Multi-Agent Off-Policy TDC with Near-Optimal Sample and Communication Complexities

Anonymous authors
Paper under double-blind review

Abstract

The finite-time convergence of off-policy temporal difference (TD) learning has been comprehensively studied recently. However, such a type of convergence has not been established for off-policy TD learning in the multi-agent setting, which covers broader reinforcement learning applications and is fundamentally more challenging. This work develops a decentralized TD with correction (TDC) algorithm for multi-agent off-policy TD learning under Markovian sampling. In particular, our algorithm avoids sharing the actions, policies and rewards of the agents, and adopts mini-batch sampling to reduce the sampling variance and communication frequency. Under Markovian sampling and linear function approximation, we proved that the finite-time sample complexity of our algorithm for achieving an ϵ-accurate solution is in the order of $\mathcal{O}\big(\frac{M\ln\epsilon^{-1}}{\epsilon(1-\sigma_2)^2}\big)$, where M denotes the total number of agents and σ_2 is a network parameter. This matches the sample complexity of the centralized TDC. Moreover, our algorithm achieves the optimal communication complexity $\mathcal{O}\big(\frac{\sqrt{M}\ln\epsilon^{-1}}{1-\sigma_2}\big)$ for synchronizing the value function parameters, which is order-wise lower than the communication complexity of the existing decentralized TD(0). Numerical simulations corroborate our theoretical findings.

1 Introduction

Multi-agent reinforcement learning (MARL) is an emerging technique that has broad applications in control Yanmaz et al. (2017); Chalaki & Malikopoulos (2020), wireless sensor networks Krishnamurthy et al. (2008); Yuan et al. (2020), robotics Yan et al. (2013), etc. In MARL, agents cooperatively interact with an environment and follow their own policies to collect local rewards. In particular, policy evaluation is a fundamental problem in MARL that aims to learn a multi-agent value function associated with the policies of the agents. This motivates the development of convergent and communication-efficient multi-agent TD learning algorithms.

For single-agent on-policy evaluation (i.e., samples are collected by target policy), the conventional TD(0) algorithm Sutton (1988); Sutton & Barto (2018) and Q-learning algorithm Dayan (1992) have been developed with asymptotic convergence guarantee. Recently, their finite-time (i.e., non-asymptotic) convergence has been established under Markovian sampling and linear approximation Bhandari et al. (2018); Zou et al. (2019). However, these algorithms may diverge in the off-policy setting Baird (1995), where samples are collected by a different behavior policy. To address this important issue, a family of gradient-based TD (GTD) algorithms were developed for off-policy evaluation with asymptotic convergence guarantee Sutton et al. (2008; 2009); Maei (2011). In particular, the TD with gradient correction (TDC) algorithm has been shown to have superior performance and its finite-time convergence has been established recently under Markovian sampling Xu et al. (2019); Gupta et al. (2019); Kaledin et al. (2020).

For multi-agent on-policy evaluation, various decentralized TD learning algorithms have been developed. For example, the finite-time convergence of decentralized TD(0) was established with i.i.d samples Wai et al. (2018); Doan et al. (2019) and Markovian samples Sun et al. (2020), respectively, under linear function approximation, and an improved result is further obtained in Wang et al. (2020) by leveraging gradient tracking. However, these algorithms do not apply to the off-policy setting. In the existing literature, decentralized off-policy TD learning has been studied only in simplified settings, for example, agents obtain

independent MDP trajectories Macua et al. (2014); Stanković & Stanković (2016); Cassano et al. (2020) or share their behavior and target policies with each other Cassano et al. (2020), and the data samples are either i.i.d. or have a finite sample size. These MARL settings either are impractical or reveal the agents' policies that may be sensitive. Therefore, we want to ask the following question:

- *Q1: Can we develop a decentralized off-policy TD algorithm for MARL with interdependent agents and avoids sharing local actions, policies and rewards of the agents?*

In fact, developing such a desired decentralized off-policy TD learning algorithm requires overcoming two major challenges. First, to perform decentralized off-policy TD learning, all the agents need to obtain a global importance sampling ratio (see Section 3.2). In Cassano et al. (2020), the authors obtained this ratio by sharing all local policies among the agents, which may lead to information leakage. Therefore, we aim to develop a safer scheme to synchronize the global importance sampling ratio among the agents without sharing any sensitive local information. Second, the existing decentralized TD-type algorithm achieves a communication complexity (number of communication rounds) of $\mathcal{O}\big(\big(\epsilon^{-1}+\frac{\sqrt{M}}{\sqrt{\epsilon}(1-\sigma_2)}\big)\ln\epsilon^{-1}\big)$ for networks with M agents and parameter σ_2 (See Assumption 5 in Section 4 for the definition of σ_2) Sun et al. (2020). This induces much communication overhead when the target accuracy ϵ is small. Hence, we want to ask the following theoretical question:

- *Q2: Can we develop a decentralized off-policy TD learning algorithm that achieves a near-optimal finite-time sample complexity and a near-optimal communication complexity under Markovian sampling?*

In this work, we provide affirmative answers to these questions by developing a decentralized TDC algorithm that avoids sharing the sensitive information of the agents and achieves the near-optimal sample complexity as well as a significantly reduced communication complexity. We summarize our contributions as follows.

1.1 Summary of Contribution

To perform multi-agent off-policy evaluation, we develop a decentralized TDC algorithm with linear function approximation. In every iteration, agents perform two-timescale TDC updates locally and exchange model parameters with their neighbors. In particular, our algorithm adopts the following designs to avoid sharing agents' local sensitive information and reduce communication load.

- We let the agents perform local averaging on their local importance sampling ratios to obtain approximated global importance sampling ratios.
- All the agents use a mini-batch of samples to update their model parameters in each iteration. The mini-batch sampling reduces the sampling variance and the communication frequency, leading to an improved communication complexity over that of the existing decentralized TD(0).
- After the decentralized TDC iterations, our algorithm performs additional local averaging steps to achieve a global consensus on the model parameters. This turns out to be critical for achieving the near-optimal complexity bounds.

Theoretically, we analyze the finite-time convergence of this decentralized TDC algorithm with Markovian samples and show that it attains a fast linear convergence. The overall sample complexity for achieving an ϵ-accurate solution is in the order of $\mathcal{O}\big(\frac{M\ln\epsilon^{-1}}{\epsilon(1-\sigma_2)^2}\big)$. When there is a single agent ($M=1$), this sample complexity result matches that of centralized TDC Xu et al. (2019) and matches the theoretical lower bound $\mathcal{O}(\epsilon^{-1})$ Kaledin et al. (2020) up to a logarithm factor. In addition, the sample complexity is proportional to M, which matches the theoretical lower bound of decentralized strongly convex optimization Scaman et al. (2017). Moreover, the communication complexity of our algorithm for synchronizing value function parameters is in the order of $\mathcal{O}\big(\frac{\sqrt{M}\ln\epsilon^{-1}}{1-\sigma_2}\big)$, which is significantly lower than the communication complexity $\mathcal{O}\big(\epsilon^{-1}+\frac{\sqrt{M}}{\sqrt{\epsilon}(1-\sigma_2)}\ln\epsilon^{-1}\big)$ of the decentralized TD(0) Sun et al. (2020) and matches the communication complexity lower bound Scaman et al. (2017).

Technically, our analysis is a nontrivial generalization of the analysis of centralized off-policy TDC to the decentralized case. In particular, our analysis establishes tight bounds of the consensus error caused by

synchronizing the global importance sampling ratio, especially under the Markovian sampling where the data samples are correlated. Moreover, we strategically bound the estimation error of the global importance sampling logarithm-ratio. Please refer to the proof sketch at the end of Section 4 for details.

1.2 Other Related Work

Centralized policy evaluation. TD(0) with linear function approximation Sutton (1988) is popular for on-policy evaluation. The asymptotic and non-asymptotic convergence results of TD(0) have been established in Sutton (1988); Dayan (1992); Jaakkola et al. (1993); Gordon (1995); Baird (1995); Tsitsiklis & Van Roy (1997); Tadić (2001); Hu & Syed (2019) and Korda & La (2015); Liu et al. (2015); Bhandari et al. (2018); Dalal et al. (2018b); Lakshminarayanan & Szepesvari (2018); Wang et al. (2019); Srikant & Ying (2019); Xu et al. (2020b) respectively. Sutton et al. (2009) proposed TDC for off-policy evaluation. The finite-sample convergence of TDC has been established in Dalal et al. (2018a; 2020) with i.i.d. samples and in Xu et al. (2019); Gupta et al. (2019); Kaledin et al. (2020) with Markovian samples.

Decentralized policy evaluation. Mathkar & Borkar (2016) proposed the decentralized TD(0) algorithm. The asymptotic and non-asymptotic convergence rate of decentralized TD have been obtained in Borkar (2009) and Sun et al. (2020); Wang et al. (2020) respectively. Exisitng decentralized off-policy evaluation studies considered simplified settings. Macua et al. (2014); Stanković & Stanković (2016) obtained asymptotic result for decentralized off-policy evaluation where the agents obtained independent MDPs. Cassano et al. (2020) obtained linear convergence rate also with independent MDPs by applying variance reduction and extended to the case where the individual behavior policies and the joint target policy are shared among the agents.

Decentralized policy control. Decentralized policy control is also an important MARL problem where the goal is to learn the optimal policy for each agent. Many algorithms have been proposed for decentralized policy control, including policy gradient Chen et al. (2021) and actor-critic Qu et al. (2020); Zhang et al. (2021).

2 Policy Evaluation in Multi-Agent RL

In this section, we introduce multi-agent reinforcement learning (MARL) and define the policy evaluation problem. Consider a fully decentralized multi-agent network that consists of M agents. The network topology is specified by an undirected graph $\mathcal{G} = (\mathcal{M}, \mathcal{E})$, where $\mathcal{M} = \{1, 2, \cdots, M\}$ denotes the set of agents and $\mathcal{E}$ denotes the set of communication links. In MARL, the agents interact with a dynamic environment through a multi-agent Markov decision process (MMDP) specified as $\{\mathcal{S}, \{\mathcal{A}^{(m)}\}_{m=1}^M, P, \{R^{(m)}\}_{m=1}^M, \gamma\}$. To elaborate, $\mathcal{S}$ denotes a global state space that is shared by all the agents, $\mathcal{A}^{(m)}$ corresponds to the action space of agent m, P is the state transition kernel and $R^{(m)}$ denotes the reward function of agent m. All the state and action spaces have finite cardinality. $\gamma \in (0, 1]$ is a discount factor.

At any time t, assume that all the agents are in the global state $s_t \in \mathcal{S}$. Then, each agent m takes a certain action $a_t^{(m)} \in \mathcal{A}^{(m)}$ following its own stationary policy $\pi^{(m)}$, i.e., $a_t^{(m)} \sim \pi^{(m)}(\cdot|s_t)$. After all the actions are taken, the global state transfers to a new state s_{t+1} according to the transition kernel P, i.e., $s_{t+1} \sim P(\cdot|s_t, a_t)$ where $a_t := \{a_t^{(m)}\}_{m=1}^M$. At the same time, each agent m receives a local reward $R_t^{(m)} := R^{(m)}(s_t, a_t, s_{t+1})$ from the environment for this action-state transition. Throughout the MMDP, each agent m has access to only the global state $\{s_t\}_t$ and its own actions $\{a_t^{(m)}\}_t$ and rewards $\{R_t^{(m)}\}_t$. The goal of policy evaluation in MARL is to evaluate the following value function associated with all the local policies $\pi := \{\pi^{(m)}\}_{m=1}^M$ for any global state s.

$$V^\pi(s) = \mathbb{E}\Big[\sum_{t=0}^{+\infty} \gamma^t \Big(\frac{1}{M}\sum_{m=1}^M R_t^{(m)}\Big)\Big| s_0 = s, \pi\Big]. \tag{1}$$

A popular algorithm for policy evaluation in MARL is the decentralized TD(0) Sun et al. (2020). Specifically, consider a popular linear function approximation of the value function $V_\theta(s) := \theta^\top \phi(s)$, where $\theta \in \mathbb{R}^d$ contains

the model parameters and $\phi(s)$ is a feature vector that corresponds to the state s. In decentralized TD(0), each agent m collects a single Markovian sample $\{s_t, a_t^{(m)}, s_{t+1}, R_t^{(m)}\}$ at time t and updates its own model parameters $\theta_t^{(m)}$ with learning rate $\alpha > 0$ as follows.

$$\theta_{t+1}^{(m)} = \sum_{m' \in \mathcal{N}_m} V_{m,m'} \theta_t^{(m')} + \alpha\big(A_t \theta_t^{(m)} + b_t^{(m)}\big), \tag{2}$$

where $\mathcal{N}_m$ denotes the neighborhood of agent m, V corresponds to a doubly stochastic communication matrix and $A_t = \phi(s_t)(\gamma\phi(s_{t+1}) - \phi(s_t))^\top$, $b_t^{(m)} = R_t^{(m)}\phi(s_t)$. The above update rule applies the local TD error to update the parameters and synchronize the parameters among neighboring agents through the network.

3 Two-Timescale Decentralized TDC for Off-Policy Evaluation

3.1 Centralized TDC

In this subsection, we review the centralized TD with gradient correction (TDC) algorithm Sutton et al. (2009). In RL, the agent may not have enough samples that are collected following the target policy π. Instead, it may have some data samples that are collected under a different behavior policy π_b. Therefore, in this *off-policy* setting, the agent would like to utilize the historical data to help evaluate the value function V^π associated with the target policy π.

In Sutton et al. (2009), a family of gradient-based TD (GTD) learning algorithms have been proposed for off-policy evaluation. In particular, the TDC algorithm has been shown to have superior performance. To explain, consider the linear approximation $V_\theta(s) = \theta^\top \phi(s)$ and suppose the state space includes states $s_1, ..., s_n$, we can define a total value function $V_\theta := [V_\theta(s_1), ..., V_\theta(s_n)]^\top$. The goal of TDC is to minimize the following mean square projected Bellman error (MSPBE).

$$\text{MSPBE}(\theta) := \mathbb{E}_{\mu_b} \|V_\theta - \Pi T^\pi V_\theta\|^2,$$

where μ_b is the stationary distribution induced by π_b, T^π is the Bellman operator and Π is a projection operator onto the space of linear models. Given the i-th sample (s_i, a_i, s_{i+1}, R_i) obtained by the behavior policy, we define the following terms

$$\begin{aligned} &\rho_i := \frac{\pi(a_i|s_i)}{\pi_b(a_i|s_i)}, \quad b_i := \rho_i R_i \phi(s_i), \quad A_i := \rho_t \phi(s_i)(\gamma\phi(s_{i+1}) - \phi(s_i))^\top, \\ &B_i := -\gamma\rho_t \phi(s_{i+1})\phi(s_i)^\top, \quad C_i := -\phi(s_i)\phi(s_i)^\top, \end{aligned} \tag{3}$$

where ρ_t is referred to as the *importance sampling ratio*. Then, with learning rates $\alpha, \beta > 0$ and initialization parameters θ_0, w_0, the-two timescale off-policy TDC algorithm takes the following recursive updates for $t = 0, 1, 2, ...$

$$\text{(TDC):} \quad \begin{cases} \theta_{t+1} = \theta_t + \alpha(A_t\theta_t + b_t + B_t w_t), \\ w_{t+1} = w_t + \beta(A_t\theta_t + b_t + C_t w_t). \end{cases} \tag{4}$$

3.2 Decentralized Mini-batch TDC

In this subsection, we propose a decentralized TDC algorithm for off-policy evaluation in MARL. In the multi-agent setting, without loss of generality, we assume that each agent m has a target policy $\pi^{(m)}$ and its samples are collected by a different behavior policy $\pi_b^{(m)}$. In particular, if agent m is on-policy, then we have $\pi_b^{(m)} = \pi^{(m)}$. In this *multi-agent off-policy* setting, the agents aim to utilize the data collected by the behavior policies $\pi_b = \{\pi_b^{(m)}\}_{m=1}^M$ to help evaluate the value function V^π associated with the target policies $\pi = \{\pi^{(m)}\}_{m=1}^M$.

However, directly generalizing the centralized TDC algorithm to the decentralized setting will encounter several challenges. First, the centralized TDC in eq. (4) consumes one sample per-iteration and achieves the sample complexity $O(\epsilon^{-1}\log\epsilon^{-1})$ Xu et al. (2019). Therefore, the corresponding decentralized TDC would

perform one local communication per-iteration and is expected to have a communication complexity in the order of $O(\epsilon^{-1}\log\epsilon^{-1})$, which induces *large communication overhead*. Second, in the multi-agent off-policy setting, every agent m has a local importance sampling ratio $\rho_i^{(m)} := \pi^{(m)}(a_i^{(m)}|s_i)/\pi_b^{(m)}(a_i^{(m)}|s_i)$. However, to correctly perform off-policy updates, every agent needs to know all the other agents' local importance sampling ratios in order to obtain the **global importance sampling ratio** $\rho_i := \prod_{m=1}^M \rho_i^{(m)}$. To address these challenges, we next propose a decentralized TDC algorithm that takes mini-batch stochastic updates.

