb{E}\left\{\left\|\frac{1}{B}\sum_{b=1}^{B}\left(\boldsymbol{g}_i(\tau_{i,b}^0|\bar{\boldsymbol{\theta}}^0) - \nabla V_i(\bar{\boldsymbol{\theta}}^0)\right)\right\|_2^2\right\} \\
&= \frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\left\{\left\|\boldsymbol{g}_i(\tau_{i,b}^0|\bar{\boldsymbol{\theta}}^0) - \nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\|_2^2\right\} \\
&\quad + \frac{1}{B^2}\sum_{b\neq b'}\mathbb{E}\left\{\left\langle \boldsymbol{g}_i(\tau_{i,b}^0|\bar{\boldsymbol{\theta}}^0) - \nabla V_i(\bar{\boldsymbol{\theta}}^0), \boldsymbol{g}_i(\tau_{i,b'}^0|\bar{\boldsymbol{\theta}}^0) - \nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\rangle\right\} \\
&\overset{(a)}{=} \frac{1}{B^2}\sum_{b=1}^{B}\mathbb{E}\left\{\left\|\boldsymbol{g}_i(\tau_{i,b}^0|\bar{\boldsymbol{\theta}}^0) - \nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\|_2^2\right\} \\
&\overset{(b)}{\leq} \frac{\nu_i^2}{B},
\end{aligned} \tag{52}$$

where step (a) is due to the fact that $\{\tau_{i,b}^0\}_{b=1}^B$ are independent trajectories and step (b) follows from Assumption 7. Substituting (52) into (51) yields that

$$\begin{aligned}
\mathbb{E}\left\{\left\|\boldsymbol{y}^1 - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^1\right\|_2^2\right\} &\leq \frac{\rho^2}{B}\sum_{i=1}^{n}\nu_i^2 + \rho^2\sum_{i=1}^{n}\left\|\nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\|_2^2 \\
&= \frac{n\rho^2\bar{\nu}^2}{B} + \rho^2\sum_{i=1}^{n}\left\|\nabla V_i(\bar{\boldsymbol{\theta}}^0)\right\|_2^2,
\end{aligned}$$

which completes the proof of (31).

5.4.2 Proof of (32)

Following the gradient tracking update in (27), we have

$$\begin{aligned}
&\mathbb{E}\left\{\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2\right\} \\
&= \mathbb{E}\left\{\left\|(\boldsymbol{W}\otimes\boldsymbol{I}_d)\left(\boldsymbol{y}^t + \boldsymbol{v}^t - \boldsymbol{v}^{t-1}\right) - \frac{1}{n}\left(\mathbf{1}_n\mathbf{1}_n^\mathsf{T}\otimes\boldsymbol{I}_d\right)(\boldsymbol{W}\otimes\boldsymbol{I}_d)\left(\boldsymbol{y}^t + \boldsymbol{v}^t - \boldsymbol{v}^{t-1}\right)\right\|_2^2\right\} \\
&= \mathbb{E}\left\{\left\|\left(\left(\boldsymbol{W} - \frac{1}{n}\boldsymbol{J}_n\right)\otimes\boldsymbol{I}_d\right)\left(\boldsymbol{y}^t + \boldsymbol{v}^t - \boldsymbol{v}^{t-1}\right)\right\|_2^2\right\} \\
&\leq \left(1 + \frac{1-\rho^2}{2\rho^2}\right)\mathbb{E}\left\{\left\|\left(\left(\boldsymbol{W} - \frac{1}{n}\boldsymbol{J}_n\right)\otimes\boldsymbol{I}_d\right)\boldsymbol{y}^t\right\|_2^2\right\} \\
&\quad + \left(1 + \frac{2\rho^2}{1-\rho^2}\right)\mathbb{E}\left\{\left\|\left(\left(\boldsymbol{W} - \frac{1}{n}\boldsymbol{J}_n\right)\otimes\boldsymbol{I}_d\right)\left(\boldsymbol{v}^t - \boldsymbol{v}^{t-1}\right)\right\|_2^2\right\} \\
&= \frac{1+\rho^2}{2\rho^2}\mathbb{E}\left\{\left\|\left(\left(\boldsymbol{W} - \frac{1}{n}\boldsymbol{J}_n\right)\otimes\boldsymbol{I}_d\right)\left(\boldsymbol{y}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^t\right)\right\|_2^2\right\}
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1+\rho^2}{1-\rho^2}\mathbb{E}\left\{\left\|\left(\left(\boldsymbol{W}-\frac{1}{n}\boldsymbol{J}_n\right)\otimes\boldsymbol{I}_d\right)(\boldsymbol{v}^t-\boldsymbol{v}^{t-1})\right\|_2^2\right\}\\
\le{}& \frac{1+\rho^2}{2}\mathbb{E}\left\{\left\|\boldsymbol{y}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^t\right\|_2^2\right\}+\frac{(1+\rho^2)\rho^2}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{v}^t-\boldsymbol{v}^{t-1}\right\|_2^2\right\}, \qquad (53)
\end{aligned}
$$