To elaborate, note that ρ_i can be rewritten as

$$\rho_i = \exp\Big(M \cdot \tfrac{1}{M}\textstyle\sum_{m=1}^M \ln \rho_i^{(m)}\Big).$$

Therefore, all the agents just need to obtain the average $\frac{1}{M}\sum_{m=1}^M \ln\rho_i^{(m)}$, which can be computed via local communication of the logarithm-ratios $\{\ln\rho_i^{(m)}\}_{m=1}^M$ for L rounds. Specifically, every agent m initializes $\widetilde{\rho}_{i,0}^{(m)} = \ln\rho_i^{(m)}$ and for iterations $\ell = 0, ..., L-1$ do

$$\widetilde{\rho}_{i,\ell+1}^{(m)} = \sum_{m'\in\mathcal{N}_m} V_{m,m'}\widetilde{\rho}_{i,\ell}^{(m')}, \tag{5}$$

$$(\text{Output}):\ \widehat{\rho}_i^{(m)} = \exp(M\cdot\widetilde{\rho}_{i,L}^{(m)}). \tag{6}$$

In Corollary 2 (see the appendix), we prove that all of these local estimates $\{\widehat{\rho}_i^{(m)}\}_{m=1}^M$ converge exponentially fast to the desired quantity ρ_i as L increases. Then, every agent m performs the following two-timescale TDC updates

$$\theta_{t+1}^{(m)} = \sum_{m'\in\mathcal{N}_m} V_{m,m'}\theta_t^{(m')} + \frac{\alpha}{N}\sum_{i=tN}^{(t+1)N-1}\big(A_i^{(m)}\theta_t^{(m)} + \widetilde{b}_i^{(m)} + B_i^{(m)}w_t^{(m)}\big), \tag{7}$$

$$w_{t+1}^{(m)} = \sum_{m'\in\mathcal{N}_m} V_{m,m'}w_t^{(m')} + \frac{\beta}{N}\sum_{i=tN}^{(t+1)N-1}\big(A_i^{(m)}\theta_t^{(m)} + \widetilde{b}_i^{(m)} + C_i w_t^{(m)}\big), \tag{8}$$

where $A_i^{(m)}, B_i^{(m)}, \widetilde{b}_i^{(m)}$ are defined by replacing the global variables ρ_i and R_i involved in $A_i, B_i, b_i^{(m)}$ (see eq. (3)) with local variables $\widehat{\rho}_i^{(m)}$ and $R_i^{(m)}$ respectively. To summarize, every TDC iteration of Algorithm 1 consumes N Markovian samples, and requires two vector communication rounds for synchronizing the parameter vectors $\theta_t^{(m)}, w_t^{(m)}$, and L scalar communication rounds for estimating the global importance sampling ratio ($\{\rho_{i,\ell}^{(m)} : i = tN, \ldots, (t+1)N-1, m\in\mathcal{M}\}$ are shared in the ℓ-th communication round). We summarize these update rules in Algorithm 1. Moreover, after the decentralized TDC updates, the agents perform additional T' local averaging steps to reach a global consensus on the model parameters.

4 Finite-Time Analysis of Decentralized TDC

In this section, we analyze the finite-time convergence of Algorithm 1. Denote μ_{π_b} as the stationary distribution of the Markov chain $\{s_t\}_t$ induced by the collection of agents' behavioral policies π_b. Throughout the analysis, we define the following expected quantities.

$$A := \mathbb{E}_{\pi_b}[A_t],\ B := \mathbb{E}_{\pi_b}[B_t],\ C := \mathbb{E}_{\pi_b}[C_t],\ b^{(m)} := \mathbb{E}_{\pi_b}\big[b_t^{(m)}\big],\ \overline{b}_t := \tfrac{1}{M}\textstyle\sum_{m=1}^M b_t^{(m)},\ \overline{b} := \mathbb{E}_{\pi_b}\big[\overline{b}_t\big],$$

where $\mathbb{E}_{\pi_b}$ denotes the expectation when $s_t\sim\mu_{\pi_b}$, $a_t^{(m)}\sim\pi_b^{(m)}(s_t)$ and $s_{t+1}\sim P(\cdot|s_t,a_t)$. It is well-known that the optimal model parameter is $\theta^* = -A^{-1}\overline{b}$ Xu et al. (2020b); Xu & Liang (2020). We make the following standard assumptions.

Assumption 1. *There exist constants $\nu > 0$ and $\delta\in(0,1)$ such that for all $t\geq 0$,*

$$\sup_{s\in\mathcal{S}} d_{TV}\left(\mathbb{P}_{\pi_b}\left(s_t \mid s_0 = s\right), \mu_{\pi_b}\right) \leq \nu\delta^t, \tag{9}$$

where d_{TV} denotes the total-variation distance.

Algorithm 1 Decentralized mini-batch TDC.

Input: Batch size N, iterations T, T', learning rates α, β.
Initialize: $\theta_0^{(m)}, w_0^{(m)}$ for all agents $m \in \mathcal{M}$.
for *iteration* $t = 0, 1, \ldots, T-1$ **do**
  Each agent collects N Markovian samples and computes their local importance sampling ratios $\rho_t^{(m)}$.
  for *agent* $m \in \mathcal{M}$ *in parallel* **do**
    Agent m estimates global importance sampling ratios for the N samples via eqs. (5) and (6) and performs the updates in eqs. (7) and (8).
  end
end
for *iteration* $t = T, T+1, \ldots, T+T'-1$ **do**
  for *agent* $m \in \mathcal{M}$ *in parallel* **do**
    $\theta_{t+1}^{(m)} = \sum_{m' \in \mathcal{N}_m} V_{m,m'} \theta_t^{(m')}$.
  end
end
Output: $\{\theta_{T+T'}^{(m)}\}_{m=1}^M$.

Assumption 2. *The matrices A and C are invertible.*

Assumption 3. *The feature vectors satisfy $\|\phi(s)\| \leq 1, \forall s$.*

Assumption 4. *There exist $R_{\max}, \rho_{\max} > 0$ such that for all $m \in \mathcal{M}$: $\max_{s,a,s'} R^{(m)}(s,a,s') < R_{\max}$ and $\max_{s,a^{(m)}} \rho^{(m)}(s,a^{(m)}) < \rho_{\max}$.*

Assumption 5. *The communication matrix V is doubly stochastic, and its second largest singular value satisfies $\sigma_2 \in [0,1)$.*

Assumption 1 has been widely adopted in the existing literature Bhandari et al. (2018); Xu et al. (2019); Xu & Liang (2020); Shaocong et al. (2020; 2021). It holds for all homogeneous Markov chains with finite state-space and all uniformly ergodic Markov chains. Assumptions 2 – 4 are widely adopted in the analysis of TD learning algorithms Xu et al. (2019); Xu & Liang (2020). In particular, Assumption 2 implies that $\lambda_1 := -\lambda_{\max}(A^\top C^{-1} A) > 0$, $\lambda_2 := -\lambda_{\max}(C) > 0$. Assumption 3 can always hold by normalizing the feature vectors. Assumption 4 holds for any uniformly lower bounded behavior policy, i.e., $\inf_{(s,a,m)} \pi_b^{(m)}(s,a) > 0$, which ensures that every state-action pair (s,a) is visited infinitely often. Assumption 5 is standard in decentralized optimization Singh et al. (2020); Saha et al. (2020) and TD learning Sun et al. (2020); Wang et al. (2020).

We obtain the following finite-time error bound for Algorithm 1 with Markovian samples.

Theorem 1. *Let Assumptions 1–5 hold. Run Algorithm 1 for T iterations with learning rates $\alpha \leq \min\{\mathcal{O}(\beta), \mathcal{O}(\frac{1-\sigma_2}{\sqrt{M}})\}$, $\beta \leq \mathcal{O}(1)$, batch size $N \geq \max\{\mathcal{O}(1), \mathcal{O}(\frac{\beta}{\alpha})\}$ and $L \geq \mathcal{O}(\frac{\ln M + M \ln \rho_{\max}}{\ln \sigma_2^{-1}})$ (see eq. (35)-(38)). Then, we have*

$$\mathbb{E}\big[\big\|\overline{\theta}_T - \theta^*\big\|^2\big] \leq \Big(1 - \frac{\alpha\lambda_1}{6}\Big)^T \big(\big\|\overline{\theta}_0 - \theta^*\big\|^2 + \|\overline{w}_0 - w_0^*\|^2\big) + \mathcal{O}\Big(\frac{\beta}{N\alpha} + \frac{\beta\sigma_2^{L/4} 2^T}{M}\Big). \tag{10}$$

Furthermore, after T' iterations of local averaging, the local models of all agents satisfy that: for all $m = 1, ..., M$,

$$\mathbb{E}\big[\|\theta_{T+T'}^{(m)} - \overline{\theta}_T\|^2\big] \leq \sigma_2^{2T'} \mathcal{O}\Big(1 + \frac{M^4\beta\alpha}{(1-\sigma_2)^2} + \frac{M\beta\alpha\sigma_2^{L/4} 2^T}{1-\sigma_2}\Big). \tag{11}$$

Consequently, by choosing $\alpha = \mathcal{O}(\frac{1-\sigma_2}{\sqrt{M}})$, $\beta = \mathcal{O}(1)$, $T = \mathcal{O}\big(\frac{\sqrt{M}\ln\epsilon^{-1}}{1-\sigma_2}\big)$, $N = \mathcal{O}\big(\frac{\sqrt{M}}{\epsilon(1-\sigma_2)}\big)$, $L = \mathcal{O}\big(\frac{\sqrt{M}\ln\epsilon^{-1}}{(1-\sigma_2)^2} + \frac{M}{1-\sigma_2}\big)$, $T' = \mathcal{O}\big(\frac{1}{1-\sigma_2}\ln\frac{M}{\epsilon(1-\sigma_2)}\big)$, we obtain that $\mathbb{E}(\|\theta_{T+T'}^{(m)} - \theta^*\|^2) \leq \epsilon$ for all m. The overall communication complexity for synchronizing $\theta_t^{(m)}$ is $T + T' = \mathcal{O}\big(\frac{\sqrt{M}\ln\epsilon^{-1}}{1-\sigma_2}\big)$, and the total sample complexity is $NT = \mathcal{O}\big(\frac{M\ln\epsilon^{-1}}{\epsilon(1-\sigma_2)^2}\big)$.

The above theorem shows that our decentralized TDC achieves the sample complexity $\mathcal{O}\big(\frac{M\ln\epsilon^{-1}}{\epsilon(1-\sigma_2)^2}\big)$, which, in the centralized setting ($M=1$, $\sigma_2=0$), matches that of centralized TDC for Markovian samples Xu et al. (2019) and matches the theoretical lower bound $\mathcal{O}(\epsilon^{-1})$ given in Kaledin et al. (2020) up to a logarithm factor. In addition, the sample complexity is proportional to M, which matches the theoretical lower bound of decentralized strongly convex optimization in Scaman et al. (2017). Importantly, the communication complexity $\mathcal{O}\big(\frac{\sqrt{M}\ln\epsilon^{-1}}{1-\sigma_2}\big)$ is substantially lower than the communication complexity $\mathcal{O}\big(\big(\epsilon^{-1}+\frac{\sqrt{M}}{\sqrt{\epsilon}(1-\sigma_2)}\big)\ln\epsilon^{-1}\big)$ of decentralized TD(0) Sun et al. (2020) [1] . Intuitively, this is because our algorithm adopts mini-batch sampling that significantly reduces the communication frequency, since communication occurs after collecting a mini-batch of samples to compute the mini-batch updates. Moreover, the communication complexity has a logarithm dependence $\ln\epsilon^{-1}$ on the target accuracy, and this matches the theoretical lower bound of decentralized strongly convex optimization in Scaman et al. (2017).

Taking a deeper look, Theorem 1 shows that the average model $\overline{\theta}_T$ converges to a small neighborhood of the optimal solution θ^* at a fast linear convergence rate (10) that matches the convergence rate of centralized TDC Xu et al. (2019); Xu & Liang (2020). In particular, the convergence error is in the order of $\mathcal{O}\big(\frac{\beta}{N\alpha}+\frac{\beta\sigma_2^{L/4}2^T}{M}\big)$, which can be driven arbitrarily close to zero by choosing a sufficiently large mini-batch size N and communication rounds L (with T fixed), and choosing constant-level learning rates α, β. Moreover, the T' steps of extra local model averaging further help all the agents achieve a small consensus error at a linear convergence rate (11). Eqs. (10) and (11) together ensure the fast convergence of all the local model parameters. We want to point out that the T' local averaging steps are critical for establishing fast convergence of local model parameters. Specifically, without the T' local averaging steps, the consensus error $\mathbb{E}\big[\|\theta_T^{(m)}-\overline{\theta}_T\|^2\big]$ would be in the order of at least $\mathcal{O}(1)$, which is constant-level and hence cannot guarantee the local model parameters converge arbitrarily close to the true solution.

Proof sketch of Theorem 1. The proof of the theorem is a nontrivial generalization of the analysis of centralized off-policy TDC to the decentralized case. Below, we sketch the technical proof and elaborate on the technical novelties.

- **Step 1.** We first consider an ideal case where the agents can access the exact global importance sampling ratio ρ_t at iteration t. In this ideal case, every agent m can replace the estimated global importance sampling ratio $\widehat{\rho}_i^{(m)}$ involved in $A_i^{(m)}$, $B_i^{(m)}$, $\widetilde{b}_i^{(m)}$ in the update rules (7) and (8) by the exact value ρ_i. Then, with the notations defined in Appendix A, the averaged update rules (14) and (15) become

$$\widetilde{\theta}_{t+1} = \overline{\theta}_t + \alpha\big(\overline{A}_t\overline{\theta}_t + \overline{\overline{b}}_t + \overline{B}_t\overline{w}_t\big), \tag{12}$$

$$\widetilde{w}_{t+1} = \overline{w}_t + \beta\big(\overline{A}_t\overline{\theta}_t + \overline{\overline{b}}_t + \overline{C}_t\overline{w}_t\big), \tag{13}$$

  which can be seen as one step of centralized TDC. Hence, we can bound its optimization error terms $\mathbb{E}\big[\|\widetilde{w}_{t+1}-w_t^*\|^2\big]$ and $\mathbb{E}\big[\|\widetilde{\theta}_{t+1}-\theta^*\|^2\big]$.

- **Step 2.** We return to Algorithm 1 and bound its optimization error terms $\mathbb{E}\big[\|\overline{w}_{t+1}-w_t^*\|^2$, $\mathbb{E}\big[\|\overline{w}_{t+1}-w_{t+1}^*\|^2$ and $\mathbb{E}\big[\|\overline{\theta}_{t+1}-\theta^*\|^2$. This is done by bounding the gap between the centralized updates (12) and (13) (with exact ρ_t) and the decentralized updates (14) and (15) (with inexact ρ_t). The key is to establish Corollary 2, which strategically controls the gap between the inexact global importance sampling ratio $\widehat{\rho}_t^{(m)}$ and the exact value ρ_t.

  To elaborate, note that the locally-averaged importance sampling ratios $\widetilde{\rho}_{t,\ell}^{(m)}$ in eq. (5) exponentially converges to the value $\ln\rho_t$. However, the initial gap $|\ln\rho_t^{(m)}-\ln\rho_t|$ can be numerically large since $\rho_t^{(m)}$ may be a numerically small positive number. To avoid such divergence issue, our proof discusses two complementary cases. Case 1: the quantity $\rho_{\min}:=\min_{m\in\mathcal{M}}\rho_t^{(m)}\in[\sigma_2^{L/2},\rho_{\max}]$. In this case, the proof is straightforward as the initial gap is bounded. Case 2: the quantity $\rho_{\min}\in(0,\sigma_2^{L/2}]$. In this case, we show that the locally-averaged logarithm-ratio $\widetilde{\rho}_{t,L}^{(m)}$ is below a large negative number $\mathcal{O}\big(\frac{\ln\rho_{\min}}{M}\big)\ll 0$ (See eq. (63)). which implies that both the global importance sampling ratio ρ_t and its estimation $\widehat{\rho}_t^{(m)}$ are close to zero. In both cases, $\{\widehat{\rho}_t^{(m)}\}_{m=1}^M$ converge exponentially fast to ρ_t as L increases. This prove eq. (10).

[1] Sun et al. (2020) does not report communication complexity, so we calculated it based on their finite-time error bound.

- **Step 3.** Finally, we prove the consensus error (11). Although the consensus error exponentially decays during the T' extra local average steps in Algorithm 1, it is non-trivial to bound the initial consensus error $\|\Delta\Theta_T\|_F$ of the T' local average iterations (see eq. (34)), which is caused by the T decentralized TDC steps. To bound this error, note that each decentralized TDC step consists of both local averaging and TDC update, which makes the consensus error $\|\Delta\Theta_t\|_F$ diminishes geometrically fast with a noise term $\sum_{m=1}^{M}\|h_m\|$ (see eq. (30)). Such a noise term is induced by the TDC update and hence its bound depends on both the consensus error and the model estimation error in eq. (10). We need to apply these correlated bounds iteratively for T iterations to bound the initial consensus error $\|\Delta\Theta_T\|_F$.

5 Experiments

5.1 Simulated Multi-Agent Networks

We simulate a multi-agent MDP with 10 decentralized agents. The shared state space contains 10 states and each agent can take 2 actions. All behavior policies are uniform policies (i.e., each agent takes all actions with equal probability), and the target policies are obtained by first perturbing the corresponding behavior policies with Gaussian noises sampled from $\mathcal{N}(0, 0.05)$ and then performing a proper normalization. The entries of the transition kernel and the reward functions are independently generated from the uniform distribution on $[0, 1]$ (with proper normalization for the transition kernel). We generate all state features with dimension 5 independently from the standard Gaussian distribution and normalize them to have unit norm. The discount factor is $\gamma = 0.95$.

We consider two types of network topologies: a fully connected network with communication matrix V having diagonal entries 0.8 and off-diagonal entries $1/45$, and a ring network with communication matrix V having diagonal entries 0.8 and entries 0.1 for adjacent agents. We implement and compare two algorithms in these networks: the decentralized TD(0) with batch size $N = 1$ (used in Sun et al. (2020)) and our decentralized TDC with batch sizes $N = 10, 20, 50, 100$.