where the third line is due to the element inequality that $\|\boldsymbol{a}+\boldsymbol{b}\|_2^2 \le (1+c)\|\boldsymbol{a}\|_2^2+(1+c^{-1})\|\boldsymbol{b}\|_2^2$ with $c=\frac{1-\rho^2}{2\rho^2}$ for any $\boldsymbol{a}$ and $\boldsymbol{b}$. Moreover, we have the following relationship:

$$
\begin{aligned}
\mathbb{E}\left\{\left\|\boldsymbol{v}^t-\boldsymbol{v}^{t-1}\right\|_2^2\right\} \le{}& \left(8(1-\beta)^2L_g^2+8(1-\beta)^2C_g^2C_\omega^2+4\beta^2L_g^2\right)\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1}\right\|_2^2\right\}\\
&+4n\beta^2\bar{\nu}^2+4\beta^2\sum_{i=1}^n\mathbb{E}\left\{\left\|\nabla V_i(\boldsymbol{\theta}_i^{t-1})-\boldsymbol{v}_i^{t-1}\right\|_2^2\right\}, \qquad (54)
\end{aligned}
$$

which has been shown in (61) of Jiang et al. (2022). Substituting (54) into (53) yields that

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\boldsymbol{y}^{t+1}-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^{t+1}\right\|_2^2\right\}\\
&\le \frac{1+\rho^2}{2}\mathbb{E}\left\{\left\|\boldsymbol{y}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^t\right\|_2^2\right\}+\frac{(1+\rho^2)\rho^2}{1-\rho^2}\Bigg(12\Phi^2\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1}\right\|_2^2\right\}+4n\beta^2\bar{\nu}^2\\
&\qquad+4\beta^2\sum_{i=1}^n\mathbb{E}\left\{\left\|\nabla V_i(\boldsymbol{\theta}_i^{t-1})-\boldsymbol{v}_i^{t-1}\right\|_2^2\right\}\Bigg)\\
&=\frac{1+\rho^2}{2}\mathbb{E}\left\{\left\|\boldsymbol{y}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^t\right\|_2^2\right\}+\frac{4n\beta^2\bar{\nu}^2\rho^2(1+\rho^2)}{1-\rho^2}\\
&\qquad+\frac{12\Phi^2(1+\rho^2)\rho^2}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1}\right\|_2^2\right\}+\frac{4\beta^2\rho^2(1+\rho^2)}{1-\rho^2}\sum_{i=1}^n\mathbb{E}\left\{\left\|\nabla V_i(\boldsymbol{\theta}_i^{t-1})-\boldsymbol{v}_i^{t-1}\right\|_2^2\right\}\\
&\le \frac{1+\rho^2}{2}\mathbb{E}\left\{\left\|\boldsymbol{y}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^t\right\|_2^2\right\}+\frac{8n\beta^2\bar{\nu}^2}{1-\rho^2}\\
&\qquad+\frac{24\Phi^2}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1}\right\|_2^2\right\}+\frac{8\beta^2}{1-\rho^2}\sum_{i=1}^n\mathbb{E}\left\{\left\|\nabla V_i(\boldsymbol{\theta}_i^{t-1})-\boldsymbol{v}_i^{t-1}\right\|_2^2\right\}, \qquad (55)
\end{aligned}
$$

where the first inequality follows from that $8(1-\beta)^2L_g^2+8(1-\beta)^2C_g^2C_\omega^2+4\beta^2L_g^2 \le 12(L_g^2+C_g^2C_\omega^2) := 12\Phi^2$ for $0\le\beta\le 1$ and the last line is due to $\rho<1$. Furthermore, the term $\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1}\right\|_2^2\right\}$ can be bounded as follows:

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\boldsymbol{\theta}^{t-1}\right\|_2^2\right\}\\
&=\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t+\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t-1}+\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t-1}-\boldsymbol{\theta}^{t-1}\right\|_2^2\right\}\\
&\le 3\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\right\}+3\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1}-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\}+3\mathbb{E}\left\{\left\|\mathbf{1}_n\otimes(\bar{\boldsymbol{\theta}}^t-\bar{\boldsymbol{\theta}}^{t-1})\right\|_2^2\right\}\\
&=3\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\right\}+3\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1}-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\}+3n\eta^2\mathbb{E}\left\{\left\|\bar{\boldsymbol{d}}^{t-1}\right\|_2^2\right\}\\
&\overset{(a)}{\le} 6\rho^2\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1}-\mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\}+\frac{12\eta^2\rho^2}{\mu_F^2}\mathbb{E}\left\{\left\|\boldsymbol{y}^t-\mathbf{1}_n\otimes\bar{\boldsymbol{y}}^t\right\|_2^2\right\}+\frac{3G^2\eta^2\rho^2}{\mu_F^2}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\}
\end{aligned}
$$