Effect of Batch size: We test these algorithms with varying batch size N and compare their sample and communication complexities. We set learning rate $\alpha = 0.2$ for the decentralized TD(0) and $\alpha = 0.2 * N$, $\beta = 0.002 * N$ for our decentralized TDC with varying batch sizes $N = 10, 20, 50, 100$. Both algorithms use $L = 3$ communication rounds for synchronizing $\widehat{\rho}_t^{(m)}$. All algorithms are repeated 100 times using a fixed set of 100 MDP trajectories, each of which has 20k Markovian samples.

We first implement these algorithms in the fully connected network. Figure 1 plots the relative convergence error $\|\overline{\theta}_t - \theta^*\|/\|\theta^*\|$ v.s. sample complexity (tN) and communication complexity (t). For each curve, its upper and lower envelopes denote the 95% and 5% percentiles of the 100 convergence errors, respectively. It can be seen that our decentralized TDC with different batch sizes achieve almost the same sample complexity as that of the decentralized TD(0), demonstrating the sample-efficiency of our algorithms. On the other hand, our decentralized TDC require much less communication complexities than the decentralized TD(0), and the required communication becomes lighter as batch size increases. All these results match our theoretical analysis well.

We further implement these algorithms in the ring network. The comparison results are exactly the same as those in Figure 1, since the update rule of $\overline{\theta}_t$ does not rely on the network topology under exact global importance sampling.

Effect of Communication Rounds: We test our decentralized TDC using varying communication rounds $L = 1, 3, 5, 7$. We use a fixed batch size $N = 100$ and set learning rates $\alpha = 5$, $\beta = 0.05$, and repeat each algorithm 100 times using the set of 100 MDP trajectories. We also implement the decentralized TDC with exact global importance sampling ratios as a baseline. Figure 2 plots the relative convergence error v.s. communication complexity in the fully-connected network (Left) and ring network (Right). It can be seen that in both networks, the asymptotic convergence error of the decentralized TDC with inexact ρ decreases as the number of communication rounds L for synchronizing the global importance sampling ratio increases. In particular, with $L = 1$, decentralized TDC diverges asymptotically due to inaccurate estimation of the

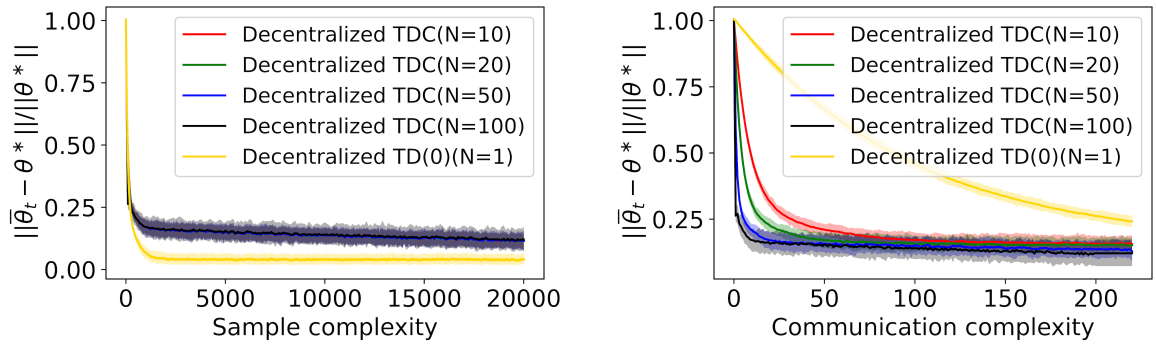

Figure 1: Comparison between decentralized TDC with varying batch sizes and decentralized TD(0).

global importance sampling ratio. As L increases to more than 5, the convergence error is as small as that under exact global importance sampling.

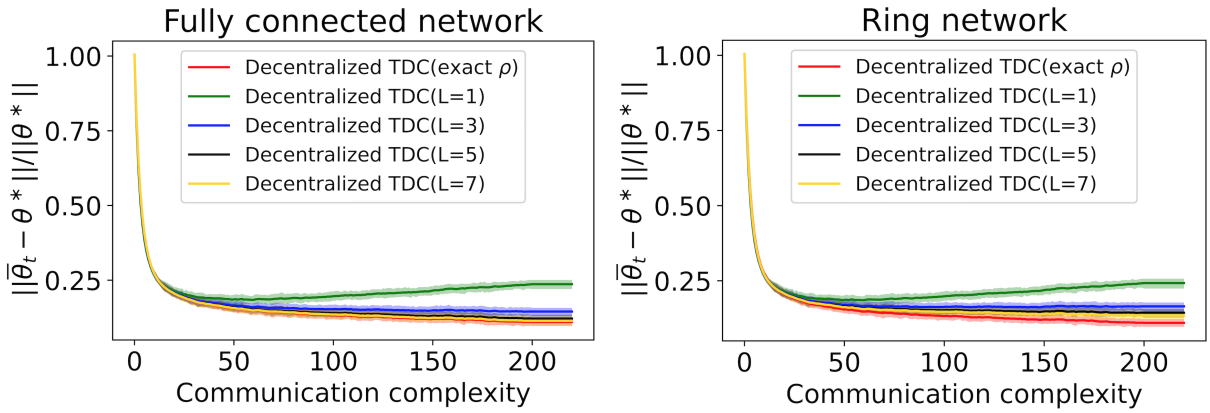

Figure 2: Effect of communication rounds L on asymptotic convergence error.

We further plot the maximum relative consensus error among all agents $\max_m \|\theta_t^{(m)} - \overline{\theta}_t\|/\|\overline{\theta}^*\|$ v.s. communication complexity (t) in the fully-connected network (Left) and ring network (Right) in Figure 3, where the tails in both figures correspond to the extra $T' = 20$ local model averaging steps. In both networks, one can see that the consensus error decreases as L increases, and the extra local model averaging steps are necessary to achieve consensus. Moreover, it can be seen that the consensus errors achieved in the fully connected network are slightly smaller than those achieved in the ring network, as denser connections facilitate achieving the global consensus.

5.2 Two-Agent Cliff Navigation Problem

In this subsection, we test our algorithms in solving a two-agent Cliff Navigation problem Qiu et al. (2021) in a grid-world environment. This problem is adapted from its single-agent version (see Example 6.6 of Sutton & Barto (2018)). As illustrated in Figure 4, two agents start from the starting point "S" on a 3×4 grid and aim to reach the destination "D". Here, global state is defined as the joint location of the two agents, and there are in total $(3 \times 4)^2 = 144$ global states. In most states, an agent can choose to move up, down, left or right by one step and receives -1 reward. However, once an agent falls into the cliff "X", it will return to the starting point "S" and receive -100 reward. When an agent reaches "D", it will always stay at "D", and receives 0 reward if the other agent also reaches/stays at "D", or receives -0.5 reward otherwise. If an agent

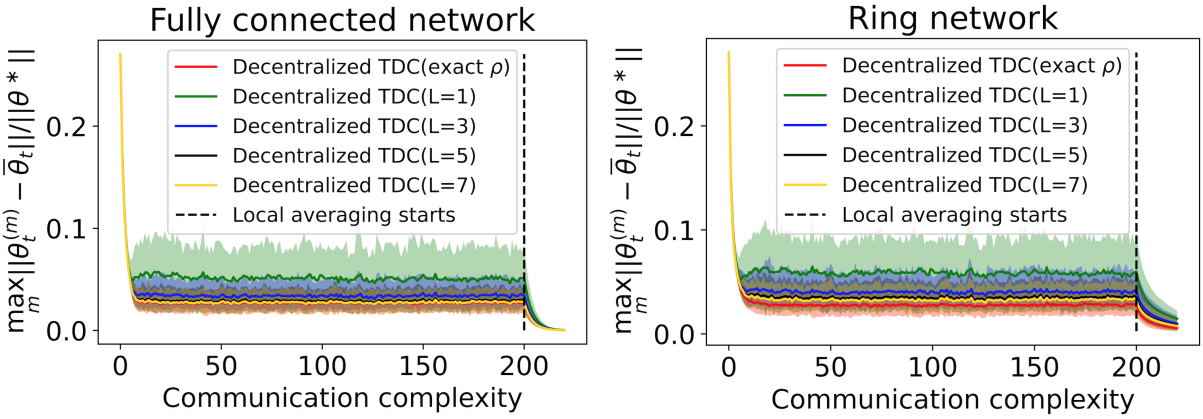

Figure 3: Effect of communication rounds L on consensus error.

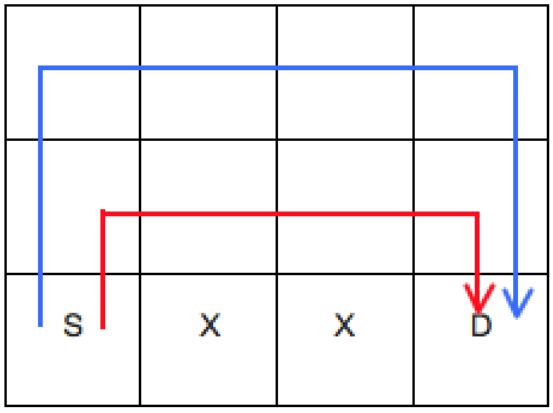

Figure 4: Two-agent cliff navigation. ("S", "X", "D" denote starting point, cliff and destination respectively. The optimal path is shown in red.)

is not at "X" or "D" and selects a direction that points outside the grid, then it stays in the previous location and receives -1 reward.

We apply the aforementioned algorithms with different batchsizes N to solve this problem. The hyperparameters and the ways to generate behavior policy and target policy are the same as the previous simulation experiment, except that $L = 3$ and the communication matrix V has diagonal entries 0.7 and off-diagnoals 0.3. All algorithms are repeated 100 times using a fixed set of 100 MDP trajectories, each of which has 20k Markovian samples. Figure 5 plots the relative convergence error $\|\bar{\theta}_t - \theta^*\|/\|\theta^*\|$ v.s. sample complexity (tN) and communication complexity (t). We can see that compared with the decentralized TD(0), our decentralized TDC with different batch sizes achieve comparable sample complexities and much lower communication complexities. Moreover, the required communication becomes lighter as batch size increases. These properties are similar to those shown in the simulation and thus have generality.

6 Conclusion

In this paper, we develop a sample-efficient and communication-efficient decentralized TDC algorithm for multi-agent off-policy evaluation. Our algorithm synchronizes the local importance sampling ratios among the agents and adopts mini-batch stochastic updates to save communication. In particular, it avoids sharing agents' sensitive local information. We prove that the proposed decentralized TDC algorithms achieve a

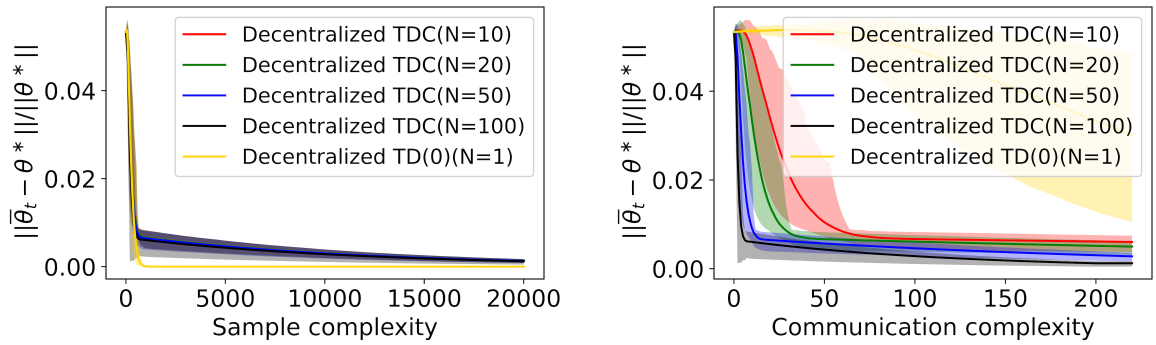

Figure 5: Comparison between decentralized TDC with varying batch sizes and decentralized TD(0).

near-optimal sample complexity as well as an optimal communication complexity that improves over the existing decentralized TD(0). In the future, we expect that our algorithm can serve as a fundamental component in the design of advanced policy optimization algorithms for MARL.

References

Leemon Baird. Residual algorithms: Reinforcement learning with function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 30–37, 1995.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Proc. Conference on Learning Theory (COLT)*, volume 75, pp. 1691–1692, 2018.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. 2009.

Lucas Cassano, Kun Yuan, and Ali H Sayed. Multi-agent fully decentralized value function learning with linear convergence rates. *IEEE Transactions on Automatic Control*, 2020.

Behdad Chalaki and Andreas A Malikopoulos. A hysteretic q-learning coordination framework for emerging mobility systems in smart cities. *ArXiv:2011.03137*, 2020.

Tianyi Chen, Kaiqing Zhang, Georgios B Giannakis, and Tamer Basar. Communication-efficient policy gradient methods for distributed reinforcement learning. *IEEE Transactions on Control of Network Systems*, 2021.

Gal Dalal, Balazs Szorenyi, Gugan Thoppe, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Proc. Conference on Learning Theory (COLT)*, 2018a.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for td (0) with function approximation. In *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, volume 32, 2018b.

Gal Dalal, Balazs Szorenyi, and Gugan Thoppe. A tale of two-timescale reinforcement learning with the tightest finite-time bound. In *Proc. Association for the Advancement of Artificial Intelligence (AAAI)*, volume 34, pp. 3701–3708, 2020.

Peter Dayan. The convergence of td (λ) for general λ. *Machine learning*, 8(3-4):341–362, 1992.

Thinh Doan, Siva Maguluri, and Justin Romberg. Finite-time analysis of distributed TD(0) with linear function approximation on multi-agent reinforcement learning. In *Proc. International Conference on Machine Learning (ICML)*, volume 97, pp. 1626–1635, 09–15 Jun 2019.

Geoffrey J Gordon. Stable function approximation in dynamic programming. In *Machine Learning Proceedings 1995*, pp. 261–268. 1995.

Harsh Gupta, R. Srikant, and Lei Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 4704–4713, 2019.

Bin Hu and Usman Ahmed Syed. Characterizing the exact behaviors of temporal difference learning algorithms using markov jump linear system theory. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8479–8490, 2019.

Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, volume 6, pp. 703–710, 1993.

Maxim Kaledin, Eric Moulines, Alexey Naumov, Vladislav Tadic, and Hoi-To Wai. Finite time analysis of linear two-timescale stochastic approximation with markovian noise. In *Proc. Conference on Learning Theory (COLT)*, pp. 2144–2203, 2020.

Nathaniel Korda and Prashanth La. On td (0) with function approximation: Concentration bounds and a centered variant with exponential convergence. In *Proc. International Conference on Machine Learning (ICML)*, pp. 626–634, 2015.

Vikram Krishnamurthy, Michael Maskery, and George Yin. Decentralized adaptive filtering algorithms for sensor activation in an unattended ground sensor network. *IEEE Transactions on Signal Processing*, 56 (12):6086–6101, 2008.

Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 1347–1355, 2018.

Bo Liu, Ji Liu, Mohammad Ghavamzadeh, Sridhar Mahadevan, and Marek Petrik. Finite-sample analysis of proximal gradient td algorithms. In *Proc. Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 504–513, 2015.

Sergio Valcarcel Macua, Jianshu Chen, Santiago Zazo, and Ali H Sayed. Distributed policy evaluation under multiple behavior strategies. *IEEE Transactions on Automatic Control*, 60(5):1260–1274, 2014.

Hamid Reza Maei. *Gradient temporal-difference learning algorithms*. PhD thesis, University of Alberta, 2011.

Adwaitvedant Mathkar and Vivek S Borkar. Distributed reinforcement learning via gossip. *IEEE Transactions on Automatic Control*, 62(3):1465–1470, 2016.

Wei Qiu, Xinrun Wang, Runsheng Yu, Rundong Wang, Xu He, Bo An, Svetlana Obraztsova, and Zinovi Rabinovich. Rmix: Learning risk-sensitive policies for cooperative reinforcement learning agents. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Guannan Qu and Na Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2017.

Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning of localized policies for multi-agent networked systems. In *Learning for Dynamics and Control (L4DC)*, pp. 256–266, 2020.

Rajarshi Saha, Stefano Rini, Milind Rao, and Andrea Goldsmith. Decentralized optimization over noisy, rate-constrained networks: How to agree by talking about how we disagree. *ArXiv:2010.11292*, 2020.

Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for smooth and strongly convex distributed optimization in networks. In *Proc. International Conference on Machine Learning (ICML)*, volume 70, pp. 3027–3036, 2017.

Ma Shaocong, Zhou Yi, and Zou Shaofeng. Variance-reduced off-policy tdc learning: Non-asymptotic convergence analysis. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Ma Shaocong, Chen Ziyi, Zhou Yi, and Zou Shaofeng. Greedy-{gq} with variance reduction: Finite-time analysis and improved complexity. In *Proc. International Conference on Learning Representations (ICLR)*, 2021.

Navjot Singh, Deepesh Data, Jemin George, and Suhas Diggavi. Squarm-sgd: Communication-efficient momentum sgd for decentralized optimization. *ArXiv:2005.07041*, 2020.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and td learning. In *Proc. Conference on Learning Theory (COLT)*, pp. 2803–2830, 2019.

Miloš S Stanković and Srdjan S Stanković. Multi-agent temporal-difference learning with linear function approximation: Weak convergence under time-varying network topologies. In *Proc. American Control Conference (ACC)*, pp. 167–172, 2016.

Jun Sun, Gang Wang, Georgios B Giannakis, Qinmin Yang, and Zaiyue Yang. Finite-sample analysis of decentralized temporal-difference learning with linear function approximation. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 4485–4495, 2020.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

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

Richard S Sutton, Csaba Szepesvári, and Hamid Reza Maei. A convergent o(n) algorithm for off-policy temporal-difference learning with linear function approximation. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, volume 21, pp. 1609–1616, 2008.

Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proc. International Conference on Machine Learning (ICML)*, pp. 993–1000, 2009.

Vladislav Tadić. On the convergence of temporal-difference learning with linear function approximation. *Machine learning*, 42(3):241–267, 2001.

John N Tsitsiklis and Benjamin Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE transactions on automatic control*, 42(5):674–690, 1997.

Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, and Mingyi Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 9672–9683, 2018.

Gang Wang, Bingcong Li, and Georgios B Giannakis. A multistep lyapunov approach for finite-time analysis of biased stochastic approximation. *ArXiv:1909.04299*, 2019.

Gang Wang, Songtao Lu, Georgios Giannakis, Gerald Tesauro, and Jian Sun. Decentralized td tracking with linear function approximation and its finite-time analysis. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Tengyu Xu and Yingbin Liang. Sample complexity bounds for two timescale value-based reinforcement learning algorithms. *ArXiv:2011.05053*, 2020.

Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy td learning: Non-asymptotic analysis over markovian samples. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10634–10644, 2019.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020a.

Tengyu Xu, Zhe Wang, Yi Zhou, and Yingbin Liang. Reanalysis of variance reduced temporal difference learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2020b.

Zhi Yan, Nicolas Jouandeau, and Arab Ali Cherif. A survey and analysis of multi-robot coordination. *International Journal of Advanced Robotic Systems*, 10(12):399, 2013.

Evşen Yanmaz, Markus Quaritsch, Saeed Yahyanejad, Bernhard Rinner, Hermann Hellwagner, and Christian Bettstetter. Communication and coordination for drone networks. In *Proc. International Conference on Ad Hoc Networks*, pp. 79–91, 2017.

Mingqi Yuan, Qi Cao, Man-on Pun, and Yi Chen. Towards user scheduling for 6g: A fairness-oriented scheduler using multi-agent reinforcement learning. *ArXiv:2012.15081*, 2020.

Junyu Zhang, Amrit Singh Bedi, Mengdi Wang, and Alec Koppel. Marl with general utilities via decentralized shadow reward actor-critic. *ArXiv:2106.00543*, 2021.

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. In *Proc. Advances in Neural Information Processing Systems*, pp. 8665–8675, 2019.

A Notations and Filtration

A.1 Notations to rewrite update rules in Algorithm 1

We introduce the following additional notations.

$\overline{\theta}_t = \frac{1}{M}\sum_{m=1}^M \theta_t^{(m)}$, $\overline{w}_t = \frac{1}{M}\sum_{m=1}^M w_t^{(m)}$ $b_t^{(m)} = \rho_t R_t^{(m)}\phi(s_t)$, $\overline{b}_t = \frac{1}{M}\sum_{m=1}^M b_t^{(m)}$,

$\overline{A}_t = \frac{1}{N}\sum_{i=tN}^{(t+1)N-1} A_i$, $\overline{B}_t = \frac{1}{N}\sum_{i=tN}^{(t+1)N-1} B_i$, $\overline{C}_t = \frac{1}{N}\sum_{i=tN}^{(t+1)N-1} C_i$, $\overline{b}_t^{(m)} = \frac{1}{N}\sum_{i=tN}^{(t+1)N-1} b_i^{(m)}$,

$\overline{\overline{b}}_t = \frac{1}{N}\sum_{i=tN}^{(t+1)N-1} \overline{b}_i = \frac{1}{M}\sum_{m=1}^M \overline{b}_t^{(m)}$, $\overline{A}_t^{(m)} = \frac{1}{N}\sum_{i=tN}^{(t+1)N-1} A_i^{(m)}$, $\overline{B}_t^{(m)} = \frac{1}{N}\sum_{i=tN}^{(t+1)N-1} B_i^{(m)}$,

$\overline{\widetilde{b}}_t^{(m)} = \frac{1}{N}\sum_{i=tN}^{(t+1)N-1} \widetilde{b}_i^{(m)}$.

Then, by averaging the update rules (7) and (8) over m, we obtain the following update rules of the model average $\overline{\theta}_t, \overline{w}_t$.

$$\overline{\theta}_{t+1} = \overline{\theta}_t + \frac{\alpha}{M}\sum_{m=1}^M \big(\overline{A}_t^{(m)}\theta_t^{(m)} + \overline{\overline{b}}_t^{(m)} + \overline{B}_t^{(m)} w_t^{(m)}\big), \tag{14}$$

$$\overline{w}_{t+1} = \overline{w}_t + \frac{\beta}{M}\sum_{m=1}^M \big(\overline{A}_t^{(m)}\theta_t^{(m)} + \overline{\overline{b}}_t^{(m)} + \overline{C}_t w_t^{(m)}\big). \tag{15}$$

A.2 Filtration

Define the filtration $\mathcal{F}_t = \sigma\big(\{s_{t'}, a_{t'}\}_{t'=1}^{tN-1} \cup \{s_{tN}\}\big)$. Then,

$$\overline{A}_t, \overline{B}_t, \overline{C}_t, \overline{b}_t^{(m)}, \overline{\overline{b}}_t, \overline{A}_t^{(m)}, \overline{B}_t^{(m)}, \overline{\overline{b}}_t^{(m)} \in \mathcal{F}_{t+1}/\mathcal{F}_t, \qquad \theta_t^{(m)}, \overline{\theta}_t, w_t^{(m)}, \overline{w}_t, w_t^* \in \mathcal{F}_t/\mathcal{F}_{t-1}.$$

B Proof of Theorem 1

We first bound the following tracking errors of $\widetilde{w}_{t+1}$, $\widetilde{\theta}_{t+1}$ obtained by the centralized update rules (13) and (12) respectively.

$$\mathbb{E}\big[\|\widetilde{w}_{t+1} - w_t^*\|^2\big|\mathcal{F}_t\big]$$

$$= \|\overline{w}_t - w_t^*\|^2 + 2\beta \underbrace{(\overline{w}_t - w_t^*)^\top \mathbb{E}\big[\overline{A}_t\overline{\theta}_t + \overline{\overline{b}}_t + \overline{C}_t\overline{w}_t\big|\mathcal{F}_t\big]}_{(I)} + \beta^2 \underbrace{\mathbb{E}\big[\big\|\overline{A}_t\overline{\theta}_t + \overline{\overline{b}}_t + \overline{C}_t\overline{w}_t\big\|^2\big|\mathcal{F}_t\big]}_{(II)}, \tag{16}$$

$$\begin{aligned}&\mathbb{E}\big[\|\widetilde{\theta}_{t+1} - \theta^*\|^2\big|\mathcal{F}_t\big]\\ &= \|\overline{\theta}_t - \theta^*\|^2 + 2\alpha \underbrace{\big(\overline{\theta}_t - \theta^*\big)^\top \mathbb{E}\big[\overline{A}_t\overline{\theta}_t + \overline{\overline{b}}_t + \overline{B}_t\overline{w}_t\big|\mathcal{F}_t\big]}_{(III)} + \alpha^2 \underbrace{\mathbb{E}\big[\big\|\overline{A}_t\overline{\theta}_t + \overline{\overline{b}}_t + \overline{B}_t\overline{w}_t\big\|^2\big|\mathcal{F}_t\big]}_{(IV)}.\end{aligned} \tag{17}$$

The above four terms (I)-(IV) are respectively bounded below.

$$\begin{aligned}(I) &= (\overline{w}_t - w_t^*)^\top \mathbb{E}\big[\overline{A}_t\overline{\theta}_t + \overline{\overline{b}}_t + \overline{C}_t\overline{w}_t\big|\mathcal{F}_t\big]\\
&\overset{(i)}{=} 2(\overline{w}_t - w_t^*)^\top \mathbb{E}\big[\overline{C}_t - C\big|\mathcal{F}_t\big](\overline{w}_t - w_t^*) + 2(\overline{w}_t - w_t^*)^\top C(\overline{w}_t - w_t^*) + 2(\overline{w}_t - w_t^*)^\top\\
&\quad \mathbb{E}\big[(\overline{A}_t - \overline{C}_tC^{-1}A)\overline{\theta}_t + \overline{\overline{b}}_t - \overline{C}_tC^{-1}b\big|\mathcal{F}_t\big]\\
&\overset{(ii)}{\le} 2\big\|\mathbb{E}\big[\overline{C}_t - C\big|\mathcal{F}_t\big]\big\|_F\big\|\overline{w}_t - w_t^*\big\|^2 - 2\lambda_2\big\|\overline{w}_t - w_t^*\big\|^2 + \lambda_2\|\overline{w}_t - w_t^*\|^2 + \frac{3}{\lambda_2}\mathbb{E}\big[\big\|\overline{\overline{b}}_t - \overline{C}_tC^{-1}b\big\|^2\big|\mathcal{F}_t\big]\\
&\quad + \frac{3}{\lambda_2}\mathbb{E}\big[\big\|(\overline{A}_t - \overline{C}_tC^{-1}A)\big\|_F^2\big|\mathcal{F}_t\big]\big\|\overline{\theta}_t - \theta^*\big\|^2 + \frac{3}{\lambda_2}\mathbb{E}\big[\big\|\overline{A}_t - \overline{C}_tC^{-1}A\big\|_F^2\big|\mathcal{F}_t\big]\big\|\theta^*\big\|^2\\
&\overset{(iii)}{\le} \Big(\frac{4\nu\rho_{\max}}{N(1-\delta)} - \lambda_2\Big)\big\|\overline{w}_t - w_t^*\big\|^2 + \frac{96\rho_{\max}^2(\nu+1)}{N\lambda_2(1-\delta)}\Big(1 + \frac{1}{\lambda_2}\Big)^2\big(\big\|\overline{\theta}_t - \theta^*\big\|^2 + \big\|\theta^*\big\|^2 + R_{\max}^2\big)\end{aligned} \tag{18}$$

where (i) uses the notation that $w_t^* = -C^{-1}(A\overline{\theta}_t + b)$, (ii) uses $\lambda_2 = -\lambda_{\max}(C)$ and the inequality that $2a_1^\top a_2 \le \sigma^{-1}\|a_1\|^2 + \sigma\|a_2\|^2$ for any $a_1, a_2 \in \mathbb{R}^d$ and $\sigma > 0$, and applies Jensen's inequality to convex functions $\|\cdot\|$ and $\|\cdot\|^2$ and uses eq. (44), (iv) uses eqs. (47) and (48).

$$\begin{aligned}(II) &= \mathbb{E}\big[\big\|\overline{A}_t\overline{\theta}_t + \overline{\overline{b}}_t + \overline{C}_t\overline{w}_t\big\|^2\big|\mathcal{F}_t\big]\\
&\overset{(i)}{=} \mathbb{E}\big[\big\|(\overline{A}_t - \overline{C}_tC^{-1}A)(\overline{\theta}_t - \theta^*) + (\overline{A}_t - \overline{C}_tC^{-1}A)\theta^* + \overline{\overline{b}}_t + \overline{C}_t(\overline{w}_t - w_t^*) - \overline{C}_tC^{-1}b\big\|^2\big|\mathcal{F}_t\big]\\
&\overset{(ii)}{\le} 4\mathbb{E}\big[\|\overline{A}_t - \overline{C}_tC^{-1}A\|_F^2\big|\mathcal{F}_t\big]\|\overline{\theta}_t - \theta^*\|^2 + 4\|\overline{w}_t - w_t^*\|^2\\
&\quad + 4\mathbb{E}\big[\|\overline{A}_t - \overline{C}_tC^{-1}A\|_F^2\big|\mathcal{F}_t\big]\|\theta^*\|^2 + 4\mathbb{E}\big[\|\overline{\overline{b}}_t - \overline{C}_tC^{-1}b\|^2\big|\mathcal{F}_t\big]\\
&\overset{(iii)}{\le} \frac{128\rho_{\max}^2(\nu+1)}{N(1-\delta)}\Big(1 + \frac{1}{\lambda_2}\Big)^2\big(\|\overline{\theta}_t - \theta^*\|^2 + \|\theta^*\|^2 + R_{\max}^2\big) + 4\big\|\overline{w}_t - w_t^*\big\|^2\end{aligned} \tag{19}$$

where (i) uses the notation that $w_t^* = -C^{-1}(A\overline{\theta}_t + b)$, (ii) uses $\|\sum_{k=1}^4 a_k\|^2 \le 4\sum_{k=1}^4\|a_k\|^2$ for any $a_1, a_2, a_3, a_4 \in \mathbb{R}^d$ and eq. (41), (iii) uses eqs. (47) and (48).

$$\begin{aligned}(III) &= \big(\overline{\theta}_t - \theta^*\big)^\top \mathbb{E}\big[\overline{A}_t\overline{\theta}_t + \overline{\overline{b}}_t + \overline{B}_t\overline{w}_t\big|\mathcal{F}_t\big]\\
&\overset{(i)}{=} 2\big(\overline{\theta}_t - \theta^*\big)^\top \mathbb{E}\big[(\overline{A}_t - \overline{B}_tC^{-1}A - A^\top C^{-1}A)(\overline{\theta}_t - \theta^*) + A^\top C^{-1}A(\overline{\theta}_t - \theta^*)\\
&\quad + (\overline{A}_t - \overline{B}_tC^{-1}A - A^\top C^{-1}A)\theta^* + \overline{B}_t(\overline{w}_t - w_t^*) + \overline{\overline{b}}_t - b - (\overline{B}_t - B)C^{-1}b\big|\mathcal{F}_t\big]\\
&\overset{(ii)}{\le} 2\mathbb{E}\big[\big\|\overline{A}_t - \overline{B}_tC^{-1}A - A^\top C^{-1}A\big\|\big|\mathcal{F}_t\big]\big\|\overline{\theta}_t - \theta^*\big\|^2 - 2\lambda_1\big\|\overline{\theta}_t - \theta^*\big\|^2 + \lambda_1\|\overline{\theta}_t - \theta^*\|^2 + \big\|\overline{B}_t(\overline{w}_t - w_t^*)\big\|^2\\
&\quad + \frac{4}{\lambda_1}\mathbb{E}\big[\big\|\overline{A}_t - \overline{B}_tC^{-1}A - A^\top C^{-1}A\big\|^2\big\|\theta^*\big\|^2 + \big\|\overline{\overline{b}}_t - b\big\|^2 + \big\|(\overline{B}_t - B)C^{-1}b\big\|^2\big|\mathcal{F}_t\big]\\
&\overset{(iii)}{\le} \Big(\frac{64\rho_{\max}^2(\nu+1)}{N(1-\delta)}\Big(1 + \frac{\rho_{\max}}{\lambda_2}\Big)^2 - \lambda_1\Big)\big\|\overline{\theta}_t - \theta^*\big\|^2 + \frac{4\rho_{\max}^2}{\lambda_1}\big\|\overline{w}_t - w_t^*\big\|^2\\
&\quad + \frac{32\rho_{\max}^2(\nu+1)}{N\lambda_1(1-\delta)}\Big(4\Big(1 + \frac{\rho_{\max}}{\lambda_2}\Big)^2\big\|\theta^*\big\|^2 + R_{\max}^2 + \frac{\rho_{\max}R_{\max}}{\lambda_2}\Big)\end{aligned} \tag{20}$$

where (i) uses the notations that $w_t^* = -C^{-1}(A\overline{\theta}_t + b)$ and that $b = -A\theta^*$, and the relation that $C - B = A^\top$, (ii) uses the notation that $\lambda_1 = -\lambda_{\max}(A^\top C^{-1}A)$ and the inequality that $2a_1^\top a_2 \le \sigma^{-1}\|a_1\|^2 + \sigma\|a_2\|^2$ for

any $a_1, a_2 \in \mathbb{R}^d$ and $\sigma > 0$, and applies Jensen's inequality to the convex functions $\|\cdot\|$ and $\|\cdot\|^2$, (iii) uses eqs. (40), (42), (43), (45), (46) and (49).

$$
\begin{aligned}
(IV) &= \mathbb{E}\big[\big\|\overline{A}_t\overline{\theta}_t + \overline{\overline{b}}_t + \overline{B}_t\overline{w}_t\big\|^2\big|\mathcal{F}_t\big] \\
&\overset{(i)}{=} \mathbb{E}\big[\big\|(\overline{A}_t - \overline{B}_tC^{-1}A)(\overline{\theta}_t - \theta^*) + \overline{\overline{b}}_t - b + \overline{B}_t(\overline{w}_t - w_t^*) \\
&\quad + (\overline{A}_t - \overline{B}_tC^{-1}A - A^\top C^{-1}A)\theta^* - (\overline{B}_t - B)C^{-1}b\big\|^2\big|\mathcal{F}_t\big] \\
&\overset{(ii)}{\le} 10\mathbb{E}\big[\big\|\overline{A}_t\big\|_F^2 + \big\|\overline{B}_tC^{-1}A\big\|_F^2\big|\mathcal{F}_t\big]\big\|\overline{\theta}_t - \theta^*\big\|^2 + 5\mathbb{E}\big[\big\|\overline{A}_t - \overline{B}_tC^{-1}A - A^\top C^{-1}A\big\|_F^2\big|\mathcal{F}_t\big]\big\|\theta^*\big\|^2 \\
&\quad + 5\mathbb{E}\big[\big\|\overline{\overline{b}}_t - b\big\|^2\big|\mathcal{F}_t\big] + 5\mathbb{E}\big[\big\|\overline{B}_t\big\|_F^2\big|\mathcal{F}_t\big]\big\|\overline{w}_t - w_t^*\big\|^2 + 5\mathbb{E}\big[\big\|\overline{B}_t - B\big\|_F^2\big|\mathcal{F}_t\big]\big\|C^{-1}b\big\|^2 \\
&\overset{(iii)}{\le} 40\rho_{\max}^2\Big(1 + \frac{\rho_{\max}}{\lambda_2}\Big)^2\big\|\overline{\theta}_t - \theta^*\big\|^2 + 5\rho_{\max}^2\big\|\overline{w}_t - w_t^*\big\|^2 \\
&\quad + \frac{160\rho_{\max}^2(\nu+1)}{N(1-\delta)}\Big(1 + \frac{\rho_{\max}}{\lambda_2}\Big)^2\big(\big\|\theta^*\big\|^2 + R_{\max}^2\big)
\end{aligned} \tag{21}
$$

where (i) uses the notations that $w_t^* = -C^{-1}(A\overline{\theta}_t + b)$ and that $b = -A\theta^*$, and the relation that $C - B = A^\top$, (ii) uses the inequality that $\|\sum_{k=1}^K a_k\|^2 \le K\sum_{k=1}^K \|a_k\|^2$ for any $a_k \in \mathbb{R}^d$ and eqs. (40), (45), (46) and (49), (iii) uses eqs. (39), (40), (42) and (43) and $(1 + \rho_{\max}^2/\lambda_2^2) \le (1 + \rho_{\max}/\lambda_2)^2$. Substituting the above terms (18)-(21) into eqs. (16) and (17) gives the following upper bounds of $\mathbb{E}\big[\|\widetilde{w}_{t+1} - w_t^*\|^2\big|\mathcal{F}_t\big]$ and $\mathbb{E}\big[\|\widetilde{\theta}_{t+1} - \theta^*\|^2\big|\mathcal{F}_t\big]$.