$$
\begin{aligned}
&\quad + 3\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\} + 3n\eta^2\mathbb{E}\left\{\left\|\bar{\boldsymbol{d}}^{t-1}\right\|_2^2\right\} \\
&\overset{(b)}{\le} 6\rho^2\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\} + \frac{12\eta^2\rho^2}{\mu_F^2}\mathbb{E}\left\{\left\|\boldsymbol{y}^{t} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t}\right\|_2^2\right\} + \frac{3G^2\eta^2\rho^2}{\mu_F^2}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\} \\
&\quad + 3\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\} + 3\eta^2\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\} \\
&\overset{(c)}{\le} 9\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\} + \frac{12\eta^2\rho^2}{\mu_F^2}\mathbb{E}\left\{\left\|\boldsymbol{y}^{t} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t}\right\|_2^2\right\} + \frac{6G^2\eta^2}{\mu_F^2}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\},
\end{aligned} \tag{56}
$$

where step (a) is due to (30), step (b) follows from the fact that $\left\|\bar{\boldsymbol{d}}^{t-1}\right\|_2^2 = \left\|\frac{1}{n}\sum_{i=1}^n \boldsymbol{d}_i^{t-1}\right\|_2^2 \le \frac{1}{n}\sum_{i=1}^n \left\|\boldsymbol{d}_i^{t-1}\right\|_2^2 = \frac{1}{n}\left\|\boldsymbol{d}^{t-1}\right\|_2^2$, and step (c) holds since $\rho < 1$ and $\mu_F \le G$. Substituting (56) into (55) yields that

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\boldsymbol{y}^{t+1} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t+1}\right\|_2^2\right\} \\
&\le \frac{1+\rho^2}{2}\mathbb{E}\left\{\left\|\boldsymbol{y}^{t} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t}\right\|_2^2\right\} + \frac{8n\beta^2\bar{\nu}^2}{1-\rho^2} + \frac{8\beta^2}{1-\rho^2}\sum_{i=1}^n \mathbb{E}\left\{\left\|\nabla V_i(\boldsymbol{\theta}_i^{t-1}) - \boldsymbol{v}_i^{t-1}\right\|_2^2\right\} \\
&\quad + \frac{24\Phi^2}{1-\rho^2}\Bigg(9\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\} \\
&\qquad + \frac{12\eta^2\rho^2}{\mu_F^2}\mathbb{E}\left\{\left\|\boldsymbol{y}^{t} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t}\right\|_2^2\right\} + \frac{6G^2\eta^2}{\mu_F^2}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\}\Bigg) \\
&= \left(\frac{1+\rho^2}{2} + \frac{24\Phi^2}{1-\rho^2}\cdot\frac{12\eta^2\rho^2}{\mu_F^2}\right)\mathbb{E}\left\{\left\|\boldsymbol{y}^{t} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t}\right\|_2^2\right\} + \frac{8n\beta^2\bar{\nu}^2}{1-\rho^2} \\
&\quad + \frac{8\beta^2}{1-\rho^2}\sum_{i=1}^n \mathbb{E}\left\{\left\|\nabla V_i(\boldsymbol{\theta}_i^{t-1}) - \boldsymbol{v}_i^{t-1}\right\|_2^2\right\} \\
&\quad + \frac{216\Phi^2}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\} + \frac{144\Phi^2G^2\eta^2}{\mu_F^2(1-\rho^2)}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\} \\
&\le \frac{3+\rho^2}{4}\mathbb{E}\left\{\left\|\boldsymbol{y}^{t} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t}\right\|_2^2\right\} + \frac{8n\beta^2\bar{\nu}^2}{1-\rho^2} + \frac{8\beta^2}{1-\rho^2}\sum_{i=1}^n \mathbb{E}\left\{\left\|\nabla V_i(\boldsymbol{\theta}_i^{t-1}) - \boldsymbol{v}_i^{t-1}\right\|_2^2\right\} \\
&\quad + \frac{216\Phi^2}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\} + \frac{144\Phi^2G^2\eta^2}{\mu_F^2(1-\rho^2)}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\} \\
&= \frac{3+\rho^2}{4}\mathbb{E}\left\{\left\|\boldsymbol{y}^{t} - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^{t}\right\|_2^2\right\} + \frac{8n\beta^2\bar{\nu}^2}{1-\rho^2} + \frac{8\beta^2}{1-\rho^2}\mathbb{E}\left\{\left\|\widetilde{\nabla V}(\boldsymbol{\theta}^{t-1}) - \boldsymbol{v}^{t-1}\right\|_2^2\right\} \\
&\quad + \frac{216\Phi^2}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\} + \frac{144\Phi^2G^2\eta^2}{\mu_F^2(1-\rho^2)}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\},
\end{aligned}
$$

where the second inequality is due to $\eta \le \frac{\mu_F(1-\rho^2)}{24\sqrt{2}\Phi}$, i.e., $\frac{1+\rho^2}{2} + \frac{24\Phi^2}{1-\rho^2}\cdot\frac{12\eta^2\rho^2}{\mu_F^2} \le \frac{3+\rho^2}{4}$. Now the proof is complete.

5.5 Proof of Lemma 17

Applying (34) to (32) yields that

$$
\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left\{\left\|\boldsymbol{y}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^t\right\|_2^2\right\} \\
&\le \frac{4}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{y}^1 - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^1\right\|_2^2\right\} + \frac{4T}{1-\rho^2}\cdot\frac{8n\beta^2\bar{\nu}^2}{1-\rho^2} \\
&\quad + \frac{4}{1-\rho^2}\sum_{t=0}^{T-2}\left(\frac{8\beta^2}{1-\rho^2}\mathbb{E}\left\{\left\|\widetilde{\nabla V}(\boldsymbol{\theta}^t) - \boldsymbol{v}^t\right\|_2^2\right\}\right. \\
&\qquad \left. + \frac{216\Phi^2}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t\right\|_2^2\right\} + \frac{144\Phi^2 G^2\eta^2}{\mu_F^2(1-\rho^2)}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\}\right) \\
&\le \frac{4}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{y}^1 - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^1\right\|_2^2\right\} + \frac{32nT\beta^2\bar{\nu}^2}{(1-\rho^2)^2} + \frac{576\Phi^2G^2\eta^2}{\mu_F^2(1-\rho^2)^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\} \\
&\quad + \frac{864\Phi^2}{(1-\rho^2)^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t\right\|_2^2\right\} + \frac{32\beta^2}{(1-\rho^2)^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\widetilde{\nabla V}(\boldsymbol{\theta}^t) - \boldsymbol{v}^t\right\|_2^2\right\} \\
&\overset{(a)}{\le} \frac{4}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{y}^1 - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^1\right\|_2^2\right\} + \frac{32nT\beta^2\bar{\nu}^2}{(1-\rho^2)^2} \\
&\quad + \frac{576\Phi^2G^2\eta^2}{\mu_F^2(1-\rho^2)^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\} + \frac{864\Phi^2}{(1-\rho^2)^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t\right\|_2^2\right\} \\
&\quad + \frac{32\beta^2}{(1-\rho^2)^2}\left(\frac{n\bar{\nu}^2}{\beta B} + 2n\beta T\bar{\nu}^2\right. \\
&\qquad \left. + \frac{12n\eta^2\Phi^2}{\beta}\sum_{t=0}^{T-1}\mathbb{E}\left\{\left\|\bar{\boldsymbol{d}}^t\right\|_2^2\right\} + \frac{24\Phi^2}{\beta}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t\right\|_2^2\right\}\right) \\
&\overset{(b)}{\le} \frac{4}{1-\rho^2}\left(\frac{n\rho^2\bar{\nu}^2}{B} + \rho^2\left\|\widetilde{\nabla V}(\boldsymbol{\theta}^0)\right\|_2^2\right) + \frac{32nT\beta^2\bar{\nu}^2}{(1-\rho^2)^2} + \frac{576\Phi^2G^2\eta^2}{\mu_F^2(1-\rho^2)^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\} \\
&\quad + \frac{864\Phi^2}{(1-\rho^2)^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t\right\|_2^2\right\} \\
&\quad + \frac{32\beta^2}{(1-\rho^2)^2}\left(\frac{n\bar{\nu}^2}{\beta B} + 2n\beta T\bar{\nu}^2\right. \\
&\qquad \left. + \frac{12n\eta^2\Phi^2}{\beta}\sum_{t=0}^{T-1}\mathbb{E}\left\{\left\|\bar{\boldsymbol{d}}^t\right\|_2^2\right\} + \frac{24\Phi^2}{\beta}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t\right\|_2^2\right\}\right) \\
&= \frac{4}{1-\rho^2}\left(\frac{n\rho^2\bar{\nu}^2}{B} + \rho^2\left\|\widetilde{\nabla V}(\boldsymbol{\theta}^0)\right\|_2^2\right) + \frac{32nT\beta^2\bar{\nu}^2}{(1-\rho^2)^2} + \frac{32\beta^2}{(1-\rho^2)^2}\left(\frac{n\bar{\nu}^2}{\beta B} + 2n\beta T\bar{\nu}^2\right) \\
&\quad + \left(\frac{864\Phi^2}{(1-\rho^2)^2} + \frac{32\beta^2}{(1-\rho^2)^2}\cdot\frac{24\Phi^2}{\beta}\right)\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t\right\|_2^2\right\}
\end{aligned}
$$