$$
\begin{aligned}
&\mathbb{E}\big[\|\widetilde{w}_{t+1} - w_t^*\|^2\big|\mathcal{F}_t\big] \\
&\le \Big[1 + 2\beta\Big(\frac{4\nu\rho_{\max}}{N(1-\delta)} - \lambda_2\Big) + 4\beta^2\Big]\big\|\overline{w}_t - w_t^*\big\|^2 \\
&\quad + \frac{64\rho_{\max}^2(\nu+1)}{N(1-\delta)}\Big(1 + \frac{1}{\lambda_2}\Big)^2\Big(\frac{3\beta}{\lambda_2} + 2\beta^2\Big)\big(\|\overline{\theta}_t - \theta^*\|^2 + \|\theta^*\|^2 + R_{\max}^2\big) \\
&\overset{(i)}{\le} \Big(1 - \frac{\beta\lambda_2}{2}\Big)\big\|\overline{w}_t - w_t^*\big\|^2 + \frac{320\beta\rho_{\max}^2(\nu+1)}{N\lambda_2(1-\delta)}\Big(1 + \frac{1}{\lambda_2}\Big)^2\big(\|\overline{\theta}_t - \theta^*\|^2 + \|\theta^*\|^2 + R_{\max}^2\big),
\end{aligned} \tag{22}
$$

where (i) uses $N \ge \frac{8\nu\rho_{\max}}{\lambda_2(1-\delta)}$ and $\beta \le \min\big(\frac{\lambda_2}{8}, \frac{1}{\lambda_2}\big)$.

$$
\begin{aligned}
&\mathbb{E}\big[\|\widetilde{\theta}_{t+1} - \theta^*\|^2\big|\mathcal{F}_t\big] \\
&\le \Big(1 + \frac{128\alpha\rho_{\max}^2(\nu+1)}{N(1-\delta)}\Big(1 + \frac{\rho_{\max}}{\lambda_2}\Big)^2 - 2\alpha\lambda_1 + 40\alpha^2\rho_{\max}^2\Big(1 + \frac{\rho_{\max}}{\lambda_2}\Big)^2\Big)\big\|\overline{\theta}_t - \theta^*\big\|^2 \\
&\quad + \alpha\rho_{\max}^2\Big(\frac{8}{\lambda_1} + 5\alpha\Big)\big\|\overline{w}_t - w_t^*\big\|^2 + \frac{64\alpha\rho_{\max}^2(\nu+1)}{N\lambda_1(1-\delta)}\Big(4\Big(1 + \frac{\rho_{\max}}{\lambda_2}\Big)^2\big\|\theta^*\big\|^2 + R_{\max}^2 + \frac{\rho_{\max}R_{\max}}{\lambda_2}\Big) \\
&\quad + \frac{160\alpha^2\rho_{\max}^2(\nu+1)}{N(1-\delta)}\Big(1 + \frac{\rho_{\max}}{\lambda_2}\Big)^2\big(\big\|\theta^*\big\|^2 + R_{\max}^2\big) \\
&\overset{(i)}{\le} \Big(1 - \frac{\alpha\lambda_1}{2}\Big)\big\|\overline{\theta}_t - \theta^*\big\|^2 + \frac{13\alpha\rho_{\max}^2}{\lambda_1}\big\|\overline{w}_t - w_t^*\big\|^2 + \frac{224\alpha\rho_{\max}^2(\nu+1)}{N\lambda_1(1-\delta)}\Big(1 + \frac{\rho_{\max}}{\lambda_2}\Big)^2\big(4\big\|\theta^*\big\|^2 + 2R_{\max}^2 + 1\big),
\end{aligned} \tag{23}
$$

where (i) uses $N \ge \frac{128\rho_{\max}^2(\nu+1)}{\lambda_1(1-\delta)}\big(1 + \frac{\rho_{\max}}{\lambda_2}\big)^2$, $\alpha \le \min\big[\frac{\lambda_1}{40\rho_{\max}^2}\big(1 + \frac{\rho_{\max}}{\lambda_2}\big)^{-2}, \frac{1}{\lambda_1}\big]$ and $\frac{\rho_{\max}R_{\max}}{\lambda_2} \le \big(1 + \frac{\rho_{\max}}{\lambda_2}\big)^2 + R_{\max}^2$.

With the above upper bounds (22) and (23), we derive the upper bounds of $\mathbb{E}\big[\|\overline{w}_{t+1} - w_t^*\|^2\big|\mathcal{F}_t\big]$, $\mathbb{E}\big[\|\overline{w}_{t+1} - w_{t+1}^*\|^2\big|\mathcal{F}_t\big]$ and $\mathbb{E}\big[\|\overline{\theta}_{t+1} - \theta^*\|^2\big|\mathcal{F}_t\big]$ as follows.

$$
\begin{aligned}
&\mathbb{E}\big[\|\overline{w}_{t+1} - w_t^*\|^2\big|\mathcal{F}_t\big] \\
&\overset{(i)}{\le} \Big(1 + \frac{1}{6/(\beta\lambda_2) - 3}\Big)\mathbb{E}\big(\|\widetilde{w}_t - w_t^*\|^2\big|\mathcal{F}_t\big) + \Big(1 + \frac{6}{\beta\lambda_2} - 3\Big)\mathbb{E}\big(\big\|\overline{w}_t - \widetilde{w}_t\big\|^2\big|\mathcal{F}_t\big)
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(ii)}{\leq} \frac{6-2\beta\lambda_2}{6-3\beta\lambda_2}\Big[\Big(1-\frac{\beta\lambda_2}{2}\Big)\big\|\overline{w}_t - w_t^*\big\|^2 + \frac{320\beta\rho_{\max}^2(\nu+1)}{N\lambda_2(1-\delta)}\Big(1+\frac{1}{\lambda_2}\Big)^2\big(\|\overline{\theta}_t-\theta^*\|^2 + \|\theta^*\|^2 + R_{\max}^2\big)\Big] \\
&\quad + \frac{6}{\beta\lambda_2}\mathbb{E}\Big[\Big\|\frac{\beta}{M}\sum_{m=1}^{M}\big[(\overline{A}_t^{(m)}-\overline{A}_t)\theta_t^{(m)} + \bar{\tilde{b}}_t^{(m)} - \overline{b}_t^{(m)}\big]\Big\|^2\Big|\mathcal{F}_t\Big] \\
&\overset{(iii)}{\leq} \Big(1-\frac{\beta\lambda_2}{3}\Big)\big\|\overline{w}_t - w_t^*\big\|^2 + \frac{4}{3}\frac{320\beta\rho_{\max}^2(\nu+1)}{N\lambda_2(1-\delta)}\Big(1+\frac{1}{\lambda_2}\Big)^2\big(\|\overline{\theta}_t-\theta^*\|^2 + \|\theta^*\|^2 + R_{\max}^2\big) \\
&\quad + \frac{6\beta}{\lambda_2}\mathbb{E}\Big[\frac{1}{M}\sum_{m=1}^{M}\Big(\big\|(\overline{A}_t^{(m)}-\overline{A}_t)\theta_t^{(m)} + \bar{\tilde{b}}_t^{(m)} - \overline{b}_t^{(m)}\big\|^2\Big)\Big|\mathcal{F}_t\Big] \\
&\overset{(iv)}{\leq} \Big(1-\frac{\beta\lambda_2}{3}\Big)\|\overline{w}_t - w_t^*\|^2 + \frac{427\beta\rho_{\max}^2(\nu+1)}{N\lambda_2(1-\delta)}\Big(1+\frac{1}{\lambda_2}\Big)^2\big(\|\overline{\theta}_t-\theta^*\|^2 + \|\theta^*\|^2 + R_{\max}^2\big) \\
&\quad + \frac{12\beta\sigma_2^{L/4}}{M\lambda_2}\Big(16\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\| + \|w_0^{(m)}\| + R_{\max})^2[1+\beta(2\rho_{\max}+3)]^{2t} + R_{\max}^2\Big),
\end{aligned} \tag{24}
$$

where (i) uses the inequality that $\|a_1+a_2\|^2 \leq (1+\sigma)\|a_1\|^2 + (1+\sigma^{-1})\|a_2\|^2$ for any $a_1, a_2 \in \mathbb{R}^d$ and $\sigma > 0$, (ii) uses eqs. (15), (13) and (22), (iii) applies Jensen's inequality to the convex function $\|\cdot\|^2$ and uses $\beta \leq \frac{1}{\lambda_2}$, (iv) uses the condition that $\beta \leq \frac{1}{\lambda_2}$ which implies that $1 + \frac{1}{6/(\beta\lambda_2)-3} \leq 2$, the inequality that $\|a_1+a_2\|^2 \leq 2\|a_1\|^2 + 2\|a_2\|^2$ for any $a_1, a_2 \in \mathbb{R}^d$, and eqs. (50), (52) and (66).

$$
\begin{aligned}
&\mathbb{E}\big[\|\overline{w}_{t+1} - w_{t+1}^*\|^2\big|\mathcal{F}_t\big] \\
&\overset{(i)}{\leq} \Big(1+\frac{1}{2[3/(\beta\lambda_2)-1]}\Big)\mathbb{E}\big[\|\overline{w}_{t+1} - w_t^*\|^2\big|\mathcal{F}_t\big] + \big[1+2\big(3/(\beta\lambda_2)-1\big)\big]\mathbb{E}\big[\|w_{t+1}^* - w_t^*\|^2\big|\mathcal{F}_t\big] \\
&\overset{(ii)}{\leq} \frac{6/(\beta\lambda_2)-1}{2[3/(\beta\lambda_2)-1]}\Big[\Big(1-\frac{\beta\lambda_2}{3}\Big)\|\overline{w}_t - w_t^*\|^2 + \frac{427\beta\rho_{\max}^2(\nu+1)}{N\lambda_2(1-\delta)}\Big(1+\frac{1}{\lambda_2}\Big)^2\big(\|\overline{\theta}_t-\theta^*\|^2 + \|\theta^*\|^2 + R_{\max}^2\big) \\
&\quad + \frac{12\beta\sigma_2^{L/4}}{M\lambda_2}\Big(16\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\| + \|w_0^{(m)}\| + R_{\max})^2[1+\beta(2\rho_{\max}+3)]^{2t} + R_{\max}^2\Big)\Big] \\
&\quad + \frac{6}{\beta\lambda_2}\mathbb{E}\big[\|C^{-1}A(\overline{\theta}_{t+1} - \overline{\theta}_t)\|^2\big|\mathcal{F}_t\big] \\
&\overset{(iii)}{\leq} \Big(1-\frac{\beta\lambda_2}{6}\Big)\|\overline{w}_t - w_t^*\|^2 + \frac{534\beta\rho_{\max}^2(\nu+1)}{N\lambda_2(1-\delta)}\Big(1+\frac{1}{\lambda_2}\Big)^2\big(\|\overline{\theta}_t-\theta^*\|^2 + \|\theta^*\|^2 + R_{\max}^2\big) \\
&\quad + \frac{15\beta\sigma_2^{L/4}}{M\lambda_2}\big(17\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\| + \|w_0^{(m)}\| + R_{\max})^2[1+\beta(2\rho_{\max}+3)]^{2t}\big) \\
&\quad + \frac{24\alpha^2\rho_{\max}^2}{\beta\lambda_2^3}\underbrace{\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\big(\overline{A}_t^{(m)}\theta_t^{(m)} + \bar{\tilde{b}}_t^{(m)} + \overline{B}_t^{(m)}w_t^{(m)}\big)\Big\|^2\Big|\mathcal{F}_t\Big]}_{(V)},
\end{aligned} \tag{25}
$$

where (i) uses the inequality that $\|a_1+a_2\|^2 \leq (1+\sigma)\|a_1\|^2 + (1+\sigma^{-1})\|a_2\|^2$ for any $a_1, a_2 \in \mathbb{R}^d$ and $\sigma > 0$, (ii) uses eq. (24 as well as the notation that $w_t^* = -C^{-1}(A\overline{\theta}_t + b)$, (iii) uses $\beta \leq \frac{1}{\lambda_2}$ (this implies that $\frac{6/(\beta\lambda_2)-1}{2[3/(\beta\lambda_2)-1]} \leq \frac{5}{4}$) and eqs. (14, (39 and (43. The above term (V) can be upper bounded as follows.

$$
\begin{aligned}
(V) &= \mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\big(\overline{A}_t^{(m)}\theta_t^{(m)} + \bar{\tilde{b}}_t^{(m)} + \overline{B}_t^{(m)}w_t^{(m)}\big)\Big\|^2\Big|\mathcal{F}_t\Big] \\
&\overset{(i)}{\leq} 2\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\big((\overline{A}_t^{(m)}-\overline{A}_t)\theta_t^{(m)} + (\bar{\tilde{b}}_t^{(m)} - \tilde{b}_t^{(m)}) + (\overline{B}_t^{(m)}-\overline{B}_t)w_t^{(m)}\big)\Big\|^2\Big|\mathcal{F}_t\Big] \\
&\quad + 2\mathbb{E}\Big[\Big\|\frac{1}{M}\sum_{m=1}^{M}\big(\overline{A}_t\theta_t^{(m)} + \overline{b}_t^{(m)} + \overline{B}_t w_t^{(m)}\big)\Big\|^2\Big|\mathcal{F}_t\Big]
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(ii)}{\leq} \frac{2}{M}\sum_{m=1}^{M}\mathbb{E}\big[\big\|(\overline{A}_t^{(m)}-\overline{A}_t)\theta_t^{(m)}+(\overline{\widetilde{b}}_t^{(m)}-\widetilde{b}_t^{(m)})+(\overline{B}_t^{(m)}-\overline{B}_t)w_t^{(m)}\big)\big\|^2\big|\mathcal{F}_t\big]+2\mathbb{E}\big(\big\|\overline{A}_t\overline{\theta}_t+\overline{\widetilde{b}}_t+\overline{B}_t\overline{w}_t\big\|^2\big|\mathcal{F}_t\big)\\
&\overset{(iii)}{\leq} \frac{6}{M}\sum_{m=1}^{M}\mathbb{E}\big[\big\|\overline{A}_t^{(m)}-\overline{A}_t\big\|^2\big\|\theta_t^{(m)}\big\|^2+\big\|\overline{\widetilde{b}}_t^{(m)}-\widetilde{b}_t^{(m)}\big\|^2+\big\|\overline{B}_t^{(m)}-\overline{B}_t\big\|^2\big\|w_t^{(m)}\big\|^2\big|\mathcal{F}_t\big]\\
&\quad+2\Big(40\rho_{\max}^2\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\big\|\overline{\theta}_t-\theta^*\big\|^2+5\rho_{\max}^2\big\|\overline{w}_t-w_t^*\big\|^2+\frac{160\rho_{\max}^2(\nu+1)}{N(1-\delta)}\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\big(\big\|\theta^*\big\|^2+R_{\max}^2\big)\Big)\\
&\overset{(iv)}{\leq} \frac{6\sigma_2^{L/4}}{M}\Big(20\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\|+\|w_0^{(m)}\|+R_{\max})^2[1+\beta(2\rho_{\max}+3)]^{2t}+R_{\max}^2\Big)\\
&\quad+2\Big(40\rho_{\max}^2\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\big\|\overline{\theta}_t-\theta^*\big\|^2+5\rho_{\max}^2\big\|\overline{w}_t-w_t^*\big\|^2+\frac{160\rho_{\max}^2(\nu+1)}{N(1-\delta)}\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\big(\big\|\theta^*\big\|^2+R_{\max}^2\big)\Big)
\end{aligned}
$$

where (i) uses the inequality that $\|a_1+a_2\|^2\leq 2\|a_1\|^2+2\|a_2\|^2$ for any $a_1,a_2\in\mathbb{R}^d$, (ii) applies Jensen's inequality to the convex function $\|\cdot\|^2$, (iii) uses eq. (21) and the inequality that $\|a_1+a_2+a_3\|^2\leq 3\sum_{k=1}^3\|a_k\|^2$ for any $a_1,a_2,a_3\in\mathbb{R}^d$, (iv) uses eqs. (50), (51), (52) and (66). Substituting the above inequality into eq. (25) yields that