$$
\begin{aligned}
&\quad + \frac{576\Phi^2 G^2\eta^2}{\mu_F^2(1-\rho^2)^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\} + \frac{32\beta^2}{(1-\rho^2)^2}\cdot\frac{12n\eta^2\Phi^2}{\beta}\sum_{t=0}^{T-1}\mathbb{E}\left\{\left\|\bar{\boldsymbol{d}}^t\right\|_2^2\right\} \\
&\overset{(c)}{\le} \frac{4}{1-\rho^2}\left(\frac{n\rho^2\bar{\nu}^2}{B} + \rho^2\left\|\widetilde{\nabla V}(\boldsymbol{\theta}^0)\right\|_2^2\right) + \frac{32nT\beta^2\bar{\nu}^2}{(1-\rho^2)^2} + \frac{32\beta^2}{(1-\rho^2)^2}\left(\frac{n\bar{\nu}^2}{\beta B} + 2n\beta T\bar{\nu}^2\right) \\
&\quad + \left(\frac{864\Phi^2}{(1-\rho^2)^2} + \frac{32\beta^2}{(1-\rho^2)^2}\cdot\frac{24\Phi^2}{\beta}\right)\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\right\} \\
&\quad + \left(\frac{576\Phi^2 G^2\eta^2}{\mu_F^2(1-\rho^2)^2} + \frac{384\beta\eta^2\Phi^2}{(1-\rho^2)^2}\right)\sum_{t=0}^{T-1}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\} \\
&\le \frac{4}{1-\rho^2}\left(\frac{n\rho^2\bar{\nu}^2}{B} + \rho^2\left\|\widetilde{\nabla V}(\boldsymbol{\theta}^0)\right\|_2^2\right) + \frac{32nT\beta^2\bar{\nu}^2}{(1-\rho^2)^2} + \frac{32\beta^2}{(1-\rho^2)^2}\left(\frac{n\bar{\nu}^2}{\beta B} + 2n\beta T\bar{\nu}^2\right) \\
&\quad + \frac{1632\Phi^2}{(1-\rho^2)^2}\sum_{t=0}^{T}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\right\} + \frac{960\Phi^2 G^2\eta^2}{\mu_F^2(1-\rho^2)^2}\sum_{t=0}^{T-1}\mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\},
\end{aligned}
$$

where step (a) is due to (37), step (b) follows from (31), step (c) holds since $\left\|\bar{\boldsymbol{d}}^t\right\|_2^2 \le \frac{1}{n}\left\|\boldsymbol{d}^t\right\|_2^2$, and the last inequality is due to $\beta < 1$, i.e.,

$$
\frac{864\Phi^2}{(1-\rho^2)^2} + \frac{32\beta^2}{(1-\rho^2)^2}\cdot\frac{24\Phi^2}{\beta} = \frac{864\Phi^2}{(1-\rho^2)^2} + \frac{768\Phi^2\beta}{(1-\rho^2)^2} \le \frac{1632\Phi^2}{(1-\rho^2)^2}.
$$

Thus we complete the proof.

5.6 Proof of Lemma 18

Due to (29), it can be seen that

$$
\begin{aligned}
&\mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^t\right\|_2^2\right\} \\
&\le \frac{1+\rho^2}{2}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\} \\
&\qquad + \frac{4\eta^2}{\mu_F^2(1-\rho^2)}\mathbb{E}\left\{\left\|\boldsymbol{y}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^t\right\|_2^2\right\} + \frac{G^2\eta^2}{\mu_F^2(1-\rho^2)}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\} \\
&= \frac{1+\rho^2}{2}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\} + \frac{1+\rho^2}{2}\cdot\frac{2}{1+\rho^2}\cdot\frac{\eta^2\kappa_F^2}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\} \\
&\qquad + \frac{4\eta^2}{\mu_F^2(1-\rho^2)}\mathbb{E}\left\{\left\|\boldsymbol{y}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^t\right\|_2^2\right\} \\
&\le \frac{1+\rho^2}{2}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\} + \frac{1+\rho^2}{2}\cdot\frac{2}{1-\rho^2}\cdot\frac{\eta^2\kappa_F^2}{1-\rho^2}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\} \\
&\qquad + \frac{4\eta^2}{\mu_F^2(1-\rho^2)}\mathbb{E}\left\{\left\|\boldsymbol{y}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^t\right\|_2^2\right\} \\
&= \frac{1+\rho^2}{2}\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{t-1} - \mathbf{1}_n\otimes\bar{\boldsymbol{\theta}}^{t-1}\right\|_2^2\right\} \\
&\qquad + \frac{1+\rho^2}{2}\cdot\frac{2\eta^2\kappa_F^2}{(1-\rho^2)^2}\mathbb{E}\left\{\left\|\boldsymbol{d}^{t-1}\right\|_2^2\right\} + \frac{4\eta^2}{\mu_F^2(1-\rho^2)}\mathbb{E}\left\{\left\|\boldsymbol{y}^t - \mathbf{1}_n\otimes\bar{\boldsymbol{y}}^t\right\|_2^2\right\} \qquad (57)
\end{aligned}
$$