$$
\begin{aligned}
&\mathbb{E}\big[\|\overline{w}_{t+1}-w_{t+1}^*\|^2\big|\mathcal{F}_t\big]\\
&\leq\Big(1-\frac{\beta\lambda_2}{6}+\frac{120\alpha^2\rho_{\max}^4}{\beta\lambda_2^3}\Big)\|\overline{w}_t-w_t^*\|^2+\Big[\frac{534\beta\rho_{\max}^2(\nu+1)}{N\lambda_2(1-\delta)}\Big(1+\frac{1}{\lambda_2}\Big)^2+\frac{1920\alpha^2\rho_{\max}^4}{\beta\lambda_2^3}\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\Big]\|\overline{\theta}_t-\theta^*\|^2\\
&\quad+\frac{3840\rho_{\max}^2(\nu+1)}{N(1-\delta)}\Big[\frac{\beta}{\lambda_2}\Big(1+\frac{1}{\lambda_2}\Big)^2+\frac{\alpha^2\rho_{\max}^2}{\beta\lambda_2^3}\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\Big]\big(\|\theta^*\|^2+R_{\max}^2\big)\\
&\quad+\frac{255\sigma_2^{L/4}}{M\lambda_2}\Big(\beta+\frac{12\alpha^2\rho_{\max}^2}{\beta\lambda_2^2}\Big)\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\|+\|w_0^{(m)}\|+R_{\max})^2[1+\beta(2\rho_{\max}+3)]^{2t}.
\end{aligned}\tag{26}
$$

$$
\begin{aligned}
&\mathbb{E}\big[\|\overline{\theta}_{t+1}-\theta^*\|^2\big|\mathcal{F}_t\big]\\
&\overset{(i)}{\leq}\Big(1+\frac{1}{6/(\alpha\lambda_1)-3}\Big)\mathbb{E}\big[\widetilde{\theta}_{t+1}-\theta^*\big|\mathcal{F}_t\big]+\Big(1+\frac{6}{\alpha\lambda_1}-3\Big)\mathbb{E}\big[\overline{\theta}_{t+1}-\widetilde{\theta}_{t+1}\big|\mathcal{F}_t\big]\\
&\overset{(ii)}{\leq}\frac{6-2\alpha\lambda_1}{6-3\alpha\lambda_1}\Big[\Big(1-\frac{\alpha\lambda_1}{2}\Big)\big\|\overline{\theta}_t-\theta^*\big\|^2+\frac{13\alpha\rho_{\max}^2}{\lambda_1}\big\|\overline{w}_t-w_t^*\big\|^2\\
&\quad+\frac{224\alpha\rho_{\max}^2(\nu+1)}{N\lambda_1(1-\delta)}\big(4\big\|\theta^*\big\|^2+2R_{\max}^2+1\big)\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\Big]\\
&\quad+\frac{6}{\alpha\lambda_1}\mathbb{E}\Big[\Big\|\frac{\alpha}{M}\sum_{m=1}^{M}\big[(\overline{A}_t^{(m)}-\overline{A}_t)\theta_t^{(m)}+\overline{\widetilde{b}}_t^{(m)}-\overline{b}_t^{(m)}+(\overline{B}_t^{(m)}-\overline{B}_t)w_t^{(m)}\big]\Big\|^2\Big|\mathcal{F}_t\Big]\\
&\overset{(iii)}{\leq}\Big(1-\frac{\alpha\lambda_1}{3}\Big)\big\|\overline{\theta}_t-\theta^*\big\|^2+\frac{18\alpha\rho_{\max}^2}{\lambda_1}\big\|\overline{w}_t-w_t^*\big\|^2+\frac{300\alpha\rho_{\max}^2(\nu+1)}{N\lambda_1(1-\delta)}\big(4\big\|\theta^*\big\|^2+2R_{\max}^2+1\big)\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\\
&\quad+\frac{6\alpha}{M\lambda_1}\sum_{m=1}^{M}\mathbb{E}\big[\big\|(\overline{A}_t^{(m)}-\overline{A}_t)\theta_t^{(m)}+\overline{\widetilde{b}}_t^{(m)}-\overline{b}_t^{(m)}+(\overline{B}_t^{(m)}-\overline{B}_t)w_t^{(m)}\big\|^2\big|\mathcal{F}_t\big]\\
&\overset{(iv)}{\leq}\Big(1-\frac{\alpha\lambda_1}{3}\Big)\big\|\overline{\theta}_t-\theta^*\big\|^2+\frac{18\alpha\rho_{\max}^2}{\lambda_1}\big\|\overline{w}_t-w_t^*\big\|^2+\frac{300\alpha\rho_{\max}^2(\nu+1)}{N\lambda_1(1-\delta)}\big(4\big\|\theta^*\big\|^2+2R_{\max}^2+1\big)\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\\
&\quad+\frac{18\alpha\sigma_2^{L/4}}{M\lambda_1}\Big[20\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\|+\|w_0^{(m)}\|+R_{\max})^2[1+\beta(2\rho_{\max}+3)]^{2t}+R_{\max}^2\Big],
\end{aligned}\tag{27}
$$

where (i) uses the inequality that $\|a_1+a_2\|^2\leq(1+\sigma)\|a_1\|^2+(1+\sigma^{-1})\|a_2\|^2$ for any $a_1,a_2\in\mathbb{R}^d$ and $\sigma>0$, (ii) uses eqs. (14), (12) and (23), (iii) uses $\alpha\leq 1/\lambda_1$ and applies Jensen's inequality to the convex function $\|\cdot\|^2$, and (iv) uses the inequality that $\|a_1+a_2+a_3\|^2\leq 3\sum_{k=1}^3\|a_k\|^2$ for any $a_1,a_2,a_3\in\mathbb{R}^d$ and then uses eqs. (50)-(52).

Taking expectation on both sides of eqs. (26) and (27) and summing up the two inequalities yields that

$$
\begin{aligned}
&\mathbb{E}(\|\overline{\theta}_{t+1}-\theta^*\|^2)+\mathbb{E}(\|\overline{w}_{t+1}-w^*_{t+1}\|^2)\\
&\overset{(i)}{\leq}\Big(1-\frac{\beta\lambda_2}{6}+\frac{120\alpha^2\rho^4_{\max}}{\beta\lambda_2^3}+\frac{18\alpha\rho^2_{\max}}{\lambda_1}\Big)\mathbb{E}(\|\overline{w}_t-w^*_t\|^2)\\
&\quad+\Big[1-\frac{\alpha\lambda_1}{3}+\frac{534\beta\rho^2_{\max}(\nu+1)}{N\lambda_2(1-\delta)}\Big(1+\frac{1}{\lambda_2}\Big)^2+\frac{1920\alpha^2\rho^4_{\max}}{\beta\lambda_2^3}\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\Big]\mathbb{E}(\|\overline{\theta}_t-\theta^*\|^2)\\
&\quad+\frac{3840\rho^2_{\max}(\nu+1)}{N(1-\delta)}\Big[\frac{\beta}{\lambda_2}\Big(1+\frac{1}{\lambda_2}\Big)^2+\frac{\alpha^2\rho^2_{\max}}{\beta\lambda_2^3}\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2+\frac{\alpha}{\lambda_1}\Big]\big(\|\theta^*\|^2+R^2_{\max}+1\big)\\
&\quad+\frac{255\sigma_2^{L/4}}{M\lambda_2}\Big(\beta+\frac{12\alpha^2\rho^2_{\max}}{\beta\lambda_2^2}+\frac{\alpha\lambda_2}{\lambda_1}\Big)\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\|+\|w_0^{(m)}\|+R_{\max})^2[1+\beta(2\rho_{\max}+3)]^{2t}\\
&\overset{(i)}{\leq}\Big(1-\frac{\alpha\lambda_1}{6}\Big)\big[\mathbb{E}(\|\overline{\theta}_t-\theta^*\|^2)+\mathbb{E}(\|\overline{w}_t-w^*_t\|^2)\big]+\frac{6000\beta\rho^2_{\max}(\nu+1)}{N\lambda_2(1-\delta)}\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\big(\|\theta^*\|^2+R^2_{\max}+1\big)\\
&\quad+\frac{574\beta\sigma_2^{L/4}}{M\lambda_2}\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\|+\|w_0^{(m)}\|+R_{\max})^2[1+\beta(2\rho_{\max}+3)]^{2t},
\end{aligned}
$$

where (i) uses the conditions that $N\geq\frac{6408\beta\rho^2_{\max}(\nu+1)}{\alpha\lambda_1\lambda_2(1-\delta)}\big(1+\frac{1}{\lambda_2}\big)^2$,
$\alpha\leq\min\big(\frac{\beta\lambda_1\lambda_2^3}{23040\rho^4_{\max}},\frac{\beta\lambda_2^2}{53\rho^2_{\max}},\frac{\beta\lambda_1\lambda_2}{432\rho^2_{\max}},\frac{\beta\lambda_2}{2\lambda_1},\frac{\beta\lambda_1}{2\lambda_2},\frac{\beta\lambda_2}{4\rho_{\max}}\big)$. Iterating the inequality above yields that

$$
\begin{aligned}
&\mathbb{E}(\big\|\overline{\theta}_T-\theta^*\big\|^2+\big\|\overline{w}_T-w^*\big\|^2)\\
&\leq\Big(1-\frac{\alpha\lambda_1}{6}\Big)^T\big(\big\|\overline{\theta}_0-\theta^*\big\|^2+\|\overline{w}_0-w^*_0\|^2\big)\\
&\quad+\sum_{t=0}^{T-1}\Big(1-\frac{\alpha\lambda_1}{6}\Big)^{T-1-t}\Big[\frac{6000\beta\rho^2_{\max}(\nu+1)}{N\lambda_2(1-\delta)}\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\big(\|\theta^*\|^2+R^2_{\max}+1\big)\\
&\quad+\frac{574\beta\sigma_2^{L/4}}{M\lambda_2}\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\|+\|w_0^{(m)}\|+R_{\max})^2[1+\beta(2\rho_{\max}+3)]^{2t}\Big]\\
&\overset{(i)}{\leq}\Big(1-\frac{\alpha\lambda_1}{6}\Big)^T\big(\big\|\overline{\theta}_0-\theta^*\big\|^2+\|\overline{w}_0-w^*_0\|^2\big)+\frac{36000\beta\rho^2_{\max}(\nu+1)}{\alpha N\lambda_1\lambda_2(1-\delta)}\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2\big(\|\theta^*\|^2+R^2_{\max}+1\big)\\
&\quad+\frac{574\beta\sigma_2^{L/4}2^T}{M\lambda_2}\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\|+\|w_0^{(m)}\|+R_{\max})^2\\
&=\Big(1-\frac{\alpha\lambda_1}{6}\Big)^T\big(\big\|\overline{\theta}_0-\theta^*\big\|^2+\|\overline{w}_0-w^*_0\|^2\big)+\mathcal{O}\Big(\frac{\beta}{N\alpha}+\frac{\beta\sigma_2^{L/4}2^T}{M}\Big),
\end{aligned}
\tag{28}
$$

where (i) uses the definitions of c_{10}, c_{11} in Appendix A and the conditions that $\alpha\leq\frac{1}{\lambda_1}$ and $\beta\leq\frac{2}{5(2\rho_{\max}+3)}$ which respectively imply that $1-\frac{\alpha\lambda_1}{6}\geq\frac{5}{6}$ and that $1+\beta(2\rho_{\max}+3)\leq\sqrt{2}$. This proves eq. (10).

Next, we prove eq. (11). Note that the local model averaging iterations can be rewritten into the matrix-vector form as $\Theta_{t+1}=V\Theta_t$ where $T\leq t\leq T+T'$ and $\Theta_t\overset{\triangle}{=}[\theta_t^{(1)};\theta_t^{(2)};\ldots;\theta_t^{(M)}]^\top$. Hence, it can be derived from Lemma C.3 that

$$
\|\Delta\Theta_{T+T'}\|_F=\|\Delta V^{T'}\Theta_T\|_F=\|V^{T'}\Delta\Theta_T\|_F\leq\sigma_2^{T'}\|\Delta\Theta_T\|_F. \tag{29}
$$

Hence, we only need to obtain an upper bound of $\mathbb{E}\|\Delta\Theta_T\|^2$. Subtracting eq. (14) from the local update rule yields that for any $0\leq t\leq T-1$,

$$
\begin{aligned}
\theta_{t+1}^{(m)}-\overline{\theta}_{t+1}=&\sum_{m'\in\mathcal{N}_m}V_{m,m'}(\theta_t^{(m')}-\overline{\theta}_t)+\frac{M-1}{M}\alpha\big(\overline{A}_t^{(m)}\theta_t^{(m)}+\overset{\approx}{b}{}_t^{(m)}+\overline{B}_t^{(m)}w_t^{(m)}\big)\\
&-\frac{\alpha}{M}\sum_{m'=1,m'\neq m}^{M}\big(\overline{A}_t^{(m')}\theta_t^{(m')}+\overset{\approx}{b}{}_t^{(m')}+\overline{B}_t^{(m')}w_t^{(m')}\big).
\end{aligned}
$$

This can be rewritten into the following matrix-vector form,

$$\Delta\Theta_{t+1} = V\Delta\Theta_t + [h_1; h_2; \ldots; h_M]^\top,$$

where $h_m \triangleq \frac{M-1}{M}\alpha\big(\overline{A}_t^{(m)}\theta_t^{(m)} + \overline{\overline{b}}_t^{(m)} + \overline{B}_t^{(m)}w_t^{(m)}\big) - \frac{\alpha}{M}\sum_{m'=1,m'\neq m}^{M}\big(\overline{A}_t^{(m')}\theta_t^{(m')} + \overline{\overline{b}}_t^{(m')} + \overline{B}_t^{(m')}w_t^{(m')}\big)$. The item 2 of Lemma C.3 implies that for any $0 \le t \le T-1$,

$$\|\Delta\Theta_{t+1}\|_F \le \sigma_2\|\Delta\Theta_t\|_F + \sqrt{\sum_{m=1}^{M}\|h_m\|^2} \le \sigma_2\|\Delta\Theta_t\|_F + \sum_{m=1}^{M}\|h_m\|. \tag{30}$$

Then, using eqs. (53)-(55) yields that

$$\begin{aligned}
\sum_{m=1}^{M}\|h_m\| &\le \frac{M-1}{M}(2\alpha)(2\rho_{\max}+2)\sum_{m=1}^{M}\big(\|\theta_t^{(m)}\| + R_{\max} + \|w_t^{(m)}\|\big)\\
&\overset{(i)}{\le} 4\alpha(\rho_{\max}+1)\sum_{m=1}^{M}\big(\|\theta_t^{(m)} - \overline{\theta}_t\| + \|w_t^{(m)} - \overline{w}_t\| + \|\overline{\theta}_t - \theta^*\| + \|\overline{w}_t - w_t^*\| + \|\theta^*\| + \|C^{-1}(A\overline{\theta}_t + b)\|\big)\\
&\overset{(ii)}{\le} 4\alpha(\rho_{\max}+1)\Big(\sum_{m=1}^{M}\big(\|\theta_t^{(m)} - \overline{\theta}_t\| + \|w_t^{(m)} - \overline{w}_t\|\big)\\
&\quad + M\Big(1 + \frac{2\rho_{\max}}{\lambda_2}\Big)\|\overline{\theta}_t - \theta^*\| + M\|\overline{w}_t - w_t^*\| + \frac{M\rho_{\max}}{\lambda_2}(2\|\theta^*\| + R_{\max})\Big),
\end{aligned}$$

where (i) uses the notations that $w_t^* = -C^{-1}(A\overline{\theta}_t + b)$, (ii) uses eqs. (39), (42) and (43). Hence, we obtain that

$$\begin{aligned}
\mathbb{E}(\|\Delta\Theta_{t+1}\|_F^2) &\overset{(i)}{\le} \Big(1 + \frac{\sigma_2^{-2}-1}{2}\Big)\sigma_2^2\mathbb{E}\big(\|\Delta\Theta_t\|_F^2\big) + \Big(1 + \frac{2}{\sigma_2^{-2}-1}\Big)\mathbb{E}\Big[\Big(\sum_{m=1}^{M}\|h_m\|\Big)^2\Big],\\
&\overset{(ii)}{\le} \frac{1+\sigma_2^2}{2}\mathbb{E}\big(\|\Delta\Theta_t\|_F^2\big) + \frac{48\alpha^2(1+\sigma_2^2)}{1-\sigma_2^2}(\rho_{\max}+1)^2\mathbb{E}\Big[2M\sum_{m=1}^{M}\big(\|\theta_t^{(m)} - \overline{\theta}_t\|^2 + \|w_t^{(m)} - \overline{w}_t\|^2\big)\\
&\quad + M^2\Big(1 + \frac{2\rho_{\max}}{\lambda_2}\Big)^2\big(\|\overline{\theta}_t - \theta^*\|^2 + \|\overline{w}_t - w_t^*\|^2\big) + \frac{4M^2\rho_{\max}^2}{\lambda_2^2}(\|\theta^*\| + R_{\max})^2\Big],
\end{aligned} \tag{31}$$

where (i) uses eq. (30) and the fact that $(u+v)^2 \le (1+\sigma)u^2 + (1+\sigma^{-1})v^2$ for any $u, v, \sigma \ge 0$, (ii) uses $(\sum_{i=1}^n q_i)^2 \le n\sum_{i=1}^n q_i^2$ for any $q_i \in \mathbb{R}$ and $n \in \mathbb{N}^+$. Similarly, we obtain from the update rule of $W_t = [w_t^{(1)}; w_t^{(2)}; \ldots; w_t^{(M)}]^\top \in \mathbb{R}^{M\times d}$ and eq. (15) that

$$\begin{aligned}
\mathbb{E}(\|\Delta W_{t+1}\|_F^2) \le& \frac{1+\sigma_2^2}{2}\mathbb{E}\big(\|\Delta W_t\|_F^2\big) + \frac{48\alpha^2(1+\sigma_2^2)}{1-\sigma_2^2}(\rho_{\max}+1)^2\mathbb{E}\Big[2M\sum_{m=1}^{M}\big(\|\theta_t^{(m)} - \overline{\theta}_t\|^2 + \|w_t^{(m)} - \overline{w}_t\|^2\big)\\
&+ M^2\Big(1 + \frac{2\rho_{\max}}{\lambda_2}\Big)^2\big(\|\overline{\theta}_t - \theta^*\|^2 + \|\overline{w}_t - w_t^*\|^2\big) + \frac{4M^2\rho_{\max}^2}{\lambda_2^2}(\|\theta^*\| + R_{\max})^2\Big],
\end{aligned} \tag{32}$$