for any $t \geq 0$. Applying (33) to (57) leads to that

$$
\begin{aligned}
&\sum_{t=0}^{T} \mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t\right\|_2^2\right\} \\
&\leq \frac{4G^2\eta^2}{\mu_F^2(1-\rho^2)^3} \sum_{t=0}^{T} \mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\} + \frac{8\eta^2}{\mu_F^2(1-\rho^2)^2} \sum_{t=1}^{T} \left\|\boldsymbol{y}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{y}}^t\right\|_2^2 \\
&\leq \frac{4G^2\eta^2}{\mu_F^2(1-\rho^2)^3} \sum_{t=0}^{T} \mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\} \\
&\quad + \frac{8\eta^2}{\mu_F^2(1-\rho^2)^2}\Bigg(A_1 + \frac{1632\Phi^2}{(1-\rho^2)^2} \sum_{t=0}^{T} \mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t\right\|_2^2\right\} \\
&\qquad + \frac{960\Phi^2G^2\eta^2}{\mu_F^2(1-\rho^2)^2} \sum_{t=0}^{T-1} \mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\}\Bigg),
\end{aligned}
$$

where the first inequality has used the fact that $\boldsymbol{\theta}_i^0 = \bar{\boldsymbol{\theta}}^0$ for all $i \in [n]$ and the second inequality follows from (35). Since

$$
0 < \eta < \frac{\mu_F(1-\rho^2)^3}{\kappa_F\sqrt{1632000(L^2+\Phi^2)}},
$$

it can be seen that

$$
\begin{aligned}
&\frac{8\eta^2}{\mu_F^2(1-\rho^2)^2} \cdot \frac{1632\Phi^2}{(1-\rho^2)^2} \leq \frac{8 \cdot 1632\Phi^2}{\mu_F^2(1-\rho^2)^4} \cdot \frac{\mu_F^2(1-\rho^2)^6}{1632000\kappa_F^2(L^2+\Phi^2)} \leq \frac{1}{2}, \\
&\frac{4G^2\eta^2}{\mu_F^2(1-\rho^2)^3} + \frac{8\eta^2}{\mu_F^2(1-\rho^2)^2} \cdot \frac{960\Phi^2G^2\eta^2}{\mu_F^2(1-\rho^2)^2} = \frac{4G^2\eta^2}{\mu_F^2(1-\rho^2)^3} + \frac{G^2\eta^2}{\mu_F^2(1-\rho^2)^4} \cdot \frac{8 \cdot 960\Phi^2\eta^2}{\mu_F^2} \\
&\quad \leq \frac{4G^2\eta^2}{\mu_F^2(1-\rho^2)^3} + \frac{G^2\eta^2}{\mu_F^2(1-\rho^2)^4} \cdot \frac{8 \cdot 960\Phi^2}{\mu_F^2} \cdot \frac{\mu_F^2(1-\rho^2)^6}{1632000\kappa_F^2(L^2+\Phi^2)} \\
&\quad \leq \frac{5G^2\eta^2}{\mu_F^2(1-\rho^2)^3}.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
\sum_{t=0}^{T} \mathbb{E}\left\{\left\|\boldsymbol{\theta}^t - \mathbf{1}_n \otimes \bar{\boldsymbol{\theta}}^t\right\|_2^2\right\} &\leq \frac{16A_1\eta^2}{\mu_F^2(1-\rho^2)^2} + 2\Bigg(\frac{4G^2\eta^2}{\mu_F^2(1-\rho^2)^3} \\
&\qquad + \frac{8\eta^2}{\mu_F^2(1-\rho^2)^2} \cdot \frac{960\Phi^2G^2\eta^2}{\mu_F^2(1-\rho^2)^2}\Bigg) \sum_{t=0}^{T} \mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\} \\
&\leq \frac{16A_1\eta^2}{\mu_F^2(1-\rho^2)^2} + \frac{10G^2\eta^2}{\mu_F^2(1-\rho^2)^3} \sum_{t=0}^{T} \mathbb{E}\left\{\left\|\boldsymbol{d}^t\right\|_2^2\right\},
\end{aligned}
$$

which completes the proof.

6. Conclusions

In this work, we propose a novel decentralized algorithm named MDNPG for MARL. We have established the sample complexity for local convergence of MDNPG, which achieves the best available rate. The key ingredient to our development is a new stochastic ascent inequality for non-convex objectives, which could be of independent interest. Numerical results have demonstrated the efficiency of the proposed method.

There are several interesting directions for future research. Firstly, it is natural to study the global convergence of MDNPG and extend our framework to the class of entropy-regularized natural policy gradient methods in MARL. Secondly, the Fisher information matrix in this paper is empirically estimated by sample averaging, which may incur large variance. Thus, we may also consider variance reduction for the estimation of the precondition matrix. Lastly, though importance sampling is widely used to address the varying data distribution issue when developing variance reduced policy gradient methods, there are also a few recent works (Shen et al., 2019; Salehkaleybar et al., 2022) which instead use a hessian-based technique in the single-agent setting. Therefore, it is also interesting to investigate whether importance sampling can be removed in the multi-agent setting when developing decentralized (natural) policy gradient methods.