Summing up eqs. (31) and (32) yields that

$$\begin{aligned}
&\mathbb{E}(\|\Delta\Theta_{t+1}\|_F^2 + \|\Delta W_{t+1}\|_F^2)\\
&\overset{(i)}{\le} \frac{1+\sigma_2^2}{2}\mathbb{E}\big(\|\Delta\Theta_t\|_F^2 + \|\Delta W_t\|_F^2\big) + \frac{96\alpha^2(1+\sigma_2^2)}{1-\sigma_2^2}(\rho_{\max}+1)^2\mathbb{E}\Big[2M\big(\|\Delta\Theta_t\|_F^2 + \|\Delta W_t\|_F^2\big)\\
&\quad + M^2\Big(1 + \frac{2\rho_{\max}}{\lambda_2}\Big)^2\Big(\Big(1 - \frac{\alpha\lambda_1}{6}\Big)^t\big(\big\|\overline{\theta}_0 - \theta^*\big\|^2 + \|\overline{w}_0 - w_0^*\|^2\big)\\
&\quad + \frac{36000\beta\rho_{\max}^2(\nu+1)}{\alpha N\lambda_1\lambda_2(1-\delta)}\Big(1 + \frac{\rho_{\max}}{\lambda_2}\Big)^2\big(\|\theta^*\|^2 + R_{\max}^2 + 1\big) + \frac{574\beta\sigma_2^{L/4}2^t}{M\lambda_2}\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\| + \|w_0^{(m)}\| + R_{\max})^2\Big)
\end{aligned}$$

$$+\frac{4M^2\rho_{\max}^2}{\lambda_2^2}(\|\theta^*\|+R_{\max})^2\Big]$$

$$\overset{(ii)}{\leq}\frac{2+\sigma_2^2}{3}\mathbb{E}\big(\|\Delta\Theta_t\|_F^2+\|\Delta W_t\|_F^2\big)+\frac{192M^2\alpha^2}{1-\sigma_2^2}(\rho_{\max}+1)^2\Big(1+\frac{2\rho_{\max}}{\lambda_2}\Big)^2\Big[\big(\big\|\overline{\theta}_0-\theta^*\big\|^2+\|\overline{w}_0-w_0^*\|^2\big)$$

$$+\Big(\frac{282\beta}{\alpha\lambda_2}+\frac{8M^2\rho_{\max}^2}{\lambda_2^2}\Big)\big(\|\theta^*\|^2+R_{\max}^2+1\big)+\frac{574\beta\sigma_2^{L/4}2^t}{M\lambda_2}\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\|+\|w_0^{(m)}\|+R_{\max})^2\Big)\Big], \tag{33}$$

where (i) uses eq. (28), and (ii) uses $(\|\theta^*\|+R_{\max})^2\leq 2\|\theta^*\|^2+2R_{\max}^2$, $\alpha\leq\frac{1-\sigma_2}{50\sqrt{M}(\rho_{\max}+1)}$ and $N\geq\frac{128\rho_{\max}^2(\nu+1)}{\lambda_1(1-\delta)}\big(1+\frac{\rho_{\max}}{\lambda_2}\big)^2$.

Iterating eq. (33) yields that

$$\begin{aligned}
\mathbb{E}\big(\|\Delta\Theta_T\|_F^2\big)&\leq\mathbb{E}\big(\|\Delta\Theta_T\|_F^2+\|\Delta W_T\|_F^2\big)\\
&\leq\Big(\frac{2+\sigma_2^2}{3}\Big)^T\big(\|\Delta\Theta_0\|_F^2+\|\Delta W_0\|_F^2\big)+\frac{192M^2\alpha^2}{1-\sigma_2^2}(\rho_{\max}+1)^2\Big(1+\frac{2\rho_{\max}}{\lambda_2}\Big)^2\\
&\quad+\Big[\frac{3}{1-\sigma_2}\Big(\big\|\overline{\theta}_0-\theta^*\big\|^2\|\overline{w}_0-w_0^*\|^2+\Big(\frac{282\beta}{\alpha\lambda_2}+\frac{8M^2\rho_{\max}^2}{\lambda_2^2}\Big)\big(\|\theta^*\|^2+R_{\max}^2+1\big)\Big)\\
&\quad+\frac{574\beta\sigma_2^{L/4}2^T}{M\lambda_2}\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\|+\|w_0^{(m)}\|+R_{\max})^2\Big]\\
&\leq\|\Delta\Theta_0\|_F^2+\|\Delta W_0\|_F^2+\mathcal{O}\Big[\frac{M^2\alpha^2}{(1-\sigma_2)^2}\Big(\frac{\beta}{\alpha}+M^2\Big)+\frac{M\beta\alpha^2\sigma_2^{L/4}2^T}{1-\sigma_2}\Big]\\
&\overset{(i)}{\leq}\mathcal{O}\Big(1+\frac{M^4\beta\alpha}{(1-\sigma_2)^2}+\frac{M\beta\alpha\sigma_2^{L/4}2^T}{1-\sigma_2}\Big),
\end{aligned}\tag{34}$$

where (i) uses $\alpha\leq\frac{\beta}{2\rho_{\max}+2}=\mathcal{O}(\beta)$. Substituting the above inequality into eq. (29) proves eq. (11).

To summarize, the following conditions of the hyperparameters are used in the proof of Theorem 1, including those required by Corollary 2 and lemma C.4.

$$\begin{aligned}
\alpha\leq&\min\Big(\frac{\beta}{2\rho_{\max}+2},\frac{\lambda_1}{40\rho_{\max}^2}\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^{-2},\frac{1}{\lambda_1},\frac{\beta\lambda_2}{4\rho_{\max}},\frac{\beta\lambda_1\lambda_2^3}{23040\rho_{\max}^4},\frac{\beta\lambda_2^2}{53\rho_{\max}^2},\frac{\beta\lambda_1\lambda_2}{432\rho_{\max}^2},\\
&\frac{1-\sigma_2}{50\sqrt{M}(\rho_{\max}+1)},\frac{\beta\lambda_2}{2\lambda_1},\frac{\beta\lambda_1}{2\lambda_2}\Big)=\min\{\mathcal{O}(\beta),\mathcal{O}(M^{-1/2}(1-\sigma_2))\}
\end{aligned}\tag{35}$$

$$\beta\leq\min\Big(\frac{1}{\lambda_2},\frac{2}{5(2\rho_{\max}+3)}\Big)=\mathcal{O}(1)\tag{36}$$

$$N\geq\max\Big(\frac{8c_{\text{sd}}}{\lambda_2},\frac{8c_{\text{var},3}\Omega_A^2}{\lambda_1^2},\frac{12\beta c_8}{\alpha\lambda_1}\Big)=\max\{\mathcal{O}(1),\mathcal{O}(\beta/\alpha)\}.\tag{37}$$

$$L\geq\frac{12\ln M+(8M+10)\ln\rho_{\max}}{\ln\sigma_2^{-1}}=\mathcal{O}\Big(\frac{M}{1-\sigma_2}\Big)\tag{38}$$

Under the above conditions, we choose the following hyperparameter values.

$$\alpha=\mathcal{O}(M^{-1/2}(1-\sigma_2)),\beta=\mathcal{O}(1)$$

$$T=\frac{6}{\alpha\lambda_1}\ln\epsilon^{-1}=\mathcal{O}\Big(\frac{\sqrt{M}\ln\epsilon^{-1}}{1-\sigma_2}\Big)$$

$$N=\frac{\beta}{\alpha\epsilon}=\mathcal{O}\Big(\frac{\sqrt{M}}{\epsilon(1-\sigma_2)}\Big)$$

$$L=\frac{4}{\ln\sigma_2^{-1}}\Big(\ln\Big(\frac{\beta}{M\epsilon}\Big)+T\ln 2\Big)+\frac{12\ln M+(8M+10)\ln\rho_{\max}}{\ln\sigma_2^{-1}}=\mathcal{O}\Big(\frac{\sqrt{M}\ln\epsilon^{-1}}{(1-\sigma_2)^2}+\frac{M}{1-\sigma_2}\Big)\leq\mathcal{O}\Big(\frac{M\ln\epsilon^{-1}}{(1-\sigma_2)^2}\Big)$$

$$T'=\frac{1}{\ln\sigma_2^{-1}}\ln\Big(\epsilon^{-1}\mathcal{O}\Big(1+\frac{M^4\beta\alpha}{(1-\sigma_2)^2}+\frac{M\beta\alpha\sigma_2^{L/4}2^T}{1-\sigma_2}\Big)\Big)$$

$$= \frac{1}{\ln \sigma_2^{-1}} \ln\Big(\epsilon^{-1}\mathcal{O}\Big(1 + \frac{M^{3.5}}{1-\sigma_2} + \frac{\alpha M^2 \epsilon}{1-\sigma_2}\Big)\Big) = \mathcal{O}\Big(\frac{1}{1-\sigma_2}\ln\Big(\frac{M}{\epsilon(1-\sigma_2)}\Big)\Big).$$

Substituting these hyperparameters into eqs. (10) and (11) implies $\mathbb{E}(\big\|\overline{\theta}_T - \theta^*\big\|^2), \mathbb{E}(\|\theta^{(m)}_{T+T'} - \overline{\theta}_T\|^2) \le \mathcal{O}(\epsilon)$, so $\mathbb{E}(\|\theta^{(m)}_{T+T'} - \theta^*\|^2) \le 2\mathbb{E}(\|\theta^{(m)}_{T+T'} - \overline{\theta}_T\|^2) + 2\mathbb{E}(\|\overline{\theta}_T - \theta^*\|^2) \le \mathcal{O}(\epsilon)$. Therefore, the overall communication complexity for synchronizing $\theta^{(m)}_t$ is $T + T' = \mathcal{O}\big(\frac{\sqrt{M}\ln\epsilon^{-1}}{1-\sigma_2}\big)$, and the total sample complexity is $NT = \mathcal{O}\big(\frac{M\ln\epsilon^{-1}}{\epsilon(1-\sigma_2)^2}\big)$.

C Supporting Lemmas

In this section, we prove some supporting lemmas that are used throughout the analysis of Algorithm 1.

Lemma C.1. *Regarding the terms defined in Appendix A, their norms have the following upper bounds.*

$$\|A_t\|_F, \|\overline{A}_t\|_F, \|A\|_F \le 2\rho_{\max}, \tag{39}$$
$$\|B_t\|_F, \|\overline{B}_t\|_F, \|B\|_F \le \rho_{\max}, \tag{40}$$
$$\|C_t\|_F, \|\overline{C}_t\|_F, \|C\|_F \le 1, \tag{41}$$
$$\|b^{(m)}_t\|, \|\overline{b}^{(m)}_t\|, \|\overline{b}_t\|, \|\overline{\overline{b}}_t\|, \|b\| \le \rho_{\max} R_{\max}, \tag{42}$$
$$\|C^{-1}\| = \lambda_2^{-1}. \tag{43}$$

Proof. Consider any two vectors $u, v \in \mathbb{R}^d$, we have that $\|uv^\top\|_F = \sqrt{\mathrm{tr}(vu^\top uv^\top)} = \|u\|\|v\|$. Therefore, by Assumption 3, we obtain that

$$\begin{aligned}
\|A_t\|_F \le& \rho_t \|\phi(s_t)\|\|\gamma\phi(s_{t+1}) - \phi(s_t)\| \le \rho_{\max}\big[\gamma\|\phi(s_{t+1})\| + \|\phi(s_t)\|\big] \le 2\rho_{\max},\\
\|B_t\|_F \le& \gamma\rho_t\|\phi(s_{t+1})\|\|\phi(s_t)\| \le \rho_{\max},\\
\|C_t\|_F \le& \|\phi(s_t)\|^2 \le 1,\\
\|b^{(m)}_t\| \le& \rho_t R^{(m)}_t \|\phi(s_t)\| \le \rho_{\max} R_{\max}.
\end{aligned}$$

The proof for $\|A^{(m)}_t\|_F$, $\|\widetilde{b}^{(m)}_t\|$, etc. is similar.

On the other hand, by Jensen's inequality, we obtain that

$$\|A\|_F = \|\mathbb{E}_{\pi_b}[A_t]\|_F \le \mathbb{E}_{\pi_b}\|A_t\|_F \le 2\rho_{\max}, \quad \|\overline{A}_t\|_F \le \frac{1}{N}\sum_{i=tN}^{(t+1)N-1}\|A_i\|_F \le 2\rho_{\max}.$$

The proof for the other remaining matrices in eqs. (39)-(42) is similar by using the Jensen's inequality. Finally, we prove eq. (43). Note that $-C = \mathbb{E}_{\pi_b}\big(\phi(s_t)\phi(s_t)\big) \succ 0$ with $\lambda_{\min}(-C) = -\lambda_{\max}(C) = \lambda_2$. Hence, $\|C^{-1}\| = \lambda_{\max}(-C^{-1}) = \lambda^{-1}_{\min}(-C) = \lambda_2^{-1}$. □

Lemma C.2. *Suppose the MDP trajectory $\{s_t, a_t\}_{t\ge 0}$ is generated following a behavioral policy π_b where $a_t \triangleq \{a^{(m)}_t\}_m$. For any deterministic mappings $Y : \mathcal{S}\times\mathcal{A}_1\times\ldots\times\mathcal{A}_M\times\mathcal{S} \to \mathbb{R}^{p\times q}$ such that $\|Y(s,a,s')\|_F \le C_y, \forall s, s' \in \mathcal{S}, a^{(m)} \in \mathcal{A}_m$ where $a = \{a^{(m)}\}_m$, we have*

$$\Big\|\mathbb{E}\Big[\frac{1}{N}\sum_{i=tN}^{(t+1)N-1} Y(s_i,a_i,s_{i+1})\Big|\mathcal{F}_t\Big] - \overline{Y}\Big\| \le \frac{2\nu C_y}{N(1-\delta)},$$
$$\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=tN}^{(t+1)N-1} Y(s_i,a_i,s_{i+1}) - \overline{Y}\Big\|_F^2\Big|\mathcal{F}_t\Big] \le \frac{8C_y^2(\nu+1)}{N(1-\delta)},$$

where $\overline{Y} = \mathbb{E}Y(s_i,a_i,s_{i+1})$.

Note: A simplified version of the above lemma has been proposed and proved in Xu et al. (2020a), where a_i and s_{i+1} are omitted in the above inequality. We add a_i and s_{i+1} so that this lemma can be better applied to the quantities A_i, B_i, C_i and $b_i^{(m)}$ which rely on s_i as well as a_i and s_{i+1}. The proof logic is very similar to that of Xu et al. (2020a) and thus omitted here.

Corollary 1. *Regarding the terms defined in Appendix A, they have the following upper bounds.*

$$\mathbb{E}\big[\|\overline{C}_t - C\|_F \big| \mathcal{F}_t\big] \leq \frac{2\nu\rho_{\max}}{N(1-\delta)} \tag{44}$$

$$\mathbb{E}\big[\big\|\overline{B}_t - B\big\|_F^2 \big| \mathcal{F}_t\big] \leq \frac{8\rho_{\max}^2(\nu+1)}{N(1-\delta)} \tag{45}$$

$$\mathbb{E}\big[\big\|\bar{\overline{b}}_t - b\big\|^2 \big| \mathcal{F}_t\big] \leq \frac{8\rho_{\max}^2 R_{\max}^2(\nu+1)}{N(1-\delta)} \tag{46}$$

$$\mathbb{E}\big[\big\|\overline{A}_t - \overline{C}_t C^{-1}A\big\|_F^2 \big| \mathcal{F}_t\big] \leq \frac{32\rho_{\max}^2(\nu+1)}{N(1-\delta)}\Big(1+\frac{1}{\lambda_2}\Big)^2 \tag{47}$$

$$\mathbb{E}\big[\big\|\bar{\overline{b}}_t - \overline{C}_t C^{-1}b\big\|^2 \big| \mathcal{F}_t\big] \leq \frac{8\rho_{\max}^2 R_{\max}^2(\nu+1)}{N(1-\delta)}\Big(1+\frac{1}{\lambda_2}\Big)^2 \tag{48}$$

$$\mathbb{E}\big[\big\|\overline{A}_t - \overline{B}_t C^{-1}A - A^\top C^{-1}A\big\|_F^2 \big| \mathcal{F}_t\big] \leq \frac{32\rho_{\max}^2(\nu+1)}{N(1-\delta)}\Big(1+\frac{\rho_{\max}}{\lambda_2}\Big)^2, \tag{49}$$

Proof. Let $Y(s,a,s') = -\gamma\rho(s,a)\phi(s')\phi(s)^\top$ in Lemma C.2. Then it can be checked that $Y(s_t,a_t,s_{t+1}) = B_t, C_y = \rho_{\max}$, $\frac{1}{N}\sum_{i=tN}^{(t+1)N-1} Y(s_i,a_i,s_{i+1}) = \overline{B}_t$, and $\overline{Y} = \mathbb{E}_{\pi_b} Y(s_i,a_i,s_{i+1}) = B$.