Acknowledgments

This work was partially supported by the National Key R&D Program of China (Grant No. 2021YFA1003300), Natural Science Foundation of Shanghai (Grant No. 23ZR1406400), and Science and Technology Commission of Shanghai Municipality (No. 23JC1401000).

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

Shun-ichi Amari. Neural learning in structured parameter spaces-natural riemannian gradient. *Advances in Neural Information Processing Systems*, 9, 1996.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, pages 1–50, 2022.

Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6): 26–38, 2017.

J Andrew Bagnell and Jeff Schneider. Covariant policy search. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 1019–1024, 2003.

Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Dimitri Bertsekas. *Reinforcement learning and optimal control.* Athena Scientific, 2019.

Shalabh Bhatnagar, Mohammad Ghavamzadeh, Mark Lee, and Richard S Sutton. Incremental natural actor-critic algorithms. *Advances in Neural Information Processing Systems*, 20, 2007.

Léon Bottou. Stochastic gradient descent tricks. In *Neural Networks: Tricks of the Trade: Second Edition*, pages 421–436. Springer, 2012.

Stephen Boyd, Arpita Ghosh, Balaji Prabhakar, and Devavrat Shah. Randomized gossip algorithms. *IEEE Transactions on Information Theory*, 52(6):2508–2530, 2006.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Charles George Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

Lucian Busoniu, Robert Babuska, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 2021.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in Neural Information Processing Systems*, 23, 2010.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in Neural Information Processing Systems*, 32, 2019.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in Neural Information Processing Systems*, 27, 2014.

Yuhao Ding, Junzi Zhang, and Javad Lavaei. On the global convergence of momentum-based policy gradient. *arXiv preprint arXiv:2110.10116*, 2021.

Thinh Doan, Siva Maguluri, and Justin Romberg. Finite-time analysis of distributed td (0) with linear function approximation on multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 1626–1635. PMLR, 2019.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pages 1407–1416. PMLR, 2018.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

Ilyas Fatkhullin, Anas Barakat, Anastasia Kireeva, and Niao He. Stochastic policy gradient methods: Improved sample complexity for fisher-non-degenerate policies. In *International Conference on Machine Learning*, pages 9827–9869. PMLR, 2023.

R Fletcher. A new approach to variable metric algorithms. *Computer Journal*, 13(3): 317–317, 1970.

Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of Computation*, 24(109):23–26, 1970.

Matteo Hessel, Hubert Soyer, Lasse Espeholt, Wojciech Czarnecki, Simon Schmitt, and Hado van Hasselt. Multi-task deep reinforcement learning with popart. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3796–3803, 2019.

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In *International Conference on Machine Learning*, pages 4422–4433. PMLR, 2020.

Feihu Huang, Shangqian Gao, and Heng Huang. Bregman gradient policy optimization. In *International Conference on Learning Representations*, 2021.

Zhanhong Jiang, Xian Yeow Lee, Sin Yong Tan, Kai Liang Tan, Aditya Balu, Young M Lee, Chinmay Hegde, and Soumik Sarkar. Mdpgt: momentum-based decentralized policy gradient tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9377–9385, 2022.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in Neural Information Processing Systems*, 26, 2013.

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285, 1996.

Sham M Kakade. A natural policy gradient. *Advances in Neural Information Processing Systems*, 14, 2001.

Jens Kober, J Andrew Bagnell, and Jan Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. *Advances in Neural Information Processing Systems*, 12, 1999.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, pages 1–48, 2022.

Boyue Li, Shicong Cen, Yuxin Chen, and Yuejie Chi. Communication-efficient distributed optimization in networks with gradient tracking and variance reduction. In *International Conference on Artificial Intelligence and Statistics*, pages 1662–1672. PMLR, 2020.

Xiangru Lian, Ce Zhang, Huan Zhang, Cho-Jui Hsieh, Wei Zhang, and Ji Liu. Can decentralized algorithms outperform centralized algorithms? a case study for decentralized parallel stochastic gradient descent. *Advances in Neural Information Processing Systems*, 30, 2017.

Qifeng Lin and Qing Ling. Decentralized td (0) with gradient tracking. *IEEE Signal Processing Letters*, 28:723–727, 2021.

Xiaoyang Liu, Hongyang Yang, Qian Chen, Runjia Zhang, Liuqing Yang, Bowen Xiao, and Christina Wang. Finrl: A deep reinforcement learning library for automated stock trading in quantitative finance. *Available at SSRN 3737859*, 2020a.

Yanli Liu, Kaiqing Zhang, Tamer Basar, and Wotao Yin. An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods. *Advances in Neural Information Processing Systems*, 33:7624–7636, 2020b.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30, 2017.

Songtao Lu, Kaiqing Zhang, Tianyi Chen, Tamer Başar, and Lior Horesh. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8767–8775, 2021.