Applying Lemma C.2 to these equations proves eq. (45). The eqs. (44) and (46) can be proved in a similar way.

Let $Y(s,a,s') = \rho(s,a)\phi(s)[\gamma\phi(s') - \phi(s)]^\top + \gamma\rho(s,a)\phi(s')\phi(s)^\top C^{-1}A$. Then, it can be checked that $Y(s_t,a_t,s_{t+1}) = A_t - B_t C^{-1}A$, $\frac{1}{N}\sum_{i=tN}^{(t+1)N-1} Y(s_i,a_i,s_{i+1}) = \overline{A}_t - \overline{B}_t C^{-1}A$. Moreover,

$$\|Y(s,a,s')\|_F \leq \rho_{\max}(\gamma+1) + \gamma\rho_{\max}\|C^{-1}\|\|A\| \leq 2\rho_{\max} + \rho_{\max}(\lambda_2^{-1})(2\rho_{\max}) = 2\rho_{\max}(1+\rho_{\max}/\lambda_2) := C_y,$$
$$\overline{Y} = \mathbb{E}_{\pi_b} Y(s_i,a_i,s_{i+1}) = A - BC^{-1}A = A^\top C^{-1}A.$$

Applying Lemma C.2 to these equations proves eq. (49). The equations (47) and (48) can be proved in a similar way. □

Lemma C.3. *The doubly stochastic matrix V and the difference matrix $\Delta = I - \frac{1}{M}\mathbf{1}\mathbf{1}^\top$ have the following properties:*

1. $\Delta V = V\Delta = V - \frac{1}{M}\mathbf{1}\mathbf{1}^\top$
2. *For any $x \in \mathbb{R}^M$ and $n \in \mathbb{N}^+$, $\|V^n\Delta x\| \leq \sigma_2^n\|\Delta x\|$ (σ_2 is the second largest singular value of V). Hence, for any $H \in \mathbb{R}^{M\times M}$, $\|V^n\Delta H\|_F \leq \sigma_2^n\|\Delta H\|_F$*

Proof. The first item can be proved by the following two equalities.

$$\Delta V = \Big(I - \frac{1}{M}\mathbf{1}\mathbf{1}^\top\Big)V = V - \frac{1}{M}\mathbf{1}\mathbf{1}^\top V = V - \frac{1}{M}\mathbf{1}\mathbf{1}^\top$$
$$V\Delta = V\Big(I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top\Big) = V - \frac{1}{M}V\mathbf{1}\mathbf{1}^\top = V - \frac{1}{M}\mathbf{1}\mathbf{1}^\top$$

The proof of the item 2 can be seen in page 3 of Qu & Li (2017). □

Based on Lemma C.3, we obtain the following inexactness of importance sampling ratio estimation $\widehat{\rho}_t^{(m)} \approx \rho_t$.

Corollary 2. *Under Assumption 4 and choosing* $L \geq \mathcal{O}(\frac{\ln M + M \ln \rho_{\max}}{\ln \sigma_2^{-1}})$*, the estimation error of the inexact global importance sampling ratio* $\widehat{\rho}_t^{(m)}$ *satisfies* $\sum_{m=1}^M \left(\widehat{\rho}_t^{(m)} - \rho_t\right)^2 \leq \sigma_2^{L/4}$*. Therefore, the following inequalities hold.*

$$\sum_{m=1}^M \left\|A_t^{(m)} - A_t\right\|_F^2, \sum_{m=1}^M \left\|\overline{A}_t^{(m)} - \overline{A}_t\right\|_F^2 \leq 4\sigma_2^{L/4} \tag{50}$$

$$\sum_{m=1}^M \left\|B_t^{(m)} - B_t\right\|_F^2, \sum_{m=1}^M \left\|\overline{B}_t^{(m)} - \overline{B}_t\right\|_F^2 \leq \sigma_2^{L/4}, \tag{51}$$

$$\sum_{m=1}^M \left\|\widetilde{b}_t^{(m)} - b_t^{(m)}\right\|^2, \sum_{m=1}^M \left\|\widetilde{\overline{b}}_t^{(m)} - \overline{b}_t^{(m)}\right\|^2 \leq \sigma_2^{L/4} R_{\max}^2. \tag{52}$$

As a result, the following upper bounds hold.

$$\|A_t^{(m)}\|_F, \left\|\overline{A}_t^{(m)}\right\|_F \leq 2\rho_{\max} + 2 \tag{53}$$

$$\|B_t^{(m)}\|_F, \left\|\overline{B}_t^{(m)}\right\|_F \leq \rho_{\max} + 1 \tag{54}$$

$$\|\widetilde{b}_t^{(m)}\|, \|\widetilde{\overline{b}}_t^{(m)}\| \leq R_{\max}(\rho_{\max} + 1) \tag{55}$$

Proof. Eq. (5 can be rewritten into the following matrix form.

$$\left[\widetilde{\rho}_{t,L}^{(1)}; \ldots; \widetilde{\rho}_{t,L}^{(M)}\right] = V^L \left[\widetilde{\rho}_{t,0}^{(1)}; \ldots; \widetilde{\rho}_{t,0}^{(M)}\right].$$

Hence, the item 1 of Lemma C.3 yields that

$$\Delta\left[\widetilde{\rho}_{t,L}^{(1)}; \ldots; \widetilde{\rho}_{t,L}^{(M)}\right] = V^L \Delta\left[\widetilde{\rho}_{t,0}^{(1)}; \ldots; \widetilde{\rho}_{t,0}^{(M)}\right].$$

Then the item 2 of Lemma C.3 yields that

$$\left\|\Delta\left[\widetilde{\rho}_{t,L}^{(1)}; \ldots; \widetilde{\rho}_{t,L}^{(M)}\right]\right\|^2 \leq \sigma_2^{2L} \left\|\Delta\left[\widetilde{\rho}_{t,0}^{(1)}; \ldots; \widetilde{\rho}_{t,0}^{(M)}\right]\right\|^2. \tag{56}$$

Denote $\rho_{\min} := \min_{m \in \mathcal{M}} \rho_t^{(m)}$. Then Assumption 4 implies that $\widetilde{\rho}_{t,0}^{(m)} = \ln \rho_t^{(m)} \in [\ln \rho_{\min}, \ln \rho_{\max}]$. Then it can be proved by iterating eq. (5) that $\widetilde{\rho}_{t,L}^{(m)} \in [\ln \rho_{\min}, \ln \rho_{\max}]$. Hence,

$$\frac{1}{M} \ln \rho_t = \frac{1}{M} \sum_{m=1}^M \ln \rho_t^{(m)} \in [\ln \rho_{\min}, \ln \rho_{\max}] \tag{57}$$

Then eqs. (56) and (57) imply that

$$\sum_{m=1}^M \left(\widetilde{\rho}_{t,L}^{(m)} - \frac{1}{M} \ln \rho_t\right)^2 \leq \sigma_2^{2L} \sum_{m=1}^M \left(\widetilde{\rho}_{t,0}^{(m)} - \frac{1}{M} \ln \rho_t\right)^2 \leq M \sigma_2^{2L} \ln^2(\rho_{\max}/\rho_{\min}). \tag{58}$$

Hence,

$$\left|\widetilde{\rho}_{t,L}^{(m)} - \frac{1}{M} \ln \rho_t\right| \leq \sqrt{M} \sigma_2^L \ln\left(\frac{\rho_{\max}}{\rho_{\min}}\right) \overset{(i)}{\leq} \frac{1}{2M} \ln\left(\frac{\rho_{\max}}{\rho_{\min}}\right), \tag{59}$$

where (i) uses the conditions that $L \geq \frac{12 \ln M + (8M+10) \ln \rho_{\max}}{\ln(\sigma_2^{-1})}$ and $\sigma_2 \in [0, 1)$.

Hence, eqs. (57) and (58) imply that

$$\widetilde{\rho}_{t,L}^{(m)} \leq \ln \rho_{\max} + \frac{1}{2M} \ln\left(\rho_{\max}/\rho_{\min}\right). \tag{60}$$

Therefore, we obtain that

$$\begin{aligned}\sum_{m=1}^{M}\left(\widehat{\rho}_t^{(m)}-\rho_t\right)^2 &\overset{(i)}{=} \sum_{m=1}^{M}\left(e^{M\widetilde{\rho}_{t,L}^{(m)}}-e^{\ln\rho_t}\right)^2\\ &\overset{(ii)}{\leq} \sum_{m=1}^{M}\left[\max\left(e^{M\widetilde{\rho}_{t,L}^{(m)}},e^{\ln\rho_t}\right)\right]^2\left(M\widetilde{\rho}_{t,L}^{(m)}-\ln\rho_t\right)^2\\ &\overset{(iii)}{\leq} M^3\sigma_2^{2L}\rho_{\max}^{M}\sqrt{\rho_{\max}/\rho_{\min}}\ln^2(\rho_{\max}/\rho_{\min})\\ &\overset{(iv)}{\leq} M^3\sigma_2^{2L}(\rho_{\max}^{M+2.5}/\rho_{\min}^{2.5}),\end{aligned} \tag{61}$$

where (i) uses eq. (6), (ii) uses the Lagrange's Mean Value Theorem, (iii) uses eqs. (57), (58) and (60), (iv) uses the inequality that $\ln x < x$ for $x = \rho_{\max}/\rho_{\min} \geq 1$.

Since at least one of $\{\widetilde{\rho}_{t,0}^{(m)}\}_{m\in\mathcal{M}}$ equals $\ln\rho_{\min}$, we have

$$\ln\rho_t = \sum_{m=1}^{M}\widetilde{\rho}_{t,0}^{(m)} \leq \ln\rho_{\min} + (M-1)\ln\rho_{\max}. \tag{62}$$

Then, eqs. (59) and (62) imply that

$$\widetilde{\rho}_{t,L}^{(m)} \leq \frac{1}{2M}\ln\rho_{\min} + \left(1-\frac{1}{2M}\right)\ln\rho_{\max} \tag{63}$$

Hence, we conclude that

$$\sum_{m=1}^{M}\left(\widehat{\rho}_t^{(m)}-\rho_t\right)^2 \overset{(i)}{=} \sum_{m=1}^{M}\left(e^{M\widetilde{\rho}_{t,L}^{(m)}}-e^{\ln\rho_t}\right)^2 \leq \sum_{m=1}^{M}\max\left(e^{2M\widetilde{\rho}_{t,L}^{(m)}},e^{2\ln\rho_t}\right) \overset{(ii)}{\leq} M\rho_{\min}\rho_{\max}^{2M-1} \tag{64}$$

where (i) uses eq. (6), (ii) uses eqs. (62) and (63).

When $\rho_{\min} \geq \sigma_2^{L/2}$, eq. (61) implies that $\sum_{m=1}^{M}\left(\widehat{\rho}_t^{(m)}-\rho_t\right)^2 \leq M^3\rho_{\max}^{M+2.5}\sigma_2^{0.75L}$; When $\rho_{\min} < \sigma_2^{L/2} < 1$, eq. (64) implies that $\sum_{m=1}^{M}\left(\widehat{\rho}_t^{(m)}-\rho_t\right)^2 \leq M\rho_{\max}^{2M-1}\sigma_2^{L/2}$. Both imply $\sum_{m=1}^{M}\left(\widehat{\rho}_t^{(m)}-\rho_t\right)^2 \leq \sigma_2^{L/4}$ since $L \geq \frac{12\ln M+(8M+10)\ln\rho_{\max}}{\ln(\sigma_2^{-1})}$.

Then, eq. (50) can be proved as follows.

$$\sum_{m=1}^{M}\left\|A_t^{(m)}-A_t\right\|_F^2 \leq \left\|\phi(s_t)[\gamma\phi(s_{t+1})-\phi(s_t)]^\top\right\|_F^2\sum_{m=1}^{M}(\widehat{\rho}_t^{(m)}-\rho_t)^2 \leq (1+\gamma)^2\sigma_2^{L/4} \leq 4\sigma_2^{L/4} \tag{65}$$

The above inequality implies that $\sum_{m=1}^{M}\left\|\overline{A}_t^{(m)}-\overline{A}_t\right\|_F^2 = \sum_{m=1}^{M}\left\|\frac{1}{N}\sum_{i=tN}^{(t+1)N-1}(A_i^{(m)}-A_i)\right\|_F^2 \leq 4\sigma_2^{L/4}$, where $\leq$ applies Jensen's inequality to the convex function $\|\cdot\|_F^2$. Eqs. (51) and (52) can be proved similarly.

Eq. (50) implies that $\left\|A_t^{(m)}-A_t\right\|_F, \left\|\overline{A}_t^{(m)}-\overline{A}_t\right\|_F \leq 2$. Hence, eq. (53) can be proved using triangle inequality and eq. (39). Eqs. (54) and (55) can be proved similarly. □

The proof of Corollary 2 introduces a new technique, which includes discussion of two cases: $\rho_{\min} := \min_{m\in\mathcal{M}}\rho_t^{(m)}$ lies in $[\sigma^{L/2},\rho_{\max}]$ and $(0,\sigma^{L/2}]$. This is necessary as the local average is applied to $\ln\widehat{\rho}_t^{(m)}$, which may be a large negative number that cannot ensure a small consensus error for a fixed number of local average steps L.

Lemma C.4. *Under the update rules of Algorithm 1 and choosing* $L \geq \frac{12\ln M+(8M+10)\ln\rho_{\max}}{\ln(\sigma_2^{-1})}$, $\alpha \leq \frac{\beta}{2\rho_{\max}+2}$, *the parameters have the following upper bound.*

$$\max_{m\in\mathcal{M}}\|\theta_T^{(m)}\| + \max_{m\in\mathcal{M}}\|w_T^{(m)}\| \leq 2\max_{m\in\mathcal{M}}(\|\theta_0^{(m)}\|+\|w_0^{(m)}\|+R_{\max})[1+\beta(2\rho_{\max}+3)]^T. \tag{66}$$

Proof. Since $L \geq \frac{12 \ln M + (8M+10) \ln \rho_{\max}}{\ln(\sigma_2^{-1})}$, eqs. (53)-(55) hold. Hence, these equations and the update rule imply that

$\|\theta_{t+1}^{(m)}\| \leq \sum_{m' \in \mathcal{N}_m} V_{m,m'} \|\theta_t^{(m')}\| + \alpha(\rho_{\max}+1)(2\|\theta_t^{(m)}\| + R_{\max} + \|w_t^{(m)}\|)$. Taking maximum with respect to m yields that

$$\begin{aligned}\max_{m \in \mathcal{M}} \|\theta_{t+1}^{(m)}\| &\leq \max_{m \in \mathcal{M}} \sum_{m' \in \mathcal{N}_m} V_{m,m'} \max_{m'' \in \mathcal{M}} \|\theta_t^{(m'')}\| + \alpha(\rho_{\max}+1)\big(2 \max_{m \in \mathcal{M}} \|\theta_t^{(m)}\| + R_{\max} + \max_{m \in \mathcal{M}} \|w_t^{(m)}\|\big). \\ &\leq \alpha(\rho_{\max}+1)\big(2 \max_{m \in \mathcal{M}} \|\theta_t^{(m)}\| + R_{\max} + \max_{m \in \mathcal{M}} \|w_t^{(m)}\|\big) + \max_{m \in \mathcal{M}} \|w_t^{(m)}\|. \end{aligned} \tag{67}$$

Similarly, it can be obtained that

$$\max_{m \in \mathcal{M}} \|w_{t+1}^{(m)}\| \leq 2\beta(\rho_{\max}+1) \max_{m \in \mathcal{M}} \|\theta_t^{(m)}\| + (1+\beta) \max_{m \in \mathcal{M}} \|w_t^{(m)}\| + \beta R_{\max}(\rho_{\max}+1). \tag{68}$$

Adding up eqs. (67) and (68) yields that

$$\begin{aligned}&\max_{m \in \mathcal{M}} \|\theta_{t+1}^{(m)}\| + \max_{m \in \mathcal{M}} \|w_{t+1}^{(m)}\| \\ &\leq 2(\alpha+\beta)(\rho_{\max}+1) \max_{m \in \mathcal{M}} \|\theta_t^{(m)}\| + (\alpha\rho_{\max} + \alpha + \beta + 1) \max_{m \in \mathcal{M}} \|w_t^{(m)}\| + R_{\max}(\alpha+\beta)(\rho_{\max}+1) \\ &\overset{(i)}{\leq} \beta(2\rho_{\max}+3) \max_{m \in \mathcal{M}} \|\theta_t^{(m)}\| + (1.5\beta+1) \max_{m \in \mathcal{M}} \|w_t^{(m)}\| + 0.5\beta R_{\max}(2\rho_{\max}+3) \\ &\leq [1+\beta(2\rho_{\max}+3)]\big(\max_{m \in \mathcal{M}} \|\theta_t^{(m)}\| + \max_{m \in \mathcal{M}} \|w_t^{(m)}\|\big) + 0.5\beta R_{\max}(2\rho_{\max}+3),\end{aligned}$$

where (i) uses the condition that $\alpha \leq \frac{\beta}{2\rho_{\max}+2}$ and (ii) uses $\rho_{\max} \geq 1$. By iterating the inequality above and using $\max_{m \in \mathcal{M}} \|\theta_0^{(m)}\| + \max_{m \in \mathcal{M}} \|w_0^{(m)}\| \leq 2 \max_{m \in \mathcal{M}} (\|\theta_0^{(m)}\| + \|w_0^{(m)}\|)$, we prove eq. (66). □