James Martens. New insights and perspectives on the natural gradient method. *The Journal of Machine Learning Research*, 21(1):5776–5851, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Angelia Nedić, Alex Olshevsky, and Michael G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017.

Ann Nowé, Peter Vrancx, and Yann-Michaël De Hauwere. Game theory and multi-agent reinforcement learning. In *Reinforcement Learning*, pages 441–470. Springer, 2012.

Taoxing Pan, Jun Liu, and Jie Wang. D-spider-sfo: A decentralized optimization algorithm with faster convergence rate for nonconvex problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1619–1626, 2020.

Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *International Conference on Machine Learning*, pages 4026–4035. PMLR, 2018.

Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

Jan Peters, Sethu Vijayakumar, and Stefan Schaal. Reinforcement learning for humanoid robotics. In *Proceedings of the Third IEEE-RAS International Conference on Humanoid Robots*, pages 1–20, 2003.

Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Policy gradient in lipschitz markov decision processes. *Machine Learning*, 100:255–283, 2015.

Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 187(1):409–457, 2021.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

Chao Qu, Shie Mannor, Huan Xu, Yuan Qi, Le Song, and Junwu Xiong. Value propagation for decentralized networked deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Aravind Rajeswaran, Igor Mordatch, and Vikash Kumar. A game theoretic framework for model based reinforcement learning. In *International Conference on Machine Learning*, pages 7953–7963. PMLR, 2020.

Saber Salehkaleybar, Sadegh Khorasani, Negar Kiyavash, Niao He, and Patrick Thiran. Adaptive momentum-based policy gradient with second-order information. *arXiv preprint arXiv:2205.08253*, 2022.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897. PMLR, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

David F Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656, 1970.

Zebang Shen, Alejandro Ribeiro, Hamed Hassani, Hui Qian, and Chao Mi. Hessian aided policy gradient. In *International Conference on Machine Learning*, pages 5729–5738. PMLR, 2019.

Wei Shi, Qing Ling, Gang Wu, and Wotao Yin. Extra: An exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):944–966, 2015.

Haoran Sun, Songtao Lu, and Mingyi Hong. Improving the sample and communication complexity for decentralized non-convex optimization: Joint gradient estimation and tracking. In *International Conference on Machine Learning*, pages 9217–9228. PMLR, 2020.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* MIT press, 2018.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12, 1999.

Hanlin Tang, Xiangru Lian, Ming Yan, Ce Zhang, and Ji Liu. d^2: Decentralized training over decentralized data. In *International Conference on Machine Learning*, pages 4848–4856. PMLR, 2018.

Quoc Tran-Dinh, Nhan H Pham, Dzung T Phan, and Lam M Nguyen. Hybrid stochastic gradient descent algorithms for stochastic nonconvex optimization. *arXiv preprint arXiv:1905.05920*, 2019.

Quoc Tran-Dinh, Nhan H Pham, Dzung T Phan, and Lam M Nguyen. A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, 191(2):1005–1071, 2022.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575 (7782):350–354, 2019.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3):229–256, 1992.

Ran Xin, Usman Khan, and Soummya Kar. A hybrid variance-reduced method for decentralized stochastic non-convex optimization. In *International Conference on Machine Learning*, pages 11459–11469. PMLR, 2021.

Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. In *Uncertainty in Artificial Intelligence*, pages 541–551. PMLR, 2020a.

Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations*, 2020b.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. *Advances in Neural Information Processing Systems*, 33:4358–4369, 2020c.

Long Yang, Yu Zhang, Gang Zheng, Qian Zheng, Pengfei Li, Jianhang Huang, and Gang Pan. Policy optimization with stochastic mirror descent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8823–8831, 2022.

Fuqiang Yao and Luliang Jia. A collaborative multi-agent reinforcement learning anti-jamming algorithm in wireless networks. *IEEE Wireless Communications Letters*, 8(4): 1024–1027, 2019.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020.

Kun Yuan, Bicheng Ying, Xiaochuan Zhao, and Ali H Sayed. Exact diffusion for distributed optimization and learning—part i: Algorithm development. *IEEE Transactions on Signal Processing*, 67(3):708–723, 2018.

Rui Yuan, Robert M Gower, and Alessandro Lazaric. A general sample complexity analysis of vanilla policy gradient. In *International Conference on Artificial Intelligence and Statistics*, pages 3332–3380. PMLR, 2022.

Sihan Zeng, Malik Aqeel Anwar, Thinh T Doan, Arijit Raychowdhury, and Justin Romberg. A decentralized policy gradient approach to multi-task reinforcement learning. In *Uncertainty in Artificial Intelligence*, pages 1002–1012. PMLR, 2021.

Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021a.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pages 5872–5881. PMLR, 2018.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021b.

Xiaoxiao Zhao, Jinlong Lei, and Li Li. Distributed policy gradient with variance reduction in multi-agent reinforcement learning. *arXiv preprint arXiv:2111.12961*, 2021.